Lecture Notes in Computer Science 2369
Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Tokyo

Claus Fieker David R. Kohel (Eds.)

Algorithmic Number Theory

5th International Symposium, ANTS-V
Sydney, Australia, July 7-12, 2002
Proceedings

Springer

Volume Editors

Claus Fieker
David R. Kohel
University of Sydney, School of Mathematics and Statistics, F07
Sydney, NSW 2006, Australia
E-mail:{claus,kohel}@maths.usyd.edu.au

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Algorithmic number theory : 5th international symposium ; proceedings /
ANTS-V, Sydney, Australia, July 7 - 12, 2002. Claus Fieker ; David R. Kohel
(ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ;
Milan ; Paris ; Tokyo : Springer, 2002
(Lecture notes in computer science ; Vol. 2369)
ISBN 3-540-43863-7

CR Subject Classification (1998):F.2, G.2, E.3

ISSN 0302-9743
ISBN 3-540-43863-7 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Typesetting: Camera-ready by author, data conversion by Steingräber Satztechnik GmbH, Heidelberg
Printed on acid-free paper SPIN 10870350 06/3142 5 4 3 2 1 0

Preface

The Algorithmic Number Theory Symposia began in 1994 at Cornell University in Ithaca, New York to recognize the growing importance of algorithmic work in the theory of numbers. The subject of the conference is broadly construed to encompass a diverse body of mathematics, and to cover both the theoretical and practical advances in the field. They have been held every two years since: in Bordeaux (Université Bordeaux I) in 1996, Portland (Reed College) in 1998, Leiden (Universiteit Leiden) in 2000, and the present conference hosted by the Magma Computational Algebra Group at the University of Sydney.

The conference program included invited talks by Manjul Bhargava (Princeton), John Coates (Cambridge), Antoine Joux (DCSSI Crypto Lab), Bjorn Poonen (Berkeley), and Takakazu Satoh (Saitama), as well as 34 contributed talks in various areas of number theory. In addition to the mathematical program, the conference included a special dinner to honour Alf van der Poorten of Macquarie University, on the occasion of his 60th birthday.

Each paper was reviewed by at least two experts external to the program committee and the selection of papers was made on the basis of these recommendations. We express our appreciation to the 66 expert referees who provided reports on a very tight schedule. Refereeing of the submission from a member of the Magma group was organized by Joe Buhler.

The program committee thanks the generous advice from organizers of previous ANTS conferences, particularly Joe Buhler, Wieb Bosma, Hendrik Lenstra, and Bart de Smit. The conference was generously supported by the College of Science and Technology, the School of Mathematics and Statistics (both at the University of Sydney), the Australian Defence Science Technology Organisation, and eSign.

April 2002

John Cannon
Claus Fieker
David Kohel

Table of Contents

Invited Talks

Number Theory

Arithmetic Geometry

Elliptic Curves and CM

Point Counting

Cryptography

Function Fields

Discrete Logarithms and Factoring

Gröbner Bases

Complexity

Gauss Composition and Generalizations

Manjul Bhargava*

Clay Mathematics Institute and Princeton University

Abstract. We discuss several higher analogues of Gauss composition and consider their potential algorithmic applications.

1 Introduction

The class groups of quadratic fields have long held a special place in the annals of algorithmic algebraic number theory. This special place has been due in large part to the close relationship between ideal class groups of quadratic fields and integral binary quadratic forms, which allows one to reduce the study and computation of ideal classes in quadratic orders to the study of lattice points in a certain fixed three-dimensional real vector space—namely the space of binary quadratic forms over $\mathbb{R}$.

This fundamental correspondence, known classically as "Gauss composition", was discovered by Gauss almost exactly 200 years ago in his celebrated work *Disquisitiones Arithmeticae* of 1801. Even after two centuries, there is still no faster way known for computing the ideal class groups of quadratic fields than by Gauss composition.

The key feature of Gauss composition, which makes it so useful, is that one has a bijective correspondence between the arithmetic objects of interest (ideal classes of quadratic orders) with the integer points in a vector space—rather than, say, with the integer points on a high codimension variety in an affine space. The principle here is that one can readily locate all the integer points in a codimension zero region in a vector space, whereas searching for integer points on higher codimension subvarieties is extremely difficult in general, both computationally and theoretically.

Thus situations where one has a direct bijection between arithmetic objects of study and the integer points in a vector space (modulo, say, the action of a reductive group over $\mathbb{Z}$) are clearly of intrinsic interest, both from a theoretical and an algorithmic standpoint; and the question naturally arises as to whether there exist any spaces in addition to Gauss's space of binary quadratic forms that might share this remarkable property.

* I am very grateful to Professors Andrew Wiles and Peter Sarnak for all their enthusiasm and encouragement, and to Jonathan Hanke, Kiran Kedlaya, and Lenny Ng for helpful comments on an earlier draft of this paper. This work was supported by the Hertz Foundation and the Clay Mathematics Institute, and was conducted at Princeton University.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 1–8, 2002.

In [2] it was shown that, in fact, Gauss's space of binary quadratic forms is only one of at least 14 such vector spaces existing in nature whose lattice points may be put in correspondence with number fields and their class groups. A detailed treatment of these so-called "higher composition laws" will appear in [3]. The purpose of the current article is to give a short summary and announcement of these higher correspondences, and to discuss some of their potential algorithmic implications.

2 On Higher Composition Laws

The aforementioned higher correspondences generalizing Gauss composition are summarized in Table 1. Each such correspondence consists of a lattice $V_{\mathbb{Z}}$ and an arithmetic group $G_{\mathbb{Z}}$, such that the orbit space $V_{\mathbb{Z}}/G_{\mathbb{Z}}$ yields a bijective parametrization of some class $\mathscr{C}$ of number-theoretic objects.

For example, item #3 of Table 1 describes Gauss composition. Indeed, in this case, $V_{\mathbb{Z}}$ is the lattice $(\mathrm{Sym}^2\mathbb{Z}^2)^*$ of binary quadratic forms with integer coefficients, $G_{\mathbb{Z}}$ is $\mathrm{SL}_2(\mathbb{Z})$, and $V_{\mathbb{Z}}/G_{\mathbb{Z}}$ parametrizes (narrow) ideal classes in quadratic rings. As Table 1 also shows, there exist pairs $(V_{\mathbb{Z}}, G_{\mathbb{Z}})$ whose orbit spaces $V_{\mathbb{Z}}/G_{\mathbb{Z}}$ parametrize cubic rings, ideal classes in cubic rings, order 2 ideal classes in cubic rings, quartic rings, quintic rings, and more.

All 14 correspondences listed in Table 1, including Gauss's case, have the wonderful property that the maps $V_{\mathbb{Z}} \to \mathscr{C}$ are easily computed. In fact, all structure constants of the rings and modules in the fourth column can be given in terms of explicit polynomials in the coordinates of the lattice points $x \in V_{\mathbb{Z}}$. The inverse mappings $\mathscr{C} \to V_{\mathbb{Z}}/G_{\mathbb{Z}}$ can also be computed explicitly.

What this means as far as algorithms are concerned is that, rather than computing directly with the arithmetic objects in $\mathscr{C}$, one may instead compute with the points in the lattices $V_{\mathbb{Z}}$, which for many purposes proves to be much more efficient. We give some examples below.

Application 1 *(Discriminants) The discriminants of the rings occurring in the fourth column of Table 1 can be quickly evaluated in terms of the elements $x \in V_{\mathbb{Z}}$. Like the $\mathrm{SL}_2(\mathbb{Z})$-action on binary quadratic forms, each case 1–14 listed in Table 1 has the property that the action of $G_{\mathbb{Z}}$ on $V_{\mathbb{Z}}$ has a single polynomial invariant, which we call the discriminant. A beautiful calculation reveals that, in every case, this discriminant invariant coincides precisely with the discriminant of the corresponding ring in the fourth column! The fifth column of Table 1 lists the degrees k of these discriminant invariants as polynomials on $V_{\mathbb{Z}}$. For example, in Gauss's case, the discriminant D of a binary quadratic form $ax^2 + bxy + cy^2$ is simply the quadratic expression $D = b^2 - 4ac$; hence the value of k listed in Gauss's case is 2. In every case, the discriminant polynomial itself may be efficiently evaluated at any given point of $V_{\mathbb{Z}}$, and thus the discriminants of the rings occurring in $\mathscr{C}$ can also be computed efficiently.*

Application 2 *(Maximality) Criteria to test the maximality of the rings in the fourth column may be given in terms of certain simple congruence conditions* *on the corresponding points $x \in V_{\mathbb{Z}}$. Thus, sorting out which $x \in V_{\mathbb{Z}}$ correspond to maximal orders in number fields is a relatively simple process. Moreover, in the case when $x \in V_{\mathbb{Z}}$ corresponds to a maximal order $\mathcal{O}_K$, splitting behavior of primes in $\mathcal{O}_K$ can also be given in terms of simple congruence conditions on x.*

Application 3 *(Invertibility) In all the cases of Table 1 that involve ideal classes, one can write down explicit congruence conditions on $x \in V_{\mathbb{Z}}$ that determine whether a corresponding ideal class is invertible. This can be useful when one only wishes to work in the ideal class group, rather than with general ideal classes.*

Besides such basic data on discriminant, maximality, prime splitting, and invertibility, the points in the spaces $V_{\mathbb{Z}}$ also carry much additional information that is more subtle. For example, the lattice $V_{\mathbb{Z}}$ in #13 not only carries information on quartic rings, but it also carries complete information on their "cubic resolvent" rings. (Cubic resolvents are cubic rings that are related to quartic rings in a certain special way; see [2].) Similarly, $V_{\mathbb{Z}}$ in #14 not only carries information on quintic rings, but also on their sextic resolvents. In addition, it turns out that the lattice $V_{\mathbb{Z}}$ in #8 may be used to parametrize all rank 2 modules over quadratic orders, while $V_{\mathbb{Z}}$ in #7 and #12 contain information on certain special rank 3 and rank 2 modules over quadratic and cubic orders respectively (see [3]). Various other properties of the rings and ideal classes corresponding to elements $x \in V_{\mathbb{Z}}$ can also be read off quite simply from appropriate properties of x.

For these reasons, we expect that these higher correspondences should be very useful for computations, in the same way that Gauss composition has become an indispensible tool in computing with ideal class groups of quadratic fields. In particular, the correspondences should be useful in the *enumeration* of small degree number fields and their class groups, and in the construction of the relevant tables. For the latter application a theory of reduction is required, which we discuss more fully in Section 3.

Notation on Table 1. The symbol $\tilde{\mathbb{Z}}$ in #2 denotes the set of elements in $\mathbb{Z}$ congruent to 0 or 1 (mod 4). We use $(\mathrm{Sym}^2\mathbb{Z}^2)^*$ to denote the set of binary quadratic forms with integral coefficients, while $\mathrm{Sym}^2\mathbb{Z}^2$ denotes the sublattice of integral binary quadratic forms whose middle coefficients is even. Similarly, $(\mathrm{Sym}^3\mathbb{Z}^2)^*$ denotes the space of binary cubic forms with integer coefficients, while $\mathrm{Sym}^3\mathbb{Z}^2$ denotes the subset of forms whose middle two coefficients are multiples of 3. The symbol $\otimes$ is used for the usual tensor product; thus, for example, $\mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$ is the space of $2 \times 2 \times 2$ cubical integer matrices, $(\mathbb{Z}^2 \otimes \mathrm{Sym}^3\mathbb{Z}^2)^*$ is the space of pairs of ternary quadratic forms with integer coefficients, and $\mathbb{Z}^2 \otimes \mathrm{Sym}^3\mathbb{Z}^2$ is the space of pairs of integral ternary quadratic forms whose cross terms have even coefficients.

The fifth column of Table 1 gives the $\mathbb{Z}$-rank of the lattice $V_{\mathbb{Z}}$. The sixth column gives the degree k of the discriminant invariant as a polynomial on $V_{\mathbb{Z}}$. Finally, it turns out that each of the correspondences listed in Table 1 is related in a special way to some exceptional Lie group H (see [2, §6.1]). These exceptional

Table 1. Summary of Higher Composition Laws

Summary of Higher Composition Laws						
#	Lattice ($V_{\mathbb{Z}}$)	Group acting ($G_{\mathbb{Z}}$)	Parametrizes ($\mathscr{C}$)	(k)	(n)	(H)
1.	$\{0\}$	-	Linear rings	0	0	A_0
2.	$\widetilde{\mathbb{Z}}$	$\mathrm{SL}_1(\mathbb{Z})$	Quadratic rings	1	1	A_1
3.	$(\mathrm{Sym}^2\mathbb{Z}^2)^*$ (GAUSS'S LAW)	$\mathrm{SL}_2(\mathbb{Z})$	Ideal classes in quadratic rings	2	3	B_2
4.	$\mathrm{Sym}^3\mathbb{Z}^2$	$\mathrm{SL}_2(\mathbb{Z})$	Order 3 ideal classes in quadratic rings	4	4	G_2
5.	$\mathbb{Z}^2 \otimes \mathrm{Sym}^2\mathbb{Z}^2$	$\mathrm{SL}_2(\mathbb{Z})^2$	Ideal classes in quadratic rings	4	6	B_3
6.	$\mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$	$\mathrm{SL}_2(\mathbb{Z})^3$	Pairs of ideal classes in quadratic rings	4	8	D_4
7.	$\mathbb{Z}^2 \otimes \wedge^2\mathbb{Z}^4$	$\mathrm{SL}_2(\mathbb{Z}) \times \mathrm{SL}_4(\mathbb{Z})$	Ideal classes in quadratic rings	4	12	D_5
8.	$\wedge^3\mathbb{Z}^6$	$\mathrm{SL}_6(\mathbb{Z})$	Quadratic rings	4	20	E_6
9.	$(\mathrm{Sym}^3\mathbb{Z}^2)^*$	$\mathrm{GL}_2(\mathbb{Z})$	Cubic rings	4	4	G_2
10.	$\mathbb{Z}^2 \otimes \mathrm{Sym}^2\mathbb{Z}^3$	$\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$	Order 2 ideal classes in cubic rings	12	12	F_4
11.	$\mathbb{Z}^2 \otimes \mathbb{Z}^3 \otimes \mathbb{Z}^3$	$\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})^2$	Ideal classes in cubic rings	12	18	E_6
12.	$\mathbb{Z}^2 \otimes \wedge^2\mathbb{Z}^6$	$\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_6(\mathbb{Z})$	Cubic rings	12	30	E_7
13.	$(\mathbb{Z}^2 \otimes \mathrm{Sym}^2\mathbb{Z}^3)^*$	$\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$	Quartic rings	12	12	F_4
14.	$\mathbb{Z}^4 \otimes \wedge^2\mathbb{Z}^5$	$\mathrm{GL}_4(\mathbb{Z}) \times \mathrm{SL}_5(\mathbb{Z})$	Quintic rings	40	40	E_8

groups have been listed in the last column of Table 1. The list shows that the spaces underlying higher composition laws may be thought of as being roughly in one-to-one correspondence with the exceptional Lie groups.

3 Reduction Theory and Other Algorithmic Considerations

In order to develop fast algorithms to enumerate the objects listed in column 4 of Table 1, we would like to have a good reduction theory which allows the selection of convenient representatives in $V_{\mathbb{Z}}$ for the elements of $V_{\mathbb{Z}}/G_{\mathbb{Z}}$.

In cases #1, #2, #4, #9, #10, #13, #14, and the definite (i.e., negative discriminant) subcases of #3, #5, and #6, what we expect, more precisely, is a "fundamental region" $\mathcal{F}$ in the real vector space $V_{\mathbb{R}} = V_{\mathbb{Z}} \otimes \mathbb{R}$, defined by homogeneous polynomial inequalities, such that every element of $V_{\mathbb{Z}}/G_{\mathbb{Z}}$ is represented exactly once in $\mathcal{F}$. Such fundamental regions $\mathcal{F}$ can be proven to exist in all these cases from a purely logical standpoint (e.g., using the work of Tarski [15] and Seidenberg [13]). But from the standpoint of algorithmic number theory, we are not merely interested in the existence of a region $\mathcal{F}$—we would also like to be able to explicitly write it down, and have the polynomial inequalities bounding the region be as nice as possible. There is certainly an element of art to the problem.

However, once such a reduction theory is established, and a corresponding region $\mathcal{F}$ has been obtained, then the arithmetic objects in the fourth column of Table 1 in these cases can be enumerated, up to (absolute) discriminant D, simply by listing all the lattice points in the region

$$\mathcal{F}_D = \mathcal{F} \cap \{x \in V_{\mathbb{R}} : |\mathrm{Disc}(x)| < D\}, \tag{1}$$

where $\mathrm{Disc}(x)$ denotes the discriminant of the point $x \in V_{\mathbb{R}}$.

If the region $\mathcal{F}$ is reasonably nice, then, by homogeneity considerations, the time taken to list all lattice points in $\mathcal{F}_D$ should not be more than $O(D^{n/k+\epsilon})$, where n and k are as given in Table 1. Moreover, by searching only for those elements of $\mathcal{F}_D$ satisfying certain congruence conditions, one can enumerate various subclasses of these arithmetic objects, such as those involving maximal orders, or projective ideal classes, etc. Again, the time needed here should also not be more than $O(D^{n/k+\epsilon})$. Since every object of interest is represented in $\mathcal{F}$ exactly once, these algorithms would be quite close to being optimal for generating the relevant tables.

"Reduction theories" yielding such nice fundamental regions $\mathcal{F}$ are in fact known in some cases.

Example 1. The first nontrivial case, namely Gauss's case of binary quadratic forms, is due to Gauss himself. Gauss showed that any positive definite quadratic form $f(x,y) = ax^2 + bxy + cy^2$ can be uniquely transformed, by a linear substitution in $\mathrm{SL}_2(\mathbb{Z})$, into one whose coefficients satisfy

$$-a < b \leq a < c \quad \text{or} \quad 0 \leq b \leq a = c. \tag{2}$$

The region $\mathcal{F}$ defined by (2) has all the properties we require of it, and indeed has been fundamental in numerous algorithms relating to the ideal class groups of imaginary quadratic fields (see [4], [5]).

Example 2. An analogous reduction theory for binary cubic forms of positive discriminant was discovered by Hermite [10]. Hermite showed that the generic binary cubic form $ax^3 + bx^2y + cxy^2 + dy^3$ of positive discriminant can be transformed by an element of $\mathrm{GL}_2(\mathbb{Z})$ into a unique form satisfying $a > 0$ and

$$\begin{gathered} -(b^2 - 3ac) < bc - 9ad \leq b^2 - 3ac < c^2 - 3bd \\ \text{or} \quad 0 \leq bc - 9ad \leq b^2 - 3ac = c^2 - 3bd. \end{gathered} \tag{3}$$

Mathews and Berwick [11] subsequently studied the case of cubic forms of negative discriminant, and showed that the generic cubic form $ax^3+bx^2y+cxy^2+dy^3$ of negative discriminant can be uniquely transformed into a form satisfying

$$\begin{gathered} d(d-b)+a(c-a)>0, \quad ad-(a+b)(a+b+c)<0, \\ \text{and} \quad ad+(a-b)(a-b+c)>0. \end{gathered} \tag{4}$$

The correspondence between integral binary cubic forms and cubic rings, summarized in #9 of Table 1, is the only other nontrivial lattice correspondence (outside Gauss composition) that has been known previously. This remarkable connection was discovered by Delone-Faddeev in [8]; shortly thereafter, congruence conditions to determine whether a binary cubic form corresponds to a maximal order were obtained by Davenport-Heilbronn in [7]. Using this theory of Davenport-Heilbronn and the reduction theories of Hermite and Mathews-Berwick, a very fast algorithm to enumerate cubic orders and cubic fields was recently implemented by Belabas [1].

Since the lattice $\mathrm{Sym}^3\mathbb{Z}^2$ in case #4 of Table 1 is simply the dual of $(\mathrm{Sym}^3\mathbb{Z}^2)^*$ in the same vector space, the methods of Belabas could also be used to quickly enumerate order 3 ideal classes in quadratic orders.

Example 3. We discuss a method for constructing a fundamental region $\mathcal{F}$ in an important subcase of #13. Let us say an element $x \in (\mathbb{Z}^2 \otimes \mathrm{Sym}^2\mathbb{Z}^3)^*$ is *totally real* if it corresponds to an order in a totally real quartic field (under the association of Table 1). One can show that the space $V_\mathbb{Z} = (\mathbb{Z}^2 \otimes \mathrm{Sym}^2\mathbb{Z}^3)^*$ has a degree 4 map $x \mapsto Q_x$ to the space of ternary quadratic forms which is $\mathrm{SL}_3(\mathbb{Z})$-covariant, and a degree 3 map $x \mapsto f_x$ to the space of binary cubic forms which is $\mathrm{SL}_2(\mathbb{Z})$-covariant. Moreover, if x is totally real, then Q_x is a definite quadratic form. We say an element $x \in V_\mathbb{Z}$ is *reduced* if Q_x is $\mathrm{SL}_3(\mathbb{Z})$-reduced in the sense of Minkowski and f_x is $\mathrm{GL}_2(\mathbb{Z})$-reduced in the sense of Example 2. This leads to various homogeneous inequalities defining the desired fundamental region $\mathcal{F} \in V_\mathbb{R}$. These inequalities are explicitly written down in [2].

Presumably, Example 3 could be used to obtain a quasi-linear time algorithm for enumerating totally real quartic fields. In a similar manner, we would like such reduction theories to be developed in all relevant cases.

Examples 2 and 3 above were both based on finding appropriate positive definite quadratic form covariants, and defining reduction in terms of those quadratics. Indeed, many of the items of Table 1 can be handled in this way. Whether that is the best way to proceed in all cases is an open problem.

Problem 4. For each of the cases #1, #2, #4, #9, #10, #13, #14, and the definite subcases of #3, #5, and #6, develop a reduction theory analogous to those presented in Examples 1–3.

Outside the cases listed in Problem 1, there are also case #11 and the indefinite (positive discriminant) subcases of items #3, #5, and #6, which may also have significant algorithmic consequences. Although fundamental domains $\mathcal{F}$

conjecturally will not exist in these cases, we can still expect to have a codimension zero region $\mathcal{F}$ such that each element of the orbit space $V_{\mathbb{Z}}/G_{\mathbb{Z}}$ is represented in $\mathcal{F}$ at least once but only finitely many times. Moreover, we suspect that in all these cases $\mathcal{F}$ could be chosen so that $\mathcal{F}_D$ is *compact* for every D.

Example 5. For indefinite binary quadratic forms, Gauss used the following definition of reduction. An indefinite form $f(x,y) = ax^2 + bxy + cy^2$ of discriminant D is said to be *reduced* if it satisfies the inequalities

$$0 < b < \sqrt{D} \quad \text{and} \quad \sqrt{D} - b < 2a < \sqrt{D} + b. \tag{5}$$

One can check that any indefinite binary quadratic is $\mathrm{SL}_2(\mathbb{Z})$-equivalent to some form in the region $\mathcal{F}$ defined by (5). Moreover, (5) implies that $|a|, |b|, |c| < \sqrt{D}$, and hence $\mathcal{F}_D$ is compact for all D.

To enumerate all ideal classes in real quadratic orders of discriminant at most D, it again suffices to list all lattice points in the region $\mathcal{F}_D$, where $\mathcal{F}$ is given by the inequalities (5). However, since $\mathcal{F}$ is not a true fundamental domain, there is a slight additional complication in that one must then group that list of lattice points into $\mathrm{SL}_2(\mathbb{Z})$-equivalence classes. It turns out this can be done quite efficiently using the theory of "cycles" (see [4]). Hence this does not affect the running time too much, and one can determine all $\mathrm{SL}_2(\mathbb{Z})$-equivalence classes of indefinite quadratic forms of discriminant at most D in time $O(D^{3/2})$, which is very fast.

It is actually conceivable that there could be faster algorithms for this purpose, although, in our current state of knowledge, there is no algorithm that could *provably* run faster than $O(D^{3/2})$. The reason for this is that we know

$$\sum_{0<d<X} h_d \log \epsilon_d \sim \frac{\pi^2}{18\zeta(3)} D^{3/2}, \tag{6}$$

where h_d and $\log \epsilon_d$ denote the class number and regulator, respectively, of the unique quadratic order of discriminant d. This asymptotic formula was first stated by Gauss, and was subsequently proven by Siegel [14]. However, there is no way known to separate the class number and regulator in sums such as (6). Hence the best estimate currently known for

$$\sum_{0<d<X} h_d \tag{7}$$

is also $O(D^{3/2})$; it is a major and long-standing unsolved problem in number theory to improve this estimate, and any algorithm that provably ran faster than $O(D^{3/2})$ would necessarily require a spectacular theoretical breakthrough involving a separation of class number and regulator.

Barring such a breakthrough, Gauss's algorithm for enumerating ideal classes in real quadratic fields is essentially the best that one could hope for. We should like to have similarly "optimal" algorithms in the other analogous cases— case #11 is of particular interest, since it would allow for the quick computation of ideal classes in *cubic* fields.

Problem 6. For case #11 and the indefinite subcases of #3, #5, and #6, develop (a) a notion of "reduced" analogous to Example 4, and (b) a method for determining when two reduced elements are equivalent, analogous to the theory of "cycles" in the case of indefinite binary quadratic forms.

Progress on Problems 1 and 2 will be key to making full algorithmic and theoretical use of the higher correspondences discussed in Section 2. In particular, once such reduction theories are established, the correspondences of [2] and [3] listed in Table 1 should eventually yield quite efficient algorithms for enumerating and generating tables of ideal classes in cubic fields, ideal classes of order 2 in cubic fields, quartic fields (and their cubic resolvents), quintic fields (and their sextic resolvents), and many other related objects.

References

1. K. Belabas, A fast algorithm to compute cubic fields, *Math. Comp.* **66** (1997), no. 219, 1213–1237.
2. M. Bhargava, *Higher Composition Laws*, Ph. D. Thesis, Princeton University, 2001.
3. M. Bhargava, Higher composition laws I–V, in progress.
4. D. Buell, *Binary quadratic forms. Classical theory and modern computations*, Springer-Verlag, New York, 1989.
5. H. Cohen, *A course in computational algebraic number theory*, Graduate Texts in Mathematics **138**, Springer-Verlag, Berlin, 1993.
6. H. Cohen, *Advanced topics in computational number theory*, Graduate Texts in Mathematics **193**, Springer-Verlag, New York, 2000.
7. H. Davenport and H. Heilbronn, On the density of discriminants of cubic fields II, *Proc. Roy. Soc. London Ser. A* **322** (1971), no. 1551, 405–420.
8. B. N. Delone and D. K. Faddeev, *The theory of irrationalities of the third degree*, AMS Translations of Mathematical Monographs **10**, 1964.
9. C. F. Gauss, *Disquisitiones Arithmeticae*, 1801.
10. C. Hermite, Sur la réduction des formes cubiques à deux indéterminées, *C. R. Acad. Sci. Paris* **48** (1859), 351–357.
11. G. B. Mathews and W. E. H. Berwick, On the reduction of arithmetical binary cubics which have a negative determinant, *Proc. Lond. Math. Soc. (2)* **10** (1912), 48–53.
12. M. Sato and T. Kimura, A classification of irreducible prehomogeneous vector spaces and their relative invariants, *Nagoya Math. J.* **65** (1977), 1–155.
13. A. Seidenberg, A new decision method for elementary algebra, *Ann. Math.* **60** (1954), 365–374.
14. C. L. Siegel, The average measure of quadratic forms with given determinant and signature, *Ann. Math. (2)* **45** (1944), 667–685.
15. A. Tarski, *A decision method for elementary algebra and geometry*, 2nd ed., revised, Berkeley and Los Angeles, 1951.
16. D. J. Wright and A. Yukie, Prehomogeneous vector spaces and field extensions, *Invent. Math.* **110** (1992), 283–314.

Elliptic Curves — The Crossroads of Theory and Computation

John Coates

Department of Pure Mathematics and Mathematical Statistics
CMS, Wilberforce Road, Cambridge CB3 0WB, U.K.

1 Introduction

The interplay between theory and computation has been a vital force for progress throughout the long history of the arithmetic of elliptic curves. I have been fortunate to see at fairly close hand two marvellous examples of this interplay. Firstly, I remember my amazement as a student in Canberra and Paris in the mid 1960's to see the conjecture of Birch and Swinnerton-Dyer evolve from a series of brilliant numerical experiments, which revolutionized arithmetical algebraic geometry. Secondly, I remember my fascination as a young post-doc at Harvard in the beginning of the 1970's to see John Tate work on a daily basis by always mixing sophisticated theory with hand calculations of numerical examples. Of course, since this time, numerical computations have been greatly changed by the advent of ever faster computers, and the discovery of important practical applications via cryptography. Computational mathematics has rightly become a branch of mathematics in its own right. Nevertheless, the theme I want to stress in my lecture is that the ancient union between theory and computation is as potent a force as ever today. It is my strong personal view that the best computations on elliptic curves are those that lead to new insights for attacking the unsolved theoretical problems. Equally, I firmly believe that no abstract theorem about the arithmetic of elliptic curves is worth its salt unless illuminating numerical examples of it can be given.

I want to illustrate this general theme by discussing some aspects of the arithmetic of elliptic curves over the fields generated by the coordinates of their points of finite order. When the elliptic curve has complex multiplication, these division fields are essentially abelian, and a great deal is now known about the arithmetic, largely by mimicking the ideas introduced by Iwasawa to study cyclotomic fields. Thus I will only discuss today elliptic curves without complex multiplication. Thanks to a celebrated theorem of Serre [1], the division fields of elliptic curves without complex multiplication provide what is probably the simplest class of non-abelian extensions of number fields, and it seems certain that they hold some of the keys to our eventual understanding of non-abelian class field theory. Very little numerical work has been done so far on these division fields. Nevertheless, I hope to convince you today that the arithmetic of both these fields themselves and the elliptic curve over them is fertile ground for the interplay between numerical calculations and theory.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 9–19, 2002.

I am very grateful to John Cannon and the organizers of the Algorithmic Number Theory Symposium for their kind invitation to address this meeting.

2 Iwasawa Algebras

I want to briefly describe some recent joint work with Schneider and Sujatha [2], which provides the theoretical background to the questions we wish to study. Let p be a prime number, and G a compact p-adic Lie group, of positive dimension, which we will denote by d. In our examples, G will always be a Galois group arising from points of finite order on an elliptic curve without complex multiplication, but the theory works quite generally. We recall that the Iwasawa algebra $\Lambda(G)$ of G is defined by

$$\Lambda(G) = \varprojlim_{U} \mathbb{Z}_p[G/U],$$

where U runs over the open normal subgroups of G. This algebra is very important in arithmetic geometry, because any compact $\mathbb{Z}_p$-module on which G acts continuously on the left has a unique structure as a left $\Lambda(G)$-module, extending the G-action. When G is non-abelian, $\Lambda(G)$ is non-commutative, and its study seems to have been curiously neglected by the experts in non-commutative algebra.

Before describing our structure theorem for modules over $\Lambda(G)$ for a very wide class of non-commutative groups G, let me recall that, when $G = \mathbb{Z}_p^d$, $\Lambda(G)$ is isomorphic to the local ring $\mathbb{Z}_p[[T_1, \cdots, T_d]]$ of formal power series in d variables with coefficients in $\mathbb{Z}_p$. In this special case, the structure theory of finitely generated $\Lambda(G)$-modules, up to pseudo-isomorphism, is very well known (see [3], Chap. VII, §4.4, Theorems 4 and 5), and is due originally to Iwasawa and Serre. Returning to general G, we assume from now on that G is pro-p, and has no element of order p. In his thesis, Venjakob [4], [5] used ideas of Bjork [6] to define a good theory of dimension for finitely generated $\Lambda(G)$-modules. In particular, Venjakob defines a finitely generated left $\Lambda(G)$-module M to be pseudo-null if it is $\Lambda(G)$-torsion (i.e. each element of M has a non-zero annihilator in $\Lambda(G)$), and, in addition, $\mathrm{Ext}^1_{\Lambda(G)}(M, \Lambda(G)) = 0$. To prove our structure theorem, we need the stronger hypothesis that G is p-valued in the sense of Lazard [7] (this automatically implies that G is pro-p and has no element of order p). The classic example of a p-valued group is the group of matrices in $GL_n(\mathbb{Z}_p)$, which are congruent to the identity modulo p (resp. mod 4) if p is odd (resp. if $p = 2$). Moreover, if $p > n + 1$, any closed pro-p subgroup of $GL_n(\mathbb{Z}_p)$ is p-valued. Also every closed subgroup of a p-valued group is p-valued. If M is a left or right $\Lambda(G)$-module, we define its dual to be the corresponding right or left $\Lambda(G)$-module $M^* = \mathrm{Hom}_{\Lambda(G)}(M, \Lambda(G)))$. As usual, we say that M is reflexive if the natural map from M to M^{**} is an isomorphism. Here is the principal result of [2].

Theorem 1. *([2]). Let G be a p-valued compact p-adic Lie group, and let M be a finitely generated torsion $\Lambda(G)$-module. Let M_0 be the maximal pseudo-null* *submodule of M. Then there exist non-zero left ideals $L_1, \cdots, L_m$ and a $\Lambda(G)$-injection*

$$\varphi : \bigoplus_{i=1}^{m} \Lambda(G)/L_i \to M/M_0 \tag{1}$$

with Coker (φ) pseudo-null. Moreover, the ideals $L_1, \cdots, L_m$ are always reflexive.

We are grateful to Venjakob for pointing out to us that the left ideals appearing in Theorem 1 are always reflexive. We should also point out that the special case of Theorem 1 in which M/M_0 is killed by some power of p was proven earlier by Venjakob [4], [5], and Howson [8]. We give two proofs of Theorem 1 in [2], one based on the algebraic theory of microlocalization, and the other showing that it can be derived from the work of Chamarie [9], [10] on modules over maximal orders.

We now discuss some further aspects of the structure theory of torsion $\Lambda(G)$-modules, especially those which seem to be important in concrete examples arising from elliptic curves. We assume for the rest of this section that G is p-valued. Firstly, there is the important open question of whether or not the left ideals L_i appearing in Theorem 1 can be chosen to be principal (when $G = \mathbb{Z}_p^d$, this is well known to be true because $\Lambda(G)$ is a unique factorization domain). As $\Lambda(G)$ is a local ring, it would suffice to show that the L_i can be chosen to be projective $\Lambda(G)$-modules. Secondly, completely new phenomena occur in the non-commutative theory when one considers global annihilators of our modules. If M is a torsion $\Lambda(G)$-module, we define as usual its annihilator, which we denote by $\mathrm{ann}_{\Lambda(G)}(M)$, to be the set of all τ in $\Lambda(G)$ such that $\tau.M = 0$. Note that $\mathrm{ann}_{\Lambda(G)}(M)$ is automatically a two sided ideal of $\Lambda(G)$. For technical reasons explained below, it is more natural to consider the annihilator of M/M_0, where M_0 is the maximal pseudo-null submodule of M, and so we define

$$a(M) = \mathrm{a}nn_{\Lambda(G)}(M/M_0). \tag{2}$$

When G is non-commutative, the most common thing seems to be for finitely generated torsion $\Lambda(G)$-modules M to have $a(M) = 0$ (Greenberg (unpublished), and more recently Venjakob [11] have given examples of such modules), and the rare thing seems to be to find M with $a(M) \neq 0$ and $a(M)$ not containing a non-zero element of the centre of $\Lambda(G)$. Moreover, as far as modules M with $a(M) = 0$ are concerned, something even more surprising can occur. There exist, at least for certain non-commutative G, finitely generated torsion $\Lambda(G)$-modules M such that not only do we have $a(M) = 0$, but, in addition, for every $\Lambda(G)$-subquotient N of M, we have either that N is pseudo-null or $a(N) = 0$. Even though such modules M seem difficult to envisage intuitively, Hachimori and Venjakob [12] have already found examples of them occurring in the arithmetic of elliptic curves, and one cannot help speculating that they occur rather widely in Iwasawa theory.

To express the above notions more precisely, it is convenient to pass to a quotient category of the category of all finitely generated torsion $\Lambda(G)$-modules. Denoting this latter category by $C^{\circ}(G)$, we write $C^1(G)$ for the full subcategory of all pseudo-null $\Lambda(G)$-modules. Since $C^1(G)$ is closed under taking subobjects, quotients, and extensions, we can therefore form the quotient category

$$\mathfrak{M}(G) = C^\circ(G)/C^1(G),$$

and we write $Q : C^\circ(G) \to \mathfrak{M}(G)$ for the canonical functor. We define the annihilator of an element $Q(M)$ of the quotient category by

$$ann(Q(M)) = a(M). \tag{3}$$

This is well defined because a delicate lemma of Robson [13] proves that $a(M_1) = a(M_2)$ whenever $Q(M_1)$ is isomorphic to $Q(M_2)$ in $\mathfrak{M}(G)$. We say that $Q(M)$ is *bounded* if $ann(Q(M)) \neq 0$. We say that $Q(M)$ is *completely faithful* if $ann(Q(N)) = 0$ for every N in $C^\circ(G)$ such that $Q(N)$ is a non-zero subquotient of $Q(M)$. For an arbitrary M in $C^\circ(G)$, Chamarie [10] proves that there is a canonical decomposition in $\mathfrak{M}(G)$

$$Q(M) = Q(U) \oplus Q(V),$$

where $Q(U)$ is completely faithful and $Q(V)$ is bounded. The only general result known about completely faithful objects $Q(U)$ at present is that they are cyclic in $\mathfrak{M}(G)$, i.e. isomorphic to $Q(\Lambda(G)/L)$, where L is some non-zero left ideal of $\Lambda(G)$ (See [10]).

It is shown in [2] that one can define a characteristic ideal for each non-zero bounded object $Q(M)$ of $\mathfrak{M}(G)$, in perfect analogy with the commutative theory. However, we have to confess that this theory is largely academic at present, because for many of the most important groups G we simply do not know whether there exist non-zero bounded $Q(M)$ which are not annihilated by some non-zero element of the centre of $\Lambda(G)$. A key example where this question has not been settled yet is when G is the kernel of the reduction map from $SL_2(\mathbb{Z}_p)$ to $SL_2(\mathbb{F}_p)$, for any odd prime p.

3 Elliptic Curves over Division Fields

Let F be a finite extension of $\mathbb{Q}$, and E an elliptic curve defined over F with $\mathrm{End}_{\overline{\mathbb{Q}}}(E) = \mathbb{Z}$. Let $E_{p^n} (1 \leqslant n \leqslant \infty)$ denote the group of p^n-division points on E. We define

$$F_\infty = F(E_{p^\infty}), \quad G = G(F_\infty/F). \tag{4}$$

The action of G on E_{p^∞} defines an injection of G into $\mathrm{Aut}(E_{p^\infty}) \simeq GL_2(\mathbb{Z}_p)$, and, by a theorem of Serre [1], the image of G is open in $GL_2(\mathbb{Z}_p)$. We believe that there are many mysteries remaining to be discovered about the arithmetic of the finite non-abelian extensions of F contained in F_∞, as well as about the behaviour of the Mordell-Weil group and the p-primary part of the Tate-Shafarevich group of E over these extensions. It is striking that what little we can actually prove about these questions at present does indeed involve a judicious blend of theoretical arguments and numerical computations.

First, let me illustrate how little we know about the arithmetic of these division fields themselves. Let E be the elliptic curve over $\mathbb{Q}$ with equation

$$y^2 + y = x^3 - x^2. \tag{5}$$

Classically, this is the elliptic curve $X_1(11)$ of conductor 11 corresponding to the modular group $\Gamma_1(11)$. The point $(0,0)$ on E has order 5, and it is well known that it generates the Mordell-Weil group $E(\mathbb{Q})$. We define

$$L = \mathbb{Q}(E_5). \tag{6}$$

By the Weil pairing, L contains the field $\mathbb{Q}(\mu_5)$ obtained by adjoining the 5-th roots of unity to $\mathbb{Q}$, and, in fact, L is a cyclic extension of degree 5 of $\mathbb{Q}(\mu_5)$. Explicitly (see Fisher [14]) L is the splitting field of the polynomial

$$x^5 + 2x^4 + 6x^3 - 2x^2 + 4x - 1. \tag{7}$$

The primes of $\mathbb{Q}$ which ramify in L are 5 and 11. Let $P(L)$ denote the set of rational primes ℓ which split completely in L. The first few primes in $P(L)$ are $\ell = 101, 151, 941, 991, \cdots$. Moreover, we must have $\ell \equiv 1 \bmod 5$ when ℓ is in $P(L)$, since ℓ must split completely in $\mathbb{Q}(\mu_5)$. However, no simple description of the set $P(L)$ appears to be known at present.

Returning to our general situation of an elliptic curve E defined over F, we study the arithmetic of E over any intermediate field K with $F \subset K \subset F_\infty$ via the Selmer group of E over K. We recall that the Selmer group of E over K is defined by

$$S(E/K) = \mathrm{Ker}(H^1(G(\overline{\mathbb{Q}}/K), E_{p^\infty}) \to \prod_v H^1(G(\overline{K}_v/K_v), E(\overline{K}_v))),$$

where v runs over all finite places of K, and K_v denotes the union of the completions at v of all finite extensions of F contained in K. As usual, we have the exact sequence

$$0 \to E(K) \otimes \mathbb{Q}_p/\mathbb{Z}_p \to S(E/K) \to Ш(E/K)(p) \to 0,$$

where $Ш(E/K)(p)$ denotes the p-primary subgroup of the Tate-Shafarevich group of E over K. We write

$$X(E/K) = \mathrm{Hom}(S(E/K), \mathbb{Q}_p/\mathbb{Z}_p)$$

for the compact Pontrjagin dual of the discrete p-primary modules $S(E/K)$. If K is Galois over F, then the Galois group $G(K/F)$ of K over F has a natural left action on both $S(E/K)$ and $X(E/K)$, and it is easily seen that $X(E/K)$ is always a finitely generated module over the Iwasawa algebra $\Lambda(G(K/F))$. We shall mainly be interested in studying the $\Lambda(G)$-module $X(E/F_\infty)$, and especially the information it encodes about both $E(F_\infty)$ and $Ш(E/F_\infty)(p)$. Let us define

$$F^{\mathrm{cyc}} = F(\mu_{p^\infty}), \tag{8}$$

where μ_{p^∞} denotes the group of all p-power roots of unity. We put

$$H = G(F_\infty/F^{\mathrm{cyc}}), \quad \Gamma = G(F^{\mathrm{cyc}}/F). \tag{9}$$

From now on, we shall always assume that G is pro-p (theoretically, this can always be achieved by replacing F by a finite extension such as $F(E_p)$, but it should be stressed that numerically this is often disastrous, taking one far beyond the limit of fields where calculations are feasible at present), so that Γ is always pro-p, and so isomorphic to $\mathbb{Z}_p$. We shall also assume that $p \geqslant 5$, so that G is automatically p-valued. We remark that every left $\Lambda(G)$-module, which has the property that it is finitely generated over $\Lambda(H)$, must be $\Lambda(G)$-torsion. This is because $\Lambda(G)$ is not finitely generated over $\Lambda(H)$, since $G/H = \Gamma$ is infinite. The following is one of the main results of [15].

Theorem 2. *([15]). Assume that (i) $p \geqslant 5$, (ii) G is pro-p, (iii) E has good ordinary reduction at all places v of F dividing p, and (iv) $X(E/F^{cyc})$ is a finitely generated $\mathbb{Z}_p$-module. Then $X(E/F_\infty)$ is finitely generated as a $\Lambda(H)$-module, where $H = G(F_\infty/F^{cyc})$. In particular, $X(E/F_\infty)$ is a torsion $\Lambda(G)$-module.*

The next result is due to Ochi and Venjakob [16].

Theorem 3. *([16]). Assume that hypotheses (i), (ii), (iii), and (iv) of Theorem 2 are valid. Then $X(E/F_\infty)$ has strictly positive $\Lambda(H)$-rank, and its $\Lambda(H)$-torsion submodule is zero.*

The proof of Theorem 3 uses, in particular, a very pretty characterization of pseudo-null modules amongst all $\Lambda(G)$-modules which are finitely generated over $\Lambda(H)$ (see [16], [11]). If M is a $\Lambda(G)$-module which is finitely generated over $\Lambda(H)$, then M is pseudo-null if and only if it is $\Lambda(H)$-torsion.

While Theorems 2 and 3 can, in principle, be applied to a wide range of elliptic curves and primes p, I am going to limit my discussion in the rest of this lecture to what is probably the first case in nature, namely when $E = X_1(11)$ is given by (5), $p = 5$, and $F = \mathbb{Q}(\mu_5)$. This intriguing numerical example is probably an important test case for the theory in general. I am grateful to Fisher, Greenberg, Hachimori, Howson, Matsuno, Sujatha and Venjakob for many illuminating conversations about this example over the last few years. The structure of the Galois group $G = G(\mathbb{Q}(E_{5^\infty})/\mathbb{Q}(\mu_5))$ was first determined by Lang and Trotter, but it can also be established by a more direct and elementary argument (see [14]). The answer can be expressed by giving a description of the image of G in $\mathrm{Aut}(T_5(E))$, but it is a little more convenient to describe the image of G in $\mathrm{Aut}(T_5(E'))$, where $E' = X_0(11)$ is the elliptic curve

$$y^2 + y = x^3 - x^2 - 10x - 20 \tag{10}$$

(here we have used the standard notation that, for any prime p, $T_p(E) = \varprojlim E_{p^n}$). Now E and E' are isogenous over $\mathbb{Q}$, and so we have $F_\infty = \mathbb{Q}(E_{5^\infty}) = \mathbb{Q}(E'_{5^\infty})$. It is also shown in [14] that there exists a $\mathbb{Z}_5$-basis of $T_5(E')$ such that the image of G and H in $\mathrm{Aut}(T_5(E'))$ are given by

$$G = \mathrm{Ker}(GL_2(\mathbb{Z}_5) \to GL_2(\mathbb{F}_5)), \quad H = \mathrm{Ker}(SL_2(\mathbb{Z}_5) \to SL_2(\mathbb{F}_5)). \tag{11}$$

Moreover, as was first remarked to me by Greenberg, we have $S(E/\mathbb{Q}(\mu_{5^\infty})) = 0$ (see [17] for a detailed proof). Thus the hypotheses (i), (ii), (iii), and (iv) of Theorem 2 hold for E over $F = \mathbb{Q}(\mu_5)$. Thus Theorem 2 and 3 tell us that $X(E/F_\infty)$ is a finitely generated $\Lambda(H)$-module, of positive $\Lambda(H)$-rank, and with its $\Lambda(H)$-torsion submodule zero.

A first step towards elucidating the structure of $X(E/F_\infty)$ as a $\Lambda(G)$-module is given in [15], where the ideas of Hachimori and Matsuno [18] are used to prove the following. From now on, we always assume that

$$E = X_1(11), \quad p = 5, \quad F = \mathbb{Q}(\mu_5).$$

For each finite Galois extension L of F contained in F_∞, we define

$$r(L^{\mathrm{cyc}}) = 4 \cdot [L^{\mathrm{cyc}} : F^{\mathrm{cyc}}] - w(L^{\mathrm{cyc}}), \tag{12}$$

where $w(L^{\mathrm{cyc}})$ denotes the number of primes of L^{cyc} above 11.

Proposition 4. *For each finite Galois extension L of F that is contained in F_∞, $X(E/L^{cyc})$ is a free $\mathbb{Z}_5$-module of rank $r(L^{cyc})$. In particular, $E(L^{cyc})$ is a finitely generated abelian group of rank at most $r(L^{cyc})$.*

Corollary 5. *We have that $X(E/F_\infty)$ is a finitely generated $\Lambda(H)$-module of rank 4, its $\Lambda(H)$-torsion submodule is zero, but it is not a free $\Lambda(H)$-module.*

In our present state of knowledge, I only know how to prove finer results about the $\Lambda(G)$-module $X(E/F_\infty)$ and the arithmetic of E in parts of the tower F_∞ over F by appealing to a brilliant piece of classical numerical descent theory by Fisher [14].

Theorem 6. *([14]). For $L = \mathbb{Q}(E_5)$, both $E(L)$ and Ш$(E/L)(5)$ are finite. More precisely,we have*

$$E(L) = (\mathbb{Z}/5\mathbb{Z})^2, \text{Ш}(E/L)(5) = (\mathbb{Z}/5\mathbb{Z})^2. \tag{13}$$

I want to stress that $L = \mathbb{Q}(E_5)$ has degree 20 over $\mathbb{Q}$, and this numerical work is even more remarkable in that the final outcome can be checked by hand. Now if we apply Proposition 4 to L, we find

$$S(E/L^{\mathrm{cyc}}) = (\mathbb{Q}_5/\mathbb{Z}_5)^{16}. \tag{14}$$

If I might be permitted to confess personal frailty, I knew (14) was true before Theorem 6 was proven, and it led me to erroneously suspect that E had points of infinite order over L. I could not have been more wrong, as the following result shows. Write $\Gamma_L = G(L^{\mathrm{cyc}}/L)$, and fix a topological generator σ_L of Γ_L. As usual, we identify $\Lambda(\Gamma_L)$ with the formal power series ring $\mathbb{Z}_5[[T_L]]$ by mapping σ_L to $1 + T_L$. We define a polynomial in $\mathbb{Z}_5[[T_L]]$ to be *special* if it is non-zero and if all of its roots are of the form $\zeta - 1$, where ζ is some 5-power root of unity.

Theorem 7. *Let $L = \mathbb{Q}(E_5)$. Then, for all $n \geqslant 0$, $E(L(\mu_{5^{n+1}}))$ is finite and $Ш(E/L(\mu_{5^{n+1}}))(5)$ is finite of exact order 5^{16n+2}. Moreover, we have*

$$Ш(E/L^{cyc})(5) = (\mathbb{Q}_5/\mathbb{Z}_5)^{16},$$

and $X(E/L^{cyc})$ is not annihilated by any special polynomial in $\Lambda(\Gamma_L) = \mathbb{Z}[[T_L]]$.

This proof of this result was worked out by Fisher, Greenberg and myself, and uses rather classical arguments in the Iwasawa theory of elliptic curves over cyclotomic fields. We do not have time to give the detailed proof here, but note the following key points. Let $f_E(T_L)$ be the characteristic power series of $X(E/L^{\mathrm{cyc}})$. By virtue of (14), we can assume that $f_E(T_L)$ is a distinguished polynomial of degree 16. In fact $f_E(T_L)$ must be the fourth power of a distinguished polynomial $h_E(T_L)$ of degree 4, because $G(L/\mathbb{Q})$ has an irreducible representation of degree 4. Next, we use Theorem 6 together with a classical Euler characteristic result (see [17]) to conclude that $h_E(0) = 5^3 u$, where u is a unit in $\mathbb{Z}_5$. It follows that $f_E(T_L)$ can have no root in common with a special polynomial, and the proof can now be completed by a standard argument.

Theorem 7 has an interesting consequence for the structure of the $\Lambda(G)$-module $X(E/F_\infty)$. Let C be the centre of G. Of course, C is isomorphic to $1+5\mathbb{Z}_5$, embedded as multiples of the identity matrix in the identifications (11), and $G = C \times H$. Pick some topological generator σ_C of C, and let us identify as usual $\Lambda(C)$ with the formal power series ring $\mathbb{Z}_5[[T_C]]$ by mapping σ_C to $1+T_C$.

Corollary 8. *The $\Lambda(G)$-module $X(E/F_\infty)$ is not annihilated by any special polynomial in $\Lambda(C) = \mathbb{Z}_5[[T_C]]$. In particular, C acts nontrivially on $X(E/F_\infty)$.*

To prove Corollary 8, let K_∞ be the fixed field of C, and put $G_L = G(F_\infty/L)$, $H_L = G(F_\infty/L^{\mathrm{cyc}})$. It is easily seen that L is contained in K_∞, and that we have

$$F_\infty = K_\infty(\mu_{5^\infty}), \quad K_\infty \cap L^{\mathrm{cyc}} = L.$$

Hence C is mapped isomorphically onto $\Gamma_L = G_L/H_L$ under the natural surjection of G_L onto Γ_L. In particular, if Y is any G_L-module on which H_L acts trivially, we can identify the $\Lambda(C)$-action on Y with the $\Lambda(\Gamma_L)$-action, even assuming that we have chosen the topological generators σ_C and σ_L to coincide. We apply this remark to $Y = X(E/F_\infty)_{H_L}$. Now there is a $\Lambda(\Gamma_L)$-homomorphism from this Y to $X(E/L^{\mathrm{cyc}})$, with finite cokernel, which is obtained by dualizing the restriction map from $S(E/L^{\mathrm{cyc}})$ to $S(E/F_\infty)^{H_L}$ (see [15] for the proof that this restriction map has a finite kernel). It follows that the characteristic power series of $X(E/L^{\mathrm{cyc}})$ as a $\Lambda(\Gamma_L)$-module must divide the characteristic power series of Y as a $\Lambda(\Gamma_L)$-module. Now, if $X(E/F_\infty)$ were annihilated by a special polynomial in $\Lambda(C)$, the same polynomial would clearly annihilate Y. But this in turn would then imply that every root of the characteristic power series of $X(E/L^{\mathrm{cyc}})$ would have to be of the form $\zeta - 1$, where ζ is some 5-power root of unity, contradicting Theorem 7. Note that the whole of the theoretical argument proving Theorem 7 and Corollary 8 would break down if we did not know the numerical result in Theorem 6.

Corollary 8, as well as analogy with the results proven in [12], make it natural to pose the following question:
Question: Is $X(E/F_\infty)$ completely faithful as a $\Lambda(G)$-module?

Here is one interesting consequence of an affirmative answer to this question. As above, K_∞ denotes the fixed field of the centre C of G, and we put $\Omega = G/C$.

Lemma 9. *If $X(E/F_\infty)$ is a completely faithful $\Lambda(G)$-module, then $X(E/K_\infty)$ is a torsion $\Lambda(\Omega)$-module.*

Proof. Let σ_C be a topological generator of C, and put

$$W = X(E/F_\infty)/(\sigma_C - 1)X(E/F_\infty).$$

Clearly W is a quotient of $X(E/F_\infty)$ with non-zero global annihilator. Moreover, it is well known (see [8]) that we have

$$\mathrm{Ext}^1_{\Lambda(G)}(W, \Lambda(G)) = \mathrm{Hom}_{\Lambda(\Omega)}(W, \Lambda(\Omega)),$$

so that W is torsion as a $\Lambda(G)$-module if and only if it is pseudo-null as a $\Lambda(G)$-module. Now assume that $X(E/F_\infty)$ is completely faithful, whence W must be pseudo-null as a $\Lambda(G)$-module, and so torsion as a $\Lambda(\Omega)$-module. Moreover, the restriction map from $S(E/K_\infty)$ to $S(E/F_\infty)^C$ has finite kernel because $E^C_{5^\infty}$ is finite. Dualizing, we get a $\Lambda(\Omega)$-homomorphism from W to $X(E/K_\infty)$ with finite cokernel, and so $X(E/K_\infty)$ must also be $\Lambda(\Omega)$-torsion, as required.

For each $n \geqslant 0$, let $M_n = \mathbb{Q}(E'_{5^{n+1}})$, where $E' = X_0(11)$ is given by equation (10). By considering the action of $G(M_n/F)$ on $E'_{5^{n+1}}$, we obtain an isomorphism φ_n from $G(M_n/F)$ onto the kernel of the natural map from $GL_2(\mathbb{Z}/5^{n+1}\mathbb{Z})$ to $GL_2(\mathbb{Z}/5\mathbb{Z})$ (see [14]). Now the kernel of the natural map from $(\mathbb{Z}/5^{n+1}\mathbb{Z})^\times$ to $(\mathbb{Z}/5\mathbb{Z})^\times$ can be viewed as embedded in $G(M_n/F)$ via this isomorphism, and we define K_n to be the fixed field of this kernel. Thus we see that K_n is a Galois extension of F contained in F_∞ with

$$G(K_n/F) \xrightarrow{\sim} \mathrm{Ker}(PGL_2(\mathbb{Z}/5^{n+1}\mathbb{Z}) \to PGL_2(\mathbb{Z}/5\mathbb{Z})).$$

It is clear that $[K_n : F] = 5^{3n}$, and that $K_\infty = \underset{n \geqslant 0}{U} K_n$. Moreover, $[K_n^{\mathrm{cyc}} : F^{\mathrm{cyc}}] = 5^{3n}$ as $K_\infty \cap F^{\mathrm{cyc}} = F$. Finally, it is easily seen using the Tate curve for E' over $\mathbb{Q}_{11}$ that there are precisely 4×5^{2n} primes of K_n^{cyc} above 11. Hence Proposition 4 yields the following result.

Corollary 10. *For all $n \geqslant 0$, the rank of $E(K_n)$ is at most $4 \cdot 5^{3n} - 4 \cdot 5^{2n}$.*

By contrast, if $X(E/F_\infty)$ is a completely faithful $\Lambda(G)$-module, one can easily deduce the following stronger asymptotic bound from Lemma 9.

Proposition 11. *Assume that $X(E/F_\infty)$ is a completely faithful $\Lambda(G)$-module. Then there exists a constant $c > 0$ such that the rank of $E(K_n)$ is at most $c \cdot 5^{2n}$ for all $n \geqslant 0$.*

We end by briefly commenting on what is known about the existence of points of infinite order on $E = X_1(11)$ in $F_\infty = \mathbb{Q}(E_{5^\infty})$. Let E'' be the third elliptic curve of conductor 11 defined over $\mathbb{Q}$, namely

$$y^2 + y = x^3 - x^2 - 7820x - 263580. \tag{15}$$

It is well known that there exist degree 25 isogenies

$$\psi_1 : E \to E'', \quad \psi_2 : E'' \to E.$$

We write J_i for the field generated over F by the coordinates of the points in the kernel of $\psi_i (i = 1, 2)$. In fact, J_1 is the compositum of F with the maximal real subfield $\mathbb{Q}(\mu_{11})^+$ of the field of 11-th roots of unity. I am grateful to Fisher and Matsuno for informing me that J_2 is the compositum of F with the splitting field of the abelian polynomial

$$x^5 - 65x^4 + 205x^3 + 140x^2 + 25x + 1.$$

By further brilliant descent calculations, Fisher has proven that E has no points of infinite order in either of the fields J_1 or $\mathbb{Q}(E''_5) = \mathbb{Q}(\mu_5, \sqrt[5]{11})$, and that the 5-primary component of the Tate-Shafarevich group of E over both these fields is zero. However, I was shocked to learn from Matsuno some months back that he had proven that the complex L-function of E over the abelian field J_2 has a zero at $s = 1$ of exact order 4. Thus, unless the arithmetical universe is to fall apart, there must presumably be a point of infinite order on E over J_2. It presents a very interesting challenge to computational number theory to exhibit this point, thereby proving that it does exist. Finally, we remark that J_2 is not contained in the fixed field K_∞ of the centre of G. We believe that there would be great interest in deciding whether or not the complex L-functions of E over the fields K_n can have a zero at $s = 1$. At present, it is even conceivable that E has no points of infinite order in K_∞.

References

1. Serre, J.-P.: Propriétés Galoisiennes des points d'ordre fini des courbes elliptiques. Invent. Math. 15, 259–331 (1972).
2. Coates, J., Schneider, P., Sujatha, R.: Modules over Iwasawa algebras. To appear.
3. Bourbaki, N.: Algèbre Commutative. Paris, Hermann 1972.
4. Venjakob, O.: Iwasawa Theory of p-adic Lie Extensions. Thesis, Heidelberg University, 2000.
5. Venjakob, O.: On the structure of the Iwasawa algebra of a p-adic Lie group. To appear in the J. European Math. Soc.
6. Björk, J.-E.: Filtered Noetherian Rings. In Noetherian rings and their applications. Math. Survey Monographs 24, pp. 59–97, AMS 1987.
7. Lazard, M.: Groupes analytiques p-adiques. Publ. Math. IHES 26, 380–603 (1965).
8. Howson, S.: Structure of central torsion Iwasawa modules. To appear.
9. Chamarie, M.: Anneaux de Krull non-commutatifs. J. Algebra 72, 210–222 (1981).

10. Chamarie, M.: Modules sur les anneaux de Krull non-commutatifs. In Sém. d'Algèbres P. Dubreil et M.-P. Malliavin 1982, Springer Lecture Notes 1029, pp. 283–310, Springer 1983.
11. Venjakob, O.: The Weierstrass preparation theorem in non-commutative Iwasawa theory. In preparation.
12. Hachimori, U., Venjakob, O.: Completely faithful Selmer groups over Kummer extensions. In preparation.
13. Robson, J.C.: Cyclic and faithful objects in quotient categories with applications to Noetherian simple or Asano rings. In Non-commutative Ring Theory, Springer Lecture Notes 545, pp. 151–172, Springer 1976.
14. Fisher, T.: Descent calculations for the elliptic curves of conductor 11. To appear.
15. Coates, J., Howson, S.: Euler characteristics and elliptic curves II. J. Math. Soc. Japan 53, 175–235 (2001).
16. Ochi, Y., Venjakob, O.: On the structure of Selmer groups over p-adic Lie extensions. To appear in J. Alg. Geom.
17. Coates, J., Sujatha, R.: Galois cohomology of elliptic curves. TIFR-AMS Lecture Notes 88, 2000
18. Hachimori, Y., Matsuno, K.: An analogue of Kida's formula for the Selmer groups of elliptic curves. J. Alg. Geometry 8, 581–601 (1999).

The Weil and Tate Pairings as Building Blocks for Public Key Cryptosystems
(Survey)

Antoine Joux

DCSSI Crypto Lab
51, Bd de Latour Maubourg
F-75700 Paris 07 SP
France
Antoine.Joux@m4x.org

1 Introduction

Elliptic curves were first proposed as a tool for cryptography by V. Miller in 1985 [29]. Indeed, since elliptic curves have a group structure, they nicely fit as a replacement for more traditional groups in discrete logarithm based systems such as Diffie–Hellman or ElGamal. Moreover, since there is no non-generic algorithm for computing discrete logarithms on elliptic curves, it is possible to reach a high security level while using relatively short keys.

However, in [27] Menezes, Okamoto and Vanstone showed that some special elliptic curves, called supersingular curves, are weaker than general elliptic curves. On these special curves, some additional properties allow an attacker to transport the discrete logarithm problem to a finite field where more efficient algorithms are available for discrete logarithm computation. This was a concern, since supersingular elliptic curves were initially considered as good practical choices for elliptic curve systems. As a consequence the issue of choosing curves has been quite debated. According to a talk by Koblitz [22], two different answers can be given. The pragmatic answer is that any curve which has not been proved insecure can be used. This point of view leads to more efficient implementations, since it allows to choose special curves, where computations are faster (one notable example is the use of Koblitz's curves in DSS). On the other hand, the hardliner answer is that all special cases should carefully be avoided, since they might become weak with the next discovery. As a consequence, all curves should be generated randomly or using a strong pseudo-random generator. Following the generation, some additional checks should be performed (such as testing the primality of the number of points on the curve). Moreover, to convince the users of the system that the chosen curve is indeed a regular curve, it is good practice to publish the seed of the pseudo-random generator, in order to allow users to check the generation process by themselves (a similar precaution is used in DSA using SHA-1 as pseudo-random generator).

Several recent papers have shown that the additional properties of weak curves can also be used positively. Indeed, it is possible to base cryptosystems

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 20–32, 2002.

on weak elliptic curves and to turn the additional properties of the curve into additional properties of the systems. Discussing these systems and their security is the main topic of this paper. We start with some preliminaries in section 2, in section 3 we survey the known applications and finally in section 4 we look at the security issues.

2 Definition and Computation of Pairings

Informally speaking a (non degenerate) pairing is a (non constant) bilinear map from a group $\mathbb{G}_1$ to another group $\mathbb{G}_2$. When the context clearly show which groups and pairing are used, we use the notation $\langle P, Q\rangle$ to denote the pairing of two elements P and Q in $\mathbb{G}_1$. A important prerequisite for using a pairing in cryptography is that the discrete logarithm problem in $\mathbb{G}_2$ should be hard. Otherwise, the discrete logarithm problem in $\mathbb{G}_1$ and the inversion of the pairing are also easy; thus, there is no hard problem left on which to base cryptosystems. Of course, this requirement rules out many simple bilinear maps. A nice possibility is to choose $\mathbb{G}_1$ as (a large subgroup of) the group of points of an elliptic curve over $\mathbb{F}_q$. Moreover, the order of $\mathbb{G}_1$ is usually chosen to be a large prime ℓ. Note that when $\ell = q$, there exists an additive pairing that sends $\mathbb{G}_1$ to the additive group $\mathbb{G}_2 = (\mathbb{F}_q, +)$, which does not satisfy our prerequisite. In that case, as was noted in [35] and [37], discrete logarithms in $\mathbb{G}_1$ can be solved in polynomial time.

We now assume that $\ell \neq q$, then $\mathbb{G}_2$ can be chosen as a subgroup of $\mathbb{F}_{q^r}^*$ for some r. The key parameter here is the value of r, which is usually called the security parameter or the security multiplier (see [5]). If r is too large, the pairing can still be defined but it cannot be computed. However, when r is small enough, the pairing can be efficiently computed. To get a lower bound on r, we should remark that ℓ must divide $q^r - 1$. It turns out that this lower bound is in fact the right value for r. As soon as ℓ divides $q^r - 1$, some non degenerate pairing does exist. In fact, two different pairings can be defined with elliptic curve, the Weil pairing and the Tate pairing. The Weil pairing has simpler mathematic properties and has been used for cryptanalytic purposes since [27]. One of its main drawback is the fact that in some cases it does not reach the optimal value for r defined above. On the other hand, the Tate pairing always reaches this optimal value. For this reason, Frey, Müller and Rück proposed in [15] to use it as a replacement for the Weil pairing. Moreover, it is somewhat less costly from a computational point of view (see [3,17]). In cryptographic applications, the terms of Weil and Tate pairings are also used, somewhat abusively, to denote various modified pairings based upon the original ones. One of the most important modification to the pairings was proposed by Verheul in [38]. Without this modification, the Weil and Tate pairings cannot be used, but in a few exceptional cases (see [33]), to pair linearly dependent points and get a non-trivial result (such that $\langle P, P\rangle \neq 1$). With the modification, this becomes possible. For many applications, being able to pair linearly dependent points has a large added value, thus whenever possible we try to use such a modified pairing. The modification proposed by Verheul applies to supersingular curves and works by making use of the additional endomorphisms that exist on such curves. Indeed, using such an endomorphism, called a distorsion by Verheul, it is possible to send points from one subgroup of the ℓ-torsion to another. As a consequence, we can map a pair of linearly dependent points (whose non modified pairing is usually 1) to a pair of linearly independent points. Examples of supersingular curves and distorsions are given in figure 1.

Field	Curve	Distorsion	Conditions	Group order	Sec. param.
$\mathbb{F}_p$	$y^2 = x^3 + ax$	$(x, y) \mapsto (-x, iy)$ $i^2 = -1$	$p \equiv 3[4]$	p+1	2
$\mathbb{F}_p$	$y^2 = x^3 + a$	$(x, y) \mapsto (\zeta x, y)$ $\zeta^3 = 1$	$p \equiv 2[3]$	p+1	2
$\mathbb{F}_{p^2}$	$y^2 = x^3 + a$ $a \notin \mathbb{F}_p$	$(x, y) \mapsto (\omega \frac{x^p}{r^{(2p-1)/3}}, \frac{y^p}{r^{p-1}})$ $r^2 = a, r \in \mathbb{F}_{p^2}$ $\omega^3 = r, \omega \in \mathbb{F}_{p^6}$	$p \equiv 2[3]$	$p^2 - p + 1$	3
$\mathbb{F}_{3^n}$	$y^2 = x^3 + 2x + 1$	$(x, y) \mapsto (-x + r, uy)$ $u^2 = -1, u \in \mathbb{F}_{3^{2n}}$ $r^3 + 2r + 2 = 0, r \in \mathbb{F}_{3^{3n}}$	$n \equiv \pm 1[12]$	$3^n + 3^{\frac{n+1}{2}} + 1$	6
$\mathbb{F}_{3^n}$	$y^2 = x^3 + 2x + 1$	$(x, y) \mapsto (-x + r, uy)$ $u^2 = -1, u \in \mathbb{F}_{3^{2n}}$ $r^3 + 2r + 2 = 0, r \in \mathbb{F}_{3^{3n}}$	$n \equiv \pm 5[12]$	$3^n - 3^{\frac{n+1}{2}} + 1$	6
$\mathbb{F}_{3^n}$	$y^2 = x^3 + 2x - 1$	$(x, y) \mapsto (-x + r, uy)$ $u^2 = -1, u \in \mathbb{F}_{3^{2n}}$ $r^3 + 2r - 2 = 0, r \in \mathbb{F}_{3^{3n}}$	$n \equiv \pm 1[12]$	$3^n - 3^{\frac{n+1}{2}} + 1$	6
$\mathbb{F}_{3^n}$	$y^2 = x^3 + 2x - 1$	$(x, y) \mapsto (-x + r, uy)$ $u^2 = -1, u \in \mathbb{F}_{3^{2n}}$ $r^3 + 2r - 2 = 0, r \in \mathbb{F}_{3^{3n}}$	$n \equiv \pm 5[12]$	$3^n + 3^{\frac{n+1}{2}} + 1$	6

Fig. 1. Some supersingular curves and their distorsions

Another approach is to use non-supersingular curves. In that case, the pairings are still defined, however, in general, the security parameter r is so large that computations in $\mathbb{F}_{q^r}$ cannot be performed. Yet, using complex multiplication techniques, it is possible to construct curves with reasonably small values of r. The main drawback of non-supersingular curves is that, according to Verheul in [38], distorsions do not exist. As a consequence, it is usually more practical to use supersingular curves.

Both the Weil and the Tate pairing can be defined by using the notions of divisors and function fields. Very informally, the function field $K(E)$ of E is the set of rational maps in x and y modulo the equation of E (e.g. $y^2 - x^3 - ax - b$). A divisor D is an element of the free group generated by the points on E, i.e. it can be written as a finite formal sum: $D = \sum_i a_i(P_i)$, where the P_i's are points on E and the a_i's are integers. In the sequel, we will only consider divisors of degree 0, i.e. such that $\sum_i a_i = 0$.

Given any function f in $K(E)$, we can build a degree 0 divisor $div(f)$ from the zeros and poles of f simply by forming the formal sum of the zeros (with multiplicity) minus the formal sum of the poles (with multiplicity). All divisors of the form $D = div(f)$ will be called principal divisors. In the reverse direction, testing whether a degree 0 divisor $D = \sum_i a_i(P_i)$ is principal or not, can be done by evaluating $\sum a_i P_i$ on E. The result will be the point at infinity if and only if D is principal. When working with divisors and functions, the quotient of the group of divisors of degree 0 by the subgroup of principal divisors is a very important group. It contains classes of divisors and the difference of two divisors from the same class is by definition principal.

Given a function f in $K(E)$ and a point P of E, f can be evaluated at P by substituting the coordinates of P for x and y in any rational map representing f. The function f can also be evaluated at a divisor $D = \sum_i a_i(P_i)$, using the following definition:

$$f(D) = \prod_i f(P_i)^{a_i}.$$

The basic step for computing both pairing starts from a pair of ℓ–torsion points P and Q. It computes $f_P(D_Q)$ where f_P denotes a function such that:

$$div(f_P) = \ell(P) - \ell(O),$$

where O denotes the point at infinity and D_Q denotes a divisor from the class $(Q) - (O)$. We know that f_P exists, since the evaluation of $\ell(P) - \ell(O)$ on E is the point at infinity. Some choices for D_Q should be avoided, as they cause failure of the algorithm that computes $f_P(D_Q)$. In particular, $(Q) - (O)$ itself cannot be chosen. Several ways of choosing D_Q are suggested in the literature, the most popular is the method used in [27]. It works by selecting a random point R and by choosing $D_Q = (Q + R) - (R)$, it succeeds except with negligible probability. Another method is proposed in [15] and works by selecting some number k and choosing $D_Q = (kQ) - ((k-1)Q)$.

In order to get an efficient implementation, it is most important to use a technical idea first proposed by Miller [28]. The idea is that trying to write down f_P even in factored form is costly and should be avoided. Instead, all intermediate fractions should be evaluated on D_Q. Following this approach, we get an efficient algorithm for evaluating $f_P(D_Q)$. For details on how to implement this algorithm, the reader can refer to [17] in this volume or alternatively to [3].

Assume that E is an elliptic curve defined over $\mathbb{F}_q$, with $q = p^n$. Assume that ℓ divides $p^{rn} - 1$ for some reasonably small value of r. Then given two ℓ–torsion points P and Q we define their Weil pairing as:

$$w(P, Q) = f_P(D_Q)/f_Q(D_P),$$

and their Tate pairing (as in [15]) as:

$$t(P, Q) = f_P(D_Q)^{\frac{p^{rn}-1}{\ell}}.$$

Throughout the rest of this paper, we arbitrarily fix our choice of pairing and use the Tate pairing.

Whenever a distorsion Ψ, that maps ℓ–torsion points defined over $\mathbb{F}_q$ to ℓ–torsion points defined over the extension field $\mathbb{F}_{q^r}$, is available, we also define a modified pairing. Let P and Q be two ℓ–torsion points defined over $\mathbb{F}_q$, then the modified Tate pairing of P and Q is:

$$\hat{t}(P, Q) = t(P, \Psi(Q)).$$

Some of the applications described in section 3 were defined using the Weil pairing or a modified Weil pairing. However, we can easily replace them by the Tate pairing or its modified version as defined above which are faster.

Remark: In fact, the pairings can be defined not only on elliptic curves, but also on hyperelliptic curves(see [13]) and even in the more general context of abelian varieties. This direction was studied by Galbraith in [16], a recent paper by Rubin and Silverberg [32] gives more precise results. Rubin and Silverberg showed that in this case, it is possible to reach higher values for the security parameter. With abelian varieties, they generalize the definition of the security parameter as being the quotient of the degree r of the extension field by the dimension of the abelian variety involved. With elliptic curves, since the dimension is 1, we get the same value as with the definition of [5]. According to the table included in [32] it is possible to raise the security parameter to 7.5 using varieties of dimension 4 (with elliptic curves the maximum is 6). As they remark, this can be used to improve applications where the security parameter needs to be (moderately) large, such as the short signature scheme of Boneh, Lynn and Shacham(see [5] or section 3 for details).

3 Some Applications

Tripartite Diffie–Hellman. The most basic application of pairings in cryptography is the tripartite Diffie–Hellman protocol proposed in [20]. Originally, this protocol used regular pairings and required, with supersingular curves, the use of two independent points. Using modified pairings as proposed by Verheul in [38], one point suffices. The goal of the protocol is to set up a common key between three users. Without pairings, this can be done by using a conference keying protocol which requires two rounds of interaction as in [7]. With pairings, a single round of interaction is sufficient.

We now describe the protocol, using the modified pairing $\hat{t}$ from section 2. Of course, we assume that the users already know the public parameters, including the supersingular curve E defined over $\mathbb{F}_q$ and some base ℓ–torsion point P. The protocol goes as follow:

- Alice, Bob and Charlie select random integers a, b and c in $[0, \ell - 1]$.
- They respectively broadcast aP, bP and cP.
- They each obtain $\hat{t}(P, P)^{abc}$ by computing one of $\hat{t}(bP, cP)^a$, $\hat{t}(aP, cP)^b$ or $\hat{t}(aP, bP)^c$. This is used as a common secret.

Identity based encryption. Identity based encryption (IBE) is probably the nicest known application of pairings to cryptography. The concept of identity based cryptography was invented in 1984 by Shamir [36]. The basic idea is to use identities (or their images by a public transformation) as public keys and to compute the associated private key using a global secret. In his paper, he explains that concrete proposals are available for identity based signatures and he encourages the reader to look for an IBE scheme. Later on, it became clear that identity based authentication protocols and identity based signatures are quite easy to built. However, identity based encryption was found to be a much harder problem. Until recently, the only known solution was based on a paper of Maurer and Yacobi [26], and used the discrete logarithm problem modulo a RSA number. The global secret was the factorization of the RSA number, whose knowledge made discrete logarithm computations possible. Still, computing the private keys remained too costly for the system to be really practical. Quite recently, two new solutions were proposed. One of them, proposed by Cocks [14], is based on a classical cryptographic problem: deciding whether or not a number is a square in the ring defined by some RSA modulus N. This solution has a slight drawback in term of bandwidth. Indeed, for each bit of the plaintext, two numbers of the size of N are sent. The other solution is based on supersingular curves and pairing. It was proposed by Boneh and Franklin at Crypto'2001 [4].

The first requirement of this IBE scheme is a deterministic algorithm that sends an arbitrary string ID (the identity of a user) to a point Q_{ID} on the elliptic curve used by the system. This is done in two steps, by sending ID to an element of $\mathbb{F}_q$ using a cryptographic hash function G and then by finding a point. A first approach would be to send ID to the x coordinates of a point and try to find a corresponding y. However, about half of the x values do not have a corresponding y and about half of the x lead to two possible choices of y. While this could be solved by iterating the hash function until reaching a possible value for x and by choosing a rule for selecting y, it would be a cumbersome solution. In [4], a much nicer solution is proposed, using the supersingular curve $y^2 = x^3 + 1$ in $\mathbb{F}_p$, with p and ℓ primes such that $p + 1 = 6\ell$. In that case, $p \equiv 2 \pmod 3$, 3 has a multiplicative inverse modulo $p - 1$ and all elements of $\mathbb{F}_p$ have a unique cube-root. As a consequence, it is possible to find a point on the curve by first selecting its y coordinate as $y_0 = G(\mathrm{ID})$ and by computing the x coordinate x_0 as the cube root of $y^2 - 1$. After multiplying this point (x_0, y_0) by 6, we get a point Q_{ID} of order ℓ, unless (x_0, y_0) has order 6, which happens with negligible probability. The curve $y^2 = x^3 + 1$ has an extra endomorphism Ψ (see figure 1) and we can use the modified Tate pairing $\hat{t}$.

In order to compute the private key associated with a public key Q_{ID}, the key generation authority publishes as part of the system parameters two ℓ–torsion points P_{pub} and Q_{pub}. These points satisfy the relation $P_{\mathrm{pub}} = sQ_{\mathrm{pub}}$, where s is the global secret of the system. Knowing s, the private key of a user identified by the string ID is $P_{\mathrm{ID}} = sQ_{\mathrm{ID}}$.

The encryption function is very similar to ElGamal. To encrypt a (short) message M for the user identified by the string ID, perform the following steps:

1. Compute Q_{ID}
2. Select a random r in $[0; \ell - 1]$
3. Form the ciphertext pair: $(rQ_{\mathrm{pub}}, M \oplus H(\hat{t}(Q_{\mathrm{ID}}, rP_{\mathrm{pub}})))$. Here H is a cryptographic hash function.

Decryption is easy thanks to the identity:

$$\hat{t}(P_{\mathrm{ID}}, rQ_{\mathrm{pub}}) = \hat{t}(Q_{\mathrm{ID}}, rQ_{\mathrm{pub}})^s = \hat{t}(Q_{\mathrm{ID}}, rP_{\mathrm{pub}}).$$

A very interesting open problem would be to generalize IBE to allow the central authority to delegate key derivation to sub-authorities for limited subdomains. This idea of Hierarchical IBE is examined in [19] and a partial solution is proposed that works when collusion between the sub-authorities to break the scheme is limited.

Remark: In identity based encryption, the key generation authority implicitly get the capability of an escrow agent. This shows that escrowed encryption protocols arise quite naturally when using pairings. This capability of pairing based cryptography was first described in [38].

Pairings and signatures. While identity based signatures and identification protocols do not require the use of pairings, they can also be implemented by using pairings (see [8,18,31]). However, in the field of digital signatures, there are applications where the use of pairings has a much larger impact. An early proposal was described by Brands in his thesis on electronic cash [6] in 1993, without mentioning pairing. At that time, Brands did remark that in groups where DDH is easy, Chaum's undeniable signatures become regular signatures. Chaum's undeniable signatures are introduced in [11,9] and work as follow. Each user has a public/private key pair composed of the two numbers g^x and x, where g generates a group of prime order (originally this was chosen as a multiplicative subgroup of a finite field). Then any element m of the group can be signed, its signature is $s = m^x$. However, knowing the public key is not sufficient to test the signature. Instead, two zero-knowledge protocols are provided that respectively allow the signer to prove or disprove the validity of any signature. In some applications, the fact that a signature cannot be tested without knowledge and cooperation of the signer is useful to prevent uncontrolled dissemination of the signed document. However, in the context of electronic cash, Brands remarked that it would be useful if Chaum's undeniable signatures were regular signatures. Moreover, he stated that this could be done if the decision Diffie–Hellman problem were easy. Clearly, replacing the multiplicative group in the above protocol by a supersingular elliptic curve, would lead to a working solution.

Of course, in order for the above signature to be secure, the computational Diffie–Hellman problem should still be hard. Since the decision Diffie–Hellman problem is easy, these two problems should separate. This idea that separating related problems can lead to interesting applications in cryptography was formalized by Okamoto and Pointcheval [30]. They call such a separation a gap problem.

In fact, Chaum's undeniable signatures can be merged with gap problems to become regular signature schemes. Furthermore, as explained by Boneh, Lynn and Shacham in [5], if the Weil (or Tate) pairing is used to create a gap between decision Diffie–Hellman and computational Diffie-Hellman, we obtain short signatures. Their scheme works using a fixed elliptic curve defined over $\mathbb{F}_q$, with a pairing of security parameter r (i.e. such that the pairing takes values in $\mathbb{F}_{q^r}$). Let P and Q be two points on the curve such that $\langle P, Q \rangle \neq 1$ (when using a modified pairing, one can take $P = Q$). Each user has a private key s and a public key $s \cdot Q$. To sign a message M, the user sends it to a point $h(M)$ on the subgroup of the curve generated by P using a cryptographic hash function h. Then, he computes $S = s \cdot h(M)$. To verify a signature, it suffices to check that $\langle h(M), s \cdot Q \rangle = \langle S, Q \rangle$. Since a point on an elliptic curve can be represented by its x coordinate and a bit (e.g. the sign) of its y coordinate, the size of a signature is comparable with $\log_2(q)$.

In the case of short signatures, the nice shortcut for hashing to a point on the curve that was used in IBE is no longer available. Indeed, since we want a quite large security parameter, we need to use a different kind of curves for this application. It is explained in [5] that q should be a 160 bits number to prevent the computation of discrete logarithm on the elliptic curve through generic algorithm. Moreover, the size of finite field $\mathbb{F}_{q^r}$ should be of approximately 1024 bits to avoid index calculus computations of discrete logarithms in that field. Thus a security parameter of 6 seems a good compromise. Originally, the authors of [5] proposed to use some supersingular curves over $\mathbb{F}_{3^n}$, since a security parameter of 6 can be reached with supersingular elliptic curves in characteristic 3 only. With these curves, the efficiency of hashing to a point can be somewhat improved as was shown in [2]. However, due to the properties of index calculus algorithms discrete logarithm computations in small characteristic can be performed much more efficiently than in large characteristic (see [1,34,21]). As a consequence, using fields of small characteristic might be less secure than fields of large characteristic. This worry is expressed in the updated version of [5], where a solution is proposed. Since supersingular curves in large characteristic cannot have security parameter 6, Boneh, Lynn and Shacham propose to use complex multiplication in order to construct curves of cardinality $l^2 - l + 1$ over $\mathbb{F}_q$, with $q = l^2 + 1$ prime. Assuming that ℓ is a large prime dividing $l^2 - l + 1$, the ℓ–roots of unity can be embedded in $\mathbb{F}_{q^6}$ (and not in smaller extensions). However, in that case, we have seen in section 2 that according to [38] no distorsion exists. Thus, we need to work with two subgroups generated by linearly independent points. One of the points can be chosen in $\mathbb{F}_q$, while the other is defined over the extension field. As a consequence, if the signatures are to remain short, the public keys need to be chosen as (long) elements of $\mathbb{F}_{q^6}$. Since short signatures are especially useful when storage is limited, storing the public keys might become the limiting factor in some applications. A nice open problem would be to construct a short signature scheme using pairings in large characteristic that would somehow overcome this limitation.

Remark: Another variant of Chaum's signature proposed by Chaum and Pedersen in [10] can also be improved by using pairings. An application to self-blindable credentials was described by Verheul in [39].

4 Security Issues in Pairing Based Systems

When using standard cryptographic groups in discrete logarithm based cryptosystem, it is well known that the security relies on one of the three following assumptions: the hardness of the discrete logarithm problem (DL), of the computational Diffie–Hellman problem (CDH) or of the decision Diffie–Hellman problem (DDH). When dealing with cryptographic groups that admit pairings, one cannot use the same set of basic problems. Indeed, the existence of a pairing implies that DDH becomes easy. In this section, we describe the many problems which can be used when using the modified Tate pairing (i.e. when $\mathbb{G}_1$ can be generated by a single point P, such that $\langle P, P\rangle \neq 1$). In [4], Boneh and Franklin introduced a new assumption: the hardness of the Weil Diffie–Hellman (WDH) problem. Similarly, one can define the (modified) Tate Diffie–Hellman (TDH) problem as follows:

- Given (P, aP, bP, cP) for random a, b, c compute $\hat{t}(P, P)^{abc}$.

As noted in [4], the TDH assumption implies that CDH is hard in the group of points $\mathbb{G}_1$, it also implies that CDH is hard in $\mathbb{G}_2$ where pairings are taking their values. The security of the IBE scheme from [4] is based on TDH in the random oracle model, thanks to the use of the function H. When H is not used, as in the tripartite Diffie–Hellman protocol described in section 3, we need to assume the hardness of the decision problem associated with TDH, that we call DTDH. DTDH is defined as follows:

- Given (P, aP, bP, cP) a quadruplet of elements from $\mathbb{G}_1$ and $\hat{t}(P, P)^d$ an element of $\mathbb{G}_2$ for random a, b, c and d, decide whether $d = abc$.

The DTDH assumption implies DDH in $\mathbb{G}_2$ and CDH in $\mathbb{G}_1$ (remember that DDH in $\mathbb{G}_1$ is easy). The first implication can be shown by remarking that when DDH is easy in $\mathbb{G}_2$ then DTDH is also easy. Indeed, $d = abc$ if and only if $(\langle P, P\rangle, \langle aP, bP\rangle, \langle P, cP\rangle, \langle P, P\rangle^d)$ is a valid decision Diffie–Hellman instance.

Further, other related problems can be introduced to get a deeper understanding of the security of pairing based systems. Before introducing these problems, let us digress and ask the following question which arises quite naturally when looking at pairings: Can they be used as cryptanalytic tools to solve DDH in more general groups ? Indeed, if we could find a group morphism from a third group $\mathbb{G}_3$ to (one of the many possible) $\mathbb{G}_1$, deciding DDH in $\mathbb{G}_3$ would become easy. This would become extremely interesting if $\mathbb{G}_3$ could be chosen as the multiplicative subgroup of order ℓ of $\mathbb{F}_{q^r}$. Indeed, this would give a partial solution to solve DDH in finite field and would have a wide impact on many cryptographic schemes. Such an "attack" was recently proposed in [12]. It requires the construction of a special auxiliary curve, whose existence is conjectured by the authors of [12]. A recent preprint by Koblitz and Menezes [23] shows that the approach of [12] is very probably flawed, since the existence of the required auxiliary curve is unlikely. However, one might wonder about variants of this attack. In fact, we can get strong evidence against the existence of these attacks by generalizing a result of Verheul in [38] and showing that any attack of this kind would also lead to an efficient algorithm against the computational (and not only decision) Diffie–Hellman in the group $\mathbb{G}_3$. The result of Verheul was proved in the special case of the multiplicative subgroup of order $p^2 - p + 1$ in $\mathbb{F}_{p^6}$, which is sometimes called the XTR subgroup due to its relation to the XTR public key cryptosystem [24].

First of all, let us describe more precisely how the DDH attack could work. As explained in [12] and [23], we need a group morphism ϕ from $\mathbb{G}_3$ (the multiplicative subgroup of order ℓ in $\mathbb{F}_{q^r}$) to $\mathbb{G}_1$ (an additive subgroup of order ℓ of an elliptic curve defined over $\mathbb{F}_{q^r}$). We also need to consider the modified Tate pairing $\hat{t}(\cdot,\cdot)$ that solves the DDH in $\mathbb{G}_1$ by mapping pairs of points to ℓ-th roots of unity (i.e. back to $\mathbb{G}_3$). Given g, g^a, g^b and g^c in $\mathbb{G}_3$, testing whether $c = ab$ can be done as in [12] by computing $\hat{t}(\phi(g), \phi(g^c))$ and $\hat{t}(\phi(g^a), \phi(g^b))$ and testing equality. As long as ϕ is non constant and $\hat{t}$ non degenerate, we get an efficient way of testing DDH. However, given ϕ and $\hat{t}$ we can in fact do much more. Indeed, if g is a generator of the ℓ-th root of unity, then $\hat{t}(\phi(g), \phi(g))$ can be written as g^λ. Moreover, because of the non degeneracy properties, $\hat{t}(\phi(g), \phi(g)) \neq 1$ and thus $\lambda \neq 0$. Thanks to the bilinearity of $\hat{t}$, we can now check that $\hat{t}(\phi(g^a), \phi(g^b)) = g^{\lambda ab}$. If we could get rid of the constant λ then we would be computing CDH. For the sake of simplicity, assume that q is prime. In that case,

$$\lambda^{q-3} \equiv \lambda^{-2} \pmod{q}.$$

Moreover, thanks to the relation

$$\hat{t}(\phi(g^{\lambda^i}), \phi(g^{\lambda^j})) = g^{\lambda^{i+j+1}},$$

it is easy by using addition chains to compute $\Lambda = g^{\lambda^{q-3}} = g^{\lambda^{-2}}$. Remarking that $\hat{t}(\phi(g^{\lambda ab}), \phi(\Lambda)) = g^{ab}$ we can now solve the CDH problem in $\mathbb{G}_3 = \mathbb{G}_1$ (and also in $\mathbb{G}_2$) with two applications of the pairing $\hat{t}$.

As a consequence of this digression, we can now remark that the hardness of the Tate Diffie–Hellman problem implies that the Tate pairing is hard to invert when one side of the pairing is fixed. More precisely, it is hard to find a point R in $\mathbb{G}_1$ and a morphism ϕ from $\mathbb{G}_2$ to $\mathbb{G}_1$ such that for all g in $\mathbb{G}_2$:

$$\hat{t}(R, \phi(g)) = g.$$

We call this problem the fixed Tate inversion (FTI). Clearly assuming the hardness of the TDH problem is a stronger hypothesis than assuming the hardness of the FTI problem. An open question is to find an interesting pairing-based system whose hardness relies on FTI. However, it seems to be a difficult problem. A more promising approach would be to consider the problem of finding any pair of points (S, T) such that $\langle S, T \rangle = g$. Due to bilinearity of the pairing, this generalized Tate inversion (GTI) has rich self-randomization properties. As a consequence, it might be easier and worthwhile to devise cryptographic protocols above this problem.

Another relation is worth noting: the hardness of the discrete logarithm in $\mathbb{G}_2$ implies the hardness of either GTI or the discrete logarithm in $\mathbb{G}_1$. Indeed, when both GTI and discrete logarithm in $\mathbb{G}_1$ are easy, it is possible to compute discrete logarithm in $\mathbb{G}_2$. Assume that $g = \langle P, P \rangle$ and h are two elements of $\mathbb{G}_2$. In order to find α such that $h = g^{\alpha}$, we first use GTI and find two points Q and R such that $\langle Q, R \rangle = h$. Using discrete logarithm computations in $\mathbb{G}_1$, we find a and b such that $Q = aP$ and $R = bP$. Then $h = \langle aP, bP \rangle = g^{ab}$ and $\alpha = ab$.

We summarize the relations between all the complexity assumption in figure 2. Each arrow in the figure goes from a complexity assumption to a weaker one. The figure does not include the conditional and non-uniform equivalences between DL and CDH in a group that come from [25]. These equivalences hold when an auxiliary curve defined over $\mathbb{F}_\ell$ and of sufficiently smooth order is known. Note that in our case, $\mathbb{G}_1$ and $\mathbb{G}_2$ have the same cardinality ℓ and that the same auxiliary curve can serve this purpose for both groups.

$$\begin{tikzcd}
 & & \mathrm{CDH}_{\mathbb{G}_1} \arrow[rr] \arrow[dr] & & \mathrm{DL}_{\mathbb{G}_1} \arrow[d] & \\
\mathrm{DTDH} \arrow[r] \arrow[dr] & \mathrm{TDH} \arrow[ur] \arrow[dr] & \mathrm{GTI} \arrow[r] & \mathrm{FTI} \arrow[r] & \mathrm{DL}_{\mathbb{G}_2} \arrow[r, leftrightarrow] & \mathrm{DL}_{\mathbb{G}_1}\ \mathrm{or\,GTI} \\
 & \mathrm{DDH}_{\mathbb{G}_2} \arrow[r] & \mathrm{CDH}_{\mathbb{G}_2} \arrow[ur] & & &
\end{tikzcd}$$

Fig. 2. Relations between complexity assumptions in pairing cryptography

5 Conclusion

Since its introduction in [20], pairing based cryptography has become a rich area of cryptography. The key discovery that motivated most of the work in the domain is probably the identity based encryption scheme of Boneh and Franklin [4]. At this point in time, a lot of research is still underway on the topic of using pairings in cryptography. As a consequence, we can hope and expect that many more applications are forthcoming in the months and years to come.

References

1. L. M. Adleman and M. A. Huang. Function field sieve method for discrete logarithms over finite fields. In *Information and Computation*, volume 151, pages 5–16. Academic Press, 1999.
2. P. Barreto and H. Kim. Fast hashing onto elliptic curves of fields of characteristic 3. Cryptology eprint Archives `http://eprint.iacr.org`, 2001. Number 2001/096.

3. P. Barreto, H. Kim, B. Lynn, and M. Scott. Efficient algorithms for pairing-based cryptosystems. Cryptology eprint Archives `http://eprint.iacr.org`, 2002. Number 2002/008.
4. D. Boneh and M. Franklin. Identity-based encryption from the Weil pairing. In J. Kilian, editor, *Proceedings of CRYPTO'2001*, volume 2139 of *Lecture Notes in Comput. Sci.*, pages 213–229. Springer, 2001.
5. D. Boneh, B. Lynn, and H. Shacham. Short signatures from the Weil pairing. In C. Boyd, editor, *Proceedings of ASIACRYPT'2001*, volume 2248 of *Lecture Notes in Comput. Sci.*, pages 514–532. Springer, 2001. Updated version available from the authors.
6. S. Brands. An efficient off–line electronic cash system based on the representation problem. Technical Report CS–R9323, CWI, Amsterdam, 1993.
7. M. Burmester and Y. Desmedt. A secure and efficient conference key distribution system. In A. De Santis, editor, *Advances in Cryptology — EUROCRYPT'94*, volume 950 of *Lecture Notes in Comput. Sci.*, pages 275–286. Springer, 1995.
8. J. C. Cha and J. H. Cheon. An identity-based signature from gap Diffie–Hellman groups. Cryptology eprint Archives `http://eprint.iacr.org`, 2002. Number 2002/018.
9. D. Chaum. Zero-knowledge undeniable signatures (extended abstract). In Ivan B. Damgård, editor, *Advances in Cryptology - EuroCrypt '90*, volume 473 of *Lecture Notes in Comput. Sci.*, pages 458–464, Berlin, 1990. Springer-Verlag.
10. D. Chaum and T. P. Pedersen. Wallet databases with observers. In Ernest F. Brickell, editor, *Advances in Cryptology - Crypto '92*, volume 740 of *Lecture Notes in Comput. Sci.*, pages 89–105, Berlin, 1992. Springer-Verlag.
11. D. Chaum and H. van Antwerpen. Undeniable signatures. In Gilles Brassard, editor, *Advances in Cryptology - Crypto '89*, volume 435 of *Lecture Notes in Comput. Sci.*, pages 212–217, Berlin, 1989. Springer-Verlag.
12. Q. Cheng and S. Uchiyama. Nonuniform polynomial time algorithm to solve decisional Diffie–Hellman problem in finite fields under conjecture. In *CR-RSA 2002*, number 2271 in Lecture Notes in Comput. Sci., pages 290–299. Springer, 2002.
13. Y. Choie, E. Jeong, and E. Lee. Supersingular hyperelliptic curve of genus 2 over finite fields. Cryptology eprint Archives `http://eprint.iacr.org`, 2002. Number 2002/032.
14. C. Cocks. An identity based encryption scheme based on quadratic residues. *Cryptography and Coding*, 2001. To appear, preprint available at `http://www.cesg.-gov.uk/technology/id-pkc/media/ciren.pdf`.
15. G. Frey, M. Müller, and H.-G. Rück. The Tate pairing and the discrete logarithm applied to elliptic curve cryptosystems. *IEEE Transactions on Information Theory*, 45(5):1717–1718, 1999.
16. S. D. Galbraith. Supersingular curves in cryptography. In C. Boyd, editor, *Proceedings of ASIACRYPT'2001*, volume 2248 of *Lecture Notes in Comput. Sci.*, pages 495–513. Springer, 2001.
17. S. D. Galbraith, K. Harrison, and D. Soldera. Implementing the Tate pairing. In *This Volume*, 2002.
18. F. Hess. Exponent groups signature schemes and efficient identity based signature schemes based on pairings. Cryptology eprint Archives `http://eprint.iacr.org`, 2002. Number 2002/012.
19. J. Horwitz and B. Lynn. Toward hierarchical identity-based encryption. To appear at Eurocrypt 2002., May 2002.

20. A. Joux. A one round protocol for tripartite Diffie–Hellman. In Wieb Bosma, editor, *Proceedings of the ANTS-IV conference*, volume 1838 of *Lecture Notes in Comput. Sci.*, pages 385–394. Springer, 2000.
21. A. Joux and L. Lercier. The function field sieve is quite special. In *This Volume*, 2002.
22. N. Koblitz. Elliptic curve cryptography: Which curves to use? Transparencies available at `http://www.ipam.ucla.edu/publications/cry2002/cry2002_nkoblitz.pdf`, January 2002. Talk given at the IPAM Cryptography Workshop.
23. N. Koblitz and A. Menezes. Obstacles to the torsion-subgroup attack on the decision Diffie–Hellman problem. Technical Report CORR 2002-05, CACR, 2002. Available at `http://www.cacr.math.uwaterloo.ca/tech_reports.html`.
24. A. Lentra and E. Verheul. The XTR public key system. In Mihir Bellare, editor, *Proceedings of CRYPTO'2000*, volume 1880 of *Lecture Notes in Comput. Sci.*, pages 1–19. Springer, 2000.
25. U. Maurer and S. Wolf. The relationship between breaking the Diffie–Hellman protocol and computing discrete logarithms. *SIAM J. Comput.*, 28(5):1689–1721, 1999.
26. U. M. Maurer and Y. Yacobi. Non-interative public-key cryptography. In Donald W. Davies, editor, *Advances in Cryptology - EuroCrypt '91*, volume 547 of *Lecture Notes in Comput. Sci.*, pages 498–507, Berlin, 1991. Springer-Verlag.
27. A. Menezes, T. Okamoto, and S. Vanstone. Reducing elliptic curve logarithms to logarithms in a finite field. *IEEE Transaction on Information Theory*, 39:1639–1646, 1993.
28. V. Miller. Short programs for functions on curves. Unpublished manuscript, 1986.
29. V. Miller. Use of elliptic curves in cryptography. In H. Williams, editor, *Advances in Cryptology — CRYPTO'85*, volume 218 of *Lecture Notes in Comput. Sci.*, pages 417–428. Springer, 1986.
30. T. Okamoto and D. Pointcheval. The gap problems: a new class of problems for the security of cryptographic primitives. In *Public Key Cryptography, PKC 2001*, volume 1992 of *Lecture Notes in Comput. Sci.*, pages 104–118. Springer, 2001.
31. K. Paterson. ID–based signatures from pairings on elliptic curves. Cryptology eprint Archives `http://eprint.iacr.org`, 2002. Number 2002/004.
32. K. Rubin and A. Silverberg. The best and worst of supersingular abelian varieties in cryptology. Cryptology eprint Archives `http://eprint.iacr.org`, 2002. Number 2002/006.
33. H. G. Rück and K. Nguyen. A comparison of the Weil and Tate pairing. preprint.
34. O. Schirokauer. The special function field sieve. Preprint.
35. I. Semaev. Evaluation of discrete logarithms in a group of p-torsion points of an elliptic curve in characteristic p. *Mathematics of Computation*, 67:353–356, 1998.
36. A. Shamir. Identity-based cryptosystems and signature schemes. In G. R. Blakley and David Chaum, editors, *Advances in Cryptology: Proceedings of Crypto '84*, volume 196 of *Lecture Notes in Comput. Sci.*, pages 47–53, Berlin, 1985. Springer-Verlag.
37. N. Smart. The discrete logarithm problem on elliptic curves of trace one. *Journal of Cryptology*, 12(3):193–196, 1999.
38. E. Verheul. Evidence that XTR is more secure than supersingular elliptic curve cryptosystems. In B. Pfizmann, editor, *Proceedings of EUROCRYPT'2001*, volume 2045 of *Lecture Notes in Comput. Sci.*, pages 195–210. Springer, 2001.
39. E. Verheul. Self-blindable credential certificates from the Weil pairing. In C. Boyd, editor, *Proceedings of ASIACRYPT'2001*, volume 2248 of *Lecture Notes in Comput. Sci.*, pages 533–551. Springer, 2001.

Using Elliptic Curves of Rank One towards the Undecidability of Hilbert's Tenth Problem over Rings of Algebraic Integers

Bjorn Poonen*

Department of Mathematics, University of California, Berkeley, CA 94720-3840, USA,
poonen@math.berkeley.edu

Abstract. Let $F \subseteq K$ be number fields, and let $\mathcal{O}_F$ and $\mathcal{O}_K$ be their rings of integers. If there exists an elliptic curve E over F such that $\mathrm{rk}, E(F) = \mathrm{rk}, E(K) = 1$, then there exists a diophantine definition of $\mathcal{O}_F$ over $\mathcal{O}_K$.

1 Introduction

D. Hilbert asked, as Problem 10 of his famous list of 23 problems posed to the mathematical community in 1900, for an algorithm to decide, given a polynomial equation $f(x_1, \dots, x_n) = 0$ with coefficients in the ring $\mathbf{Z}$ of integers, whether there exists a solution with $x_1, \dots, x_n \in \mathbf{Z}$. In Hilbert's time, there was no formal definition of algorithm, but presumably what he had in mind was a mechanical procedure that a human could in principle carry out, given sufficient paper, pencils, erasers, and time, following a set of strict rules requiring no insight or ingenuity on the part of the human. In the 1930s, several rigorous models of computation were proposed as a substitute for the informal notion of "mechanical procedure" as above (the λ-definable functions of A. Church and S. Kleene, the recursive functions of K. Gödel and J. Herbrand, and the logical computing machines of A. Turing). These models, as well as others developed later, were shown to be equivalent; this gave credence to the *Church-Turing thesis*, which is the belief that every mechanical procedure can be carried out by a Turing machine. Therefore, the modern interpretation of Hilbert's Tenth Problem is that it asks whether a Turing machine can decide the existence of solutions.

J. Matijasevič [Mat70], building on earlier work by M. Davis, H. Putnam, and J. Robinson [DPR61] showed that there is no such Turing machine. To describe their work in more detail, we need a few definitions. A subset S of $\mathbf{Z}^n$ is called *listable* or *recursively enumerable* if there is an algorithm (Turing machine) such that S is exactly the set of $a \in \mathbf{Z}^n$ that are eventually printed by the algorithm. A subset S of $\mathbf{Z}^n$ is said to be *diophantine*, or to admit a *diophantine definition*, if there is a polynomial $p(a_1, \dots, a_n, x_1, \dots, x_m) \in \mathbf{Z}[a_1, \dots, a_n, x_1, \dots, x_m]$ such that

$$S = \{\, a \in \mathbf{Z}^n : (\exists x_1, \dots, x_m \in \mathbf{Z})\; p(a_1, \dots, a_n, x_1, \dots, x_m) = 0 \,\}.$$

* This research was supported by NSF grant DMS-9801104, and a Packard Fellowship.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 33–42, 2002.

For example, the subset $\mathbf{Z}_{\geq 0} := \{0, 1, 2, \dots\}$ of $\mathbf{Z}$ is diophantine, since for $a \in \mathbf{Z}$, we have

$$a \in \mathbf{Z}_{\geq 0} \iff (\exists x_1, x_2, x_3, x_4 \in \mathbf{Z})\ x_1^2 + x_2^2 + x_3^2 + x_4^2 = a.$$

One can show using "diagonal arguments" that there exists a listable subset L of $\mathbf{Z}$ whose complement is not listable. It follows that for this L, there is no algorithm that takes as input an integer a and decides in a finite amount of time whether a belongs to L; in other words, membership in L is undecidable.

Diophantine subsets of $\mathbf{Z}^n$ are listable: given p, one can write a computer program with an outer loop with B running through $1, 2, \dots$, and an inner loop in which one tests the finitely many $(a_1, \dots, a_n, x_1, \dots, x_m) \in \mathbf{Z}^{n+m}$ satisfying $|a_i|, |x_j| \leq B$ for all i and j, and prints $(a_1, \dots, a_n)$ if $p(a_1, \dots, a_n, x_1, \dots, x_m) = 0$. Davis [Dav53] conjectured conversely that all listable subsets of $\mathbf{Z}^n$ were diophantine, and this is what Matijasevič eventually proved. In particular, the set L is diophantine. Hence a positive answer to Hilbert's Tenth Problem would imply that membership in L is decidable. But membership in L is undecidable, so Hilbert's Tenth Problem is undecidable too; that is, there is no algorithm that takes as input a polynomial $p \in \mathbf{Z}[x_1, \dots, x_n]$, and decides whether $p(x_1, \dots, x_n) = 0$ has a solution in integers.

More generally, if R is any commutative ring with 1, one can define what it means for a subset of R^n to be *diophantine over* R, by replacing $\mathbf{Z}$ by R everywhere. Similarly one can speak of *Hilbert's Tenth Problem over* R provided that one has fixed some encoding of elements of R as finite strings of symbols from a finite alphabet, so that polynomials over R can be the input to a Turing machine. For some rings R (for example, uncountable rings) such an encoding may not be possible. In this case one should modify the problem, by specifying a countable subset $\mathcal{P}$ of the set of all polynomials over R and an encoding of elements of $\mathcal{P}$ as finite strings of symbols, and then asking whether there exists a Turing machine that takes as input a polynomial $p \in \mathcal{P}$ and decides whether $p(x_1, \dots, x_n) = 0$ has a solution over R. For example, K. Kim and F. Roush [KR92] proved that Hilbert's Tenth Problem over the purely transcendental function field $\mathbf{C}(t_1, t_2)$ is undecidable when one takes $\mathcal{P}$ to be the set of polynomials with coefficients in $\mathbf{Z}[t_1, t_2]$. Usually it is not necessary to specify exactly how the elements of $\mathcal{P}$ are encoded, since usually given any two reasonable encodings, a Turing machine can convert between the two.

Perhaps the most important unsolved question in this area is Hilbert's Tenth Problem over the field $\mathbf{Q}$ of rational numbers. The majority view seems to be that it should be undecidable. To prove this, it would suffice to show that the subset $\mathbf{Z}$ of $\mathbf{Q}$ is diophantine over $\mathbf{Q}$. On the other hand, B. Mazur has suggested that perhaps for any variety X over $\mathbf{Q}$, the topological closure of $X(\mathbf{Q})$ in $X(\mathbf{R})$ has at most finitely many connected components; if this is true, no such diophantine definition of $\mathbf{Z}$ over $\mathbf{Q}$ exists. See [Maz94] and the more recent articles [CZ00] and [Phe00] for further discussion.

The function field analogue, namely Hilbert's Tenth Problem over the function field k of a curve over a finite field, is known to be undecidable. The first result of this type is due to T. Pheidas [Phe91], who proved this for $k = \mathbf{F}_q(t)$ with q odd. His argument was adapted and generalized by C. Videla [Vid94] for $k = \mathbf{F}_q(t)$ with q even, by A. Shlapentokh [Shl92] for other function fields of odd characteristic, and finally by K. Eisenträger [Eis] for the remaining function fields of characteristic 2. Analogues are known also for many function fields over infinite fields of positive characteristic: see [Shl00a] and [Eis].

For more results concerning Hilbert's Tenth Problem, see the book [DL+00], and especially the survey articles [PZ00] and [Shl00b] therein. Since the publication of that book, undecidability of Hilbert's Tenth Problem has been proved also for function fields of curves X over formally real fields k_0 with $X(k_0)$ nonempty [MB] (in fact this is just one application of his results), and for function fields of surfaces over real closed or algebraically closed fields of characteristic zero [Eis].

2 Hilbert's Tenth Problem over Rings of Integers

In this article, our goal is to prove a result towards Hilbert's Tenth Problem over rings of integers. If F is a number field, let $\mathcal{O}_F$ denote the integral closure of $\mathbf{Z}$ in F. There is a known diophantine definition of $\mathbf{Z}$ over $\mathcal{O}_F$ for the following number fields:

1. F is totally real [Den80].
2. F is a quadratic extension of a totally real number field [DL78].
3. F has exactly one conjugate pair of nonreal embeddings [Phe88], [Shl89].

In particular, Hilbert's Tenth Problem over $\mathcal{O}_F$ is undecidable for such fields F.

It is conjectured [DL78] that for *every* number field F, there is a diophantine definition of $\mathbf{Z}$ over $\mathcal{O}_F$. Our main theorem gives evidence for this conjecture, by reducing it to a plausible conjecture about the existence of certain elliptic curves.

Before stating our theorem, let us recall the Mordell-Weil Theorem, which states that if E is an elliptic curve over a number field F, then the abelian group $E(F)$ is finitely generated. Let $\operatorname{rk} E(F)$ denote the rank of $E(F)$.

Theorem 1. *Let $F \subseteq K$ be number fields, and let $\mathcal{O}_F$ and $\mathcal{O}_K$ be their rings of integers. Suppose that there exists an elliptic curve E over F such that $\operatorname{rk} E(F) = \operatorname{rk} E(K) = 1$. Then there exists a diophantine definition of $\mathcal{O}_F$ over $\mathcal{O}_K$.*

Most of the rest of this paper is devoted to the proof of Theorem 1. But for now, we mention its application to Hilbert's Tenth Problem.

Corollary 2. *Under the hypotheses of Theorem 1, if in addition F is of one of the types of number fields listed above for which a diophantine definition of $\mathbf{Z}$ over $\mathcal{O}_F$ is known, then Hilbert's Tenth Problem over $\mathcal{O}_K$ is undecidable.*

Proof. Theorem 1 reduces the undecidability over $\mathcal{O}_K$ to the undecidability over $\mathcal{O}_F$. □

J. Denef, at the end of [Den80], sketches a simple proof of Theorem 1 in the case where K is totally real and $F = \mathbf{Q}$. In fact, he is also able to treat some totally real algebraic extensions K of *infinite* degree over $\mathbf{Q}$. But his proof technique does not seem to generalize easily to fields that are not totally real.

Our proof of Theorem 1 is similar to that of an older result, the theorem of [DL78], which uses a 1-dimensional torus (a Pell equation) in place of the elliptic curve. We have been inspired also by the exposition of the "weak version of the vertical method" in [Shl00b] and by the ideas in [Phe00].

2.1 Preliminaries on Diophantine Sets over $\mathcal{O}_K$

The subset $\mathcal{O}_K - \{0\}$ of $\mathcal{O}_K$ is diophantine over $\mathcal{O}_K$: see Proposition 1(c) of [DL78]. We have a surjective map $\mathcal{O}_K \times (\mathcal{O}_K - \{0\}) \to K$ taking (a, b) to a/b. If $S \subseteq K^n$ is diophantine over K, then the inverse image of S under $(\mathcal{O}_K \times (\mathcal{O}_K - \{0\}))^n \to K^n$ is diophantine over $\mathcal{O}_K$. In this case, we will also say that S is diophantine over $\mathcal{O}_K$. It follows that in constructing diophantine definitions over $\mathcal{O}_K$, there is no harm in using equations with some variables taking values in $\mathcal{O}_K$ and other variables taking values in K.

Given $t \in K^\times$, define the *denominator ideal* $\operatorname{den}(t) = \{\, b \in \mathcal{O}_K : bt \in \mathcal{O}_K \,\}$ and the *numerator ideal* $\operatorname{num}(t) = \operatorname{den}(t^{-1})$. Also define $\operatorname{num}(0)$ to be the zero ideal. These ideals behave in the obvious way upon extension of the field.

Lemma 3.

1. *For fixed $m, n \in \mathbf{Z}_{\geq 0}$, the set of $(x_1, \ldots, x_m, y_1, \ldots, y_n)$ in K^{m+n} such that the fractional ideal $(x_1, \ldots, x_m)$ divides the fractional ideal $(y_1, \ldots, y_n)$ is diophantine over $\mathcal{O}_K$.*
2. *The set of $(t, u) \in K^\times \times K^\times$ such that* $\operatorname{den}(t) \mid \operatorname{den}(u)$ *is diophantine over $\mathcal{O}_K$.*
3. *The set of $(t, u) \in K^\times \times K$ such that* $\operatorname{den}(t) \mid \operatorname{num}(u)$ *is diophantine over $\mathcal{O}_K$.*
4. *The set of $(t, u) \in \mathcal{O}_K \times K^\times$ such that $t \mid$* $\operatorname{den}(u)$ *is diophantine over $\mathcal{O}_K$.*

Proof. Statement 1 is clear, since the condition is that there exist $c_{ij} \in \mathcal{O}_K$ such that $y_j = \sum_i c_{ij} x_i$ for each j. Statement 2 follows from statement 1, since $\operatorname{den}(t) \mid \operatorname{den}(u)$ if and only if the fractional ideal $(u, 1)$ divides $(t, 1)$. Statements 3 and 4 follow from statement 2: namely,

$$\begin{aligned} \operatorname{den}(t) \mid \operatorname{num}(u) &\iff u = 0 \text{ or } (\exists v)(uv = 1 \text{ and } \operatorname{den}(t) \mid \operatorname{den}(v)), \\ t \mid \operatorname{den}(u) &\iff (\exists v)(tv = 1 \text{ and } \operatorname{den}(v) \mid \operatorname{den}(u)). \end{aligned}$$

□

2.2 Bounds from Divisibility in $\mathcal{O}_K$

Let $n = [K : \mathbf{Q}]$ and $s = [K : F]$. Fix $\alpha \in \mathcal{O}_K$ such that $\{1, \alpha, \ldots, \alpha^{s-1}\}$ is a basis for K over F. Let $D \in \mathcal{O}_F$ denote the discriminant of this basis. If I is an ideal in $\mathcal{O}_K$, let $N_{K/\mathbf{Q}}(I) \in \mathbf{Z}_{\geq 0}$ denote its norm.

Lemma 4. *There is a positive integer $c > 0$ depending only on F, K, and α such that the following holds. Let $I \subset \mathcal{O}_K$ be a nonzero nonunit ideal, and let $\mu \in \mathcal{O}_K$. Write $\mu = \sum_{i=0}^{s-1} a_i \alpha^i$ with $a_i \in F$. Suppose that $\mu(\mu+1)\cdots(\mu+n) \mid I$. Then $N_{K/\mathbf{Q}}(Da_i) \le N_{K/\mathbf{Q}}(I)^c$.*

Proof. This is essentially Section 1.2 of [Shl00b]. The only differences are that we have specialized by taking $l_i = -i$, and we have generalized by replacing the element y by an ideal I: this does not affect the proof. □

The following is similar to Lemma 2.5 in [Shl00b].

Lemma 5. *There exists a constant $c' > 0$ depending only on F and K such that the following holds: Let I be a nonzero ideal of $\mathcal{O}_F$. Suppose $\mu \in \mathcal{O}_K$ and $w \in \mathcal{O}_F$. Write $\mu = \sum_{i=0}^{s-1} a_i \alpha^i$ with $a_i \in F$. Suppose $N_{K/\mathbf{Q}}(Da_i) < c' N_{K/\mathbf{Q}}(I)$ for all i, and $\mu \equiv w \pmod{I\mathcal{O}_K}$. Then $\mu \in \mathcal{O}_F$.*

Proof. Choose ideals $J_1, \dots, J_h \subseteq \mathcal{O}_F$ representing the elements of the class group of F, and choose $c' > 0$ such that $c' N_{K/\mathbf{Q}}(J_j) < 1$ for all j. Choose j such that $J_j I^{-1}$ is principal, generated by $z \in F^\times$, say. Since $\mu \equiv w \pmod{I\mathcal{O}_K}$, we have

$$z(\mu - w) = z(a_0 - w) + (za_1)\alpha + \cdots + (za_{s-1})\alpha^{s-1} \in \mathcal{O}_K.$$

By Lemma 4.1 of [Shl00b] (an elementary lemma about discriminants), $Dza_i \in \mathcal{O}_F$ for $i = 1, 2, \dots, s-1$. On the other hand,

$$|N_{K/\mathbf{Q}}(Dza_i)| = |N_{K/\mathbf{Q}}(Da_i) N_{K/Q}(z)| < c' N_{K/\mathbf{Q}}(I) \frac{N_{K/Q}(J_j)}{N_{K/Q}(I)} < 1,$$

by definition of c', so $Dza_i = 0$. Thus $a_i = 0$ for $i = 1, 2, \dots, s-1$. Hence $\mu \in \mathcal{O}_F$. □

2.3 Denominators of x-Coordinates of Points on an Elliptic Curve

We assume that an elliptic curve E as in Theorem 1 exists. Thus E is defined over F, and $\operatorname{rk} E(F) = \operatorname{rk} E(K) = 1$. Hence E has a Weierstrass model of the form $y^2 = x^3 + ax + b$ and we may assume $a, b \in \mathcal{O}_F$. Let O denote the point at infinity on E, which is the identity of $E(F)$.

For each nonarchimedean place $\mathfrak{p}$ of K, let $K_\mathfrak{p}$ denote the completion of K at $\mathfrak{p}$. and let $\mathbf{F}_\mathfrak{p}$ denote the residue field. Reducing coefficients modulo $\mathfrak{p}$ yields a possibly singular curve

$$E_\mathfrak{p} := \operatorname{Proj} \frac{\mathbf{F}_\mathfrak{p}[X,Y,Z]}{(Y^2Z - X^3 - \bar{a}XZ^2 - \bar{b}Z^3)}$$

over $\mathbf{F}_\mathfrak{p}$. Let $E_\mathfrak{p}^{\text{smooth}}$ denote the smooth part of $E_\mathfrak{p}$. Let $E_0(K_\mathfrak{p})$ be the set of points in $E(K_\mathfrak{p})$ whose reduction mod $\mathfrak{p}$ lies in $E_\mathfrak{p}^{\text{smooth}}(\mathbf{F}_\mathfrak{p})$.

Lemma 6.

1. *$E_0(K_\mathfrak{p})$ is a subgroup of $E(K_\mathfrak{p})$.*
2. *$E^{\text{smooth}}_\mathfrak{p}(\mathbf{F}_\mathfrak{p})$ is an abelian group under the usual chord-tangent law.*
3. *Reduction modulo $\mathfrak{p}$ gives a surjective group homomorphism $\operatorname{red}_\mathfrak{p} : E_0(K_\mathfrak{p}) \to E^{\text{smooth}}_\mathfrak{p}(\mathbf{F}_\mathfrak{p})$.*
4. *Both $E_0(K_\mathfrak{p})$ and $E_1(K_\mathfrak{p}) := \ker(\operatorname{red}_\mathfrak{p})$ are of finite index in $E(K_\mathfrak{p})$.*

Proof. For the first three statements, see Proposition VII.2.1 in [Sil92]. We have not assumed that our Weierstrass model is minimal at $\mathfrak{p}$, so our definition of E_0 is different from the standard one in [Sil92], but this does not matter in the proofs. To prove statement 4, observe that $E_0(K_\mathfrak{p})$ and $E_1(K_\mathfrak{p})$ are open subgroups of the compact group $E(K_\mathfrak{p})$ in the $\mathfrak{p}$-adic topology. □

From now on, $r \in \mathbf{Z}_{\geq 1}$ is assumed to be a multiple of $\#E(K)_{\text{tors}}$, of the index $(E(K) : E(F))$, and of the index $(E(K_\mathfrak{p}) : E_0(K_\mathfrak{p}))$ for each bad nonarchimedean place $\mathfrak{p}$. Then $rE(K)$ is a subgroup of $E(F)$ that is free of rank 1, and $rE(K)$ is contained in $E_0(K_\mathfrak{p})$ for every $\mathfrak{p}$.

We will need a diophantine approximation result. First we define the norm $\| \ \|_v : K \to \mathbf{R}_{\geq 0}$ for each place v of K; it will be characterized by its values on $a \in \mathcal{O}_K$. If v is nonarchimedean and $a \in \mathcal{O}_K - \{0\}$, then $\|a\|_v := q^{-v(a)}$ where q is the size of the residue field, and the discrete valuation v is normalized to take values in $\mathbf{Z}$. If v is real, then $\|a\|_v$ is the standard absolute value of the image of a under $K \to \mathbf{R}$. If v is complex, then $\|a\|_v$ is the square of the standard absolute value of the image of a under $K \to \mathbf{C}$. Define the naive logarithmic height of $a \in K$ by

$$h(a) := \sum_{\text{places } v \text{ of } K} \log \max\{\|a\|_v, 1\}.$$

If one sums over only the nonarchimedean places v, one obtains $\log N_{K/\mathbf{Q}} \operatorname{den}(a)$.

Proposition 7. *Let X be a smooth, projective, geometrically integral curve over K of genus ≥ 1. Fix a place v of K. Let ϕ be a nonconstant rational function on X. Let $P_1, P_2, \ldots$ be a sequence of distinct points in $X(K)$. For sufficiently large m, P_m is not a pole of ϕ, so $z_m := \phi(P_m)$ belongs to K. Then*

$$\lim_{m\to\infty} \frac{\log \|z_m\|_v}{h(z_m)} = 0.$$

Proof. See Section 7.4 of [Ser97]. □

Lemma 8. *The following holds if r is sufficiently large: If $P \in rE(K) - \{O\}$ and $m \in \mathbf{Z} - \{-1, 0, 1\}$, then*

$$\log N_{K/\mathbf{Q}} \operatorname{den}(x(mP)) \geq \frac{9}{10} m^2 \log N_{K/\mathbf{Q}} \operatorname{den}(x(P)) > 0;$$

in particular $\operatorname{den}(x(mP)) \neq \operatorname{den}(x(P))$ and $\operatorname{den}(x(P)) \neq (1)$.

Proof. Let P_1 be a generator of $rE(K)$. The theory of the canonical height in Chapter 8, Section 9 of [Sil92] implies that there is a real number $\hat{h}(P_1) > 0$ (namely, the canonical height of P_1, suitably normalized) such that $h(x(mP_1)) = m^2\hat{h}(P_1) + O(1)$, where the implied constant is independent of $m \in \mathbf{Z}$. Proposition 7 applied to each archimedean v, with $X = E$ and $\phi = x$, shows that if we forget to include the (finitely many) archimedean places in the sum defining h, we obtain

$$\log N_{K/\mathbf{Q}} \operatorname{den}(x(mP_1)) = (1 - o(1))h(x(mP_1)) = (1 - o(1))m^2\hat{h}(P_1)$$

as $|m| \to \infty$. The results follow for large r. □

Of course, there is nothing special about 9/10; any real number in the interval $(1/4, 1)$ would have done just as well.

2.4 Divisibility of Denominators

From now on, we suppose that r is large enough that Lemma 8 holds.

Lemma 9. *Let $P, P' \in rE(K) - \{O\}$. Then* $\operatorname{den}(x(P)) \mid \operatorname{den}(x(P'))$ *if and only if P' is an integral multiple of P.*

Proof. We first show that for any ideal $I \subseteq \mathcal{O}_K$, the set

$$G_I := \{\, Q \in rE(K) : I \mid \operatorname{den}(x(Q)) \,\}$$

is a subgroup of $rE(K)$. (By convention, we consider O to be an element of G_I.) Since an intersection of subgroups is a subgroup, it suffices to prove this when $I = \mathfrak{p}^n$ for some prime $\mathfrak{p}$ and some $n \in \mathbf{Z}_{\geq 1}$. Let $\mathcal{O}_\mathfrak{p}$ be the completion of $\mathcal{O}_K$ at $\mathfrak{p}$. Let $\mathcal{F} \in \mathcal{O}_K[[z_1, z_2]]$ denote the formal group of E with respect to the parameter $z := -x/y$, as in Chapter 4 of [Sil92]. Then there is an isomorphism $\mathcal{F}(\mathfrak{p}\mathcal{O}_\mathfrak{p}) \simeq E_1(K_\mathfrak{p})$, given by $z \mapsto (x(z), y(z))$ where $x(z) = z^{-2} + \dots$ and $y(z) = -z^{-3} + \dots$ are Laurent series with coefficients in $\mathcal{O}_K$. It follows that $G_{\mathfrak{p}^n}$ is the set of points in $rE(K)$ lying in the image of $\mathcal{F}(\mathfrak{p}^{\lceil n/2 \rceil}\mathcal{O}_\mathfrak{p})$. In particular $G_{\mathfrak{p}^n}$ is a subgroup of $rE(K)$.

The "if" part of the lemma follows from the preceding paragraph. Now we prove the "only if" part. Let $G = G_{\operatorname{den}(x(P))}$. Then G is a subgroup of $rE(K) \simeq \mathbf{Z}$, so G is free of rank 1. Let Q be a generator of G. By definition of G, we have $P \in G$, so P is a multiple of Q. By the "if" part already proved, $\operatorname{den}(x(Q)) \mid \operatorname{den}(x(P))$. On the other hand, $Q \in G$, so $\operatorname{den}(x(P)) \mid \operatorname{den}(x(Q))$ by definition of G. Thus $\operatorname{den}(x(Q)) = \operatorname{den}(x(P))$. By Lemma 8, $Q = \pm P$. If $\operatorname{den}(x(P)) \mid \operatorname{den}(x(P'))$, then $P' \in G = \mathbf{Z}Q = \mathbf{Z}P$. □

Lemma 10. *If $I \subseteq \mathcal{O}_K$ is a nonzero ideal, then there exists $P \in rE(K) - \{O\}$ such that $I \mid \operatorname{den}(x(P))$.*

Proof. We use the notation of the previous proof. It suffices to show that $G_{\mathfrak{p}^n}$ is nontrivial. This holds since the image of $\mathcal{F}(\mathfrak{p}^{\lceil n/2 \rceil}\mathcal{O}_\mathfrak{p})$ under $\mathcal{F}(\mathfrak{p}\mathcal{O}_\mathfrak{p}) \simeq E_1(K_\mathfrak{p})$ is an open subgroup of $E(K_\mathfrak{p})$, hence of finite index. □

Lemma 11. *Suppose $P \in rE(K) - \{O\}$ and $m \in \mathbf{Z} - \{0\}$. Let $t = x(P)$ and $t' = x(mP)$. Then* $\operatorname{den}(t) \mid \operatorname{num}((t/t' - m^2)^2)$.

Proof. Suppose that $\mathfrak{p}$ is a prime dividing $\operatorname{den}(t)$. Let $v_{\mathfrak{p}} : K_{\mathfrak{p}} \to \mathbf{Z} \cup \{\infty\}$ denote the discrete valuation associated to $\mathfrak{p}$. Then $n := v_{\mathfrak{p}}(z(P))$ is positive. Since $x = z^{-2} + \dots$ is a Laurent series with coefficients in $\mathcal{O}_K$, we have $x(P) \in z(P)^{-2}(1+\mathfrak{p}^n\mathcal{O}_{\mathfrak{p}})$. Using the formal group, we see that $z(mP) \in mz(P)+\mathfrak{p}^{2n}\mathcal{O}_{\mathfrak{p}}$; in particular $v_{\mathfrak{p}}(z(mP)) \geq n$, so $x(mP) \in z(mP)^{-2}(1 + \mathfrak{p}^n\mathcal{O}_{\mathfrak{p}})$. Thus

$$\frac{t}{t'} = \frac{x(P)}{x(mP)} \in \left(\frac{z(mP)}{z(P)}\right)^2 (1 + \mathfrak{p}^n\mathcal{O}_{\mathfrak{p}}) \, .$$

But $\frac{z(mP)}{z(P)} \in m + \mathfrak{p}^n\mathcal{O}_{\mathfrak{p}}$, so $t/t' \in m^2 + \mathfrak{p}^n\mathcal{O}_{\mathfrak{p}}$, so $\mathfrak{p}^n \mid \operatorname{num}(t/t' - m^2)$. On the other hand, $\mathfrak{p}^{2n}$ is the exact power of $\mathfrak{p}$ dividing $\operatorname{den}(t)$. Applying this argument to every $\mathfrak{p}$ proves $\operatorname{den}(t) \mid \operatorname{num}((t/t' - m^2)^2)$. □

2.5 Diophantine Definition of $\mathcal{O}_F$ over $\mathcal{O}_K$

Lemma 12. *With hypotheses as in Theorem 1, there exists a subset $S \subseteq \mathcal{O}_K$ such that S is diophantine over $\mathcal{O}_K$ and $\{ m^2 : m \in \mathbf{Z}_{\geq 1} \} \subseteq S \subseteq \mathcal{O}_F$.*

Proof. Let c and c' be the constants of Lemmas 4 and 5, respectively. By Lemma 8, if $\ell \in \mathbf{Z}_{\geq 1}$ is sufficiently large, then

$$c' N_{K/\mathbf{Q}} \operatorname{den}(x(\ell P_0))^{1/2} > N_{K/\mathbf{Q}} \operatorname{den}(x(P_0)^c)$$

for all $P_0 \in rE(K) - \{O\}$. Fix such an ℓ.

Let S be the set of $\mu \in \mathcal{O}_K$ such that there exist $P_0, P', P' \in rE(K) - \{O\}$ and $t_0, t, t' \in F$ such that

1. $P = \ell P_0$
2. $t_0 = x(P_0)$, $t = x(P)$, $t' = x(P')$
3. $(\mu + 1)(\mu + 2) \dots (\mu + n) \mid \operatorname{den}(t_0)$
4. $\operatorname{den}(t) \mid \operatorname{den}(t')$
5. $\operatorname{den}(t) \mid \operatorname{num}((t/t' - \mu)^2)$

It follows from Lemma 3 that S is diophantine over $\mathcal{O}_K$.

Suppose $m \in \mathbf{Z}_{\geq 1}$. We wish to show that $\mu := m^2$ belongs to S. By Lemma 10, there exists $P_0 \in rE(K) - \{O\}$ such that $(\mu + 1)(\mu + 2) \dots (\mu + n) \mid \operatorname{den}(x(P_0))$. Let $P = \ell P_0$ and $P' = mP$. Let $t_0 = x(P_0)$, $t = x(P)$, and $t' = x(P')$. Then conditions (1), (2), and (3) in the definition of S are satisfied, and (4) and (5) follow from Lemmas 9 and 11, respectively. Hence $m^2 \in S$.

Now suppose that $\mu \in S$. We wish to show that $\mu \in \mathcal{O}_F$. Fix P_0, P, P', t_0, t, t' satisfying (1) through (5). By (4) and Lemma 9, $P' = mP$ for some nonzero $m \in \mathbf{Z}$. By Lemma 11, $\operatorname{den}(t) \mid \operatorname{num}((t/t' - m^2)^2)$. On the other hand, (5) says that $\operatorname{den}(t) \mid \operatorname{num}((t/t' - \mu)^2)$. Therefore $\operatorname{den}(t)^{1/2} \mid \operatorname{num}(\mu - m^2) = (\mu - m^2)$. (Note that each prime of $\mathcal{O}_F$ or of $\mathcal{O}_K$ that appears in $\operatorname{den}(t)$ must occur to an even power, since t is the x-coordinate of a point on $y^2 = x^3 + ax + b$. Hence $\operatorname{den}(t)^{1/2}$ is a well-defined ideal.) Write $\mu = \sum_{i=0}^{s-1} a_i \alpha^i$ with $a_i \in F$. By (3) and Lemma 4, $N_{K/\mathbf{Q}}(Da_i) \leq N_{K/\mathbf{Q}}(\operatorname{den}(t_0))^c$. By definition of ℓ, we have $N_{K/\mathbf{Q}}(\operatorname{den}(t_0))^c < c' N_{K/\mathbf{Q}} \operatorname{den}(t)^{1/2}$. Combining these shows that the hypotheses of Lemma 5 hold for $w = m^2$ and $I = \operatorname{den}(t)^{1/2}$ (as an ideal in $\mathcal{O}_F$). Thus $\mu \in \mathcal{O}_F$. □

Proof of Theorem 1. Let S be the set given by Lemma 12. Then $S_1 := \{\, s - s' : s, s' \in S \,\}$ contains all odd integers at least 3, because of the identity $(m+1)^2 - m^2 = 2m + 1$. Next, $S_2 := S_1 \cup \{\, 4 - s : s \in S_1 \,\}$ contains all odd integers, and $S_3 := S_2 \cup \{\, s + 1 : s \in S_2 \,\}$ contains $\mathbf{Z}$. Let $\beta_1, \dots, \beta_b$ be a $\mathbf{Z}$-basis for $\mathcal{O}_F$. Then $S_4 := \{\, a_1\beta_1 + \cdots + a_b\beta_b : a_1, \dots, a_b \in S_3 \,\}$ contains $\mathcal{O}_F$.

But $S \subseteq \mathcal{O}_F$, so $S_i \subseteq \mathcal{O}_F$ for $i = 1, 2, 3, 4$. In particular, $S_4 = \mathcal{O}_F$. Also, S is diophantine over $\mathcal{O}_K$, so each S_i is diophantine over $\mathcal{O}_K$. In particular, $\mathcal{O}_F = S_4$ is diophantine over $\mathcal{O}_K$. □

2.6 Questions

1. Is it true that for every number field K, there exists an elliptic curve E over $\mathbf{Q}$ such that $\operatorname{rk} E(\mathbf{Q}) = \operatorname{rk} E(K) = 1$? The author would conjecture so. If so, then Hilbert's Tenth Problem over $\mathcal{O}_K$ is undecidable for every number field K.
2. Can one weaken the hypotheses of Theorem 1 and give a diophantine definition of $\mathcal{O}_F$ over $\mathcal{O}_K$ using any elliptic curve E over K with $\operatorname{rk} E(K) = 1$, not necessarily defined over F? Such elliptic curves may be easier to find. But our proof of Theorem 1 seems to require the fact that E is defined over F and has $\operatorname{rk} E(F) = 1$, since Lemma 5 fails if the ideal I of $\mathcal{O}_F$ is instead assumed to be an ideal of $\mathcal{O}_K$.
3. Can one prove an analogue of Theorem 1 in which the elliptic curve is replaced by an abelian variety?

References

CZ00. Gunther Cornelissen and Karim Zahidi, *Topology of Diophantine sets: remarks on Mazur's conjectures*, Hilbert's tenth problem: relations with arithmetic and algebraic geometry (Ghent, 1999), Amer. Math. Soc., Providence, RI, 2000, pp. 253–260.

Dav53. Martin Davis, *Arithmetical problems and recursively enumerable predicates*, J. Symbolic Logic **18** (1953), 33–41.

Den80. J. Denef, *Diophantine sets over algebraic integer rings. II*, Trans. Amer. Math. Soc. **257** (1980), no. 1, 227–236.

DL78. J. Denef and L. Lipshitz, *Diophantine sets over some rings of algebraic integers*, J. London Math. Soc. (2) **18** (1978), no. 3, 385–391.

DL+00. Jan Denef, Leonard Lipshitz, Thanases Pheidas, and Jan Van Geel (eds.), *Hilbert's tenth problem: relations with arithmetic and algebraic geometry*, American Mathematical Society, Providence, RI, 2000, Papers from the workshop held at Ghent University, Ghent, November 2–5, 1999.

DPR61. Martin Davis, Hilary Putnam, and Julia Robinson, *The decision problem for exponential diophantine equations*, Ann. of Math. (2) **74** (1961), 425–436.

Eis. Kirsten Eisenträger, Ph. D. thesis, University of California, Berkeley, in preparation.

KR92. K. H. Kim and F. W. Roush, *Diophantine undecidability of* $\mathbf{C}(t_1, t_2)$, J. Algebra **150** (1992), no. 1, 35–44.

Mat70. Ju. V. Matijasevič, *The Diophantineness of enumerable sets*, Dokl. Akad. Nauk SSSR **191** (1970), 279–282.

Maz94. B. Mazur, *Questions of decidability and undecidability in number theory*, J. Symbolic Logic **59** (1994), no. 2, 353–371.

MB. Laurent Moret-Bailly, paper in preparation, extending results presented in a lecture 18 June 2001 at a conference in honor of Michel Raynaud in Orsay, France.

Phe88. Thanases Pheidas, *Hilbert's tenth problem for a class of rings of algebraic integers*, Proc. Amer. Math. Soc. **104** (1988), no. 2, 611–620.

Phe91. Thanases Pheidas, *Hilbert's tenth problem for fields of rational functions over finite fields*, Invent. Math. **103** (1991), no. 1, 1–8.

Phe00. Thanases Pheidas, *An effort to prove that the existential theory of* $\mathbf{Q}$ *is undecidable*, Hilbert's tenth problem: relations with arithmetic and algebraic geometry (Ghent, 1999), Amer. Math. Soc., Providence, RI, 2000, pp. 237–252.

PZ00. Thanases Pheidas and Karim Zahidi, *Undecidability of existential theories of rings and fields: a survey*, Hilbert's tenth problem: relations with arithmetic and algebraic geometry (Ghent, 1999), Amer. Math. Soc., Providence, RI, 2000, pp. 49–105.

Ser97. Jean-Pierre Serre, *Lectures on the Mordell-Weil theorem*, third ed., Friedr. Vieweg & Sohn, Braunschweig, 1997, Translated from the French and edited by Martin Brown from notes by Michel Waldschmidt, With a foreword by Brown and Serre.

Shl89. Alexandra Shlapentokh, *Extension of Hilbert's tenth problem to some algebraic number fields*, Comm. Pure Appl. Math. **42** (1989), no. 7, 939–962.

Shl92. Alexandra Shlapentokh, *Hilbert's tenth problem for rings of algebraic functions in one variable over fields of constants of positive characteristic*, Trans. Amer. Math. Soc. **333** (1992), no. 1, 275–298.

Shl00a. Alexandra Shlapentokh, *Hilbert's tenth problem for algebraic function fields over infinite fields of constants of positive characteristic*, Pacific J. Math. **193** (2000), no. 2, 463–500.

Shl00b. Alexandra Shlapentokh, *Hilbert's tenth problem over number fields, a survey*, Hilbert's tenth problem: relations with arithmetic and algebraic geometry (Ghent, 1999), Amer. Math. Soc., Providence, RI, 2000, pp. 107–137.

Sil92. Joseph H. Silverman, *The arithmetic of elliptic curves*, Springer-Verlag, New York, 1992, Corrected reprint of the 1986 original.

Vid94. Carlos R. Videla, *Hilbert's tenth problem for rational function fields in characteristic* 2, Proc. Amer. Math. Soc. **120** (1994), no. 1, 249–253.

On p-adic Point Counting Algorithms for Elliptic Curves over Finite Fields

Takakazu Satoh

Department of Mathematics, Faculty of Science,
Saitama University, Urawa, Saitama 338-8570, Japan
tsatoh@rimath.saitama-u.ac.jp

Abstract. Let p be a prime and let $q := p^N$. Let E be an elliptic curve over $\mathbf{F_q}$. We are interested in efficient algorithms to compute the order of the group $E(\mathbf{F_q})$ of $\mathbf{F_q}$-rational points of E. An l-adic algorithm, known as the SEA algorithm, computes $\#E(\mathbf{F_q})$ with $O((\log q)^{4+\varepsilon})$ bit operations (with fast arithmetic) and $O((\log q)^2)$ memory. In this article, we survey recent advances in p-adic algorithms. For a fixed small p, the computational complexity of the known fastest p-adic point counting algorithm is $O(N^{3+\varepsilon})$ in time and $O(N^2)$ in space. If we accept some precomputation depending only on p and N or a certain restriction on N, the time complexity is reduced to $O(N^{2.5+\varepsilon})$ still with $O(N^2)$ space requirement.

1 Introduction

Let p be a prime and $N \in \mathbf{N}$, let $q := p^N$. Let $\mathbf{F}_q$ be the finite field of q elements. Our problem is to find a fast algorithm to compute the number of $\mathbf{F}_q$-rational points of a given elliptic curve $E/\mathbf{F}_q$. In other words, we seek a fast algorithm to compute the trace of the q-th power Frobenius endomorphism Fr_q since $\#E(\mathbf{F}_q) = 1 + q - \mathrm{Tr}(\mathrm{Fr}_q)$. We can consider a similar problem for wider classes of objects such as hyperelliptic curves, Abelian varieties or arbitrary algebraic varieties. However, we shall mainly study algorithms for elliptic curves.

The first polynomial time (with respect to $\log q$) algorithm was found by Schoof[49]. Let μ be a constant such that the multiplication of two n bit integers can be carried out with $O(n^\mu)$ bit operations and that[†1] a multiplication of two polynomials of degree n is performed in $O(n^\mu)$ arithmetic operations over their coefficient ring. Then the running time of Schoof's algorithm is $O((\log q)^{3\mu+2})$. Elkies and Atkin (cf. Elkies[17] and Schoof[50]) made significant practical improvements and the resulting method is now called the SEA algorithm. The running time of the SEA algorithm is heuristically estimated as $O((\log q)^{2\mu+2})$ bit operations.[†2] The key idea of the SEA algorithm is to compute $\mathrm{Tr}(\mathrm{Fr}_q)$ mod l

[†1] In an actual implementation, different algorithms may be used for polynomial multiplications and integer multiplications. However, we assume that they are the same for simplicity.

[†2] Under the Generalized Riemann hypothesis(GRH) it can be proved that the largest prime l used in the Elkies' algorithm is $O((\log q)^{2+\varepsilon})$ for any $\varepsilon > 0$. See Ap-

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 43–66, 2002.

for various small primes $l(\neq p)$. By Hasse's inequality $|\mathrm{Tr}\,\mathrm{Fr}_q| \leq 2\sqrt{q}$, we can recover $\mathrm{Tr}(\mathrm{Fr}_q)$ using the Chinese Remainder Theorem. Couveignes and Morain[13] obtained an algorithm to compute $\mathrm{Tr}(\mathrm{Fr}_q) \bmod l^n$ for small values of l^n (but in theory it works for all $n \in \mathbf{N}$). Thus indeed the SEA algorithm is an "l-adic" method.

On the other hand, p-adic methods attempt to construct (in some suitable sense) a p-adic lift of the Frobenius endomorphism to characteristic zero. Such an idea goes back to Dwork's proof[15] of the rationality of the zeta function of a variety over a finite field. Wan[56, Cor. 5.3] proposed an algorithm which computes the zeta function of an arbitrary hypersurface over a finite field, modulo p^m for small p^m. However its growth rate is exponential with respect to m.[†3] So, it is not feasible to count the number of points on elliptic curves using this algorithm when N is large. The challenge for p-adic point counting algorithms for elliptic curves is as follows: By 1996, Couveignes[11, 12] and Lercier[35] had already extended the SEA algorithm to small characteristic cases. Their (heuristic) complexities are $O((\log q)^{2\mu+2})$. The goal is to construct a faster algorithm.

We fix a (small in practice) prime p and study computational complexities as $N \to \infty$. The first p-adic algorithm for elliptic curve point counting which (at least asymptotically) runs faster than the SEA algorithm was obtained in [45]. The main strategy is to lift E to an elliptic curve over a field of characteristic zero so that $\mathrm{Fr}_q \in \mathrm{End}(E)$ also lifts to an endomorphism of the lifted curve. Such a lift is called the canonical lift of E. Although this algorithm was not refined — it requires $O(N^3)$ memory, and works only for $p \geq 5$ — its time complexity is $O(N^{2\mu+1})$. Shortly afterwards, Fouquet, Gaudry, Harley[19] generalized this algorithm to the cases $p = 2$ and $p = 3$. Independently, Skjernaa[52] obtained a different algorithm for $p = 2$. The most difficult part of the calculation is to compute the kernel of the dual of Frobenius, for which we need a totally different algorithm from that applicable to an odd p. Vercauteren, Preneel, Vandewalle[55] reduced the space complexity to $O(N^2)$. Here, the Kronecker relation is important. A fast norm computation algorithm in Satoh, Skjernaa, Taguchi[46] makes the O-constant in the time complexity much smaller. On the other hand, Harley et al.[24] developed an algorithm for $p = 2$ based on the arithmetic-geometric mean(AGM). This is a very simple and fast algorithm.[†4] Combining these results, the computational complexity of elliptic curve point counting is $O(N^{2\mu+1})$ in time and $O(N^2)$ in space with quite reasonable O-constants. We can now compute the number of $\mathbf{F}_q$-rational points of a randomly given elliptic curve

pendix A. This implies that the time complexity of Elkies' algorithm is bounded by $O((\log q)^{3\mu+2+\varepsilon})$, whereas that of Schoof's algorithm is $O((\log q)^{3\mu+2})$. However, in practice, Elkies' algorithm runs much faster than Schoof's algorithm and numerical experiments support the above heuristic. Therefore, we use $O((\log q)^{2\mu+2})$ as a benchmark time complexity for Elkies' algorithm.

[†3] Later, Lauder and Wan[34] constructed a polynomial time algorithm for an arbitrary variety. See Section 5.

[†4] To the best knowledge of the author as of March 2002, the AGM method is the fastest algorithm which works for all N without precomputation.

over $\mathbf{F}_q$ for $q \approx 2^{15000}$ or more. When $q \approx 2^{200}$, the algorithm terminates in about a second.

The rest of the paper is organized as follows: After introducing some notation, we review the computational complexity of arithmetic operations. In Section 2 we describe the algorithm based on the canonical lift. In Section 3, we review the AGM point counting algorithm. Section 4 describes how the fast evaluation of the inverse of the Frobenius substitution reduces the run-time of the algorithm described in Section 2. Algorithms for more general classes of varieties are briefly summarized in Section 5.

1.1 Notation

Throughout this paper, $q = p^N$, K is the (unique up to isomorphism) unramified extension of degree N over $\mathbf{Q}_p$ and R is its valuation ring. Since K is unramified, the prime p is still a prime element of R. In general, π stands for a reduction modulo p map (of numbers, polynomials, curves, etc.). Let $\sigma \in \mathrm{Gal}(K/\mathbf{Q}_p)$ be the Frobenius substitution. Since σ is an isometry over K, it induces a ring automorphism of $R/p^m R$ for each $m \in \mathbf{N}$, which is also denoted by σ. By definition, $\sigma(x) = x^p$ for $x \in \mathbf{F}_q \cong R/pR$. The p^m-th Frobenius endomorphism is denoted by Fr_{p^m}. Hence, for an elliptic curve E defined over a field of characteristic p, $\sigma(E) = \mathrm{Fr}_p(E)$. However, the Frobenius substitution (Galois action) should not be confused with the lift of Frobenius endomorphism (rational map) for elliptic curves over K. The multiplicative p-adic valuation $|\cdot|_p$ is normalized as $|p|_p = \frac{1}{p}$. The additive valuation ord_p with respect to p is normalized as $\mathrm{ord}_p p = 1$.

The point at infinity of an elliptic curve given by the Weierstrass equation is denoted by $\mathcal{O}$. We use $-X/Y$ as a local parameter at $\mathcal{O}$. For elliptic curves E_1 and E_2 over a field k, the Abelian group of isogenies (defined over the algebraic closure of k) from E_1 to E_2 with addition by value is denoted by $\mathrm{Isog}(E_1, E_2)$. Let $f \in \mathrm{Isog}(E_1, E_2)$ and τ_i be the local parameter of E_i at $\mathcal{O}$ for $i = 1, 2$. Then we have the expansion

$$f^*(\tau_2) = c_1\tau_1 + c_2\tau_1^2 + \cdots .$$

We call c_1 the leading coefficient of f and denote it by $\mathrm{lc}(f)$.

1.2 Complexity for Ring Operations

Let A be a commutative ring (with the identity element). Let μ be as described in the Introduction. Hence $\mu = 2$ if we use a naive multiplication algorithm and $\mu = \log_2 3$ if we use the Karatsuba algorithm[27] (see Aho, Ullman, Hopcroft[1, §2.6] or Cohen[9, §3.1.2]). Asymptotically, we can take $\mu = 1 + \varepsilon$ for any $\varepsilon > 0$ if we use the Schönhage-Strassen algorithm[48] for integer multiplications and the Cantor-Kaltofen algorithm[6] for polynomial multiplications.[†5] Let $F(X) \in A[X]$

[†5] Actually, this algorithm works for an arbitrary (not necessarily commutative, associative) algebra. In the case that a small prime (in practice, either 2 or 3) is invertible in A, we can make some simplification to the Cantor-Kaltofen algorithm, which makes the algorithm about twice as fast.

be a monic polynomial of degree n. The ideal generated by $F(X)$ is denoted by $\langle F(X)\rangle$. Then a multiplication in $A[X]/\langle F(X)\rangle$ is performed with $O(n^\mu)$ ring operations of A. To see this, it is enough to show that the remainder $\mathrm{rem}(H,F)$ of the division H/F for $H \in A[X]$ with $\deg H \le 2n-2$ is obtained with $O(n^\mu)$ ring operations.[†6] This is implicit in Aho, Hopcroft, Ullman[1, §8.3]. Explicitly, for $H \in A[X]$ satisfying $\deg H \le 2n-2$,

$$\mathrm{rem}(H,F) = H - (((H/X^n)Z)/X^{n-2})F \tag{1.1}$$

where $Z := X^{2n-2}/F$.

As to inversion, we limit ourselves to the case $A = B/I^M$ where B is a local ring and I is the maximal ideal of B. We also assume that a ring operation of B/I^M amounts to $O(M^\mu)$ field operations of B/I. Then, computation of a^{-1} for $a \in (A[X]/\langle F(X)\rangle)^\times$ amounts to $O(n^\mu M^\mu + n^\mu \log n)$ field operations of B/I. In the case of $M = 1$, this can be carried out using an asymptotically fast GCD algorithm, say, [1, §8.9]. For $M \ge 2$, we can lift an inverse element modulo $F(X)\cdot I^{[(M+1)/2]}$ to an inverse element modulo $F(X)\cdot I^M$.[†7]

Now we can estimate the time complexity of arithmetic operations (namely, ring operations and an inversion of an invertible element) over $R/p^M R$. For simplicity, we assume that $M^\mu \ge \log N$ holds.[†8] There exists $\theta \in R^\times$ such that $\mathbf{Q}_p(\theta) = K$. Let $F \in \mathbf{Z}_p[X]$ be the monic minimal polynomial of θ. Then, $R = \mathbf{Z}_p[\theta]$ and $R/p^M R = B_M[X]/\langle F(X) \bmod p^M\rangle$ with $B_M = \mathbf{Z}_p/p^M\mathbf{Z}_p = \mathbf{Z}/p^M\mathbf{Z}$. Hence, an arithmetic operation over $R/p^M R$ amounts to $O((NM)^\mu)$ bit operations.

2 Canonical Lift Method

The canonical lift method is based on the following observation. Assume we can lift $E/\mathbf{F}_q$ to $\tilde{E}/K$ so that $\mathrm{Fr}_q \in \mathrm{End}(E)$ lifts to some $\varphi \in \mathrm{End}(\tilde{E})$. Then $\mathrm{Tr}(\mathrm{Fr}_q) = \mathrm{Tr}(\varphi)$. On the other hand, $\mathrm{lc}(\varphi)$ which lies in a field of characteristic zero gives enough information to compute $\mathrm{Tr}(\varphi)$. Computing the lift of Fr_q still needs a long computational time, but Fr_q is the N-fold iteration of the Fr_p whose lifting should be much easier (since p is small).

However, not every lift $\tilde{E}$ admits the lift of Fr_q. Given an ordinary elliptic curve $E/\mathbf{F}_q$, we call an elliptic curve $E^\uparrow/K$ the canonical lift of E if $\mathrm{End}(E) \cong \mathrm{End}(E^\uparrow)$. This is a special case of a deep theory due to Lubin, Serre, Tate[36] (see also Messing[38], especially its Appendix). The canonical lift of an ordinary

[†6] In many cases, F is a low weight polynomial, i.e., the number of non-zero coefficients of F is very small. Then, a naive division performs remainder computation with $O(n)$ ring operations of A.

[†7] For f, $g \in A[X]$ satisfying $fg \equiv 1 \bmod F\cdot I^{[(M+1)/2]}$, we see $f\cdot\mathrm{rem}(g(2-fg),F) \equiv 1 \bmod F\cdot I^M$. We note that in case of $M = \Omega(n)$, the naive Euclid algorithm is applicable to obtain $g \bmod p$ without changing the groth rate of the complexity of inversion.

[†8] In an application to elliptic curve point counting algorithms, $M = N/2 + O(1)$. Hence this condition holds except for tiny N.

elliptic curve exists and it is unique up to an isomorphism. For two ordinary elliptic curves E_1 and E_2,

$$\mathrm{Isog}(E_1, E_2) \cong \mathrm{Isog}(E_1^{\uparrow}, E_2^{\uparrow}). \tag{2.1}$$

We denote by $f^{\uparrow}$ the isogeny from $E_1^{\uparrow}$ to $E_2^{\uparrow}$ corresponding to $f \in \mathrm{Isog}(E_1, E_2)$ in (2.1).

Put $E^{(i)} := \sigma^i(E)$ and denote the dual isogeny of $\mathrm{Fr}_p \in \mathrm{Isog}(E^{(i-1)}, E^{(i)})$ by $V_p^{(i)}$, which is called the Verschiebung. Then, by (2.1), each $V_p^{(i)}$ lifts to $V_p^{(i)\uparrow}$. Let $V_q \in \mathrm{End}(E)$ be the dual of $\mathrm{Fr}_q \in \mathrm{End}(E)$. Since

$$V_q^{\uparrow} = V_p^{(1)\uparrow} \circ V_p^{(2)\uparrow} \circ \cdots \circ V_p^{(N)\uparrow},$$

it is clear that $\mathrm{lc}(V_q^{\uparrow}) = \prod_{i=1}^{N} \mathrm{lc}(V_p^{(i)\uparrow})$. On the other hand, $V_q^2 - \mathrm{Tr}(V_q)V_q + q = 0$ lifts to $V_q^{\uparrow 2} - \mathrm{Tr}(V_q)V_q^{\uparrow} + q = 0$, which implies $\mathrm{lc}(V_q^{\uparrow})^2 - \mathrm{Tr}(V_q)\mathrm{lc}(V_q^{\uparrow}) + q = 0$. Since E is ordinary, $V_p^{(i)}$ and V_q are separable and thus $\mathrm{lc}(V_p^{(i)\uparrow})$ and $\mathrm{lc}(V_q^{\uparrow})$ belong to $R^{\times}$. Therefore,

$$\mathrm{Tr}(\mathrm{Fr}_q) = \mathrm{Tr}(V_q) = \mathrm{lc}(V_q^{\uparrow}) + \frac{q}{\mathrm{lc}(V_q^{\uparrow})} \equiv \mathrm{lc}(V_q^{\uparrow}) \bmod q,$$

from which we see that $\mathrm{lc}(V_q^{\uparrow}) \bmod p^{N/2+O(1)}$ suffices to determine $\mathrm{Tr}(\mathrm{Fr}_q)$.

Before proceeding further, we note that in fact we can avoid the use of the above high-powered algebraic geometry. For the purpose of point counting, we can assume $j(E) \notin \mathbf{F}_{p^2}$.[†9] Otherwise, $k := \mathbf{F}_p(j(E))$ is either $\mathbf{F}_p$ or $\mathbf{F}_{p^2}$. Let $r := \#k$. We can construct E_0/k which is isomorphic to E over $\mathbf{F}_q$. Hence, letting $c_n := \mathrm{Tr}\,\mathrm{Fr}_{r^n}|_{E_0}$, we obtain $\mathrm{Tr}\,\mathrm{Fr}_q|_{E_0}$ (which is also $\mathrm{Tr}\,\mathrm{Fr}_q|_E$) by the recurrence formula

$$c_n = c_1 c_{n-1} - r c_{n-2}$$

with initial values $c_0 = 2$ and $c_1 = r + 1 - \#E_0(k)$. See Blake, Seroussi, Smart[3, Cor. VI.2] or Silverman[51, §V.2].

From now on, we assume that $j(E) \notin \mathbf{F}_{p^2}$. Let Φ_p be the p-th modular polynomial. Recall that two elliptic curves $\mathcal{E}$ and $\mathcal{E}'$ over K are p-isogenous if and only if $\Phi_p(j(\mathcal{E}), j(\mathcal{E}')) = 0$.

Theorem 1 ([45, Prop. 3.4]). *Assume $j(E) \notin \mathbf{F}_{p^2}$. Then, the system of equations*

$$\begin{cases} \Phi_p(Z_0, Z_1) = 0, \ldots, \Phi_p(Z_{N-1}, Z_0) = 0, \\ \pi(Z_0) = j(E), \ \ \pi(Z_1) = j(E^{(1)}), \ \ \ldots, \ \ \pi(Z_{N-1}) = j(E^{(N-1)}) \end{cases} \tag{2.2}$$

has a unique solution, which lies in R^N.

[†9] In particular, this implies that E is ordinary.

Theorem 2 (Skjernaa[52, Theorem 2.1]). *Let $E/\mathbf{F}_q$ satisfy $j(E) \notin \mathbf{F}_{p^2}$. Let E' and E'' be arbitrary lifts of E and $E^{(1)}$, respectively. Assume there exists a p-isogeny between E' and E''. Then $\mathrm{Fr}_p \in \mathrm{Isog}(E, E^{(1)})$ lifts to $\mathrm{Isog}(E', E'')$.*

Thus, the solution Z_i of (2.2) must be $j(E^{(i)\uparrow})$. Let $\mathcal{E}_i$ be an elliptic curve over K with $j(\mathcal{E}_i) = Z_i$. Even without a knowledge of canonical lifts, the above theorems ensure that $\mathrm{Fr}_p \in \mathrm{Isog}(E^{(i-1)}, E^{(i)})$ lifts to an element of $\mathrm{Isog}(\mathcal{E}_{i-1}, \mathcal{E}_i)$, which is $\mathrm{Fr}_p^\uparrow$. Then, $V_p^{(i)\uparrow}$ is the dual of $\mathrm{Fr}_p^\uparrow$.

Let E' be a quadratic twist of E. Then, $\mathrm{Tr}\,\mathrm{Fr}_q|_{E'} = -\mathrm{Tr}\,\mathrm{Fr}_q|_E$. Hence, without loss of generality, we may assume that E is given as follows:[†10]

$$\begin{array}{ll} Y^2 + XY = X^3 + j(E)^{-1} & (p=2), \\ Y^2 = X^3 + X^2 - j(E)^{-1} & (p=3), \\ Y^2 = X^3 + 3\gamma X + 2\gamma & \left(p \geq 5, \gamma = \frac{j(E)}{1728 - j(E)}\right). \end{array}$$

Then,

$$\mathrm{Tr}\,\mathrm{Fr}_q \equiv \begin{cases} 1 \bmod 4 & (p=2), \\ 1 \bmod 3 & (p=3), \\ N_{\mathbf{F}_q/\mathbf{F}_p}(h_E) \bmod p & (p \geq 5), \end{cases} \tag{2.3}$$

where h_E is the coefficient of X^{p-1} in $(X^3 + 3\gamma X + 2\gamma)^{(p-1)/2}$ (cf. Silverman[51, proof of Theorem V.4.1(a)] for $p \neq 2$, Blake, Seroussi, Smart[3, Lemma III.4] for $p = 2$). Now we can give an outline of the algorithm. (For simplicity, we assume that N is large enough so that $M \leq N$ in (0).)

(0) Let M be the minimal integer satisfying $p^M > 4\sqrt{q}$. (Note $M = N/2 + O(1)$.)
(1) Compute $j(E^{(i-1)\uparrow})$ and $j(E^{(i)\uparrow}) \bmod p^{M+O(1)}$ for some i. (The O-constant depends on p and an algorithm in (2).)
(2) Compute $c := \mathrm{lc}(V_p^{(i)\uparrow} : E^{(i)\uparrow} \to \sigma^{-1}(E^{(i)\uparrow}))^2$.
(3) Compute $t' := \sqrt{N_{K/\mathbf{Q}_p}(c)}$; the sign of the square root is determined by (2.3).
(4) return $t \in \mathbf{Z}$ satisfying $t \equiv t' \bmod p^M$ and $|t| < 2\sqrt{q}$.

In (1) and (2), any value of i will do as long as $j(E^{(i-1)\uparrow})$ and $j(E^{(i)\uparrow})$ have necessary precision. In what follows, we describe Steps (1)-(3) in some detail.

2.1 Construction of Canonical Lifts

In [45], the canonical lift is constructed by solving (2.2) using the multivariate Newton iterative root finding algorithm, which requires $O(N^{2\mu+1})$ bit operations and $O(N^3)$ space. Vercauteren, Preneel and Vandewalle[55] reduced the growth rate of the space complexity to $O(N^2)$. Although the time complexity is still $O(N^{2\mu+1})$, according to [55], it runs faster than the method of [45] by a factor of 1.5. The key point of their method is the following theorem.

[†10] See Blake, Seroussi, Smart[3, §III.3] or Enge[18, §3.10].

Theorem 3 (Vercauteren et al.[55, §2]). *Let $x \in R$ satisfy $x \equiv j(E^{\uparrow}) \bmod p^i$ with $i \in \mathbf{N}$. Then there exists a unique $y \in R$ such that $y \equiv x^p \bmod p$ and $\Phi_p(x, y) = 0$. Moreover, we have $y \equiv j(E^{(1)\uparrow}) \bmod p^{i+1}$.*

Using the Kronecker relation, we see that if $x \notin \mathbf{F}_{p^2}$ and $y \equiv x^p \bmod p$ then $\partial_X \Phi_p(x, y) \not\equiv 0 \bmod p$ and $\partial_Y \Phi_p(x, y) \equiv 0 \bmod p$. Hence $\frac{dy}{dx} \equiv 0 \bmod p$ when x and y change according to $\Phi_p(x, y) = 0$. Therefore, one might expect that the error $\left|y - j(E^{(1)\uparrow})\right|_p$ is less than $\frac{1}{p}\left|x - j(E^{\uparrow})\right|_p$. This is proved by virtue of the Taylor expansion of Φ_p. What is important here is that two j invariants of p-isogenous curves are related by an analytic function (in fact by the modular polynomial Φ_p).†11 The resulting algorithm is described below. For later use, we compute the j-invariants of two adjacent canonical lifts.

Algorithm 1. Computing the j-invariants of canonical lifts.
Input: $M \in \mathbf{N}$, an elliptic curve $E/\mathbf{F}_q$ satisfying $j(E) \notin \mathbf{F}_{p^2}$.
Output: $j(E^{(M-1)\uparrow}) \bmod p^M$ and $j(E^{(M)\uparrow}) \bmod p^M$.
Procedure:

```
1: $x :=$ any lift of $j(E)$ to $R$
2: for $(i := 1 \ ; \ i < M \ ; \ i := i + 1)$ {
3:     find $y \in R$ satisfying $\Phi_p(x, y) \equiv 0 \bmod p^{i+1}$ and $y \equiv x^p \bmod p$.
4:     $x := y$ ;
5: }
6: find $y \in R$ satisfying $\Phi_p(x, y) \equiv 0 \bmod p^M$ and $y \equiv x^p \bmod p$.
7: return $x$ and $y$ ;
```

At Steps 3 and 6, we use Newton's root finding algorithm. Then the running time of the above algorithm is $O(M^{\mu+1}N^{\mu})$. The space complexity is clearly $O(MN)$.

2.2 Computing the Leading Coefficient of the Verschiebung

For notational simplicity, assume that we have obtained J_0 and J_1 of j-invariants of canonical lifts of E and $E^{(1)}$, respectively. We omit the superscript (1) in $V_p^{(1)}$. The purpose of this section is to compute $\mathrm{lc}(V_p^{\uparrow})^2$.

First we consider the case $p \geq 5$. We use

$$Y^2 = X^3 + A_i X + B_i \ \text{ where } \ A_i := \frac{3J_i}{1728 - J_i}, \ \ B_i := \frac{2J_i}{1728 - J_i}$$

as the Weierstrass model of $E^{(i)\uparrow}$ for $i = 0$, 1. Assume that we have obtained

$$H(X) := \prod_{P \in (\mathrm{Ker} V_p^{\uparrow} - \{\mathcal{O}\})/\pm 1} (X - \xi(P)) \tag{2.4}$$

[†11] Indeed, an injective map $f \in \mathrm{Map}(R, R)$ may well have zero derivative. The following example is due to Dieudonné[14, §8]. Fix the set S of complete representatives of R/pR. Define f by $f\left(\sum_{n=0}^{\infty} a_n p^n\right) := \sum_{n=0}^{\infty} a_n p^{2n}$ where $a_n \in S$. Clearly, f is injective. Since $|f(x+h) - f(x)|_p \leq |h|_p^2$, it is also obvious that f' vanishes identically.

where $\xi(P)$ is the x-coordinate of P. Then by Vélu's formulae[54], we can express the Weierstrass model $Y^2 = X^3 + \alpha X + \beta$ of $E' := E^{\uparrow}/\mathrm{Ker} V_p^{\uparrow}$ by A_1, B_1 and coefficients of H. Vélu's formulae also give the explicit form of $u \in \mathrm{Isog}(E^{(1)\uparrow}, E')$ but the fact $\mathrm{lc}(u) = 1$ is enough for our purpose. By construction, $\mathrm{Ker} u = \mathrm{Ker} V_p^{\uparrow}$. Hence there exists $\lambda \in \mathrm{Isog}(E', E^{\uparrow})$ satisfying $V_p^{\uparrow} = \lambda \circ u$ by Silverman[51, III.4.11].

$$\begin{tikzcd}
E^{(1)\uparrow} \arrow[rr, "V_p^{\uparrow}"] \arrow[dr, "u"'] & & E^{\uparrow} \\
& E' \arrow[ur, "\exists\lambda"'] &
\end{tikzcd} \tag{2.5}$$

Note that all the curves appearing in (2.5) are defined over a field of characteristic zero. Therefore, all the isogenies are separable. Comparing degrees, we see that λ is an isomorphism. Hence there exists $\gamma \in K^{\times}$ so that $\lambda(X, Y) = (\gamma^2 X, \gamma^3 Y)$. Comparing the Weierstrass forms, we have $\gamma^2 = \frac{\alpha/\beta}{A_0/B_0}$. On the other hand $\mathrm{lc}(V_p^{\uparrow}) = \mathrm{lc}(u)\mathrm{lc}(\lambda) = \gamma^{-1}$. Thus, we obtain the desired value $\mathrm{lc}(V_p^{\uparrow})^2$.

So, the problem is how to find $H(X)$ in (2.4). Let K^{ur} be the maximal unramified extension of K and R^{ur} its valuation ring. In general, we denote the p-th division polynomial of an elliptic curve $\mathcal{E}$ by $\Psi_p(X, \mathcal{E})$. In the case of odd p, the following lemma is crucial.

Lemma 1 ([45, Cor. 3.3]). *Let $p \geq 3$. Then* $\mathrm{Ker} V_p^{\uparrow} = E^{(1)\uparrow}[p] \cap E^{(1)\uparrow}(R^{\mathrm{ur}})$.

Hence H is the unique monic polynomial of degree $\frac{p-1}{2}$ such that H divides $\Psi_p(X, E^{(1)\uparrow})$ and such that $\pi(H)$ is square free. Since $E^{(1)}$ is ordinary, $\mathrm{Ker} V_p = E^{(1)}[p]$ and $\Psi_p(X, E^{(1)})$ is of inseparable degree p by Cassels[7, Theorem I]. Therefore, $\pi(H(X)) = \Psi_p(X, E^{(1)})^{1/p}$. Thus we cannot apply Hensel's lemma to lift $\pi(H)$ to a factor of $\Psi_p(X, E^{(1)\uparrow})$ because $\pi(H(X))$ and $\frac{\pi(\Psi_p(X, E^{(1)\uparrow}))}{\pi(H(X))}$ are not co-prime. We need the following modified version of Hensel's lemma.

Lemma 2 ([45, §2]). *Let p be an odd prime. For a given $U \in R[X]$ whose reduction modulo p is inseparable, put $t := \mathrm{ord}_p \frac{dU}{dX}$. Let $h \in R[X]$ be a monic polynomial satisfying the following conditions:*

(1) $\pi(h)$ is square free.
(2) $\pi(h)$ is relatively prime to $\pi\left(p^{-t} \frac{dU}{dX}\right)$.
(3) There exists $g \in R[X]$ and $u \in \mathbf{N}$ such that $\mathrm{ord}_p(U - gh) \geq u + t$.

Then we can lift $\pi(h)$ to a monic factor H of U such that $H \equiv h \bmod p$.

Since $E^{(1)\uparrow}$ is the canonical lift of $E^{(1)}$, we can prove that $U := \Psi_p(X, E^{(1)\uparrow})$ and any lift h of $\Psi_p(X, E^{(1)})^{1/p} \in \mathbf{F}_q[X]$ satisfies the above conditions ([45, Lemma 3.8]). The complexity of the above process is $O((MN)^{\mu})$ in time and $O(MN)$ in space.

In the case of $p = 3$, the algorithm is almost the same. However, we use the Weierstrass equation $Y^2 = X^3 + a_2 X^2 + a_6$. See Fouquet, Gaudry, Harley[19, §7].

Again, the complexity of the above process is $O((MN)^{\mu})$ in time and $O(MN)$ in space.

However, in the case of $p = 2$, there is an essential difficulty with the above method: Lemma 1 no longer holds for $p = 2$. Indeed, there are two non-trivial points in $E^{(1)\uparrow}[2] \cap E^{(1)\uparrow}(R^{\mathrm{ur}})$ whereas $\mathrm{Ker}V_2^{\uparrow}$ has only one non-trivial point. In order to choose the correct point, we utilize Diagram (2.5). Let Q be the non-trivial point in $\mathrm{Ker}V_2^{\uparrow}$. Since λ is an isomorphism, $j(E^{(1)\uparrow}/\langle Q\rangle) = j(E^{\uparrow})$. The problem is how to compute the X-coordinate $\xi(Q)$ of Q in deterministic polynomial time. There are two methods.

The method of Fouquet, Gaudry, Harley[19] is to find the root of the 2-division polynomial using Newton's root finding algorithm with the *correct initial value.* They use $Y^2 + XY = X^3 + A_1$ for the Weierstrass model of $E^{(1)\uparrow}$ where $A_1 \in R$ is determined so that its j-invariant is $j(E^{(1)\uparrow})$. Newton's root finding algorithm is used here, too. Let S be the unique non-trivial point of $E^{(1)\uparrow}[2] \cap \mathrm{Ker}\pi$. Then, $E^{(1)\uparrow}[2] \cap E^{(1)\uparrow}(R^{\mathrm{ur}}) = \{\mathcal{O}, Q, Q+S\}$. Note $P \in E^{(1)\uparrow}[2]$ if and only if $\psi(\xi(P)/2) = 0$ where $\psi(X) = 8X^3 + X^2 + A_1$. From this, we see $\mathrm{ord}_p\xi(S) = -2$ and hence $j(E^{(1)\uparrow}/\langle Q\rangle) \not\equiv j(E^{(1)\uparrow}/\langle Q+S\rangle) \bmod 8$. With some computations, they proved that $\xi(Q) = 2z$ where $z \in R^{\times}$ is the root $\psi(X) = 0$ obtained by Newton's root finding algorithm taking the initial value $j(E^{\uparrow})^{-1} \bmod 4$. Note that $j(E^{(1)\uparrow}) \bmod 2^M$ is sufficient to obtain $z \bmod 2^M$. Then Vélu's formulae yield

$$\mathrm{lc}(V_p^{\uparrow})^2 = \frac{1 - 504z + 19008A_1}{1 + 240(z + 12z^2)(1 + 864A_1)}. \tag{2.6}$$

Note z, $A_1 \in R$.

On the other hand, Skjernaa[52] gives an explicit formula for $\xi(Q)$. Take

$$y^2 + xy = x^3 - \frac{36}{j(E^{(i)\uparrow}) - 1728}x - \frac{1}{j(E^{(i)\uparrow}) - 1728}$$

as the Weierstrass model of $E^{(i)\uparrow}$. Put $J_i := j(E^{(i)\uparrow})$ and let $Y^2 + XY = X^3 + \alpha X + \beta$ be the Weierstrass model of $E^{(1)\uparrow}/\langle Q\rangle$ obtained by Vélu's formulae. Explicitly,

$$\alpha = -\frac{36}{J_1 - 1728} - 5t, \quad \beta = -\frac{1}{J_1 - 1728} - (1 + 7\xi(Q))t \tag{2.7}$$

where $t := 3\xi(Q)^2 - \frac{36}{J_1 - 1728} + \frac{\xi(Q)}{2}$. Then $j(E^{(1)\uparrow}/\langle Q\rangle) = j(E^{\uparrow})$ explicitly yields a polynomial $u \in \mathbf{Z}[J_0, J_1][z]$ satisfying $u(\xi(Q)/2) = 0$. On the other hand, $Q \in E^{(1)\uparrow}[2]$ implies $v(\xi(Q)/2) = 0$ where

$$v(z) := 8(J_1 - 1728)z^3 + (J_1 - 1728)z^2 - 72z - 1. \tag{2.8}$$

Evaluating $\gcd(u, v)$,[†12] one finds

$$\frac{\xi(Q)}{2} = -\frac{J_0^2 + 195120J_0 + 4095J_1 + 660960000}{8(J_0^2 - J_1(512J_0 - 372735) + 563760J_0 + 8981280000)}.$$

[†12] This explains why we work symbolically over $\mathbf{Z}[J_0, J_1]$, not numerically over $R/p^M R$. Because $(R/p^m R)[X]$ is not a UFR for $m \geq 2$, the notion of the gcd is lost here.

However, in order to evaluate $\frac{\xi(Q)}{2} \bmod 2^M$, we need $J_0 \bmod 2^{M+12}$ and $J_1 \bmod 2^{M+12}$. See Skjernaa[52, Lem. 5.1] for details. Eventually, (2.7) gives

$$\mathrm{lc}(V_2^{\uparrow})^2 = \frac{1-48\alpha}{1+864\beta-72\alpha}. \tag{2.9}$$

Note $\frac{\xi(Q)}{2} \in R^\times$ by (2.8). Hence α, $\beta \in R$ by (2.7).

The computational complexities of both methods are $O((MN)^\mu)$ in time and $O(MN)$ in space.

2.3 Norm Computation

The norm computation, which looks quite simple, is in fact a troublesome task. Let $a \in R^\times$ and assume we know $a \bmod p^M$. Our problem is how to compute $N_{K/\mathbf{Q}_p}(a) \bmod p^M$ efficiently. We keep in mind that $M = N/2 + O(1)$ in the context of point counting of elliptic curves. Let $\theta \in R^\times$ be a generator of $K/\mathbf{Q}_p$ and F the monic minimal polynomial of θ over $\mathbf{Q}_p$. There exists $A(X) \in R[X]$ such that $\deg A < N$ and $A(\theta) = a$. Then $N_{K/\mathbf{Q}_p}(a)$ is the resultant of A and F. One might expect that the resultant algorithm using pseudo remainder sequences (e.g. Cohen[9, Algorithm 3.3.7]) work. There are at least two problems: First, pseudo divisions give rise to coefficient explosion. We have to know the precision of intermediate arithmetic operations to ensure that the result is accurate mod p^M. Another problem is that even if we could bound the precision of the intermediate process by $O(M)$, to compute the pseudo remainder sequence, one needs $O(N^2M^\mu)$ bit operations. This is still slow in practical applications. Here we present an "analytic" algorithm from [46].

First assume $\mathrm{ord}_p(a-1) > \frac{1}{p-1}$. Then

$$N_{K/\mathbf{Q}_p}(a) = \exp(\mathrm{Tr}_{K/\mathbf{Q}_p}(\log a)). \tag{2.10}$$

Note exp and log in (2.10) converge under this assumption.[†13] The dominant step when evaluating the right hand side is the evaluation of log. The straightforward evaluation of $\log a = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}(a-1)^n$ would need $O(M)$ multiplications over R/p^MR. Put $m := [\sqrt{M}]$ for simplicity. Then $\mathrm{ord}_p(a^{p^m}-1) > m + \frac{1}{p-1}$ and $a^{p^m} \bmod p^{M+m}$ is well defined. Here $O(m)$ multiplications over $R/p^{M+m}R$ are necessary to compute a^{p^m}. We can obtain $\log(a^{p^m}) \bmod p^{M+m}$ with $O(m)$ multiplications over $R/p^{M+m}R$. Then, $(\log a) \bmod p^M$ is given by $p^{-m}(\log a^{p^m} \bmod p^{M+m})$. Since $m = O(\sqrt{M})$, we need $O(M^{\mu+1/2}N^\mu)$ bit operations and $O(MN)$ space to evaluate $N_{K/\mathbf{Q}_p}(a)$ when a is close to unity.

[†13] The p-adic exponential function and the p-adic logarithm function are defined by the power series $\exp(x) := \sum_{n=0}^{\infty} \frac{x^n}{n!}$ and $\log(y) = \sum_{n=1}^{\infty} \frac{(-1)^n}{n}(y-1)^n$, respectively. For basic properties, see e.g. Koblitz[32, Chap. 4]. We also need the following fact to prove (2.10): Let $F(X) \in \mathbf{Q}_p[[X]]$ and $a \in K$. Assume $F(a)$ converges. Then for any $\rho \in \mathrm{Gal}(K/\mathbf{Q}_p)$ we have $\rho(F(a)) = F(\rho(a))$. This follows from continuity of ρ.

Remark 1. It was pointed out by R. Harley that if we accept $O(M^{4/3}N)$ space complexity, then the time complexity is reduced to $O(M^{\mu+1/3}N^{\mu})$. Indeed, instead of $[\sqrt{M}]$, we put $m := [M^{1/3}]$. We compute x^{p^m} with $O(m)$ multiplications and then evaluate the first $O(M^{2/3})$ terms of the expansion of log with $O(M^{1/3})$ multiplications and $O(M^{1/3})$ storage over $R/p^{M+m}R$ by, say, the Paterson-Stockmeyer algorithm[43].

Let us consider the case of general $a \in R^{\times}$. Let $T \in \mathrm{Map}(\mathbf{F}_q, R)$ be the Teichmüller lifting map. For an odd p, we utilize[†14]

$$N_{K/\mathbf{Q}_p}(a) = T(N_{\mathbf{F}_q/\mathbf{F}_p}(\alpha))N_{K/\mathbf{Q}_p}(T(\alpha^{-1})a).$$

where $\alpha := \pi(a)$. Since $\mathrm{ord}_p(T(\alpha^{-1})a - 1) \geq 1 > \frac{1}{p-1}$, we use (2.10) for $T(\alpha^{-1})a$. The best method to compute $T(\alpha)$ depends on M.[†15] In the case of small M (say, $M < N$), we use the following algorithm, whose complexity is $O(\max(N^{2\mu}, M^{\mu+1}N^{\mu}))$ in time and $O(MN)$ in space.

Algorithm 2. Teichmüller lift by powering.
Input: $a \in \mathbf{F}_q$
Output: $T(a) \bmod p^M$
Procedure:

```
$x := a^{\mathrm{rem}(N-M+1,N)}$ ;
for ($i := 1$ ; $i < M$ ; $i := i+1$) {
    lift $x$ to $R/p^{i+1}R$
    $x := x^p \bmod p^{i+1}$ ;
}
return $x$ ;
```

In the case of large M, we find the root of $X^{1-q} - 1 = 0$[†16] by applying Newton's root finding algorithm with initial value α. This amounts to $O(M^{\mu}N^{\mu+1})$ bit operations with $O(MN)$ space.

In the case of $p = 2$, it is not necessarily true that $\mathrm{ord}_p(T(\alpha^{-1})a - 1) > 1$. However, either (2.6) or (2.9) shows $\mathrm{lc}(V_2^{\uparrow})^2 \equiv 1 \bmod 8$. So, as far as point counting is concerned, we can simply evaluate[†17] (2.10) at $a = \mathrm{lc}(V_i^{\uparrow})^2$.

In conclusion, the time complexity of norm computation for point counting on elliptic curves is $O(N^{2\mu+1/2})$ for $p = 2$ and $O(N^{2\mu+1})$ for $p \geq 3$. The space complexity is $O(N^2)$ in both cases.

[†14] Note $N_{K/\mathbf{Q}_p}(T(\alpha)) = T(N_{\mathbf{F}_q/\mathbf{F}_p}(\alpha))$ for $\alpha \in \mathbf{F}_q$.

[†15] Of course the break-even point is implementation dependent. However, for the application to point counting of elliptic curves, the repeated p-th powering seems to be faster.

[†16] The iteration process to solve $X^{q-1} - 1 = 0$ is $x \leftarrow \frac{(q-2)x^{q-1}+1}{(q-1)x^{q-2}}$ while that to solve $X^{1-q} - 1 = 0$ is $x \leftarrow x - \frac{1}{1-q}(x - x^q)$ which does not contain a division by an element of R.

[†17] We can do this even better by using $\log\frac{1+x}{1-x} = \sum_{n=1}^{\infty}\frac{x^{2n-1}}{2n-1}$. Note division by the odd number $2n-1$ does not lose 2-adic precision.

3 Arithmetic Geometric Mean

In [24], Harley announced the point counting algorithm based on the arithmetic geometric mean(AGM). Although its computational complexity is $O(N^2)$ in space and $O(N^{2\mu+1})$ in time, the O-constants are much smaller than those for the algorithm described in the previous section. In practice, the one variable version of the AGM method runs much faster than a naive implementation of the two variable AGM iteration. However, for simplicity, we shall work with the two variable AGM. See Harley et al.[25] for details. It should be noted that the techniques of AGM based point counting are the subject of a U.S. patent(pending) by ArgoTech.

For real numbers $a \geq b > 0$, put

$$\mathcal{M}(a,b) := \left(\frac{a+b}{2}, \sqrt{ab}\right).$$

Given $a_0 \geq b_0 > 0$, define two sequences $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ by

$$(a_{n+1}, b_{n+1}) := \mathcal{M}(a_n, b_n).$$

Then, $\lim\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty} b_n$ (both limits exist). This common value is called the AGM of a_0 and b_0. The AGM is closely related to elliptic curves. Some of them go back to Gauss. See e.g. Borwein and Borwein[4].

In the rest of this section, we will only consider the case $p = 2$. So, K is the unramified extension of $\mathbf{Q}_2$ of degree N and $q = 2^N$. For $a \in 1 + 8R$, we denote the unique element $b \in 1 + 4R$ satisfying $b^2 = a$ by $\sqrt{a}$. Then, given a, $b \in R^\times$ with $\frac{b}{a} \in 1 + 8R$, we see that $a' := \frac{a+b}{2}$ and $b' := a\sqrt{\frac{b}{a}}$ also belong to $R^\times$ and that $\frac{b'}{a'} \in 1 + 8R$. (Moreover, if $a \in 1 + 4R$ and $b \in 1 + 4R$, then $a' \in 1 + 4R$ and $b' \in 1 + 4R$.) Hence, as is in the real case, we can repeat the AGM process. Put

$$\mathcal{M}(a,b) := \left(\frac{a+b}{2}, a\sqrt{\frac{b}{a}}\right).$$

Let $E_{a,b}$ be the curve $y^2 = x(x-a^2)(x-b^2)$. Note that $E_{a,b}$ is not a minimal Weierstrass model in general. The following lemma gives a Weierstrass model of $\pi(E_{a,b})$.

Lemma 3. *Let a, $b \in 1 + 4R$ satisfy $\frac{b}{a} \equiv 1 \bmod 8$. Define $\gamma \in R$ by $3\gamma^2 - 2(a^2+b^2)\gamma + a^2b^2 = 0$ and $\gamma \equiv 1 \bmod 8$. Then, the change of variables $(X,Y) \to \left(\frac{X-\gamma}{4}, \frac{Y-(X-\gamma)}{8}\right)$ transforms $E_{a,b}$ to $Y^2 + XY = X^3 + rX^2 + s$ with $r \in 2R$ and $s \in R^\times$, which is a minimal Weierstrass model of $E_{a,b}$. Moreover $s \equiv \left(\frac{b-a}{8a}\right)^2 \bmod 2$.*

Using the AGM, we can obtain $j(E^\uparrow)$ quickly as follows. First, we observe a relation between the AGM and a 2-isogeny.

Proposition 1. *Let a, $b \in R^\times$ and $\frac{b}{a} \in 1 + 8R$. Then the map $\mathcal{F}$ defined by*

$$\mathcal{F} \,:\, (x, y) \to \left(\frac{1}{4}\frac{y^2}{x^2} + \left(\frac{a+b}{2}\right)^2, -\frac{1}{8}\frac{y(a^2b^2 - x^2)}{x^2} \right) \tag{3.1}$$

is a 2-isogeny from $E_{a,b}$ to $E_{\mathcal{M}(a,b)}$ whose kernel is $\langle(0,0)\rangle$. In particular,

$$\Phi_2(j(E_{\mathcal{M}(a,b)}), j(E_{a,b})) = 0. \tag{3.2}$$

Proof. Let $Q_{a,b}$ be the elliptic curve defined by $y^2 = x^3 + 2(a^2+b^2)x^2 + (a^2 - b^2)^2 x$. As is described in Silverman[51, III.4.5], the map defined by $(x, y) \to \left(\frac{y^2}{x^2}, \frac{y(a^2b^2-x^2)}{x^2}\right)$ is a 2-isogeny from $E_{a,b}$ to $Q_{a,b}$ whose kernel is $\langle(0,0)\rangle$. Then, the curve $Q_{a,b}$ is isomorphic to $E_{\mathcal{M}(a,b)}$ with respect to the map $(x, y) \to \left(\frac{x}{4} + \left(\frac{a+b}{2}\right)^2, -\frac{y}{8}\right)$. □

Let $c \in \mathbf{F}_q^\times$. Let E be the elliptic curve defined by $y^2 + xy = x^3 + c$. Take any lift $u \in R$ of $c^{1/2}$ $(= c^{2^{N-1}})$ and put $a_0 := 1 + 4u$, $b_0 := 1 - 4u$. Then, $\pi(E_{a_0,b_0}) \cong E$ by Lemma 3. Define two sequences $\{a_n\}_{n=0}^\infty$ and $\{b_n\}_{n=0}^\infty$ as in the real case: $(a_{n+1}, b_{n+1}) := \mathcal{M}(a_n, b_n)$. A straightforward computation shows $j(E_{\mathcal{M}(a,b)}) \equiv j(E_{a,b})^2 \bmod 2$ for any a, $b \in R$ with $\frac{b}{a} \in 1 + 8R^\times$. Therefore, $j(E_{a_n,b_n}) \equiv j(\sigma^n(E)^\uparrow) \bmod 2^{n+1}$ by (3.2) and Theorem 3.

Remark 2. Two sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$ converge provided that $\frac{b_0}{a_0} \in 1 + 16R$ by Henniart, Mestre[26]. In our case, $\frac{b_0}{a_0} \in 1 + 8R^\times$ and they do not converge. Only $j(E_{a_n,b_n}) - j(\sigma^n(E)^\uparrow)$ converges to zero as $n \to \infty$.

The AGM also provides us with a very efficient way to compute $\mathrm{lc}(V_2^\uparrow)$. Assume $j(E_{a,b}) = j(E^\uparrow)$. Then $\sigma(E_{a,b})$ is a Weierstrass model of $E^{(1)\uparrow}$. By Proposition 1, there exists an isomorphism $u \,:\, E_{\mathcal{M}(a,b)} \to \sigma(E_{a,b})$ satisfying $\mathrm{Fr}_2^\uparrow = u \circ \mathcal{F}$ where $\mathcal{F}$ is defined by (3.1). Hence $\mathrm{lc}(V_2^\uparrow) = \mathrm{lc}(\widehat{\mathcal{F}})\mathrm{lc}(u)^{-1}$. We know an explicit formula for $\widehat{\mathcal{F}}$ (see Silverman[51, III.4.5] again). The tricky part is the computation of $\mathrm{lc}(u)$. This is accomplished with some diagram chasing and we have $\mathrm{lc}(V_2^\uparrow) = \pm\frac{\sigma(a)}{(a+b)/2}$. Actually, we have only approximate values of a and b. So, we need to determine how much precision is necessary to retrieve $\mathrm{Tr}(\mathrm{Fr}_q)$. The result is as follows:

Theorem 4. *Let $m \geq 3$. Assume a, $b \in R^\times$ satisfies $\frac{b}{a} \in 1+8R^\times$ and $j(E_{a,b}) \equiv j(E^\uparrow) \bmod 2^m$. Set $(\alpha_1, \beta_1) := \mathcal{M}(a, b)$ and $(\alpha_2, \beta_2) := \mathcal{M}(\alpha_1, \beta_1)$. Then*

$$\mathrm{Tr}(\mathrm{Fr}_q^\uparrow) \equiv N_{K/\mathbf{Q}_2}\left(\frac{\alpha_1}{\alpha_2}\right) \bmod 2^{\min(N, m+2)}.$$

Summing up, we obtain the following algorithm. For simplicity, we assume $N \geq 6$ in order that $M \leq N$ in the following algorithm.

Algorithm 3. Computing $\operatorname{Tr}\operatorname{Fr}_q$ by AGM.
Input: An elliptic curve $y^2 + xy = x^3 + c$ ($c \in \mathbf{F}_q^\times$).[†18]
Output: $\operatorname{Tr}(\operatorname{Fr}_q)$
Procedure:

```
1: $u :=$ any lift of $c^{1/2}$ to $R$ ;
2: $a := 1 + 4u$ ; $b := 1 - 4u$ ;
3: $M := \lceil N/2 \rceil + 2$ ;
4: for ($i := 0$ ; $i < M - 2$ ; $i := i + 1$) {
5:     $(a, b) := \mathcal{M}(a, b)$ ;
6: }
7: $s := (a + b)/2$ ;
8: return $t \in \mathbf{Z}$ satisfying $t \equiv N_{K/\mathbf{Q}_2}\left(\frac{a}{s}\right) \bmod 2^M$ and $|t| < 2\sqrt{q}$ ;
```

4 Inverse Frobenius Substitution

In this section, we observe that fast evaluation of the Frobenius substitution on $R/p^M R$ with $M \in \mathbf{N}$ improves the algorithm described in Section 2. In order to evaluate the Frobenius substitution, our algorithm utilizes a root of unity. It computes $\sigma(x)$ for $x \in R/p^M R$ with $O((MN)^\mu)$ bit operations and precomputation (which depends only on K). The resulting point counting algorithm runs in $O(N^{2\mu+0.5})$ bit operations with $O(N^2)$ memory (not including precomputation).

Let $\overline{\theta}$ be a generator of $\mathbf{F}_q/\mathbf{F}_p$ and $f \in \mathbf{F}_p[X]$ its monic minimal polynomial. We recall that in practice f is chosen to be a low weight polynomial. Take a lift $F \in R[X]$ of f such that the weight of F is equal to the weight of f. Then, $R \cong \mathbf{Z}_p[X]/\langle F \rangle$. As before, we denote the Teichmüller lifting map by T. Put $\psi := T(\overline{\theta})$ and let G be its monic minimal polynomial. Then, we have another realization of R, namely, $R \cong \mathbf{Z}_p[X]/\langle G \rangle$. In general, G is a dense polynomial. As was described in Section 1.2, this implies that a multiplication over $\mathbf{Z}_p[X]/\langle G \rangle$ is about three times slower than that of $\mathbf{Z}_p[X]/\langle F \rangle$. However, we can easily compute the action of σ^{-1} on $\mathbf{Z}_p[X]/\langle G \rangle \cong \mathbf{Z}_p[\psi]$. Explicitly, for a given $\gamma := \sum_{i=0}^{N-1} c_i \psi^i \in \mathbf{Z}_p[\psi]$ it is true that

$$\sigma^{-1}(\gamma) = \sum_{j=0}^{p-1} \left(\sum_{0 \le pi+j < N} c_{pi+j} \psi^i \right) H_j(\psi)$$

where

$$H_j(X) := \operatorname{rem}(X^{jp^{N-1}}, G), \tag{4.1}$$

[†18] For $c \in \mathbf{F}_4^\times$, one cannot apply Theorem 3. However, P. Gaudry pointed out that this algorithm works also for $c \in \mathbf{F}_4^\times$. The correctness for such c is proved by a more careful analysis on derivatives of the modular polynomial.

hence $H_j(\psi) = \sigma^{-1}(\psi^j)$. In this section, we assume that we have precomputed G and $H_1, \ldots, H_{p-1}$.[†19] Thus, the complexity of computing $\sigma^{-1}(\cdot) \bmod p^M$ is $O((MN)^\mu)$ in time and $O(MN)$ in space (not including precomputation).

Let us see how the Frobenius substitution reduces the time complexity of the Teichmüller lifting. Let $a \in \mathbf{F}_q$. Assume we omit the first step of Algorithm 2. Then it terminates with the output $\sigma^{M-1}(T(a))$. The purpose of Step 1 is to compensate the action of σ. This can also be done as follows:

Algorithm 4. Computing the Teichmüller lift with the Frobenius substitution (naive version).
Input: $a \in \mathbf{F}_q$, $M \in \mathbf{N}$
Output: $T(a) \bmod p^M$
Procedure:

```
$x_1 := a$ ;
for $(i := 1 ; i < M ; i := i + 1)$ {
    lift $x_i$ to $R/p^{i+1}R$
    $x_{i+1} := \sigma^{-1}(x_i^p)$ ;
}
return $x_M$ ;
```

This algorithm is slower than Algorithm 2 unless M is very small. However, observe that $x_i \equiv T(a) \bmod p^i$ holds for each i. During the computation of x_{i+1}, we obtain x_i^{p-1} and x_i^p. Then the Taylor expansion of x^p around $x = x_i$ gives $x_{i+1}^p \bmod p^{i+2}$ with only one multiplication over $R/p^{i+2}R$. More specifically, put $\delta_{i,j} := x_i - x_j$ for $i \geq j$. We have $\delta_{i,j} \equiv 0 \bmod p^j$. On the other hand,

$$\delta_{i+1,j} = \sigma^{-1}((\delta_{i,j} + x_j)^p) - x_j \equiv p\sigma^{-1}(x_j^{p-1}\delta_{i,j}) + (\sigma^{-1}(x_j^p) - x_j) \bmod p^{2j}.$$

Letting $d_{i,j} := p^{-j}\delta_{i,j} \in R$ and $z_j := p^{-j}(\sigma^{-1}(x_j^p) - x_j) \in R$, we obtain

$$d_{i+1,j} \equiv p\sigma^{-1}(x_j^{p-1})\sigma^{-1}(d_{i,j}) + z_j \bmod p^j. \tag{4.2}$$

Let $j \leq i < k \leq 2j$. In order to obtain $\delta_{i+1,j} \bmod p^k$ for $k \leq 2j$, we have only to perform arithmetic operations in the right hand side of (4.2) over $R/p^{k-j}R$, where complexities of arithmetic operations are much smaller than those of R/p^kR. Now we state our algorithm. We introduce a new parameter W.

Algorithm 5. Computing the Teichmüller lift with the Frobenius substitution
Input: $a \in \mathbf{F}_q$, $M \in \mathbf{N}$, $W \in \mathbf{N}$
Output: $T(a) \bmod p^M$
Procedure:

```
$x :=$ lift of $a$ to $R/p^WR$.
```

[†19] For small N (say, $N < 200$), we can obtain $G(Y) \bmod p^M$ by computing $\prod_{i=0}^{N-1}(Y - \psi^{p^i})$ in $((\mathbf{Z}/p^M\mathbf{Z})[X]/\langle F \rangle)[Y]$ with $O(M^\mu N^{\mu+2})$ bit operations and $O(N^2M)$ memory. Since the precomputation is required once for each N (and f), this naive method is not a problem. For large values of N, see Appendix B.

```
 2: for ($i := 0$ ; $i < W - 1$ ; $i := i + 1$) {
 3:     $x := \sigma^{-1}(a^p)$ ;
 4: }
 5: $\Delta := px^{p-1} \bmod p^W$ ;
 6:
 7: for ($m := 1$ ; $mW < M$ ; $m := m + 1$) {
 8:     lift $x$ to $R/p^{(m+1)W}R$
 9:     $d := 0$ ;
10:     $z := \sigma^{-1}(x^p) - x \bmod p^{(m+1)W}$ ;
11:     $z := (p^{-mW}z) \bmod p^W$ ;
12:     for ($i := 0$ ; $i < W$ ; $i := i + 1$) {
13:         $d := \Delta * \sigma^{-1}(d) + z \bmod p^W$ ;
14:     }
15:     $x := x + p^{mW}d$ ;
16: }
17: return $x$ ;
```

Note that all the arithmetic operations in Step 13 are performed mod p^W. The running time of the above algorithm is

$$O(\max(W(NW)^{\mu}, (M/W)(MN)^{\mu}, M(NW)^{\mu})).$$

Taking $W := O(M^{\mu/(\mu+1)})$, we see it runs in $O(M^{\mu+1/(\mu+1)}N^{\mu})$ bit operations. The same idea applies to the p-th modular polynomial Φ_p. But we need to modify the above algorithm so that it works with a two variable polynomial. See [46] for details. The time complexity of the resulting algorithm for computing the j-invariant of the canonical lift modulo p^M is $O(M^{\mu+1/(\mu+1)}N^{\mu})$ bit operations. Consequently, if we accept precomputation of the minimal polynomial G of ψ and the polynomials defined in (4.1), we obtain a point counting algorithm whose complexity is $O(N^{2\mu+0.5})$ in time and $O(N^2)$ in space.

Remark 3. It is pointed out by P. Gaudry that the above method is applicable to a variant of the modular equation (at least for $p \leq 5$) in Borwein and Borwein[4, Chap. 4] which is closely related to classical elliptic integrals. This further reduces the time complexity by a constant factor.

Remark 4. The actual choice of W depends on the particular implementation. For a cryptographic application (i.e. when $p = 2$ and $N < 300$), using the CPU word size as W, regardless of N, would often give the best results. In this case, we can perform each ring operation over $\mathbf{Z}/p^W\mathbf{Z}$ without multi-precision integer arithmetic.

Remark 5. In theory, the growth rate of the time complexity of the above algorithm is $O(N^{2.5+\varepsilon})$ if we adopt FFT based multiplications. However, if we apply the Schönhage-Strassen algorithm for integer multiplication and the Cantor-Kaltofen algorithm for polynomial multiplication, p^W should be sufficiently large so that the running time of the Schönhage-Strassen algorithm behaves almost linearly with respect to W. This implies N should be greater than something like 10^6, which seems to be a non-feasible size. Alternatively, we use a similar technique to Schönhage[47, §2], where a polynomial is encoded to a large integer.[†20] This increases memory requirements a bit, but it lowers the value of N where FFT based multiplications become efficient.[†21]

Remark 6. Note that the dominant step of the resulting point counting algorithm is an evaluation of logarithm involved in the norm computation. Recently, Kim et al.[29] proposed an algorithm using the Gaussian normal basis(GNB) to represent elements of R. Such a basis does not necessarily exist. But if $\mathbf{F}_q/\mathbf{F}_p$ has a GNB, then $K/\mathbf{Q}_p$ also has a GNB and we can evaluate the inverse of the Frobenius substitution with $O(N)$ bit operations without precomputations. Moreover, the norm computation can be done with $O(\log N)$ multiplications over $R/p^M R$.[†22] Hence, in the case where a GNB exists, the time complexity of elliptic curve point counting is $O(N^{2\mu+1/(\mu+1)})$ bit operations.

5 Point Counting for non Elliptic Curves

We now consider the problem of counting $\mathbf{F}_q$-rational points on non elliptic curves. For results on l-adic methods up to 1996, see Poonen[44, §5]. In theory, Schoof's algorithm generalizes to polynomial time algorithms, but their efficiency is questionable. Nevertheless, there are some implementation details for hyperelliptic curves of genus two. Harley and Gaudry[23] computed the number of $\mathbf{F}_q$-rational points of the Jacobians of hyperelliptic curves of genus two[†23] for $q = 3^{30}$ ($\log_2 q \approx 47.5$) and for $q = p = 10^{19} + 51$ ($\log_2 q \approx 63.1$). Matsuo, Chao, Tsujii[37] made some improvements to the Harley-Gaudry method and performed the point counting for $q = (2^{20} - 5)^4$ ($\log_2 q \approx 80.0$). The actual running time is about 26 days with 12GB RAM.

Let us consider p-adic algorithms. Although the notion of a canonical lift is well formulated in the category of ordinary Abelian varieties, it seems difficult to construct canonical lifts of general Abelian varieties. However, there are at least two p-adic algorithms to lift the Frobenius morphism to certain cohomology groups. In addition to these methods, Harley et al.[25] constructed the genus two AGM algorithm for point counting of hyperelliptic curves of genus two over finite fields of characteristic two.

[†20] This technique goes back at least to Exercise 4 of Knuth[30, §4.6]

[†21] We notice that this technique is quite efficient even for not so large N (say, $N \geq 3000$). See Fouquet, Gaudry and Harley[19, §2.4] and Gaudry and Gürel[22, §4.3].

[†22] Under the GNB representation of R, however, a multiplication in $R/p^M R$ would need $O(N^2 M^\mu)$ bit operations, which makes the resulting algorithm too slow. See [29] on how to avoid this difficulty.

[†23] Note that the magnitude of the number of $\mathbf{F}_q$-rational points are about twice of the size of the base field for genus two curves.

Kedlaya[28] constructed a p-adic algorithm to compute the zeta function of an arbitrary hyperelliptic curve over a finite field of odd characteristic. This method computes the action of the Frobenius morphism on the Monsky-Washnitzer cohomology: Monsky and Washnitzer[41], Monsky[39, 40]. (See Koblitz[31, Chap. III] for a quick introduction.) This method is generalized to so-called superelliptic curves by Gaudry and Gürel[22]. The computational complexity of these algorithms is $O(N^{3\mu+\varepsilon})$ in time and $O(N^{3+\varepsilon})$ in space for fixed genus.

Lauder and Wan[34] constructed another algorithm based on exponential sums and Dwork's trace formula. If we apply this algorithm in a straightforward manner to hyperelliptic curves (of a fixed genus), its complexity is $O(N^{3\omega+2\mu}\log N)$ in time and $O(N^8)$ in space. Here ω is the exponent of the number of ring operations in a matrix multiplication.[†24] Although these complexities look large, note that this algorithm works for arbitrary algebraic varieties. In [33], they also constructed an algorithm for Artin-Schreier curves defined by $Y^p - Y = f(X)$ with $f(X) \in \mathbf{F}_q[X, X^{-1}]$. If we fix p and the largest absolute value of powers of X appearing in f, then its time complexity is $O(N^{3\mu+\varepsilon})$ and its space complexity is $O(N^{3+\varepsilon})$. In the case of $p = 2$, this can be used to compute the zeta functions of hyperelliptic curves given by $Y^2 + X^m Y = h(X)$ where $0 \le m < \deg h$.

Recall that the AGM point counting algorithm uses neither modular polynomials nor Vélu's formulae. In the case of $p = 2$, Harley, Gaudry and Mestre designed an AGM point counting algorithm for ordinary hyperelliptic curves of genus two. This algorithm is based on Bost and Mestre[5] where a sequence of (2,2)-isogenous hyperelliptic curves of genus two is constructed using the AGM. The computational complexity is $O(N^3)$ in time and $O(N^2)$ in space. See Harley et al.[25] and Gaudry[21]. The result is impressive: the time to compute the number of rational points for $g = 2$ and $q = 2^{4000}$ is 144 hours with an Alpha/750MHz.

The time complexity of an algorithm which requires $\Omega(N^3)$ memory cannot be $o(N^3)$ even if we accept some precomputation or some restrictions on N. One can naturally ask whether it is possible to design a point counting algorithm for, say, hyperelliptic curves of an arbitrary genus with $o(N^3)$ time complexity. Note, in the case of elliptic curves, we made the assumption that the j-invariants of a given curve do not belong to $\mathbf{F}_{p^2}$. Can we obtain a faster algorithm if we limit ourselves to ordinary curves (or Abelian varieties)? The genus two AGM point counting algorithm suggests that there still should be many improvements in this area (possibly including algorithms for elliptic curves).

[†24] So, $\omega = 3$ for naive multiplications, $\omega = \log_8 7$ for the Strassen algorithm[53] (see also [1, Chap. 6]), $\omega = 2.376$ for the Coppersmith-Winograd algorithm[10].

6 Appendix A: Theoretical Upper Bound for the Running Time of Elkies' Algorithm (joint work with S. Galbraith)

In this appendix, we prove that the largest prime used in the Elkies algorithm is $O((\log q)^{2+\varepsilon})$ for any $\varepsilon > 0$ under GRH. This implies that the time complexity of Elkies' algorithm is $O((\log q)^{3\mu+2+\varepsilon})$.[†25]

In this section, l always stands for prime numbers. Let $E/\mathbf{F}_q$ be an elliptic curve and let χ_E be the Kronecker symbol associated to the quotient field of $\mathrm{End}(E)$. The estimate of the cardinality of $\{\, l \,:\, l < L,\ \chi_E(l) \neq -1 \,\}$ seems to be difficult. Ankeny[2] studied the least quadratic non-residue, but the results of [2] do not seem to give estimates on the second least quadratic non-residue and so on. However, in order to estimate the time complexity of Elkies' algorithm, what we really need is the growth rate of $\prod\limits_{l<L,\chi_E(l)\neq -1} l$ as $L \to \infty$.

Recall that there exist constants c_1, c_2, c_3 such that $c_1 L \le \sum_{l \le L} \log l \le c_2 L$ for all $L \ge c_3$ by Chebyshev's estimate[8].[†26]

Theorem 5. *Let $\varepsilon > 0$. There exist constants c_4 and c_5 depending only on ε with the following property. For any real primitive character χ modulo d where $d \ge c_4$ and for all $L \ge (\log d)^{2+\varepsilon}$,*

$$\sum_{l \le L, \chi(l) \neq -1} \log l \ge c_5 L.$$

Proof. Without loss of generality, we may assume $0 < \varepsilon < 2$. Put $X := c_6^{-1} L$ where c_6 is a constant whose value is determined later. We have

$$\begin{aligned}\sum_{l \le L, \chi(l) \neq -1} \log l \ge & \tfrac{1}{2} \sum_{l \le L} (1 + \chi(l)) \log l \ge \tfrac{1}{2} \sum_{l \le L} (1 + \chi(l)) e^{-l/X} \log l \\ = & \tfrac{1}{2} \left(\sum_{l \le L} e^{-\frac{l}{X}} \log l + \sum_{l} \chi(l) e^{-\frac{l}{X}} \log l - \sum_{l > L} \chi(l) e^{-\frac{l}{X}} \log l \right) \\ = & \tfrac{1}{2} (S_1 + S_2 - S_3).\end{aligned}$$

Then,

$$S_1 \ge \sum_{l \le L/2} e^{-l/X} \log l \ge e^{-L/2X} c_1 \frac{L}{2}$$

[†25] Frey[20, Th. 3.8] states the same result, which is based on the observations by K. Murty and R. Murty communicated in Feb. 2000. But to the best knowledge of the author, their proof is not published. Independently, the author and S. Galbraith discussed the running time of Elkies' algorithm and obtained the following elementary proof in May 2000.

[†26] This can be deduced from the prime number theorem. But, in fact, the prime number theorem is proved via $\lim\limits_{L \to \infty} \frac{1}{L} \sum_{l \le L} \log l = 1$. See e.g. Edwards[16, Chap. 4].

for all $L \geq 2c_3$. By Ankeny[2, Theorem 1], there exist constants c_7, c_8, c_9, c_{10} such that

$$|S_2| \leq c_7 \left(X^{1/2} \frac{\log X \log d}{\log \log d} + \frac{\log d}{\log X} \right) + c_8 X^{1/3} \log d \ \text{ for all } \ X \geq c_9, d \geq c_{10}.$$

For $L \geq (\log d)^{2+\varepsilon}$ (i.e. $\log d \leq (c_6 X)^{1/(2+\varepsilon)}$),

$$|S_2| < c_7 c_6^{1/(2+\varepsilon)} \left(X^{\frac{1}{2+\varepsilon}+\frac{1}{2}} \log X + X^{1/(2+\varepsilon)} \right) + c_8 c_6^{1/(2+\varepsilon)} X^{\frac{1}{3}+\frac{1}{2+\varepsilon}}$$

provided $\log \log d \geq 1$. Hence, there exist constants c_{11} and c_{12} such that $|S_2| < c_{11} L^{1-\varepsilon/9}$ for all $L \geq \max(c_{12}, (\log d)^{2+\varepsilon})$ and $d \geq \max(c_{10}, e^e)$.

Now, we estimate S_3. This is already done in Ankeny's work. Put $\theta(u) = \sum_{L<l\leq u} \log l$. Then,

$$|S_3| \leq \frac{1}{X} \int_L^\infty \theta(u) e^{-u/X} du \leq \frac{c_2}{X} \int_L^\infty u e^{-u/X} du = c_2 e^{-L/X}(L+X).$$

Thus $|S_3| \leq c_2 e^{-c_6} \left(\frac{1+c_6}{c_6} \right) L$. Choose c_6 so that

$$c_5 := \frac{1}{3} e^{-c_6/2} \left(\frac{c_1}{2} - c_2 (1 + c_6^{-1}) e^{-c_6/2} \right) > 0.$$

(Note that c_6 is independent of ε.) Thus $S_1 - S_3 > 3c_5 L$ for $L \geq 2c_3$.

Summing up, we see $S_1 + S_2 - S_3 > 2c_5 L$ for $L \geq \max(c_{13}, (\log d)^{2+\varepsilon})$ with a suitable constant c_{13}. Put $c_4 := \max(c_{10}, e^e, \exp(c_{13}^{1/(2+\varepsilon)}))$. Then $d > c_4$ implies $(\log d)^{2+\varepsilon} > c_{13}$. □

Corollary 1. *Let ε and c_{14} be arbitrary positive real numbers. Then, there exists a constant c_{15} satisfying*

$$\sum_{\chi_E(l) \neq -1, l \leq (\log q)^{2+\varepsilon}} \log l \geq c_{14} \log q$$

for all $q \geq c_{15}$ and all elliptic curves $E/\mathbf{F}_q$.

Proof. Let d be the discriminant of the quotient field of End(E). For $|d| < c_4$, the assertion follows from the prime number theorem for arithmetic progressions. Otherwise, the assertion comes from the above theorem. □

Remark 7. For a fundamental discriminant $d < 0$ of an imaginary quadratic field, let l_d be the least prime which does not remain prime in $\mathbf{Q}(\sqrt{d})$. Under the GRH, there exists a constant $c_{16} > 0$ such that there exists infinitely many d satisfying $l_d > c_{16} \log d \log \log d$. This follows from a similar proof to Montgomery[42, Th. 13.5].

7 Appendix B: A Minimal Polynomial of a Root of Unity

Let $\overline{\theta} \in \mathbf{F}_q$ be a generator of $\mathbf{F}_q/\mathbf{F}_p$ and let $\psi \in R$ be the Teichmüller lift of $\overline{\theta}$. Let G be the monic minimal polynomial of ψ. Here, we present an algorithm to compute $G \bmod p^M$ with $O(M^\mu N^{\mu+1})$ bit operations and $O(NM)$ memory. Let $F(X) := \sum_{n=0}^{N} a_n X^n \in \mathbf{Z}[X]$ be a monic lift of the monic minimal polynomial of $\overline{\theta}$ such that $0 \leq a_n < p$ for $0 \leq n \leq N$. Let $\theta \in R$ be the unique root of $F(X) = 0$ satisfying $\pi(\theta) = \overline{\theta}$. Put $\mathcal{P} := \{f \in \mathbf{Z}_p[X] \,:\, \deg f < N\}$. If $f \in \mathcal{P}$ and $\mathrm{ord}_p f(\psi) \geq i$, then f is divisible by p^i. Note $\mathrm{ord}_p f(\theta) = \mathrm{ord}_p f(\psi)$ because $\theta \equiv \psi \bmod p$. Hence we can define $C \in \mathcal{P}$ by $\theta = C(\psi)$ and $A \in \mathcal{P}$ by $\psi = A(\theta)$. Again, using $\theta \equiv \psi \bmod p$, we see that $F(X) \equiv G(X) \bmod p$ and that $X \equiv C(X) \bmod p$. Our strategy is to successively construct better approximations of G and C.

For f, $g \in \mathbf{Z}_p[X]$ and a monic $h \in \mathbf{Z}_p[X]$, we define $f \underset{h}{\circ} g \in \mathbf{Z}_p[X]$ by $(f \underset{h}{\circ} g)(X) := \mathrm{rem}(f(g(X)), h(X))$. Hence, $(f \underset{G}{\circ} C)(\psi) = f(\theta)$ and $(f \underset{F}{\circ} A)(\theta) = f(\psi)$. Assume we have obtained a monic polynomial $G_1 \in \mathbf{Z}_p[X]$ of degree N and $C_1 \in \mathcal{P}$ satisfying $G_1(\psi) \equiv 0 \bmod p^i$ and $\theta \equiv C_1(\psi) \bmod p^i$ with some $i \in \mathbf{N}$. Then the polynomial $V := G_1 \underset{F}{\circ} A$ satisfies $V(\theta) = G_1(\psi) \equiv 0 \bmod p^i$, hence V is divisible by p^i. If we know C and G, we can represent $V(\theta)$ in terms of ψ and adjust G_1. Namely, set $U := V \underset{G}{\circ} C$. Then $U(\psi) = V(\theta)$ and thus G is obtained as $G_1 - U$.[†27] Actually we have only C_1 and G_1. Nevertheless, $V \underset{G_1}{\circ} C_1 \equiv V \underset{G}{\circ} C \bmod p^{2i}$ and this implies $G \equiv G_1 - V \underset{G_1}{\circ} C_1 \bmod p^{2i}$. Note that C is characterized by $F(C(\psi)) = 0$ and $C(X) \equiv X \bmod p$. We can compute $C_2 \in \mathcal{P}$ satisfying $F(C_2(\psi)) \equiv 0 \bmod p^{2i}$ from C_1 by Newton's iterative root finding algorithm. Namely, define $C_2 \in \mathcal{P}$ so that $C_2(\psi) = C_1(\psi) - F(C_1(\psi))F'(C_1(\psi))^{-1}$ in $\mathbf{Z}_p[\psi]$. Repeating this process, we obtain approximations to G and C with arbitrary precision. The explicit algorithm is as follows. During execution, we keep track of $S \in \mathcal{P}$ satisfying $S(\psi) \equiv F'(C(\psi))^{-1} \bmod p^i$.

Algorithm 6. Computing the minimal polynomial.
Input: $F(X) \in \mathbf{Z}[X]$, described as above, $M \in \mathbf{N}$.
Output: $G(X) \bmod p^M$.
Procedure:

```
$\psi := T(\overline{\theta}) \bmod p^M$ ; // use an algorithm in Section 2.3.[†28]
Define $A \in \mathcal{P}$ such that $A(\theta) = \psi$.
$C(X) := X$ ; $G := F$ ; $i := 1$ ;
Take $S \in \mathcal{P}$ so that $\pi(S)(\theta) = (\pi(F')(\theta))^{-1}$ in $\mathbf{F}_q$.
while ($i < M$) {
    $V := G \underset{F}{\circ} A$ ;
    $Z := X^{2N-2}/G$ ; $U := V \underset{G}{\circ} C$ ;
    $G := G - U$ ; $Z := X^{2N-2}/G$ ;
```

[†27] Note that $\deg U < N$ and that $G_1 - U$ is a monic polynomial of degree N.

[†28] At this moment, we cannot use Algorithm 5.

```
 9:     Adjust $S$ so that $\mathrm{rem}(S * (F' \underset{G}{\circ} C) - 1, G) \equiv 0 \bmod p^{2i}$.
10:     $C := C - \mathrm{rem}((F \underset{G}{\circ} C) * S, G)$ ;
11:     Adjust $S$ again so that $\mathrm{rem}(S * (F' \underset{G}{\circ} C) - 1, G) \equiv 0 \bmod p^{2i}$.
12:     $i := 2 * i$ ;
13: }
14: return $G$ ;
```

In Step 7 and Step 8, Z is necessary to compute a remainder mod G by (1.1). Step 9 and Step 11 actually perform $S := \mathrm{rem}(S * \mathrm{rem}(2 - S * (F' \underset{G}{\circ} C), G), G)$.

Acknowledgments

The author would like to thank Dr. Pierrick Gaudry, Robert Harley, and Berit Skjernaa for valuable comments on an earlier draft of this paper. Special thanks to Dr. Steven Galbraith for agreeing to include the materials in Appendix A.

References

1. Aho, A. V., Hopcroft, J. E., Ullman, J. D.: "The design and analysis of computer algorithms". Reading, Mass.: Addison-Wesley pub. 1974.
2. Ankeny, N.C.: The least quadratic non residue. *Ann. of Math.* **55** (1952) 65-72.
3. Blake, I.F., Seroussi, G., Smart, N.P.: "Elliptic curves in cryptography". London Math. Soc. Lecture Note Series, 265. Cambridge: Cambridge U.P. 1999.
4. Borwein, J.-M., Borwein, P.-B.: "Pi and the AGM". Canadian Math. Soc. series of monographs and Adv. texts., New York: Wiley-Interscience Pub. 1987.
5. Bost, J.-B., Mestre, J.-F.: Moyenne arithmético-géométrique et périodes des courbes de genre 1 et 2. *Gaz. Math.* **38** (1988) 36-64.
6. Cantor, D. G., Kaltofen, E.: On fast multiplication of polynomials over arbitrary algebras. *Acta Inform.* **28** (1991) 693-701.
7. Cassels, J. W. S.: A note on the division values of $\wp(u)$. *Proc. Cambridge Philos. Soc.* **45** (1949) 167-172.
8. Chebyshev, P.L.: Mémoire sur les nombres premiers. *J. Math. Pures Appl.* **17** (1852) 366-390 (Œuvres, I-5).
9. Cohen, H.: "A course in computational algebraic number theory". GTM, 138. Berlin: Springer-Verlag 1993.
10. Coppersmith, D., Winograd, S.: Matrix multiplication via arithmetic progressions. *J. Symbolic Comput.* **9** (1990) 251-280.
11. Couveignes, J.-M.: "Quelques calculs en théorie des nombres". Université de Bordeaux I: Thèse 1994.
12. Couveignes, J.-M.: Computing l-isogenies using the p-torsion, Algorithmic number theory (Telence, 1996), Lecture Notes in Comput. Sci., **1122**, Berlin: Springer, 1996.
13. Couveignes, J.-M., Morain, F.: Schoof's algorithm and isogeny cycles, Algorithmic number theory (Ithaca, NY, 1994), Lect. Notes in Comput. Sci., **877**, 43-58, Berlin: Springer, 1994.
14. Dieudonné, J.: Sur les fonctions continues p-adiques. *Bull. Sci. Math.* **68** (1944) 79-85.

15. Dwork, B.: On the rationality of the zeta functions of an algebraic variety. *Amer. J. Math.* **82** (1960) 631-648.
16. Edwards, H.M.: "Riemann's zeta function". New York and London: Academic Press 1974.
17. Elkies, N.D.: Elliptic and modular curves over finite fields and related computational issues, Computational perspectives on number theory (Chicago, IL, 1995), AMS/IP Stud. Adv. Math., **7**, 21-76, Providence, RI: AMS, 1998.
18. Enge, A.: "Elliptic curves and their applications to cryptography: An introduction". Boston, Dordrecht, London: Kluwer Acad. Pub. 1999.
19. Fouquet, M., Gaudry, P., Harley, R.: An extension of Satoh's algorithm and its implementation. *J. Ramanujan Math. Soc.* **15** (2000) 281-318.
20. Frey, G.: Applications of arithmetical geometry to cryptographic constructions, Finite fields and applications (Augsburg, 1999), 128-161, Berlin: Springer, 2001.
21. Gaudry, P.: Algorithms for counting points on curves, (2001) Slides at ECC2001, Waterloo, Oct. 31, 2001, Available at http://www.cacr.math.uwaterloo.ca/-conferences/2001/ecc/slides.html.
22. Gaudry, P., Gürel, N.: An extension of Kedlaya's algorithm for counting points of superelliptic curves, Advances in Cryptology - ASIACRYPT 2001, Lect. Notes in Comput. Sci., **2248**, 480-494, ed. Boyd, C., Berlin, Heidelbert: Springer Verlag, 2001.
23. Gaudry, P., Harley, R.: Counting points on hyperelliptic curves over finite fields, ANTS-IV, Lect. Notes in Comput. Sci., **1838**, 313-332, Springer, 2000.
24. Harley, R.: Counting points with the arithmetic-geometric mean(joint work with J.-F. Mestre and P. Gaudry), Eurocrypt 2001, Rump session, 2001.
25. Harley, R., et al.: On the generation of secure elliptic curves using an arithmetic-geometric mean iteration, (in preparation).
26. Henniart, G., Mestre, J.-F.: Moyenne arithmético-géométrique p-adique. *C.R. Acad. Sci. Paris Sér. I Math.* **308** (1989) 391-395.
27. Karatsuba, A., Ofman, Y.: Multiplication of multidigit numbers on automata. *Soviet physics doklady* **7** (1963) 595-596.
28. Kedlaya, K.: Counting points on hyperelliptic curves using Monsky-Washnitzer cohomology, (2001) Preprint, available at http://arXiv.org/abs/math/0105031.
29. Kim, H., Park, J., Cheon, J., Park, J., Kim, J., Hahn, S.: Fast elliptic curve point counting using Gaussian Normal Basis, (2001) preprint.
30. Knuth, D.E.: "Seminumerical algorithm". The art of computer programming, 2. Reading, Mass.: Addison-Wesley Pub. Co. 1969.
31. Koblitz, N.: "p-adic analysis: a short course on recent work". London Math. Soc. Lect. Note Ser., 46. Cambridge-New York: Cambridge University Press 1980.
32. Koblitz, N.: "p-adic numbers, p-adic analysis, and zeta-functions (2nd ed.)". GTM, 58. New York: Springer 1984.
33. Lauder, A., Wan, D.: Computing zeta functions of Artin-Schreier curves over finite fields, (2001) preprint.
34. Lauder, A., Wan, D.: Counting points on varieties over finite fields of small characteristic, (2001) preprint.
35. Lercier, R.: Computing isogenies in $\mathbf{F}_{2^n}$, Algorithmic number theory II(Talence, 1996), Lecture Notes in Comput. Sci., **1122**, 197-212, Berlin: Springer, 1996.
36. Lubin, J., Serre, J.-P., Tate, J.: Elliptic curves and formal groups, (1964) Mimeographed notes, available at http://www.ma.utexas.edu/users/voloch/lst.-html.
37. Matsuo, K., Chao, J., Tsujii, S.: An improved baby step giant step algorithm for point counting of hyperelliptic curves over finit fields, This volume, 2002.

38. Messing, W.: "The crystals associated to Barsotti-Tate groups: with applications to Abelian schemes". Lect. Notes in Math., 264. Berin-Heidelberg-New York: Springer 1972.
39. Monsky, P.: Formal cohomology. II. The cohomology of sequence of a pair. *Ann. of Math.* **88** (1968) 218-238.
40. Monsky, P.: Formal cohomology. III. Fixed point theorems. *Ann. of Math.* **93** (1971) 315-343.
41. Monsky, P., Washinitzer, G.: Formal cohomology. I. *Ann. of Math.* **88** (1968) 181-217.
42. Montgomery, H.L.: "Topics in multiplicative number theory". Lect. Notes in Math., 227. Berlin, Heidelberg: Springer 1971.
43. Paterson, M. S., Stockmeyer, L. J.: On the number of nonscalar multiplications necessary to evaluate polynomials. *SIAM J. Comput.* **2** (1973) 60-67.
44. Poonen, B.: Computational aspects of curves of genus at least 2, Algorithmic number theory II, Lect. Notes in Comput. Sci., **1122**, 283-306, ed. Cohen, H., Berlin: Springer, 1996.
45. Satoh, T.: The canonical lift of an ordinary elliptic curve over a finite field and its point counting. *J. Ramanujan Math. Soc.* **15** (2000) 247-270.
46. Satoh, T., Skjernaa, B., Taguchi, Y.: Fast Computation of Canonical Lifts of Elliptic curves and its Application to Point Counting, (2001) preprint.
47. Schönhage, A.: Asymptotically fast algorithms for the numerical multiplication and division of polynomials with complex coefficients, Computer algebra (Marseille, 1982), Lect. Notes in Comput. Sci., **144**, 3-15, Berlin-New York: Springer, 1982.
48. Schönhage, A., Strassen, V.: Schnelle Multiplikation grosser Zahlen. *Computing* **7** (1971) 281-292.
49. Schoof, R.: Elliptic curves over finite fields and the computation of square roots mod p. *Math. Comp.* **44** (1985) 483-494.
50. Schoof, R.: Counting points on elliptic curves over finite fields. *J. Théor. Nombres Bordeaux* **7** (1995) 219-254.
51. Silverman, J. H.: "The arithmetic of elliptic curves". GTM, 106. Berlin-Heidelberg-New York: Springer 1985.
52. Skjernaa, B.: Satoh's algorithm in characteristic 2, (2000) preprint, (to appear in Math. Comp.).
53. Strassen, V.: Gaussian elimination is not optimal. *Numer. Math.* **13** (1969) 354-356.
54. Vélu, J.: Isogénies entre courbes elliptiques. *C.R. Acad. Sc. Paris.* **273** (1971) 238-241.
55. Vercauteren, F., Preneel, B., Vandewalle, J.: A memory efficient version of Satoh's algorithm, Advances in Cryptology - Eurocrypt 2001 (Innsbruck, Austria, May 2001), Lect. Notes in Comput. Sci., **2045**, 1-13, ed. Pfitzmann, B., Berlin, Heidelberg: Springer Verlag, 2001.
56. Wan, D.: Computing zeta functions over finite fields, Finite fields: theory, applications, and algorithms (Waterloo, ON, 1997), Contemp. Math., **225**, 131-141, Providence, RI: AMS, 1999.

On Arithmetically Equivalent Number Fields of Small Degree

Wieb Bosma[1] and Bart de Smit[2]

[1] Mathematisch Instituut, Universiteit Nijmegen
Postbus 9010, 6500 GL Nijmegen, the Netherlands
bosma@sci.kun.nl
[2] Mathematisch Instituut, Universiteit Leiden
P. O. Box 9512, 2300 RA Leiden, the Netherlands
desmit@math.leidenuniv.nl

Abstract. For each integer n, let $\mathcal{S}_n$ be the set of all class number quotients $h(K)/h(K')$ for number fields K and K' of degree n with the same zeta-function. In this note we will give some explicit results on the finite sets $\mathcal{S}_n$, for small n. For example, for every $x \in \mathcal{S}_n$ with $n \leq 15$, x or x^{-1} is an integer that is a prime power dividing $2^{14} \cdot 3^6 \cdot 5^3$.

1 Introduction

In broad terms the main question on number fields we address in this article is:

to what extent does the zeta-function determine the class number?

Number fields with the same zeta-function are said to be *arithmetically equivalent*. Arithmetically equivalent number fields have many invariants in common. For instance, they have the same degree, discriminant, signature, Galois closure, maximal normal subfield, and number of roots of unity. By considering the residue of the zeta-function we see that arithmetically equivalent K and K' also satisfy $h(K)R(K) = h(K')R(K')$, where h denotes the class number and R denotes the regulator of a number field. Our first result summarizes the possibilities for $h(K)/h(K')$ for fields of degree at most 15.

r_2 n	0	1	2	3	4	5	6	7
7	2^3	–	2^2	–	–	–	–	–
8	$2^3 3^2$	–	2^2	$2^2 3$	2^2	–	–	–
11	3^5	–	–	–	3^3	–	–	–
12	$2^7 3^3 5^3$	–	2^3	2^3	$2^5 5^2$	$2^4 3^2$	$2^4 3^2 5$	–
13	3^6	–	–	–	3^4	–	–	–
14	2^{10}	–	2^5	–	2^4	2^6	2^5	2^3
15	2^{14}	–	–	–	2^{10}	–	2^8	–

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 67–79, 2002.

Theorem 1. *Let K and K' be non-isomorphic arithmetically equivalent number fields of degree $n \leq 15$. Then n is equal to one of the integers in the first column of the above table, and if the number of complex infinite primes of K is denoted by r_2, the class number quotient $h(K)/h(K')$ is equal to p^k or p^{-k}, where p is a prime number, k is a non-negative integer, and p^k divides the number given in the table for the pair (n, r_2).*

A dash in the table means that this pair (n, r_2) does not occur.

The class number quotient bounds depend on the Galois configuration and the signature in a strong sense: the conjugacy class in the Galois group of complex conjugation. Therefore we first show in Section 2 that there are exactly 19 Galois configurations of degree at most 15 that contain a pair of arithmetically equivalent fields. To produce the list of the 19 possible Galois configurations we used the classification of transitive groups up to degree 15 by Butler, McKay and Royle [2], [3], [17], and a database of subgroup-lattices in the Magma-system. A relatively easy run on the Magma-system produces the list, and shows that it is complete.

The 19 Galois configurations can also be obtained from theoretical considerations; a better description of a particular configuration is useful for two purposes: it might give clues about how to realize number fields with these Galois groups, and it can also give a humanly readable proof that they contain non-isomorphic fields with the same zeta-function, which may inspire other constructions. We will give such descriptions in Section 2.

In Section 3 we employ methods of [6] to obtain bounds on class number quotients for each configuration. The required symbolic computations are performed in Magma using the ideas of [1].

LaMacchia [15] found a family of number fields, parametrized by two rational numbers, each of which is a member of a pair of arithmetically equivalent fields of degree 7. In Section 4 we construct the other member of the pair in terms of the two parameters.

By computing class numbers for pairs in this family and by using earlier results [5] about families in degree 8 constructed with 3-torsion points on elliptic curves, we give a computational proof of the following result in Section 5, showing that some of the bounds on the class number quotients are tight.

Theorem 2. *The set of values of the class number quotient $h(K)/h(K')$ as (K, K') ranges over all pairs of arithmetically equivalent number fields of degree at most 10 that are not totally real, is*

$$\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4\}.$$

The first known instances of pairs of arithmetically equivalent number fields with different class numbers were generated using a family of fields in degree 8; see [8], and also [1]. For that family, with pairs of fields of the form $\mathbb{Q}(\sqrt[8]{a})$ and $\mathbb{Q}(\sqrt[8]{16a})$, a factor 2^2 will never appear in the class number quotient; see [6].

G. Dyer [10] found the first example of arithmetically equivalent fields in degree 12 with class number quotient 5, by using the method of [5].

2 Gassmann Triples

The goal of this section is to determine for all $n \leq 15$ all possible Galois groups of arithmetically equivalent number fields of degree n.

Let $L/\mathbb{Q}$ be a Galois extension with Galois group G, and let H and H' be subgroups of G corresponding to intermediate fields $K = L^H$ and $K' = L^{H'}$. Recall that the fields K and K' are isomorphic if and only if the G-sets $X = G/H$ and $X' = G/H'$ are isomorphic, i.e., if there is a G-action preserving bijection between them. We say that the G-sets X and X' are *linearly equivalent* if every $g \in G$ has the same number of fix points on X and on X'. It is well-known that K and K' are arithmetically equivalent if and only if X and X' are linearly equivalent, which is also equivalent to H and H' giving rise to the same permutation character $1_H^G = 1_{H'}^G$ of G; see [4], Exercises 6.3, 6.4.

By a *Gassmann triple* (G, X, X') we mean a group G acting faithfully and transitively on two finite sets X and X', so that X and X' are linearly equivalent but not isomorphic as G-sets. The degree of (G, X, X') is the cardinality of X. The Galois configurations of non-isomorphic arithmetically equivalent fields of degree n are given by the Gassmann triples of degree n up to isomorphism, where we say $(G, X, X') \cong (H, Y, Y')$ if $G \cong H$ and, viewing Y and Y' as G-sets through this group isomorphism, we have $X \cong_G Y$ and $X' \cong_G Y'$.

The question whether for given positive integer n a Gassmann triple of degree n exists has been addressed in [11], [13], [14] with the help of the classification of finite simple groups. The degrees of the Gassmann triples with a *solvable* group have been determined in [7]. Combining these results, one finds that for $n \leq 100$ a Gassmann triple of degree n exists if and only if $n \geq 7$ and

$$\begin{aligned} n \neq 9,\ & 10,\ 17,\ 19,\ 23,\ 25,\ 29,\ 34,\ 37,\ 38,\ 41,\ 43,\ 46,\ 47,\ 53,\ 58, \\ & 59,\ 61,\ 67,\ 69,\ 71,\ 74,\ 79,\ 82,\ 83,\ 86,\ 87,\ 89,\ 94,\ 95,\ 97. \end{aligned}$$

In particular we see from this list that the only Gassmann triples of degree at most 15 have degree 7, 8, 11, 12, 13, 14, or 15.

As we will see, all Gassmann triples of degree at most 15 can be directly constructed by, or at least derived from, one of the following three methods—see sections 2 and 5 of [7] for details.

(A) For a finite field $\mathbb{F}_q$ and $d \in \mathbb{Z}_{\geq 2}$ consider the vector space $V = \mathbb{F}_q{}^d$ and its $\mathbb{F}_q$-dual $V^* = \mathrm{Hom}(V, \mathbb{F}_q)$. Let S be a subgroup of $\mathbb{F}_q{}^*$ of index s, let $G = \mathrm{GL}_d(\mathbb{F}_q)/S$, and let $X = (V - \{0\})/S$ and $Y = (V^* - \{0\})/S$. If $d \geq 3$ or $s \geq 2$ then (G, X, Y) is a Gassmann triple of degree $s(q^d - 1)/(q - 1)$.

(B) Let $\mathbb{F}_q$ be a finite field of characteristic at least 7, and suppose that $q \equiv \pm 1$ modulo 5. Then $G = \mathrm{PSL}_2(\mathbb{F}_q)$ has two non-conjugate subgroups H and H' that are both isomorphic to A_5, and that are conjugate in $\mathrm{PGL}_2(\mathbb{F}_q)$. Then $(G, G/H, G/H')$ is a Gassmann triple of degree $q(q^2 - 1)/120$.

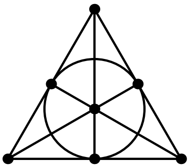
(C) Let p be a prime number, let $k > 1$ be an integer, and let $m > 1$ be a product of prime powers q that are 0 or 1 modulo p. Then there exist a Gassmann triple (G, X, X') of degree pmk with a 3-step abelian group $G = G_{p,m,k}$ of order $(pm)^k k$.

Theorem 3. *There are exactly 19 Gassmann triples (G, X, X') of degree at most* 15*, up to isomorphism. The groups G, viewed as transitive groups acting on X, are given in the table below with Butler-McKay numbering.*

deg.	no.	$\#G$	description of G	construction
7	5	168	$\mathrm{PSL}_2(\mathbb{F}_7) \cong \mathrm{PGL}_3(\mathbb{F}_2)$	(A)
8	15	32	$G_{2,2,2} \cong C_8 \rtimes V_4$	(C)
	23	48	$\mathrm{GL}_2(\mathbb{F}_3)$	(A)
11	5	660	$\mathrm{PSL}_2(\mathbb{F}_{11})$	(B)
12	26	48	$\mathrm{GL}_2(\mathbb{Z}/4\mathbb{Z}) \cap A_{12}$	(A)
	38	72	$G_{2,3,2}$	(C)
	49	96	$\mathrm{GL}_2(\mathbb{Z}/4\mathbb{Z})$	(A)
	57	96	$G_{2,2,3} \cap A_{12}$	(C)
	104	192	$G_{2,2,3}$	(C)
	124	240	$\mathrm{GL}_2(\mathbb{F}_5)/\pm 1$	(A)
13	7	5616	$\mathrm{PGL}_3(\mathbb{F}_3)$	(A)
14	10	168	$\mathrm{PGL}_3(\mathbb{F}_2)$	
	17	336	$\mathrm{PGL}_3(\mathbb{F}_2) \times C_2$	
	19	336	$\mathrm{PGL}_3(\mathbb{F}_2) \times C_2$	(A)
	52	56448	$\mathrm{PGL}_3(\mathbb{F}_2) \wr C_2$	(A)
15	15	180	$\mathrm{GL}_2(\mathbb{F}_4) \cong A_5 \times A_3$	
	21	360	$(S_5 \times S_3) \cap A_8$	
	47	2520	A_7	
	72	20160	$\mathrm{PGL}_4(\mathbb{F}_2) \cong A_8$	(A)

We explain the description of the group and the actions on the two sets X and X' degree by degree.

Degree 7 and Degree 14. Taking a 3-dimensional vector space over $\mathbb{F}_2$, we get a Gassmann triple in degree 7 from construction (A). Here the group is $G = \mathrm{GL}_3(\mathbb{F}_2) = \mathrm{PGL}_3(\mathbb{F}_2)$, and the sets X and Y are the sets of points and lines in the projective plane $\mathbb{P}^2(\mathbb{F}_2)$.

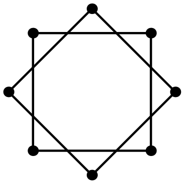


It was shown by Perlis [16] that this is the only Gassmann triple in degree 7.

In degree 14 the entries with number 19 and 52 have X equal to two copies of $\mathbb{P}^2(\mathbb{F}_2)$, where the groups are the direct product $\mathrm{PGL}_3(\mathbb{F}_2) \times C_2$ and the wreath product $\mathrm{PGL}_3(\mathbb{F}_2) \wr C_2$ respectively.

We obtain the other triples of degree 14 by adding "orientation" to the triple of degree 7. Let X be the set of points P of $\mathbb{P}^2(\mathbb{F}_2)$ together with a cyclic ordering of the three lines through P. Dually, Y is the set of lines L in $\mathbb{P}^2(\mathbb{F}_2)$ with a cyclic ordering of the three points on L. The group $\mathrm{PGL}_3(\mathbb{F}_2)$ acts naturally on X and Y, and we have a commuting action by C_2 which toggles the orientation of all points and lines. This gives the entries $(14, 10)$ and $(14, 17)$ in the table.

Degree 8. Construction (A) gives a Gassmann triple of degree 8 with group $G = \mathrm{GL}_2(\mathbb{F}_3)$. The other triple can be described with the following graph.



The plane symmetries of this graph form a dihedral subgroup D_8 of order 16 of the group of graph automorphisms. Define another graph automorphism σ by rotating one component over 180 degrees, and leaving the other component fixed. Then D_8 and σ generate a group G of graph automorphisms of order 32. The transitive actions of G on the set of vertices and on the set of edges now give a Gassmann triple of degree 8. We have $G \cong C_8 \rtimes V_4$, where the map $V_4 \to \mathrm{Aut}(C_8)$ is an isomorphism. This triple can also be obtained from construction (C) by taking $p = m = k = 2$. In fact, construction (C) was inspired by this graph theoretical example.

Degree 11. Construction (B) gives a triple of degree 11 with group $\mathrm{PSL}_2(\mathbb{F}_{11})$.

Degree 12. Construction (A) gives a triple with group $\mathrm{GL}_2(\mathbb{F}_5)/\pm 1$. We can also do construction (A) for a finite commutative local ring R rather than a finite field k. Then X is the set of elements in a free module V of rank d that are not annihilated by the maximal ideal of R, and Y is the same set in the R-linear dual of V, and $G = \mathrm{GL}_R(V)$. For $R = \mathbb{Z}/4\mathbb{Z}$ and $d = 2$ this gives entry $(12, 49)$, with $G = \mathrm{GL}_2(\mathbb{Z}/4\mathbb{Z})$ which is solvable of derived length 3, and entry $(12, 26)$ is a subgroup of index 2 acting on the same sets.

Construction (C) gives the other entries. The group $G_{2,3,2}$ has derived length 2, and the group $G_{2,2,3}$ and its subgroup $G_{2,2,3} \cap A_{12}$ have derived length 3.

Degree 13. The points in the projective plane over $\mathbb{F}_3$ together with the points in the dual projective plane form a Gassmann triple with group $\mathrm{PGL}_3(\mathbb{F}_3)$ and degree 13 by construction (A).

Degree 15. Construction (A) gives a Gassmann triple of degree 15 with group $G = \mathrm{GL}_4(\mathbb{F}_2)$. By one of the exceptional isomorphisms of simple groups [9] we have $G \cong A_8$. It turns out that we obtain other Gassmann triples by keeping the same sets, but restricting the group to the subgroup A_7, or $A_5 \times A_3$, or $(S_5 \times S_3) \cap A_8$ of A_8.

This completes the description of the 19 Gassmann triples. The second part of the proof of Theorem 3 is to show that the table is complete. The proof is based on the database of transitive groups of degree d up to 15 due to Butler, McKay and Royle [2], [3], [17]. For each transitive group G from their classification we need to determine all conjugacy classes of subgroups of index d which give rise to the same permutation character of G as a point stabilizer.

A brute force way to do this, is to find all classes of subgroups of index d and test their permutation characters. On a 1100 Mhz Athlon with 256K cache and 512 MB main memory, one can check Theorem 3 in this way with a run of Magma 2.8 of 208 seconds.

While we have no better method than brute force in general, one can often decide that a transitive group is not part of a Gassmann triple by group theoretic means. For instance, it follows from the lemmas below that neither the symmetric nor the alternating group on d letters is part of a Gassmann triple, for any d. From 1997, when the list of 19 triples was first presented at the Journées Arithmétiques in Limoges, up until the summer of 2001 when Magma 2.8 was released, these additional methods were indispensable because the routines for finding subgroups would fail on groups with a large radical index such as the alternating group on 10 letters.

Lemma 1. *Let A be the symmetric or alternating group on a finite set X. For each finite set T with trivial A-action and each A-set Y which is linearly equivalent to $X \cup T$ we have $Y \cong_A X \cup T$.*

Proof. If A is cyclic, then this is clear, so assume that the cardinality n of X is at least 3. In order to prove the lemma we first prove a weaker statement. We claim that on both $X \cup T$ and Y the group A has only one non-trivial orbit and that it has length n. To see this, note that A contains a cyclic subgroup C of order n or $n-1$, and that Y is isomorphic to $X \cup T$ as a C-set. Thus Y has an A-orbit of length n or $n-1$. Since the number of A-orbits of $X \cup T$ and Y is the same, the only case where the claim might fail is the case where Y consists of a trivial G-set, one orbit of length 2 and one orbit of length $n-1$. But then A embeds into $C_2 \times S_{n-1}$ because A acts faithfully on Y. By comparing cardinalities, and using the fact that $A_4 \not\subseteq C_2 \times S_3$ one sees that this is impossible. This proves the claim. The lemma now follows by applying the claim to A and to a point stabilizer in A of a point in X.

Lemma 2. *Let G be a finite group and X a transitive G-set and let k be a positive integer. Suppose that $X = X_1 \cup \cdots \cup X_k$ is a decomposition of X into blocks and let A be the subgroup of G of elements that fix $X_2 \cup \cdots \cup X_k$ pointwise. If A is the symmetric or alternating group on X_1 and A is non-abelian, then every G-set which is linearly equivalent to X, is G-isomorphic to X.*

Proof. We may assume that G acts faithfully on X. Let Y be a G-set which is linearly equivalent to X. For $i \in \{1, \ldots, k\}$ let A_i be the subgroup of G of elements which fix $X \setminus X_i$ pointwise. Then the A_i are the distinct conjugates of $A = A_1$. By the previous lemma, each A_i has exactly one non-trivial orbit Y_i on Y, and we have $X_i \cong_{A_i} Y_i$. It follows that the collection of all Y_i is G-stable, so that $Y_1 \cup \cdots \cup Y_k$ is a sub-G-set of Y. But since G has the same number of orbits on X and Y we have $Y = Y_1 \cup \cdots \cup Y_k$, and by counting elements we see that the Y_i are disjoint. It follows that X and Y are isomorphic over the normal subgroup $N = A_1 \times \cdots \times A_k$ of G. This means that the G-set B of bijections from X to Y contains an N-invariant element. Since A is non-abelian, the action of A on X_1 is two-transitive and $\mathrm{Aut}_A(X_1) = \{1\}$. It follows that $\#B^N = 1$. Since N is normal in G, the set B^N is a G-stable subset of B, and its unique element is a G-isomorphism from X to Y. This proves the lemma.

These lemmas tell us that the 28 largest transitive groups of degree less than 16, with orders ranging from 648000 to 1307674368000 = 15!, are not part of any Gassmann triple. The biggest group on which we use the brute force method is the 57th transitive group of degree 14, which has order 645120. The largest radical index where we apply brute force is 95040, which is the order of the simple group M_{12}, the Mathieu group in degree 12.

In all 19 Gassmann triples of degree less than 16 we found exactly two conjugacy classes of subgroups inducing the same permutation character, and they are conjugate by an outer automorphism. In other words, for these 19 triples we have $(G, X, X') \cong (G, X', X)$. This completes the proof of the Theorem.

The list of Gassmann triples of degree less than 24, based on the classification of transitive groups of degree up to 23 of A. Hulpke, was presented by the second author at a meeting in Durham in the summer of 2000. It was computed in a similar way by improving the lemmas above. A brute force run on Magma 2.8 seems to get stuck in degree 16.

3 Bounds on the Class Number Quotient

In the previous section we computed the possible Galois groups associated to a pair of non-isomorphic arithmetically equivalent fields. In this section we compute a bound on the class number quotient in each of the cases we found. To do this, we use the method explained in [6] and [1].

Let $L/\mathbb{Q}$ be a Galois extension with Galois group G, and suppose we have subgroups H, H' so that the fields $K = L^H$ and $K' = L^{H'}$ are arithmetically equivalent. Then there is an injective $\mathbb{Z}[G]$-linear map $\phi : \mathbb{Z}[G/H] \to \mathbb{Z}[G/H']$. For each subgroup J of G one has an induced map $\phi_J : \mathbb{Z}[J\backslash G/H] \to \mathbb{Z}[J\backslash G/H']$. Now let $D \subset G$ be a decomposition group at infinity. In other words, choose an embedding $L \subset \mathbb{C}$ and let D be the subgroup of order 1 or 2 of G generated by complex conjugation. For $x, y \in \mathbb{Q}$ we say that x divides y if $y \in \mathbb{Z}x$.

Proposition 1. *The class number quotient* $\dfrac{h(K)}{h(K')}$ *divides* $\dfrac{\#\mathit{Cok}(\phi_D)}{\#\mathit{Cok}(\phi_G)}$.

One gets a bound on the left hand side by computing the smallest possible value of the right hand side if one lets ϕ vary. There are some improvements on this bound, which are explained in [1]. Using these improvements we get the following table of bounds for the 19 Gassmann triples.

deg.	class number bound for given r_2								$\#G$	no.
	0	1	2	3	4	5	6	7		
7	2^3	–	2^2	–	–	–	–	–	168	5
8	2^3	–	2^2	2^2	2^2	–	–	–	32	15
	3^2	–	–	3	1	–	–	–	48	23
11	3^5	–	–	–	3^3	–	–	–	660	5
12	2^7	–	–	–	2^4	–	2^3	–	48	26
	3^3	–	–	–	–	3^2	3^2	–	72	38
	2^7	–	–	–	2^5	2^4	2^4	–	96	49
	2^4	–	2^3	–	2^2	–	2	–	96	57
	2^4	–	2^3	2^3	2^2	2^2	2	–	192	104
	5^3	–	–	–	5^2	–	5	–	240	124
13	3^6	–	–	–	3^4	–	–	–	5616	7
14	2^{10}	–	–	–	–	–	2^5	–	168	10
	2^{10}	–	–	–	–	2^6	2^5	2^3	336	17
	2^6	–	–	–	2^4	–	–	2^3	336	19
	2^6	–	2^5	–	2^4	–	–	2^3	56448	52
15	2^{10}	–	–	–	–	–	2^6	–	180	15
	2^{10}	–	–	–	–	–	2^6	–	360	21
	2^{14}	–	–	–	–	–	2^8	–	2520	47
	2^{14}	–	–	–	2^{10}	–	2^8	–	20160	72

We list the bounds by degree $[K : \mathbb{Q}] = \#X$, the number of the group in the classification, and the number r_2 of complex infinite primes of K, which is equal to the number of orbits of length 2 of the D on X. Combining the lines for a fixed degree we obtain a proof of Theorem 1.

In the table we combined results for the different subgroups D of G which give rise to the same r_2. So for specific D one can sometimes give a better bound than the one given in the table. For some of the bounds we know they can only be attained under certain strong conditions. We refer to [1], Proposition 5.2, for details.

4 A Family of Arithmetically Equivalent Fields of Degree 7

In order to test to what extent the bound in the previous section are sharp, we computed class groups for particular instances. For a good supply of arithmetically equivalent fields of degree 7 we use a family of LaMacchia [15]:

$$\begin{aligned} f_{s,t}(X) = {} & X^7 + (-6t+2)X^6 + (8t^2+4t-3)X^5 + (-s-14t^2+6t-2)X^4 \\ & + (s+6t^2-8t^3-4t+2)X^3 + (8t^3+16t^2)X^2 + (8t^3-12t^2)X - 8t^3. \end{aligned}$$

LaMacchia proved that over the function field $\mathbb{Q}(s,t)$ this polynomial is irreducible, and that its Galois group is isomorphic $G = \mathrm{GL}_3(\mathbb{F}_2)$. If we specify s and t to particular values in $\mathbb{Q}$, then the resulting polynomial in $\mathbb{Q}[X]$ might be reducible, and even if it is irreducible, then its Galois group is a subgroup of G which might not be the whole of G. But Hilbert's irreducibility theorem guarantees that there are infinitely many pairs $(a,b) \in \mathbb{Q} \times \mathbb{Q}$ for which the resulting polynomial $f_{a,b}$ in $\mathbb{Q}[X]$ is irreducible with Galois group G.

Proposition 2. *Let $a, b \in \mathbb{Q}$; if $f_{a,b}$ is irreducible in $\mathbb{Q}[X]$ then $f_{-a,b}$ is also irreducible, and the number fields of degree 7 defined by $f_{a,b}$ and $f_{-a,b}$ are arithmetically equivalent. If, moreover, $f_{a,b}$ has full Galois group $GL_3(\mathbb{F}_2)$ then these fields are not isomorphic.*

Let us consider the action of G on the 7 points of the projective plane over $\mathbb{F}_2$. The induced action on the 35 unordered triples of distinct points has two orbits: the orbit of length 7 of collinear triples, and the orbit of length 28 of non-collinear triples. The idea is that if G is the Galois group of a polynomial f over $\mathbb{Q}$ of degree 7, we can compute the polynomial P of degree 35 whose roots are all sums of three distinct roots of f. If P is a product of two irreducible polynomials P_7 and P_{28} of degree 7 and degree 28, then the field defined by P_7 is the field which is arithmetically equivalent but not isomorphic to the field defined by f.

Let us first address the issue of computing P given f. If f is monic with integer coefficients, then we could find approximations of the roots of f in $\mathbb{C}$, and then compute approximations of P. Since $P \in \mathbb{Z}[X]$ we can round off the coefficients to integers and if there is no unfortunate error blow-up then this gives the correct P.

An alternative approach uses resultants. Let us write

$$f(X) = \prod_{i=1}^{7} (X - \alpha_i).$$

For $k \in \mathbb{Q}$ with $k \neq 0$ we put

$$f_k(X) = \prod_i (X - k\alpha_i) = k^7 f(X/k).$$

Denote by R the resultant with respect to the variable T. Then we have

$$R(f_{-1}(T-X), f(T)) = \prod_{i,j}^{7} (X - \alpha_i - \alpha_j) = Q_1(X)^2 \cdot f_2(X),$$

where

$$Q_1(X) = \prod_{i<j} (X - \alpha_i - \alpha_j).$$

By computing the resultant, dividing by $f_2(X)$ and taking the square root we can thus compute $Q_1(X)$ without working in any larger fields than $\mathbb{Q}$. Similarly, we can find an expression for our polynomial P:

$$R(f_{-1}(T-X),Q_1) = \prod_{i<j}\prod_{k}(X-\alpha_i-\alpha_j-\alpha_k) = P(X)^3Q_2(X),$$

with

$$Q_2 = \prod_{i}\prod_{j\neq i}(X-\alpha_i-2\alpha_j).$$

We can find Q_2 by computing one more resultant:

$$R(f_{-1}(T-X),f_2(T)) = Q_2(X)f_3(X).$$

The three resultant equations allow one to successively compute Q_1, Q_2 and P by taking resultants, computing quotients and take a square and a cube root.

Note that the largest resultant has degree 147, but it turns out that one can do this computation for any given $f \in \mathbb{Q}[X]$ quite easily on a computer. We then find the degree 7 factor P_7 of P by a rational polynomial factorization algorithm.

If we take $f = f_{s,t}$ it would be nice to obtain P_7 as a polynomial with coefficients in $\mathbb{Q}(s,t)$, so that we do not have to go through the resultant computation for each pair of rational numbers a, b. One could try to do this with the resultant-method given above, with base-field $\mathbb{Q}(s,t)$ rather than $\mathbb{Q}$. This symbolic computation turns out not to be feasible. Instead, we compute P_7 for many values of a and b in $\mathbb{Z}$ and then interpolate. To see how this works, let us consider the polynomial P. The coefficients of P can be expressed in terms of the symmetric functions $\sigma_1,\ldots,\sigma_7$ in $\alpha_1,\ldots,\alpha_7$, where $f = X^7-\sigma_1X^6+\sigma_2X^5-\ldots+\sigma_7$. Giving each σ_i degree i we see that all coefficients of P have degree at most 35. It follows that the coefficients have at most degree 35 in s and t. In fact, since s occurs only in σ_3 and σ_4, the degree in s is at most 11. The factor P_7 of P therefore also has coefficients $c_i(s,t)$ which are polynomials of degree at most 11 and 35 in s and t. With these bounds on the degree we can now find these polynomials by interpolating. For fixed t_0 we need at least 12 values of s to determine the polynomial $c_i(t_0,s)$, and if we do this for at least 36 values of t_0 then we know $c_i(t,s)$ by interpolation. We thus computed that P_7 is equal to the polynomial

$$\begin{aligned}g_{s,t} = {} & X^7+(-18t+6)X^6+(124t^2-64t+6)X^5+(-408t^3+208t^2-4t-16)X^4\\ & +(6(t-1)s+640t^4-156t^3-116t^2+84t-27)X^3\\ & +((-36t^2+36t-12)s-384t^5-152t^4+120t^3+88t^2-34t-6)X^2\\ & +(-s^2+(48t^3-20t^2-2t-2)s-64t^5-84t^4+52t^3-8t^2-12t)X\\ & +(-8t^3-4t^2)s+384t^6+80t^5-88t^4-24t^3.\end{aligned}$$

To finish the proof of the proposition one notices that $f_{-s,t}(X)$ divides the polynomial $X^7g_{s,t}((X-1)(1+2t/X))$, which means that the field defined by $g_{s,t}$ is contained in the field defined by $f_{-s,t}$. Thus, the polynomials $f_{s,t}$ and $f_{-s,t}$ give the Galois configuration of the desired Gassmann triple (G,X,X') over the field $\mathbb{Q}(s,t)$.

If we specify s and t to rational numbers a, b, and $f_{a,b} \in \mathbb{Q}[X]$ is irreducible, then the Galois group of $f_{a,b}$ is a subgroup G_0 of G which is transitive on X. Then G_0 contains an element of order 7, so it is also transitive on X', so $f_{-a,b}$ is also irreducible. Since X is linearly equivalent to X' over G_0, the two fields are arithmetically equivalent. Since there is only one Gassmann triple for degree 7 the two fields are isomorphic if and only if $G_0 \neq G$.

5 Class Number Computations

In this section we prove Theorem 2.

The first three lines in the table of Theorem 1 show that for arithmetically equivalent fields K and K' of degree at most 10 that are not totally real, the class number quotient $h(K)/h(K')$ or its reciprocal lies in $\{1, 2, 3, 4\}$. It remains to exhibit examples to prove that all possibilities occur. In [5] examples are given of arithmetically equivalent fields K, K' of degree 8 with $h(K)/h(K') = 3$.

We use the family of polynomials $f_{s,t}$ and $f_{-s,t}$ from the previous section to generate examples for the remaining cases. Using Magma we selected the subset of 1091 pairs of integers (a, b) with $0 \leq a, |b| \leq 100$, for which the field discriminant of the number field generated by $f_{a,b}$ has less than 25 decimal digits. In 276 of these cases the fields are totally real and in all other cases there are 2 pairs of complex embeddings.

We have used h and h' to denote the class numbers of the number fields generated by the polynomials $f_{s,t}$ and $f_{-s,t}$. The table below summarizes the class number quotients found.

h/h' r_2	1	1/2	2/1	1/4	4/1	$\sum$
2	607	104	98	2	4	815
0	210	38	25	2	1	276

The last row, representing the 276 totally real fields found, is given here for comparison, and to show that no factor 8 was found in the class number quotients.

The table below lists, of the 815 pairs that are not totally real, those with class number quotients 4 and 1/4, and the smallest (in terms of discriminant) with class number quotients 1, 2 and 1/2.

Class groups (and unit groups) in Magma are computed by a method that generates relations between prime ideals of bounded norm. This is guaranteed to give the correct class number if all primes up to the Minkowski bound are taken into consideration. For fields of small discriminant, including the first example with class number quotient 4 listed in the table, the method can be used to certify the class number. It took around 7 minutes of CPU time to find the class number pair for $(62, -1)$ with Magma this way.

(a,b)	D	factorization of D	h/h'
$(8,1)$	232745536	$2^6 \cdot 1907^2$	$1/1$
$(7,-1)$	24497571289	$281^2 \cdot 557^2$	$1/2$
$(5,1)$	31811219449	$59^2 \cdot 3023^2$	$2/1$
$(62,-1)$	3770079362544784	$2^4 \cdot 15350243^2$	$4/1$
$(22,-6)$	2429680593739514347584	$2^6 \cdot 3^6 \cdot 11^4 \cdot 71^2 \cdot 101^2 \cdot 263^2$	$1/4$
$(83,4)$	3174516214584075350089	56342845283^2	$6/24$
$(81,-6)$	10630565571038999396281	$19^2 \cdot 557^2 \cdot 1697^2 \cdot 5741^2$	$8/2$
$(53,-6)$	10726579028522017397529	$3^4 \cdot 13^2 \cdot 1453^2 \cdot 609227^2$	$8/2$
$(2,-6)$	155678051656088618455296	$2^8 \cdot 3^6 \cdot 37^2 \cdot 24684721^2$	$4/1$

For large discriminants this computation is no longer feasible. In that case the Minkowski bound can be replaced by the (usually much smaller) Bach bound, at the cost of correctness only being guaranteed under assumption of the generalized Riemann hypothesis. This was used to compute the other class number pairs, each taking less than a minute.

Alternatively, some local computations with independent units together with bounds on the regulator may provide fairly fast provably correct results; cf. [8]. For this MAGMA has a built in function `pFundamentalUnits`, which we also used to verify the above class number quotients.

References

1. Wieb Bosma, Bart de Smit, *Class number relations from a computational point of view*, J. Symbolic Computation **31** (2001), 97–112.
2. G. Butler, *The transitive groups of degree fourteen and fifteen*, J. Symbolic Computation **16** (1993), 413–422.
3. G. Butler, J. McKay, *The transitive groups of degree up to eleven*, Comm. Algebra **11** (1983), 863–911.
4. J.W.S. Cassels, A. Fröhlich (eds), *Algebraic Number Theory*, Academic Press (1986) (reprint).
5. B. de Smit, *Generating arithmetically equivalent number fields with elliptic curves*, in: J.P. Buhler (Ed.), Algorithmic Number Theory (Proceedings ANTS III), Lecture Notes in Computer Science, Vol. 1423. Springer-Verlag, Berlin (1998), 392–399.
6. B. de Smit, *Brauer-Kuroda relations for S-class numbers*, Acta Arithmetica **98** (2001), 133–146.
7. B. de Smit, H.W. Lenstra, Jr., *Linearly equivalent actions of solvable groups*, J. Algebra **228** (2000), 270–285.
8. B. de Smit and R. Perlis, *Zeta functions do not determine class numbers*, Bull. Amer. Math. Soc. **31** (1994), 213–215.
9. L.E. Dickson, *Linear groups with an exposition of the Galois Field Theory*, Dover (1958), reprint of 1901 original.
10. G. Dyer, *Two arithmetically equivalent number fields with class number quotient five*, preprint, 1999, Mathematical Research Experiences for Undergraduates at Louisiana State University, LEQS.

11. W. Feit, *Some consequences of the classification of finite simple groups*, in: B. Cooperstein and G. Mason (eds.), *The Santa Cruz conference on finite groups 1979* Proc. Sympos. Pure Math. Vol. 37, Amer. Math. Soc., Providence (1980), 175–181.
12. F. Gassmann, *Bemerkungen zu der vorstehnden Arbeit von Hurwitz*, (*'Über Beziehungen zwischen den Primidealen eines algebraischen Körpers und den Substitutionen seiner Gruppe'*), Math. Z. **25** (1926), 124–143.
13. R. M. Guralnick, *Subgroups inducing the same permutation representation*, J. Algebra **81** (1983), 312–319.
14. R. M. Guralnick and D. B. Wales, *Subgroups inducing the same permutation representation, II*, J. Algebra **96** (1985), 94–113.
15. S. E. LaMacchia, *Polynomials with Galois group PSL*$(2, 7)$, Comm. Algebra **8** (1980), 983–992.
16. R. Perlis, *On the equation* $\zeta_K(s) = \zeta_{K'}(s)$, J. Number Theory **9** (1977), 342–360.
17. G. F. Royle, *The transitive groups of degree twelve*, J. Symbolic Computation **4** (1987), 255–268.

A Survey of Discriminant Counting

Henri Cohen, Francisco Diaz y Diaz, and Michel Olivier

Laboratoire A2X, U.M.R. 5465 du C.N.R.S.,
Université Bordeaux I,
351 Cours de la Libération,
33405 TALENCE Cedex, FRANCE

Abstract. We give a survey of known results on the asymptotic and exact enumeration of discriminants of number fields, both in the absolute and relative case. We give no proofs, and refer instead to the bibliography.

1 General Conjectures and Results

Let $\overline{\mathbb{Q}}$ be a fixed algebraic closure of $\mathbb{Q}$, let $K \subset \overline{\mathbb{Q}}$ be a fixed number field, which we take as our base field, and let G be a transitive permutation group on n letters. We consider the set $\mathcal{F}_n(G)$ of all extensions L/K of degree n with $L \subset \overline{\mathbb{Q}}$ such that the Galois group of the Galois closure $\tilde{L}$ of L/K viewed as a permutation group on the set of embeddings of L into $\tilde{L}$ is permutation isomorphic to G (warning: this is equal to $n/m(G)$ times the number of extensions up to K-isomorphism, where $m(G)$ is the number of K-automorphisms of L). We write

$$N_{K,n}(G,X) = |\{L \in \mathcal{F}_n(G),\ |\mathcal{N}(\mathfrak{d}(L/K))| \leq X\}|\ ,$$

where $\mathfrak{d}(L/K)$ denotes the relative ideal discriminant and $\mathcal{N}$ the absolute norm. The aim of this paper is to give a survey of results and conjectures on asymptotic and exact values of this quantity, without proof. It is usually easy to generalize the results to the case where the behavior of a *finite* number of places of K in the extension L/K is specified. In particular, if $K = \mathbb{Q}$ we will give the results and conjectures when the signature (R_1, R_2) of L is specified, with $R_1 + 2R_2 = n$. In this case, we will write $N_{R_1,R_2}(G,X)$ for the number of L as above with signature (R_1, R_2).

It is also sometimes possible to give additional main terms and rather good error terms instead of asymptotic formulas, and we will do this in some cases, but not systematically.

General conjectures on the subject have been made by several authors. In view of the available data and theorems, it seems reasonable to formulate the following precise statements (see for example [6] and [29]).

Conjecture 1. (1) For each number field K and transitive group G on n letters as above, there exist three strictly positive constants $a_K(G)$, $b_K(G)$ and $c_K(G)$ such that

$$N_{K,n}(G,X) \sim c_K(G)\, X^{a_K(G)} (\log X)^{b_K(G)-1}\ .$$

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 80–94, 2002.

(2) Furthermore, the constant $a_K(G)$ should not depend on K (hence will be denoted by $a(G)$) and should be a rational number satisfying $0 < a(G) \leq 1$, and $b_K(G)$ should be an integer greater or equal to 1, equal to 1 if $a(G) = 1$.
(3) If G is a primitive transitive group, we should have $a(G) < 1$ except when $G \simeq S_n$, in which case we should have $a(S_n) = 1$.
(4) On the contrary, if n is composite (so that there exist imprimitive groups G), there exists at least one imprimitive transitive group G such that $a(G) = 1$.
(5) The total number of extensions L/K in $\overline{\mathbb{Q}}$ of degree n and norm of relative discriminant bounded by X should be asymptotic to $c_K\, X$ for some positive constant c_K.

An even more precise version of this conjecture concerning the value of $a(G)$ has been made by G. Malle [29] as follows.

Definition 1. *For any element $g \in S_n$ different from the identity, define the index* $\operatorname{ind}(g)$ *of g by the formula*

$$\operatorname{ind}(g) = n - |\text{orbits of } g|\ .$$

We define the index $i(G)$ of a transitive subgroup G of S_n by the formula

$$i(G) = \min_{g \in G,\ g \neq 1} \operatorname{ind}(g)\ .$$

Examples:

(1) The index of a transposition is equal to 1, and this is the lowest possible index for a nonidentity element. It follows that $i(S_n) = 1$.
(2) If G is an Abelian group, and if ℓ is the smallest prime divisor of $|G|$, then it is easy to show that $i(G) = |G|(1 - 1/\ell)$.

Conjecture 2. (Malle)

(1) Strong form: for any transitive subgroup G of S_n, we have $a(G) = 1/i(G)$ in Conjecture 1.
(2) Weak form: for any transitive subgroup G of S_n, we have for all $\varepsilon > 0$ and sufficiently large X

$$c_K(G) \cdot X^{a(G)} < N_{K,n}(G, X) < X^{a(G)+\varepsilon}$$

for some strictly positive constant $c_K(G)$, with $a(G) = 1/i(G)$.

It can be shown that the statements (2), (3), and (4) about $a(G)$ in Conjecture 1 follow from (the strong form of) Malle's conjecture.

The following results give support to the conjecture (see [5], [14], [26], [34], [36]).

Theorem 1. (1) *(Wright). The strong form of Malle's conjecture is true for all Abelian groups G.*

(2) *(Cohen–Diaz–Olivier). The strong form of Malle's conjecture is true for $G = D_4$.*
(3) *(Yukie, Bhargava) The weak form of Malle's conjecture is true for $G = S_4$, and the strong form is true for $K = \mathbb{Q}$.*
(4) *(Klüners–Malle). The weak form of Malle's conjecture is true for all nilpotent groups.*

In addition, Malle gives partial results towards the statement that his conjecture is compatible with direct products and with wreath products.

In a preprint in preparation, Malle also gives a conjecture for the value of the exponent of the logarithm $b_K(G)$. For instance, in the Abelian case, the paper of Wright mentioned above proves the following theorem:

Theorem 2. *(Wright). Let G be an Abelian group, and let ℓ be the smallest prime divisor of $|G|$ (so that $a(G) = (|G|(1 - 1/\ell))^{-1}$). Denote by $B_\ell(G)$ the number of elements of G of order ℓ, which will be of the form $\ell^k - 1$ for some positive k. Then we have*

$$b_K(G) = \frac{B_\ell(G)}{[K(\zeta_\ell) : K]} ,$$

where as usual ζ_ℓ denotes a primitive ℓ-th root of unity.

The paper of Wright also claims an explicit formula for the constant $c_K(G)$, but although it is a finite expression in terms of adelic integrals, as far as the authors are aware, it has not been computed explicitly by this method apart from the case where G is of order 2. We have computed it using Kummer theory in many other cases (see [10], [16], [19], [20]).

In the case $K = \mathbb{Q}$ and Abelian groups G, after many papers mostly from authors from the former Soviet union (see the references given in [27] and [28]), Mäki has given the constant $c_{\mathbb{Q}}(G)$ explicitly for all Abelian groups G (see [27], [28]).

Finally, concerning statement (4) of the general conjecture, Malle proves the following theorem:

Theorem 3. *If n is a composite number divisible by either 2 or 3, there exists an imprimitive transitive subgroup G of S_n such that $a(G) = 1$, for which $N_{K,n}(G, X) \geq c\,X$ for some strictly positive constant c.*

He of course conjectures that this remains true for any composite n, not only those divisible by 2 or 3. In particular, if Conjecture 1 is true, this shows that, for composite n, the proportion of S_n-extensions among all extensions of degree n is strictly less than 1. Thanks to some of the abovementioned results, this is now a theorem for $n = 4$ (at least over $K = \mathbb{Q}$, but certainly also over general K).

This is in complete opposition with the situation for *polynomials*, where Hilbert's irreducibility theorem shows that "almost all" polynomials of degree n have Galois group S_n.

A final remark concerning Conjecture 1. In view of the methods which are in use to prove the known results, its validity seems very plausible for *solvable* groups G (for instance because of the use of class field theory and Kummer theory), and it seems not to be out of reach. Because of the methods used by Bhargava (higher composition laws linked to certain root systems) and Yukie (prehomogeneous vector spaces), it is also plausible for $n = 5$, i.e., for $G = A_5$ or $G = S_5$. On the other hand, for general nonsolvable groups G, it is not impossible that there do not exist such precise asymptotic formulas, but only the weak form of Malle's conjecture.

In addition to asymptotic formulas, we give *exact* results on $N_{R_1,R_2}(G,X)$ for quite large values of X (see [18] for much more complete tables). These have been obtained by a variety of methods (see [2], [7], [8], [12], [14]). The value of X that we choose as upper limit of our computations corresponds to the use of approximately 1 month of CPU time on a 1 Ghz Pentium III workstation with 1 GB of main memory. It should be emphasized that all of the exact counting methods are algorithmic, and that if the number of fields is reasonable, we can just as easily construct *tables* of extensions (see for example [11], [15]).

Notations. We denote by $m = [K : \mathbb{Q}]$ the absolute degree of the base field K, and by (r_1, r_2) the signature of K, so that $r_1 + 2r_2 = m$. The letter $\mathfrak{p}$ will always denote a prime ideal of K, and the letter p a prime number. The notation $e(\mathfrak{p})$ stands for the absolute ramification index of $\mathfrak{p}$ above the prime number below $\mathfrak{p}$. As usual, $\zeta_K(s)$ denotes the Dedekind zeta function of the number field K, and by a convenient abuse of notation, we will denote by $\zeta_K(1)$ the residue of $\zeta_K(s)$ at $s = 1$, given by the well-known formula

$$\zeta_K(1) = 2^{r_1}(2\pi)^{r_2}\frac{h(K)R(K)}{w(K)\sqrt{|d(K)|}}\ ,$$

with the usual notations of algebraic number theory.

2 Results in Small Degrees

2.1 $G = C_2$

$$N_{K,2}(C_2, X) \sim c_K(C_2)\,X \quad \text{with}$$

$$c_K(C_2) = \frac{1}{2^{r_2}}\frac{\zeta_K(1)}{\zeta_K(2)}\ .$$

This very simple result deserves to be better known, and its proof is easy, although not completely trivial. It is due to Datskovsky and Wright [23], using Shintani's theory of zeta functions of prehomogeneous vector spaces (see [31], [32], [35]), and a much simpler proof is given by the authors in [16] using Kummer theory. Of course, in particular

$$c_{\mathbb{Q}}(C_2) = \frac{1}{\zeta(2)} = \frac{6}{\pi^2} = 0.6079271018540266286632767 79\ldots$$

We have in fact the more precise result

$$N_{K,2}(C_2, X) = c_K(C_2)\, X + O(X^{\alpha})$$

for some explicit $\alpha < 1$ depending on $m = [K : \mathbb{Q}]$, and in particular $\alpha = 1/2$ for $K = \mathbb{Q}$, and even $\alpha < 1/2$ under the GRH (see for example [33]).

$$N_{2,0}(C_2, X) \sim N_{0,1}(C_2, X) \sim \frac{c_{\mathbb{Q}}(C_2)}{2}\, X\ .$$

$$\begin{aligned} N_{\mathbb{Q},2}(C_2, 10^{25}) &= 6079271018540266286517795 \\ N_{2,0}(C_2, 10^{25}) &= 3039635509270133143448215 \\ N_{0,1}(C_2, 10^{25}) &= 3039635509270133143069580\ . \end{aligned}$$

2.2 $G = C_3$

$$N_{K,3}(C_3, X) \sim \begin{cases} c_K(C_3)\, X^{1/2} \log X & \text{if } \zeta_3 \in K \\ c_K(C_3)\, X^{1/2} & \text{if } \zeta_3 \notin K\ . \end{cases}$$

Here, if $\zeta_3 \in K$ we have

$$c_K(C_3) = \frac{1}{4 \cdot 3^{r_2}} \zeta_K(1)^2 \prod_{\mathfrak{p}} \left(1 + \frac{2}{\mathcal{N}\mathfrak{p}}\right)\left(1 - \frac{1}{\mathcal{N}\mathfrak{p}}\right)^2\ .$$

On the other hand, if $\zeta_3 \notin K$, we set $K_z = K(\zeta_3) = K(\sqrt{-3})$, and we have

$$c_K(C_3) = \frac{1}{2 \cdot 3^{r_1+r_2-1}} \frac{\zeta_{K_z}(1)}{\zeta_K(2)} \prod_{\left(\frac{K_z}{\mathfrak{p}}\right)=1} \left(1 - \frac{2}{\mathcal{N}\mathfrak{p}(\mathcal{N}\mathfrak{p}+1)}\right)$$
$$\prod_{\left(\frac{K_z}{\mathfrak{p}}\right)=0} \left(1 + \frac{1}{\mathcal{N}\mathfrak{p}+1} - \frac{1}{\mathcal{N}\mathfrak{p}^{(e(\mathfrak{p})+1)/2}}\right) \prod_{\substack{\left(\frac{K_z}{\mathfrak{p}}\right)=-1 \\ \mathfrak{p} \mid 3}} \left(1 + \frac{2}{\mathcal{N}\mathfrak{p}} - \frac{2}{\mathcal{N}\mathfrak{p}^{e(\mathfrak{p})/2}}\right)\ .$$

In the above, $\left(\frac{K_z}{\mathfrak{p}}\right) = -1, 0$ or 1 means that $\mathfrak{p}$ is inert, ramified or split in the quadratic extension K_z/K. Note that only the first product is an infinite product, and that in the second product the condition $\left(\frac{K_z}{\mathfrak{p}}\right) = 0$ implies that $\mathfrak{p} \mid 3$ (it is in fact equivalent to $e(\mathfrak{p})$ being odd).

In particular,

$$\begin{aligned} c_{\mathbb{Q}}(C_3) &= \frac{11\sqrt{3}}{36\pi} \prod_{p \equiv 1 \pmod 6} \left(1 - \frac{2}{p(p+1)}\right) \\ &= 0.1585282583961420602835078 2\ldots \end{aligned}$$

As with all the other constants that we will give for *Abelian* extensions of $\mathbb{Q}$, this numerical value has been computed using well-known methods which are however difficult to find in the literature (see for example [9]).

Evidently $N_{3,0}(C_3,X)=N_{\mathbb{Q},3}(C_3,X)$ and $N_{1,1}(C_3,X)=0$.

The general result is due to the authors [16], and the result for $K=\mathbb{Q}$ is due to H. Cohn [22].

$$N_{\mathbb{Q},3}(C_3,10^{37})=N_{3,0}(C_3,10^{37})=501310370031289126\ .$$

2.3 $G=S_3$

$$N_{K,3}(S_3,X)\sim c_K(S_3)\,X\quad\text{with}$$
$$c_K(S_3)=\left(\frac{2}{3}\right)^{r_1-1}\left(\frac{1}{6}\right)^{r_2}\frac{\zeta_K(1)}{\zeta_K(3)}=\frac{2^{r_1-r_2-1}}{3^{r_1+r_2-1}}\frac{\zeta_K(1)}{\zeta_K(3)}\ .$$

In particular

$$c_{\mathbb{Q}}(S_3)=\frac{1}{\zeta(3)}=0.831907372580707468683126 28\ldots$$

$$N_{3,0}(S_3,X)\sim\frac{c_{\mathbb{Q}}(S_3)}{4}X$$
$$N_{1,1}(S_3,X)\sim\frac{3c_{\mathbb{Q}}(S_3)}{4}X\ .$$

These results over $\mathbb{Q}$ are the beautiful and difficult results of Davenport and Heilbronn [24], [25], but the result over a general number field is much deeper and is due to Datskovsky and Wright [23].

For $K=\mathbb{Q}$, Belabas in [3] proves that the error term in the asymptotic formula is at most $O(X\exp(-c(\log X\log\log X)^{1/2}))$ for any $c<1/24$. However, considering the available numerical and heuristic evidence, it seems quite plausible (Yukie (personal communication) and Roberts [30]) that for $K=\mathbb{Q}$ there is an additional main term, and that we have the much stronger conjecture

$$N_{\mathbb{Q},3}(S_3,X)=c_{\mathbb{Q}}(S_3)\,X+c'_{\mathbb{Q}}(S_3)\,X^{5/6}+o(X^{5/6})\ ,$$

with

$$c'_{\mathbb{Q}}(S_3)=\frac{4(\sqrt{3}+1)}{5\Gamma(2/3)^3}\frac{\zeta(1/3)}{\zeta(5/3)}=-1.2104509099918403959009 2077\ldots\ ,$$

and that

$$N_{3,0}(S_3,X)=\frac{c_{\mathbb{Q}}(S_3)}{4}X+\frac{c'_{\mathbb{Q}}(S_3)}{\sqrt{3}+1}X^{5/6}+o(X^{5/6})$$
$$N_{1,1}(S_3,X)=\frac{3c_{\mathbb{Q}}(S_3)}{4}X+\frac{\sqrt{3}c'_{\mathbb{Q}}(S_3)}{\sqrt{3}+1}X^{5/6}+o(X^{5/6})\ .$$

Using algorithmic methods based on the Heilbronn–Davenport theory, Belabas in [2] can for example compute exactly

$$\begin{aligned}N_{\mathbb{Q},3}(S_3,10^{11})&=81414013239\\ N_{3,0}(S_3,10^{11})&=20147321619\\ N_{1,1}(S_3,10^{11})&=61266691620\ .\end{aligned}$$

It should be possible to push these computations up to 10^{13} and perhaps 10^{14}.

2.4 $G = C_4$

To state the result, we need some definitions.

Definition 2. *Let* $\mathfrak{c} \mid 2\mathbb{Z}_K$.

(1) *We denote by* $\mathcal{I}^q$ *(resp.,* $Q_2(K)$*) the group of fractional ideals* $\mathfrak{a}$ *of* K *(resp., of elements of* K^**) which are norms from* $K(i)$ *to* K *of an ideal (resp., of an element).*

(2) *We set*

$$G^q(\mathfrak{c}^2)=\frac{\{\mathfrak{a}\in\mathcal{I}^q,\ (\mathfrak{a},\mathfrak{c})=1\}}{\{\mathfrak{q}^2\beta,\ (\mathfrak{q},\mathfrak{c})=1,\ \beta\in Q_2(K),\ \beta\equiv 1\ (\mathrm{mod}\ ^*\mathfrak{c}^2)\}}\ ,$$

and we denote by $\widehat{G^q(\mathfrak{c}^2)}$ *the group of characters of* $G^q(\mathfrak{c}^2)$.

Finally, for any ideal $\mathfrak{c} \mid 2\mathbb{Z}_K$ we denote by $h(\mathfrak{c})$ the number of prime ideals dividing $2\mathbb{Z}_K/\mathfrak{c}$ which are either unramified in $K(i)/K$, or which divide $\mathfrak{c}$ and are ramified in $K(i)/K$. Recall that $m = [K : \mathbb{Q}]$. We then have the following results (see [19]):

$$N_{K,4}(C_4,X)\sim c_K(C_4)\,X^{1/2}\quad\text{with}$$

$$c_K(C_4)=\frac{\zeta_K(1)}{\zeta_K(2)}\left(\frac{1}{2^{3m}}\sum_{\mathfrak{c}\mid 2\mathbb{Z}_K}2^{h(\mathfrak{c})}\mathcal{N}(\mathfrak{c})^2P(\mathfrak{c})S(\mathfrak{c})-\frac{1}{2^{r_2+1}}\right)\ ,$$

where

$$P(\mathfrak{c})=\frac{1}{\prod_{\mathfrak{p}\mid 2/\mathfrak{c}}(1+1/\mathcal{N}\mathfrak{p})}\prod_{\mathfrak{p}\mid(\mathfrak{c},2/\mathfrak{c})}\left(1-\frac{1}{\mathcal{N}\mathfrak{p}^3}\right)\prod_{\mathfrak{p}\mid\mathfrak{c},\ \mathfrak{p}\nmid 2/\mathfrak{c}}\left(1-\frac{2}{\mathcal{N}\mathfrak{p}^3(1+1/\mathcal{N}\mathfrak{p})}\right)\ ,$$

and

$$S(\mathfrak{c})=\sum_{\chi\in\widehat{G^q(\mathfrak{c}^2)}}\prod_{\substack{\mathfrak{p}\mid 2,\ \mathfrak{p}\nmid\mathfrak{c}\\ \mathfrak{p}\in\mathcal{I}^q}}\left(1+\frac{\chi(\mathfrak{p})}{\mathcal{N}\mathfrak{p}^{3/2}}\right)\prod_{\mathcal{N}\mathfrak{p}\equiv 1\ (\mathrm{mod}\ 4)}\left(1+\frac{2\chi(\mathfrak{p})}{\mathcal{N}\mathfrak{p}^{3/2}(1+1/\mathcal{N}\mathfrak{p})}\right)\ .$$

In the above, $\omega_K(\mathfrak{a})$ denotes the number of distinct prime ideal divisors of $\mathfrak{a}$, and $\widehat{G}$ denotes the group of characters of G.

For $K = \mathbb{Q}$, we have the more precise result

$$N_{\mathbb{Q},4}(C_4, X) = c_{\mathbb{Q}}(C_4)\, X^{1/2} + c'_{\mathbb{Q}}(C_4)\, X^{1/3} + o(X^{1/3}) ,$$

with

$$\begin{aligned} c_{\mathbb{Q}}(C_4) &= \frac{3}{\pi^2}\left(\left(1+\frac{\sqrt{2}}{24}\right)\prod_{p\equiv 1 \pmod 4}\left(1+\frac{2}{p^{3/2}+p^{1/2}}\right)-1\right) \\ &= 0.1220526732513967609226080529\ldots \\ c'_{\mathbb{Q}}(C_4) &= \frac{11-3\cdot 2^{1/3}+4\cdot 2^{2/3}}{20\pi}\frac{\zeta(2/3)}{\zeta(4/3)}\prod_{p\equiv 1 \pmod 4}\left(1+\frac{2}{p+p^{1/3}}\right)\left(\frac{1-1/p}{1+1/p}\right) \\ &= -0.1156751993942787883018548368\ldots \end{aligned}$$

The value of $c_{\mathbb{Q}}(C_4)$ is due in principle to A. Baily [1], with computational errors.

$$\begin{aligned} N_{4,0}(C_4, X) &= \frac{c_{\mathbb{Q}}(C_4)}{2}\, X^{1/2} + \frac{c'_{\mathbb{Q}}(C_4)}{2}\, X^{1/3} + o(X^{1/3}) \\ N_{2,1}(C_4, X) &= 0 \\ N_{0,2}(C_4, X) &= \frac{c_{\mathbb{Q}}(C_4)}{2}\, X^{1/2} + \frac{c'_{\mathbb{Q}}(C_4)}{2}\, X^{1/3} + o(X^{1/3}) . \end{aligned}$$

$$\begin{aligned} N_{\mathbb{Q},4}(C_4, 10^{32}) &= 1220521363354404 \\ N_{4,0}(C_4, 10^{32}) &= 610260681684841 \\ N_{0,2}(C_4, 10^{32}) &= 610260681669563 . \end{aligned}$$

2.5 $G = V_4 = C_2 \times C_2$

$$N_{K,4}(V_4, X) \sim c_K(V_4)\, X^{1/2} \log^2 X \quad \text{with}$$

$$\begin{aligned} c_K(V_4) = \frac{1}{48\cdot 4^{r_2}}\zeta_K(1)^3 \prod_{\mathfrak{p}}\left(1+\frac{3}{\mathcal{N}\mathfrak{p}}\right)\left(1-\frac{1}{\mathcal{N}\mathfrak{p}}\right)^3 \\ \prod_{\mathfrak{p}\mid 2\mathbb{Z}_K}\frac{1+\dfrac{4}{\mathcal{N}\mathfrak{p}}+\dfrac{2}{\mathcal{N}\mathfrak{p}^2}+\dfrac{1}{\mathcal{N}\mathfrak{p}^3}-\dfrac{(1-1/\mathcal{N}\mathfrak{p}^2)e(\mathfrak{p})+(1+1/\mathcal{N}\mathfrak{p})^2}{\mathcal{N}\mathfrak{p}^{e(\mathfrak{p})+1}}}{1+\dfrac{3}{\mathcal{N}\mathfrak{p}}} . \end{aligned}$$

Note that the local factor at 2 given in the preprint [13] is incorrect.

We have in fact the more precise result

$$N_{K,4}(V_4, X) = (c_K(V_4)\log^2 X + c'_K(V_4)\log X + c''_K(V_4))\, X^{1/2} + o(X^{1/2})$$

where $c'_K(V_4)$ and $c''_K(V_4)$ are explicit constants which are too complicated to be given here.

In particular,

$$\begin{aligned}
c_{\mathbb{Q}}(V_4) &= \frac{23}{960}\prod_p\left(\left(1+\frac{3}{p}\right)\left(1-\frac{1}{p}\right)^3\right)\\
&= 0.002752430222755481396638311 84\ldots\\
c'_{\mathbb{Q}}(V_4) &= 12c_{\mathbb{Q}}(V_4)\left(\gamma-\frac{1}{3}+\frac{9\log 2}{23}+4\sum_{p\geq 3}\frac{\log p}{(p-1)(p+3)}\right)\\
&= 0.05137957621042353770883347445\ldots\\
c''_{\mathbb{Q}}(V_4) &= \frac{c'_{\mathbb{Q}}(V_4)^2}{4c_{\mathbb{Q}}(V_4)}-\frac{3}{\pi^2}\\
&\quad+24c_{\mathbb{Q}}(V_4)\left(\frac{1}{6}-\gamma_1-\frac{\gamma^2}{2}-\frac{340}{529}\log^2 2-4\sum_{p\geq 3}\frac{p(p+1)\log^2 p}{(p-1)^2(p+3)^2}\right)\\
&= -0.21485834224822811751183620 61\ldots\\
c'''_{\mathbb{Q}}(V_4) &= c''_{\mathbb{Q}}(V_4)-\frac{3}{\pi^2}+\frac{7}{8\pi^2}\prod_{p\equiv 1 \pmod 4}\frac{(1+3/p)(1-1/p)}{(1+1/p)^2}\\
&= -0.44386478005469691086642198 85\ldots\ ,
\end{aligned}$$

where γ is Euler's constant,

$$\gamma_1=\lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{\log k}{k}-\frac{\log^2 n}{2n}\right)\ ,$$

and $c'''_{\mathbb{Q}}(V_4)$ will be used below. The value of $c_{\mathbb{Q}}(C_4)$ is due in principle to A. Baily [1], with computational errors.

$$\begin{aligned}
N_{4,0}(V_4,X) &= \left(\frac{c_{\mathbb{Q}}(V_4)}{4}\log^2 X+\frac{c'_{\mathbb{Q}}(V_4)}{4}\log X+\frac{c'''_{\mathbb{Q}}(V_4)}{4}\right)X^{1/2}+o(X^{1/2})\\
N_{2,1}(V_4,X) &= 0\\
N_{0,2}(V_4,X) &= \left(\frac{3}{4}c_{\mathbb{Q}}(V_4)\log^2 X+\frac{3}{4}c'_{\mathbb{Q}}(V_4)\log X+\left(c''_{\mathbb{Q}}(V_4)-\frac{c'''_{\mathbb{Q}}(V_4)}{4}\right)\right)X^{1/2}\\
&\quad+o(X^{1/2})\ .
\end{aligned}$$

$$\begin{aligned}
N_{\mathbb{Q},4}(V_4,10^{36}) &= 2295681568134760588 4\\
N_{4,0}(V_4,10^{36}) &= 568195231088342425 5\\
N_{0,2}(V_4,10^{36}) &= 1727486337046418162 9\ .
\end{aligned}$$

2.6 $G = D_4$

$$N_{K,4}(D_4, X) \sim c_K(D_4)\,X \quad \text{with}$$

$$c_K(D_4) = \sum_{\substack{[k:K]=2 \\ k\subset\overline{\mathbb{Q}}}} \frac{2^{-r_2(k)}}{\mathcal{N}(\mathfrak{d}(k/K))^2} \frac{\zeta_k(1)}{\zeta_k(2)}\ .$$

In particular,

$$c_{\mathbb{Q}}(D_4) = \frac{6}{\pi^2}\sum_D \frac{2^{-r_2(D)}}{D^2}\frac{L_D(1)}{L_D(2)} = 0.1046520224\dots\ ,$$

where the sum runs over all (positive or negative) discriminants of quadratic fields, $r_2(D) = r_2(\mathbb{Q}(\sqrt{D}))$, and $L_D(s)$ is the Dirichlet L-series attached to the character $\left(\frac{D}{n}\right)$. Note that we do not know how to compute this sum in any other way than the naive method combined with extrapolation techniques, hence we know only about 9 or 10 decimals. These results are due to the authors [14].

Denote by $c^{\pm}(D_4)$ the sum analogous to $c_{\mathbb{Q}}(D_4)$ but where the discriminants D are taken only with the given sign $+$ or $-$. For all $\varepsilon > 0$ we have

$$\begin{aligned} N_{4,0}(D_4, X) &= \frac{c^+(D_4)}{4}\,X + O(X^{3/4+\varepsilon}) \\ N_{2,1}(D_4, X) &= \frac{c^+(D_4)}{2}\,X + O(X^{3/4+\varepsilon}) \\ N_{0,2}(D_4, X) &= \left(\frac{c^+(D_4)}{4} + c^-(D_4)\right)\,X + O(X^{3/4+\varepsilon})\ , \end{aligned}$$

and

$$c^+(D_4) = 0.03942275154\dots, \quad c^-(D_4) = 0.06522927087\dots$$

In the asymptotic formula for $N_{0,2}(D_4, X)$, the term $(c^+(D_4)/4)\,X$ (respectively, $c^-(D_4)\,X$) counts the number $N_{0,2}^+(D_4, X)$ (resp., $N_{0,2}^-(D_4, X)$) of D_4-extensions having a real (resp., imaginary) quadratic subfield.

Using genus theory and more general character manipulation in a suitable way, one can compute (see [8], [14])

$$\begin{aligned} N_{\mathbb{Q},4}(D_4, 10^{17}) &= 10465196820067560 \\ N_{4,0}(D_4, 10^{17}) &= 985567460375496 \\ N_{2,1}(D_4, 10^{17}) &= 1971137479589546 \\ N_{0,2}(D_4, 10^{17}) &= 7508491880102518 \\ N_{0,2}^+(D_4, 10^{17}) &= 985567476224554 \\ N_{0,2}^-(D_4, 10^{17}) &= 6522924403877964\ . \end{aligned}$$

2.7 $G = A_4$

Set $b_K = 2$ if $\zeta_3 \in K$, and $b_K = 1$ if $\zeta_3 \notin K$. A heuristic reasoning given in [21] leads to the conjecture

$$N_{K,4}(A_4, X) \sim c_K(A_4)\, X^{1/2} \log^{b_K} X \ ,$$

where $c_K(A_4)$ is a complicated but explicit constant. For example we should have in particular

$$\begin{aligned} c_{\mathbb{Q}}(A_4) &= \lim_{N\to\infty} \frac{4}{3\zeta(3)\log 2} \sum_{\substack{K_3 \\ N<f(K_3)\le 2N}} \frac{h(K_3)R(K_3)c_2(K_3)c_r(K_3)}{f(K_3)^2} P(K_3) \\ &= 0.074\dots \ , \end{aligned}$$

with

$$P(K_3) = \prod_{p \text{ split in } K_3} \frac{(1+3/p)(1-1/p)^2}{1+1/p+1/p^2} \ ,$$

where K_3 ranges over all cyclic cubic extensions of $\mathbb{Q}$ up to isomorphism (which can easily be described explicitly), $f(K_3)$, $h(K_3)$, $R(K_3)$ denote the conductor, class number and regulator of K_3,

$$c_r(K_3) = \prod_{p \mid f(K_3)} \frac{1}{1+1/p+1/p^2}$$

and $c_2(K_3) = 11/4$ if 2 is inert in K_3, while $c_2(K_3) = 23/10$ if 2 is totally split in K_3. Note that $\zeta_{K_3}(1) = 4h(K_3)R(K_3)/f(K_3)$.

We would like to point out that contrary to what was stated in [12] and [13], we have not yet succeeded in proving that the above conjectural formula is valid. In addition, the constants $c_2(K_3)$ given in those papers are off by a factor 2, although the given numerical value for $c_{\mathbb{Q}}(A_4)$ is correct.

Conjecturally, we have similarly

$$\begin{aligned} N_{4,0}(A_4, X) &\sim c_{4,0}(A_4)\, X^{1/2} \log X \\ N_{0,2}(A_4, X) &\sim c_{0,2}(A_4)\, X^{1/2} \log X \end{aligned}$$

for other explicit constants $c_{4,0}(A_4) = 0.020\dots$ and $c_{0,2}(A_4) = 0.054\dots$.

This method is not only heuristic, since it leads to the exact (and rigorous) computation of

$$\begin{aligned} N_{\mathbb{Q},4}(A_4, 10^{16}) &= 218369252 \\ N_{4,0}(A_4, 10^{13}) &= 1417208 \\ N_{0,2}(A_4, 10^{13}) &= 3861216 \ . \end{aligned}$$

We have computed the result with signatures only up to 10^{13} because it is considerably harder.

2.8 $G = S_4$

At the time of this writing, the conjecture

$$N_{K,4}(S_4, X) \sim c_K(S_4)\, X \ ,$$

with

$$c_{K,4}(S_4) = 2\left(\frac{5}{12}\right)^{r_1}\left(\frac{1}{24}\right)^{r_2}\prod_{\mathfrak{p}}\left(1 + \frac{1}{\mathcal{N}\mathfrak{p}^2} - \frac{1}{\mathcal{N}\mathfrak{p}^3} - \frac{1}{\mathcal{N}\mathfrak{p}^4}\right) \ ,$$

is very close to being proved. More precisely, in [36], using Shintani's theory of prehomogeneous vector spaces, Yukie proves that $N_{K,4}(S_4, X) = O(X \log^2 X)$, and that the above precise conjecture is true assuming some very reasonable convergence arguments. Very possibly his proof will be completed soon. Using quite elementary although very subtle arguments, in [4] and [5], Bhargava proves the above conjecture for $K = \mathbb{Q}$, and also with signatures.

Thus, if we set

$$z(S_4) = \prod_p \left(1 + \frac{1}{p^2} - \frac{1}{p^3} - \frac{1}{p^4}\right) = 1.21669028690633093376943908 7\ldots \ ,$$

then $c_{\mathbb{Q}}(S_4) = (5/6)z(S_4)$ and

$$\begin{aligned} N_{4,0}(S_4, X) &\sim \frac{1}{12}z(S_4) \\ N_{2,1}(S_4, X) &\sim \frac{1}{2}z(S_4) \\ N_{0,2}(S_4, X) &\sim \frac{1}{4}z(S_4) \ . \end{aligned}$$

Using our Kummer-theoretic methods, we compute that

$$\begin{aligned} N_{\mathbb{Q},4}(S_4, 10^7) &= 6541232 \\ N_{4,0}(S_4, 10^7) &= 482488 \\ N_{2,1}(S_4, 10^7) &= 3958348 \\ N_{0,2}(S_4, 10^7) &= 2100396 \ . \end{aligned}$$

However Bhargava's method gives us a much more efficient way to compute these quantities exactly, so these results will certainly be superseded in the near future.

3 More General Results

As already mentioned in Section 1, it is quite plausible that one can obtain general results for all solvable groups, and perhaps also for the groups A_5 and S_5. In addition to the results and conjectures given in Section 1, the only results known to the authors are the following.

3.1 $G = C_\ell$ with ℓ Prime

The result is due to the authors (see [10], [16], [17]). Before stating it, we need some notations. Let $K_z = K(\zeta_\ell)$, $d_z = [K_z : K]$, and $q_z = (\ell-1)/d_z$. For every divisor d of d_z, we let $K_z[d]$ be the unique subextension of K_z/K such that $[K_z : K_z[d]] = d$ or, equivalently, $[K_z[d] : K] = d_z/d$. If $\mathfrak{p}$ is a prime ideal of K, we denote by $e(\mathfrak{p}_d/\mathfrak{p})$, $f(\mathfrak{p}_d/\mathfrak{p})$, and $g(\mathfrak{p}_d/\mathfrak{p})$ the ramification index, residual degree, and number, of prime ideals $\mathfrak{p}_d$ of $K_z[d]$ above $\mathfrak{p}$, so that, in particular, $e(\mathfrak{p}_d/\mathfrak{p})f(\mathfrak{p}_d/\mathfrak{p})g(\mathfrak{p}_d/\mathfrak{p}) = d_z/d$. For any integer e, we denote by $r(e)$ the unique integer such that $e \equiv r(e) \pmod{\ell-1}$ with $1 \le r(e) \le \ell-1$. Finally, denote by $\mathcal{R}$ (resp. $\mathcal{D}$) the set of prime ideals of K which are ramified (resp. totally split) in K_z/K ($\mathcal{D}$ being the set of all prime ideals of K when $\zeta_\ell \in K$). Then

$$N_{K,\ell}(C_\ell, X) \sim c_K(C_\ell)\, X^{1/(\ell-1)} \log^{q_z-1} X \ ,$$

with $c_K(C_\ell) = c_1c_2c_3c_4$ and

$$\begin{aligned}
c_1 &= \frac{\left(\prod_{d\mid d_z} \zeta_{K_z[d]}(d)^{\mu(d)}\right)^{q_z}}{\ell^{r_2+r_z}(\ell-1)^{q_z}(q_z-1)!}\ ,\\
c_2 &= \prod_{\mathfrak{p}\in\mathcal{D}} \left(\left(1+\frac{\ell-1}{\mathcal{N}\mathfrak{p}}\right)\prod_{d\mid d_z}\left(1-\frac{1}{\mathcal{N}\mathfrak{p}^d}\right)^{(\ell-1)\mu(d)/d}\right)\ ,\\
c_3 &= \left(\prod_{\mathfrak{p}\in\mathcal{R}}\prod_{d\mid d_z}\left(1-\frac{1}{\mathcal{N}\mathfrak{p}^{df(\mathfrak{p}_d/\mathfrak{p})}}\right)^{g(\mathfrak{p}_d/\mathfrak{p})\mu(d)}\right)^{q_z}\ ,\\
c_4 &= \prod_{\mathfrak{p}\mid\ell,\ \mathfrak{p}\notin\mathcal{D}}\left(1+\frac{\ell-1}{\mathcal{N}\mathfrak{p}}-\frac{\ell-1-r(e(\mathfrak{p}))(1-1/\mathcal{N}\mathfrak{p})}{\mathcal{N}\mathfrak{p}^{\lceil e(\mathfrak{p})/(\ell-1)\rceil}}\right)\ ,
\end{aligned}$$

where $r_z = 0$ if $\zeta_\ell \in K$, while $r_z = r_1 - 1$ otherwise.

In particular, for $\ell > 2$ we have $N_{\mathbb{Q},\ell}(C_\ell, X) \sim c_{\mathbb{Q},\ell}(C_\ell)\, X^{1/(\ell-1)}$ with

$$c_{\mathbb{Q}}(C_\ell) = \frac{\ell^2+\ell-1}{\ell^2(\ell-1)} \prod_{d\mid\ell-1} (\zeta_{\mathbb{Q}(\zeta_\ell)[d]}(d))^{\mu(d)} \prod_{d\mid\ell-1}\left(1-\frac{1}{\ell^d}\right)^{\mu(d)}$$
$$\prod_{p\equiv 1 \pmod{\ell}} \left(\left(1+\frac{\ell-1}{p}\right)\prod_{d\mid\ell-1}\left(1-\frac{1}{p^d}\right)^{(\ell-1)\mu(d)/d}\right)\ .$$

3.2 Nilpotent Groups

The best known result is due to Klüners–Malle [26]. They prove that the weak form of Malle's Conjecture 2 is true for a nilpotent group G in its regular representation, in other words that for all $\varepsilon > 0$ and sufficiently large X, we have

$$c_K(G)\cdot X^{a(G)} < N_{K,n}(G,X) < X^{a(G)+\varepsilon}$$

for some strictly positive constant $c_K(G)$, where $a(G)$ is the exponent given by Malle's Conjecture 2. They also prove that the same is true for more general groups, such as for example the wreath product of a nilpotent group with the cyclic group of order 2.

References

1. A. Baily, *On the density of discriminants of quartic fields*, J. reine angew. Math. **315** (1980), 190–210.
2. K. Belabas, *A fast algorithm to compute cubic fields*, Math. Comp. **66** (1997), 1213–1237.
3. K. Belabas, *On the mean 3-rank of quadratic fields*, Compositio Math. **118** (1999), 1–9.
4. M. Bhargava, *Higher composition laws*, PhD Thesis, Princeton Univ., June 2001.
5. M. Bhargava, *Gauss Composition and Generalizations*, this volume.
6. H. Cohen, *Advanced topics in computational number theory*, GTM **193**, Springer-Verlag, 2000.
7. H. Cohen, *Comptage exact de discriminants d'extensions abéliennes*, J. Th. Nombres Bordeaux **12** (2000), 379–397.
8. H. Cohen, *Enumerating quartic dihedral extensions of $\mathbb{Q}$ with signatures*, 32p., submitted.
9. H. Cohen, *High precision computation of Hardy–Littlewood constants*, preprint available on the author's web page.
10. H. Cohen, F. Diaz y Diaz and M. Olivier, *Densité des discriminants des extensions cycliques de degré premier*, C. R. Acad. Sci. Paris **330** (2000), 61–66.
11. H. Cohen, F. Diaz y Diaz and M. Olivier, *Construction of tables of quartic fields using Kummer theory*, Proceedings ANTS IV, Leiden (2000), Lecture Notes in Computer Science **1838**, Springer-Verlag, 257–268.
12. H. Cohen, F. Diaz y Diaz and M. Olivier, *Counting discriminants of number fields of degree up to four*, proceedings ANTS IV, Leiden (2000), Lecture Notes in Comp. Sci., **1838**, Springer-Verlag (2000), 269–283.
13. H. Cohen, F. Diaz y Diaz and M. Olivier, *Counting discriminants of number fields*, MSRI preprint **2000-026** (2000), 9p, available on the MSRI www server.
14. H. Cohen, F. Diaz y Diaz and M. Olivier, *Enumerating quartic dihedral extensions*, Compositio Math., 28p., to appear.
15. H. Cohen, F. Diaz y Diaz and M. Olivier, *Constructing complete tables of quartic fields using Kummer theory*, Math. Comp., 11p., to appear.
16. H. Cohen, F. Diaz y Diaz and M. Olivier, *On the density of discriminants of cyclic extensions of prime degree*, J. reine und angew. Math., 40p., to appear.
17. H. Cohen, F. Diaz y Diaz and M. Olivier, *Cyclotomic extensions of number fields*, 14p., submitted.
18. H. Cohen, F. Diaz y Diaz and M. Olivier, *Counting discriminants of number fields*, 36p., submitted.
19. H. Cohen, F. Diaz y Diaz and M. Olivier, *Counting cyclic quartic extensions of a number field*, 30p., submitted.
20. H. Cohen, F. Diaz y Diaz and M. Olivier, *Counting biquadratic extensions of a number field*, 17p., submitted.
21. H. Cohen, F. Diaz y Diaz and M. Olivier, *Counting A_4 and S_4 extensions of number fields*, 20p., in preparation.

22. H. Cohn, *The density of abelian cubic fields*, Proc. Amer. Math. Soc. **5** (1954), 476–477.
23. B. Datskovsky and D. J. Wright, *Density of discriminants of cubic extensions*, J. reine angew. Math. **386** (1988), 116–138.
24. H. Davenport and H. Heilbronn, *On the density of discriminants of cubic fields I*, Bull. London Math. Soc. **1** (1969), 345–348.
25. H. Davenport and H. Heilbronn, *On the density of discriminants of cubic fields II*, Proc. Royal. Soc. A **322** (1971), 405–420.
26. J. Klüners and G. Malle, *Counting Nilpotent Galois Extensions*, submitted.
27. S. Mäki, *On the density of abelian number fields*, Thesis, Helsinki, 1985.
28. S. Mäki, *The conductor density of abelian number fields*, J. London Math. Soc. **(2) 47** (1993), 18–30.
29. G. Malle, *On the distribution of Galois groups*, J. Number Theory, to appear.
30. D. Roberts, *Density of cubic field discriminants*, Math. Comp. **70** (2001), 1699–1705.
31. T. Shintani, *On Dirichlet series whose coefficients are class numbers of integral binary cubic forms*, J. Math. Soc. Japan **24** (1972), 132–188.
32. T. Shintani, *On zeta-functions associated with the vector space of quadratic forms*, J. Fac. Sci. Univ. Tokyo, Sec. 1a **22** (1975), 25–66.
33. G. Tenenbaum, *Introduction à la théorie analytique et probabiliste des nombres*, Cours Spécialisés SMF **1**, Société Mathématique de France, 1995.
34. D. J. Wright, *Distribution of discriminants of Abelian extensions*, Proc. London Math. Soc. (3) **58** (1989), 17–50.
35. D. J. Wright and A. Yukie, *Prehomogeneous vector spaces and field extensions*, Invent. Math. **110** (1992), 283–314.
36. A. Yukie, *Density theorems related to prehomogeneous vector spaces*, preprint.

A Higher-Rank Mersenne Problem

Graham Everest, Peter Rogers, and Thomas Ward

School of Mathematics, University of East Anglia, Norwich NR4 7TJ, U.K.
G.Everest@uea.ac.uk,
http://www.mth.uea.ac.uk/people/gre.html

Abstract. The classical Mersenne problem has been a stimulating challenge to number theorists and computer scientists for many years. After briefly reviewing some of the natural settings in which this problem appears as a special case, we introduce an analogue of the Mersenne problem in higher rank, in both a classical and an elliptic setting. Numerical evidence is presented for both cases, and some of the difficulties involved in developing even a heuristic understanding of the problem are discussed.

1 Introduction

The Mersenne problem asks if $M_n = 2^n - 1$ is prime for infinitely many values of n. Three and a half centuries after Mersenne's death this problem remains inaccessible. In addition to their position in number theory, Mersenne primes have arisen in diverse areas of mathematics, including group theory [11], ergodic theory [26] and string theory [12]. Their properties have also led some fine minds astray [2]. Wagstaff [25] modified some considerations by Gillies [13] to produce a heuristic argument of the following shape about the distribution of Mersenne primes: If various congruences satisfied by the Mersenne numbers behave like independent probabilistic events, then the number of Mersenne primes less than X should be about

$$\frac{e^{\gamma}}{\log 2} \log\log X = (2.5695\ldots) \log\log X.$$

Moreover, if $n_1, \ldots, n_r$ are the primes for which M_{n_j} is prime, then the argument predicts that

$$\frac{\log\log M_{n_j}}{j} \longrightarrow \frac{\log 2}{e^{\gamma}}. \tag{1}$$

There is little hope that this heuristic argument could ever be tightened up to become a proof, but it is certainly suggestive. For example, plotting $\log\log M_{n_j}$ against j gives an extremely close agreement with the prediction – though it is hard to attach statistical significance to a finite sample of an infinite problem. The 39 known Mersenne primes behave very much in accordance with (1) – see the Prime Pages [3] for the details. The reason so few Mersenne primes are known is that the rapid growth rate in the sequence $(2^n - 1)$ means that

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 95–107, 2002.

huge numbers must be tested for primality, and although the special shape of Mersenne numbers permits very rapid prime testing, even finding the first 39 has taken thousands of computers many years, running a distributed program.

2 Other Settings of the Mersenne Problem

One approach to the Mersenne problem is to try to see it in different contexts; several of these will be described below. A remarkable feature of the second and third of these is that for some special cases it is possible to *prove* the appearance of infinitely many primes. Our purpose here is to expand on the fourth and fifth of these, and to describe heuristic and computational evidence for the expected behaviour. There are sharp generalisations or modifications of the Mersenne problem to other specific questions (for example, see [1], [19]); we are primarily interested in naturally arising *families* of problems which may shed some light on the Mersenne problem.

2.1 Lehmer–Pierce Sequences

Fix a monic polynomial $f(x) = x^d + a_{d-1}x^{d-1} + \ldots + a_0 \in \mathbb{Z}[x]$, with factorization over $\mathbb{C}$

$$f(x) = (x - \alpha_1) \ldots (x - \alpha_d). \tag{2}$$

Following Pierce and Lehmer, associate a sequence of integers to f by defining

$$\Delta_n(f) = \prod_{i=1}^{d} |\alpha_i^n - 1| \text{ for } n \geq 1. \tag{3}$$

For the polynomial $f(x) = x - 2$ these are the Mersenne numbers. In any case, the resulting sequence is again a divisibility sequence, and an analogue of the heuristic arguments of Wagstaff may be applied to it (once generic divisibility is taken care of: $\Delta_n(f)$ is always divisible by $\Delta_1(f)$; if f is a reciprocal polynomial then $\Delta_n(f)/\Delta_1(f)$ is always a square when n is odd). The rate of growth of the sequence is determined by the *Mahler measure* of the polynomial f, and by choosing polynomials with small Mahler measure the growth rate of $\Delta_n(f)$ can be reduced dramatically. Lehmer [16] studied these sequences with the view of using them to produce large primes in novel ways. Recently, his approach was revisited using modern computing methods, together with the heuristic argument of Wagstaff. The upshot of this work is described in [6], where sequences have been found with many hundreds of primes, and a reasonable agreement with the heuristic model is found.

2.2 Primes from Dynamical Systems

The Lehmer–Pierce sequences all arise from algebraic dynamical systems in the following sense. Call a sequence $(u_n)_{n\geq 1}$ *algebraically realisable* if there is a compact group endomorphism $T : X \to X$ with the property that

$$u_n = |\mathrm{Per}_n(T)| = |\{x \in X \mid T^n(x) = x\}|.$$

Such a sequence must be a divisibility sequence in addition to being *realisable* (a general combinatorial notion expressing the property of being the periodic points for some map – see [20] for the details). The converse is not true, and only a partial characterization of algebraically realisable sequences is known.

Any divisibility sequence must satisfy $u_1|u_n$ for all n, but it seems reasonable to ask whether the quotient might be prime infinitely often. The Lehmer–Pierce sequences are a natural family of algebraically realizable sequences that are conjectured to be prime infinitely often (once this kind of generic divisibility is taken account of). It turns out that many other natural families of group automorphisms have a similar property: Example 1 shows that the even Bernoulli denominators have this property. Studying primality from this point of view gives a conjectural explanation for the infinitude of both Mersenne and Sophie-German primes within the same context. Example 2 gives some hope that such sequences might indeed be prime infinitely often.

Example 1. Let B_n be defined by

$$\frac{t}{e^t-1}=\sum_{n=0}^{\infty}B_nt^n/n!$$

Then the sequence b_n = denominator(B_{2n}) is algebraically realisable.

To see this, define $X_p=\mathbb{F}_p=\mathbb{Z}/p\mathbb{Z}$. For $p=2$ define T_p to be the identity. For $p>2$, let g_p denote an element of (multiplicative) order $(p-1)/2$. Define $T_p:X_p\to X_p$ to be the endomorphism $T_p(x)=g_px \bmod p$. Plainly $|\mathrm{Per}_n(T_p)|=p$ if and only if $p-1|2n$; for all other n, $|\mathrm{Per}_n(T_p)|=1$. The Clausen–von Staudt Theorem ([14], [15]) states that

$$B_{2n}+\sum\frac{1}{p}\in\mathbb{Z},$$

where the sum ranges over the primes p for which $p-1|2n$. Thus $|\mathrm{Per}_n(T_p)|=\max\{1,|B_{2n}|_p\}$. Now define

$$X=\prod_p X_p \text{ and } T=\prod_p T_p.$$

This shows the algebraic realisability of the Bernoulli denominators.

Notice that a prime value of b_n/b_1 can only occur if n is a Sophie-Germain prime. There are believed to be infinitely many Sophie-Germain primes but no proof is available – see [21].

The next example is a group endomorphism with a very similar shape to that of Example 1, but constructed so as to be certain that the periodic point sequence will be prime infinitely often. This example was inspired by a remark of Gerry McLaren.

Example 2. There is a group endomorphism $T:X\to X$ such that $|\mathrm{Per}_n(T)|$ takes on infinitely many distinct prime values. To see this, construct a set S

of prime numbers recursively as follows. Firstly, $2 \in S$ and a prime $p \in S$ if and only if $p-1$ is divisible by a prime $q = q_p$ which does not divide $p'-1$ for all $p' \in S$ with $p' < p$. Clearly S is infinite – otherwise all sufficiently large primes could be written as $1+p_1^{e_1}\dots p_r^{e_r}$ for some fixed set of primes $\{p_1,\dots,p_r\}$, where $e_1,\dots,e_r$ lie in $\mathbb{N}$. The number of such primes less than or equal to X is $O((\log X)^r)$, which contradicts the Prime Number Theorem.

For each prime $p \in S$, let h_p denote an element of multiplicative order $q = q_p$ in $X_p = \mathbb{F}_p$, and define an endomorphism $T_p : X_p \to X_p$ by $T_p(x) = h_p x$. Then define an endomorphism T on X by

$$X = \prod_{p\in S} X_p \text{ and } T = \prod_{p \in S} T_p.$$

Clearly $|\mathrm{Per}_{q_p}(T)| = p$ for all p, showing that the sequence $(|\mathrm{Per}_n(T)|)$ takes on infinitely many distinct prime values.

2.3 Mersenne Problem in $\mathbb{A}$-Fields

Let k be an $\mathbb{A}$-field (that is, an algebraic number field or a finite extension of a rational function field $\mathbb{F}_q(t)$ of positive characteristic) with set of places $\mathbb{P}(k)$ (see [28] for a discussion of places). Fix $\xi \in k\backslash\{0\}$, not a unit root. Then the generalized Mersenne problem asks if there is a constant $B(\xi)$ with the property that the set

$$P_n = \{\nu \in \mathbb{P}(k) \mid |\xi^n - 1|_\nu \neq 1\}$$

has no more than $B(\xi)$ elements for infinitely many n. For $k = \mathbb{Q}$ and $\xi = 2$, this is a weak form of the classical Mersenne problem (in that it only asks for infinitely many numbers $2^n - 1$ to have a uniformly bounded number of prime factors). This problem has arisen in ergodic theory [26], [27] and has the following remarkable feature: There are many cases for which it is certainly true, though the proofs are not trivial. Specifically, a consequence of Heath–Brown's work on the Artin conjecture is that $|P_n| = 2$ infinitely often for many of the positive characteristic cases (see [27] for the details).

3 A Higher-Rank Mersenne Problem

The dynamical systems alluded to above have very natural higher-rank analogues, namely the $\mathbb{Z}^d$-actions generated by d commuting automorphisms of a compact abelian group X (see [18], [22] for a discussion of these dynamical systems). For these the periodic point behaviour is very complicated (some of these problems are described in [17] in a different context), and we simply extract one simple question from the simplest example available. Does the set

$$\{3^m 2^n - 1 \mid m, n \geq 0\}$$

contain infinitely many primes? Can anything be said – even heuristically – about the quantity

$$N^-(X) = |\{(m,n) \mid 3^m 2^n - 1 \text{ is prime and } m,n \leq X\}|? \tag{4}$$

This problem will be discussed in this section, along with the same question for the quantity $N^+(X)$ associated to $3^m 2^n + 1$, which is quite different in that it certainly does not come from a pair of commuting group automorphisms.

3.1 Heuristics

The heuristic argument below takes the form of a family of successive refinements of the same basic idea. Let $N^-(X)$ be defined by (4). In the discussion below, we will essentially ignore the cases $n = 0$ (for which $3^m 2^n - 1$ is always even) and $m = 0$ (the Mersenne case) since they together contribute so few primes. The discussion leads to a prediction that

$$\frac{N^-(X)}{X} \to C^- \text{ as } X \to \infty, \tag{5}$$

where C^- is a constant. The section ends with a graph to illustrate the accuracy of the prediction. We will also exhibit a graph for primes of the form $3^m 2^n + 1$.

The Prime Number Theorem implies that the probability a large random integer K is prime is approximately $\frac{1}{\log K}$. This suggests that $N^-(X)$ is approximately

$$N_1(X) = \sum_{1 \leq m,n < X} \frac{1}{n \log 2 + m \log 3} \tag{6}$$

which is given asymptotically by the double integral

$$\int_1^X \int_1^X \frac{1}{x \log 2 + y \log 3} \,\mathrm{d}x \,\mathrm{d}y,$$

so

$$N_1(X) = DX + O(\log X),$$

where

$$D = \frac{\log 6 \log\log 6 - \log 2 \log\log 2 - \log 3 \log\log 3}{\log 2 \log 3} = 1.57\ldots.$$

3.2 Obvious Congruences

For $m, n \geq 1$, $3^m 2^n - 1$ is coprime with 6. The usual Euler factor correction suggests that we should therefore increase our estimate for $N^-(X)$ by a factor of $\frac{2}{2-1} \cdot \frac{3}{3-1} = 3$. This gives a refined heuristic: Having taken account of the Prime Number Theorem and the primes 2 and 3, we expect $N^-(X)$ to be approximated by $N_2(X)$, where

$$\frac{N_2(X)}{X} \sim 4.71\ldots$$

3.3 Less Obvious Congruences

It is tempting to continue exactly as above. Consider the prime $q = 5$ and the congruence

$$3^m 2^n - 1 \equiv 0 \bmod 5.$$

The solutions are all the pairs (m, n) which reduce mod 4 to lie in the set $\{(1,1),(2,2),(3,3),(4,4)\}$. Thus asymptotically $\frac{3}{4}$ of the numbers of the form $3^m 2^n - 1$ are not divisible by 5; on the other hand $\frac{4}{5}$ of all numbers are not divisible by 5. This suggests that the heuristic estimate taking account of the prime 5 as well should be $\frac{5}{4} \cdot \frac{3}{4} \cdot N_2(X)$, leading to the estimate

$$\frac{N_3(X)}{X} \sim 4.416\ldots.$$

It is at this point that the first substantial difficulty is encountered. The proportion of numbers of the form $3^m 2^n - 1$ that are not divisible by 5 or 7 cannot be found by emulating this calculation mod 4 and 6 separately – we have to search in residue classes mod $12 = \operatorname{lcm}(4, 6)$.

3.4 Taking Account of Primes Less than *L*

The calculation to find the correcting factor for primes q, $3 < q < L$, goes as follows. Let P_L denote the least common multiple of $q - 1$ as q runs over the primes between 3 and L. For each residue pair (j, k) in $(\mathbb{Z}/P_L\mathbb{Z})^2$, and for each such prime q, reduce (j, k) mod q and decide whether

$$3^j 2^k - 1 \equiv 0 \bmod q.$$

Delete those residue pairs that satisfy this congruence for some q; call the remaining set Q_L. Then the heuristic argument suggests that we should correct by this factor and the usual Euler factor to give

$$N_L(X) = \frac{|Q_L|}{P_L^2} \cdot \prod_{3<q<L} \frac{q}{q-1} \cdot N_2(X).$$

This has two distinct pieces: the second factor is readily estimated using Merten's Theorem [14, Th. 429] which says that

$$\frac{1}{\log L} \prod_{2 \leq q < L} \frac{q}{q-1} \to e^{\gamma}, \text{ as } L \to \infty.$$

The other factor presents computational and theoretical problems: Computationally, P_L grows very rapidly in L, and the exact calculation of $|Q_L|$ requires manipulating set-memberships which is slow. However, approximations can be made easily by simple counting arguments. It is possible that results on the higher-rank Artin problem (conditional on GRH) would give more precise information, but we have not pursued this as Q_L already arises inside a heuristic argument.

3.5 A Comparison of Heuristic and Experimental Evidence

As described above, calculating exact values for $|Q_L|$ involves searching over a set of size P_L^2 (for primes up to $L = 29$, a calculation over a set of size 55440^2 is involved). Bearing in mind the sometimes delicate balance between computation time and accuracy of results we fix L and estimate $|Q_L|/P_L^2$ by counting the number of pairs (m, n) with $m, n < X$ and $\gcd(2^m 3^n - 1, \prod_{p<L} p) > 1$, then divide by X^2. Experiments suggest that for given L this converges rapidly in X, and a good approximation is found even when X is of the order of L. For $L = 1000$ the calculation suggests the further refined heuristic

$$\frac{N_L(X)}{X} \sim 4.043\ldots$$

The experimental evidence strongly supports a conjecture of the form (5), which suggests that

$$\log L \cdot \frac{|Q_L|}{P_L^2}$$

converges as $L \to \infty$. Figure 1 shows a graph of the number $N^-(X)$ of primes of the form $3^m 2^n - 1$ with $m, n < X$ against X for values of $X \leq 1000$. The gradient of this graph is approximately $C^- = 3.7$, as compared with our most refined heuristic suggestion of $C^- = 4.043\ldots$. However, the conjectured linearity is strongly supported by this numerical data.

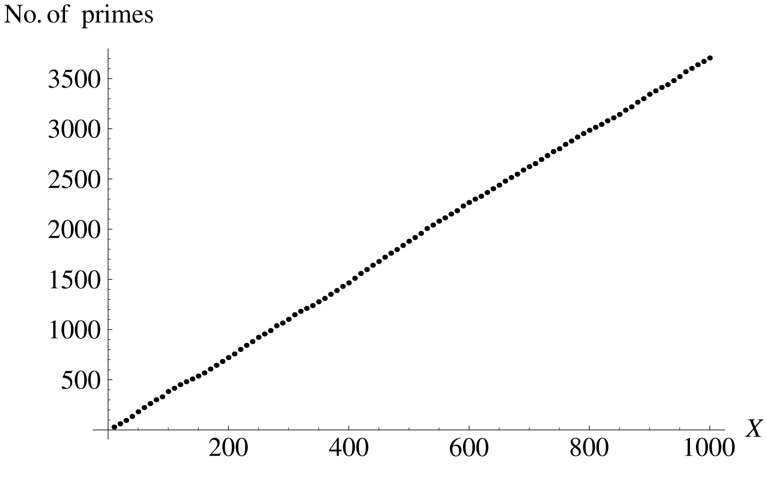


Fig. 1. Graph of $N^-(X)$ against X for $X \leq 1000$

Much of what we have said for primes of the form $3^m 2^n - 1$ can be replicated for primes of the form $3^m 2^n + 1$. That is to say, the heuristic argument above can be applied in this case also, taking into account the possible difference in the value of $|Q_L|/P_L^2$. Let $N^+(X)$ denote the number of primes of the form $3^m 2^n + 1$ with $m, n \leq X$. We expect

$$\frac{N^+(X)}{X} \to C^+, \text{ as } X \to \infty.$$

Figure 2 shows a graph of $N_+(X)$ against X for $X \leq 1000$.

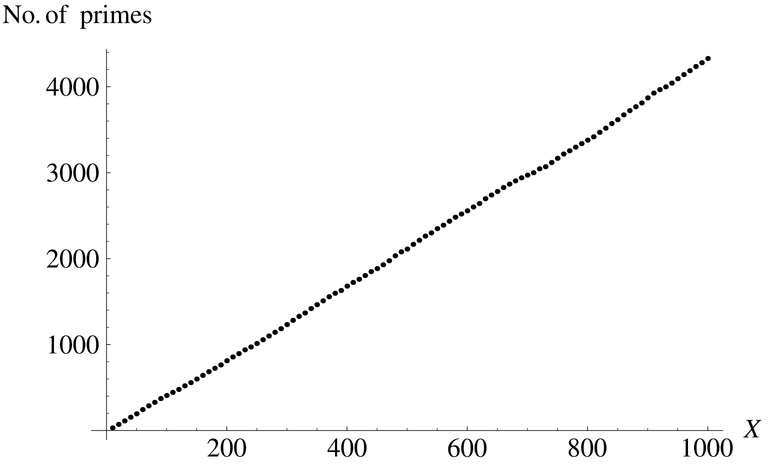


Fig. 2. Graph of $N^+(X)$ against X for $X \leq 1000$

The graph predicts the value of C^+ to be about 4.3. Comparing this with a refined heuristic calculated in an identical fashion to that above, we obtain

$$\frac{N_L^+(X)}{X} \sim 4.258\ldots$$

with $C^+ = 4.258$. This heuristic constant is *extremely* close to the experimental value, though no meaning can attach to this coincidence in light of the N^- case.

4 Elliptic Analogues

There is a dialogue between on the one hand dynamical systems and arithmetical sequences built from the circle (of which the Lehmer–Pierce sequences are the simplest example) and on the other, objects associated to elliptic curves, summarised in Table 1 (the objects on the classical side are described in [10], and on the elliptic side in [8] and [9]).

Let E denote an elliptic curve defined over the rationals (the text [24] covers all the properties of elliptic curves we use), given by a Weierstrass equation

$$y^2 + a_1 xy + a_3 y = x^3 + a_2 x^2 + a_4 x + a_6 \tag{7}$$

Table 1. Classical objects and their elliptic counterparts

classical case	elliptic case
polynomial $f \in \mathbb{Z}[x]$	point P on curve E over $\mathbb{Q}$
Mahler measure $m(f)$	canonical height $h_E(P)$
Lehmer problem	Lang's height conjecture
toral automorphism T_f	sequence of maps
$(\lvert \mathrm{Per}_n(T_f) \rvert)$ (Lehmer–Pierce)	elliptic divisibility sequence

with coefficients $a_1, \ldots, a_6 \in \mathbb{Z}$. The assumption that the curve is an elliptic curve amounts to assuming it is non-singular, that is, the discriminant does not vanish.

How might we expect to use the arithmetic of E to produce primes? Suppose E has a non-torsion rational point $Q \in E(\mathbb{Q})$. The multiples nQ for $n \in \mathbb{N}$ define a sequence of integers as follows: The x-coordinates of these points all have the shape $x(nQ) = t_n/s_n^2$ for integers s_n, t_n. These fascinating sequences were studied in [23]. We could ask whether they are likely to contain many primes - actually, it is sufficient to study s_n. The Chudnovskys did some experimental research in the 80's (see [4] and [5]) and produced some quite large prime values of s_n. Their results have been revisited recently (see [7]) in work that suggests the sequence s_n will only contain *finitely* many primes. Indeed, the sequences in [4] do not produce any additional primes when tested over a much larger range.

It seems very likely that working with translations $P + nQ$ for fixed rational points P and Q would produce similar results. Our heuristic argument depends heavily upon the growth rate of the sequence, and this would not be substantially different for nQ or $P + nQ$.

Suppose now that $E(\mathbb{Q})$ has rank > 1, and choose independent non-torsion points P and Q. Let $s(m, n) \in \mathbb{Z}$ be defined by

$$x(mP + nQ) = t(m, n)/s(m, n)^2. \tag{8}$$

In his PhD thesis the second author gives a heuristic argument, accompanied by much data, to suggest that $s(m, n)$ should take on prime values infinitely often. Indeed, the number of prime values with $|m|, |n| < X$ should be asymptotically $c \log X$, where c is a constant depending upon the finer arithmetic of E. The elliptic regulator (see below) appears in an apparently explicable fashion although the constant is also affected by the finer divisibility properties in a way that is hard to fathom. The sequences $s(m, n)$ provide large primes which can be described unambiguously in a very economical fashion, since $s(m, n)$ grows as the exponential of a positive-definite quadratic form in the variables m and n.

4.1 Heuristics in the Elliptic Case

Let $R_X = \{(m,n) \mid |m|, |n| \leq X\}$. Then the first attempt at a heuristic estimate is that the sum

$$\sum_{R_X} 1/\log s(m,n) \tag{9}$$

is the expected number of prime values of $s(m,n)$ with $(m,n) \in R_X$. Now $\log s(m,n)$ is asymptotically equivalent to a positive definite quadratic form $S(m,n)$, and the asymptotics of the sum

$$\sum_{R_X} 1/S(m,n)$$

are well known: This sum is asymptotically $(2\pi/r)\log X$, where r denotes the determinant of S (r is the *elliptic regulator* of P and Q). This asymptotic arises from comparing the sum with a suitable integral.

As before, this estimate needs refinement. If q denotes any prime then the sequence reduced mod q is periodic in both variables, with period dividing $|E(\mathbb{F}_q)|$. If follows that we can assign a (rational) probability to $s(m,n)$ not being divisible by q. Doing this for the primes $q < L$ gives approximately $c_L X^2$ elements (m,n) in R_X for which $s(m,n)$ is not divisible by primes less than L. Letting $L \to \infty$, we expect approximately $e \log X$ primes, where e depends on E but not X. It is computationally *extremely* difficult to calculate the exact probabilities for various L, but as before approximations via counting arguments are not too difficult to obtain.

4.2 Numerical Data

Figures 3 and 4 show graphs for $N_E(X)$, the number of primes $s(m,n)$ with $|m|, |n| \leq X$ against $\log X$ for two rank-2 elliptic curve E with small regulator.

The curve in Figure 3 is

$$y^2 + y = x^3 - 199x + 1092,$$

with independent rational points $P = (-13, 38)$ and $Q = (-6, 45)$ on the curve, whose regulator is $.0360\ldots$

The curve in Figure 4 is

$$y^2 + y = x^3 - 28x + 52,$$

with independent rational points $P = (-4, 10)$ and $Q = (-2, 10)$ on the curve, whose regulator is $.0813\ldots$

The numerical data is not incompatible with the heuristic suggestion of a linear relationship between $N_E(X)$ and $\log X$, but strongly suggests there are more phenomena here to understand.

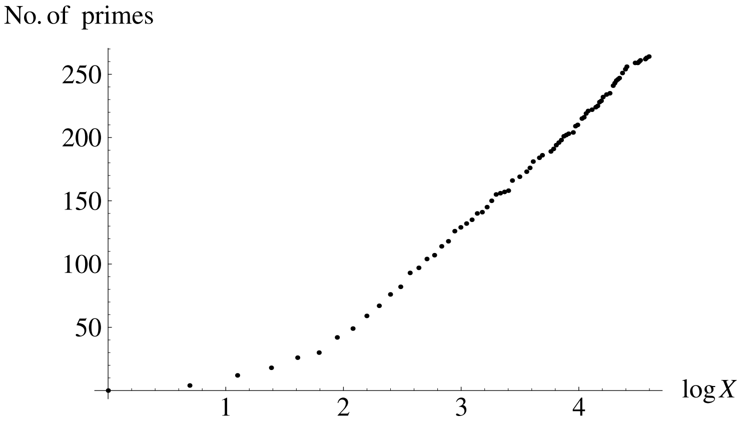


Fig. 3. Graph of $N_E(X)$ against $\log X$ for $X \leq 100$; curve $y^2 + y = x^3 - 199x + 1092$

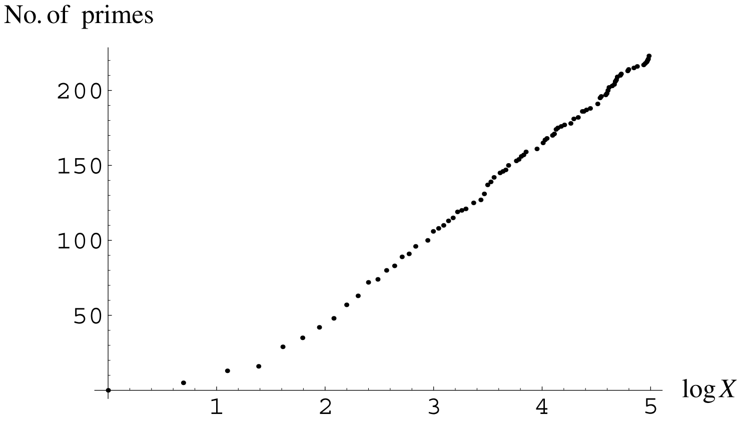


Fig. 4. Graph of $N_E(X)$ against $\log X$ for $X \leq 150$; curve $y^2 + y = x^3 - 28x + 52$

4.3 Conclusion

The classical Mersenne problem appears as a special case in many different settings. In some of these there are other cases in which prime appearance is understood. Two higher-rank analogues of the Mersenne problem are explored.

The first is a direct extension to two variables, and compelling numerical data is available concerning prime appearance.

The second occurs in an elliptic curve setting. The work of [7] suggests there are only finitely many primes in an elliptic divisibility sequence (and possibly a uniform bound on the number of primes for any elliptic divisibility sequence on curves defined over the rationals). A better elliptic analogue of the Mersenne problem therefore seems to be the study of the higher-rank sequences associated to elliptic curves of higher rank.

References

1. P. T. Bateman, J. L. Selfridge, and S. S. Wagstaff, Jr. The new Mersenne conjecture. *Amer. Math. Monthly*, 96(2):125–128, 1989.
2. P. G. Brown. A note on Ramanujan's conjectures regarding "Mersenne's primes". *Austral. Math. Soc. Gaz.*, 24(4):146–147, 1997.
3. Chris Caldwell. The prime pages. `http://www.utm.edu/research/primes/`.
4. D. V. Chudnovsky and G. V. Chudnovsky. Sequences of numbers generated by addition in formal groups and new primality and factorization tests. *Adv. in Appl. Math.*, 7(4):385–434, 1986.
5. D. V. Chudnovsky and G. V. Chudnovsky. Computer assisted number theory with applications. In *Number theory (New York, 1984–1985)*, pages 1–68. Springer, Berlin, 1987.
6. Manfred Einsiedler, Graham Everest, and Thomas Ward. Primes in sequences associated to polynomials (after Lehmer). *LMS J. Comput. Math.*, 3:125–139 (electronic), 2000.
7. Manfred Einsiedler, Graham Everest, and Thomas Ward. Primes in elliptic divisibility sequences. *LMS J. Comput. Math.*, 4:1–13 (electronic), 2001.
8. Manfred Einsiedler, Graham Everest, and Thomas Ward. Entropy and the canonical height. *J. Number Theory*, 91:256–273, 2001.
9. Graham Everest and Thomas Ward. A dynamical interpretation of the global canonical height on an elliptic curve. *Experiment. Math.*, 7(4):305–316, 1998.
10. Graham Everest and Thomas Ward. *Heights of polynomials and entropy in algebraic dynamics.* Springer-Verlag London Ltd., London, 1999.
11. Shalom Feigelstock. Mersenne primes and group theory. *Math. Mag.*, 49(4):198–199, 1976.
12. Paul H. Frampton and Thomas W. Kephart. Mersenne primes, polygonal anomalies and string theory classification. *Phys. Rev. D (3)*, 60(8):087901, 4, 1999.
13. Donald B. Gillies. Three new Mersenne primes and a statistical theory. *Math. Comp.*, 18:93–97, 1964.
14. G. H. Hardy and E. M. Wright. *An introduction to the theory of numbers.* The Clarendon Press Oxford University Press, New York, fifth edition, 1979.
15. Neal Koblitz. *p-adic numbers, p-adic analysis, and zeta-functions.* Springer-Verlag, New York, second edition, 1984.
16. D.H. Lehmer. Factorization of certain cyclotomic functions. *Ann. of Math.* 34 (1933) 461–479.
17. D. A. Lind. A zeta function for $\mathbb{Z}^d$-actions. In *Ergodic theory of* $\mathbb{Z}^d$ *actions (Warwick, 1993–1994)*, pages 433–450. Cambridge Univ. Press, Cambridge, 1996.
18. Douglas Lind, Klaus Schmidt, and Tom Ward. Mahler measure and entropy for commuting automorphisms of compact groups. *Invent. Math.*, 101(3):593–629, 1990.

19. Albert A. Mullin. Letter to the editor: “The new Mersenne conjecture” [Amer. Math. Monthly **96** (1989), no. 2, 125–128; MR 90c:11009] by P. T. Bateman, J. L. Selfridge and S. S. Wagstaff, Jr. *Amer. Math. Monthly*, 96(6):511, 1989.
20. Yash Puri and Thomas Ward. Arithmetic and growth of periodic orbits. *J. Integer Seq.*, 4(1):Article 01.2.1, 17 pp. (electronic), 2001.
21. J. H. Sampson. Sophie Germain and the theory of numbers. *Arch. Hist. Exact Sci.*, 41(2):157–161, 1990.
22. Klaus Schmidt. *Dynamical systems of algebraic origin.* Birkhäuser Verlag, Basel, 1995.
23. Rachel Shipsey. *Elliptic Divisibility Sequences.* PhD thesis, University of London, 2003.
24. Joseph H. Silverman. *The arithmetic of elliptic curves.* Springer-Verlag, New York, 1992. Corrected reprint of the 1986 original.
25. Samuel S. Wagstaff, Jr. Divisors of Mersenne numbers. *Math. Comp.*, 40(161):385–397, 1983.
26. Thomas Ward. An uncountable family of group automorphisms, and a typical member. *Bull. London Math. Soc.*, 29(5):577–584, 1997.
27. Thomas Ward. Almost all S-integer dynamical systems have many periodic points. *Ergodic Theory Dynam. Systems*, 18(2):471–486, 1998.
28. André Weil. *Basic number theory.* Springer-Verlag, New York, third edition, 1974. Die Grundlehren der Mathematischen Wissenschaften, Band 144.

An Application of Siegel Modular Functions to Kronecker's Limit Formula

Takashi Fukuda[1] and Keiichi Komatsu[2]

[1] Department of Mathematics, College of Industrial Technology,
Nihon University, 2-11-1 Shin-ei, Narashino, Chiba, Japan
fukuda@math.cit.nihon-u.ac.jp

[2] Department of Information and Computer Science,
School of Science and Engineering, Waseda University,
3-4-1 Okubo, Shinjuku, Tokyo 169, Japan
kkomatsu@mn.waseda.ac.jp

Abstract. We try to write the values of L-functions associated to some abelian extensions of $\mathbb{Q}(\exp(2\pi i/13)+\exp(6\pi i/13)+\exp(18\pi i/13))$ using units given by Siegel modular functions hoping that our trial brings some new features in algebraic number theory.

1 Introduction

In our previous papers [2] and [3], we constructed some unit groups in abelian extensions of $\mathbb{Q}(\exp(2\pi i/5))$ by the special values of Siegel modular functions. In order to obtain the above results, it was essential that $\mathbb{Q}(\exp(2\pi i/5))$ is the CM-field corresponding to the Jacobian variety of the curve $y^2 = 1 - x^5$.

Let $\zeta = \exp(2\pi i/13)$ and $\alpha = \zeta + \zeta^3 + \zeta^9$. Recently, Murabayashi, Umegaki [9] and van Wamelen [12] have showed that $\mathbb{Q}(\alpha)$ is the CM-field corresponding to the Jacobian variety of the curve $C : y^2 = x^5 - 156x^4 + 10816x^3 - 421824x^2 + 8998912x - 8042776$. In this paper, we construct unit groups in abelian extensions of $\mathbb{Q}(\alpha)$ by special values of Siegel modular functions at a CM-point corresponding to the Jacobian variety of C.

On the other hand, it is important to find a good representation of the value of L-function at one with units having well known properties, which is called Kronecker's limit formula (cf. [6], [8]). In this context, we try to write the values of L-functions associated to the above abelian fields using units given by Siegel modular functions in expectation of providing new approach to limit formulae.

2 Theorems

We begin by explaining the notations. We denote by $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ the ring of rational integers, the field of rational numbers, real numbers and complex numbers, respectively. For a positive integer n, $\mathbb{Z}^n$, $\mathbb{Q}^n$, etc. denote the module or vector space on n-dimensional column vectors with components in $\mathbb{Z}$, $\mathbb{Q}$, etc. If Y is an associative ring with identity element, $Y^\times$ denotes the group of

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 108–119, 2002.

all invertible elements of Y and $M_n(Y)$ the ring of all matrices of size n with components in Y. The identity element of $M_n(Y)$ is denoted by I_n. We write $GL_n(Y) = M_n(Y)^\times$. The transpose of a matrix α is denoted by ${}^t\alpha$. For elements $g_1, g_2, \dots, g_r$ of a group G, we denote by $\langle g_1, g_2, \dots, g_r\rangle$ the subgroup of G generated by $g_1, g_2, \dots, g_r$. For a finite algebraic extension K of k, $(K : k)$ means the degree of K over k, $N_{K/k}$ means the norm mapping of K over k and $G(K/k)$ means the Galois group of K over k when K is a Galois extension of k. If k is an algebraic number field, we denote the integer ring of k by $\mathfrak{O}_k$.

Let $\mathfrak{S}_2$ be the set of all complex symmetric matrices of degree 2 with positive definite imaginary parts. For $u \in \mathbb{C}^2$, $z \in \mathfrak{S}_2$ and $r, s \in \mathbb{R}^2$, put as usual

$$\Theta(u, z; r, s) = \sum_{x \in \mathbb{Z}^2} e\Big(\frac{1}{2}{}^t(x+r)z(x+r) + {}^t(x+r)(u+s)\Big),$$

where $e(\xi) = \exp(2\pi i\xi)$ for $\xi \in \mathbb{C}$. Let N be a positive integer. If we define

$$\Phi(z; r, s; r_1, s_1) = \frac{2\Theta(0, z; r, s)}{\Theta(0, z; r_1, s_1)}$$

for $r, s, r_1, s_1 \in \frac{1}{N}\mathbb{Z}^2$, then $\Phi(z; r, s; r_1, s_1)$ is a Siegel modular function of level $2N^2$. Let $\zeta_N = e(1/N)$ be a primitive N-th root of unity, $\zeta = \zeta_{13}$ and k the unique subfield of $\mathbb{Q}(\zeta)$ with $(k : \mathbb{Q}) = 4$. We note that k is a CM-field. Let σ be an element of the Galois group $G(\mathbb{Q}(\zeta)/\mathbb{Q})$ with $\zeta^\sigma = \zeta^2$. Then σ is a generator of $G(\mathbb{Q}(\zeta)/\mathbb{Q})$. Put $\alpha = \zeta + \zeta^3 + \zeta^9$. Since $\{\, \zeta^{\sigma^\nu} \,\}_{\nu=0}^{11}$ is a normal integral basis of $\mathbb{Q}(\zeta)$ over $/\mathbb{Q}$, $\{\, \alpha^{\sigma^\nu} \,\}_{\nu=0}^{3}$ is an integral basis of k over $\mathbb{Q}$. Let $\mathfrak{O}_k$ be the integer ring of k. For the 2-dimensional vector space $\mathbb{C}^2$, we put

$$L = \left\{ \begin{pmatrix} \xi \\ \xi^\sigma \end{pmatrix} \in \mathbb{C}^2 \mid \xi \in \mathfrak{O}_k \right\}$$

Then L is a lattice in $\mathbb{C}^2$. we put

$$\rho = \alpha - \alpha^{\sigma^2} = \sqrt{\frac{13 - 3\sqrt{13}}{2}}\, i$$

and define a Riemann form R on the complex torus $\mathbb{C}^2/L$ as follows:

$$R\Big(\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \Big) = \frac{1}{13}\Big(\rho(u_1\bar{v_1} - \bar{u_1}v_1) + \rho^\sigma(u_2\bar{v_2} - \bar{u_2}v_2)\Big)$$

for $u_i, v_i \in \mathbb{C}^2$, where $\bar{u_i}, \bar{v_i}$ mean the complex conjugates of u_i, v_i, respectively.

Moreover, for elements

$$\Omega_1 = \begin{pmatrix} \alpha \\ \alpha^\sigma \end{pmatrix}, \Omega_2 = \begin{pmatrix} \alpha^\sigma \\ \alpha^{\sigma^2} \end{pmatrix}, \Omega_3 = \begin{pmatrix} -\alpha^{\sigma 2} - \alpha^{\sigma^3} \\ -\alpha^{\sigma^3} - \alpha \end{pmatrix}, \Omega_4 = \begin{pmatrix} -\alpha^{\sigma 2} - 2\alpha^{\sigma^3} \\ -\alpha^{\sigma^3} - 2\alpha \end{pmatrix},$$

we can easily see that $\{\, \Omega_1, \Omega_2, \Omega_3, \Omega_4 \,\}$ is a free basis of L over $\mathbb{Z}$ and $(R(\Omega_i, \Omega_j)) = J$, where

$$J = \begin{pmatrix} 0 & -I_2 \\ I_2 & 0 \end{pmatrix}.$$

Hence we see that

$$z_1 = \begin{pmatrix} -\alpha^{\sigma 2} - \alpha^{\sigma^3} & -\alpha^{\sigma 2} - 2\alpha^{\sigma^3} \\ -\alpha^{\sigma^3} - \alpha & -\alpha^{\sigma^3} - 2\alpha \end{pmatrix}^{-1} \begin{pmatrix} \alpha & \alpha^{\sigma} \\ \alpha^{\sigma} & \alpha^{\sigma^2} \end{pmatrix}$$

is a CM-point of $\mathfrak{S}_2$ corresponding to the polarized abelian variety $(\mathbb{C}^2/L, R)$.

Let $k(6)$ be the ray class field of k modulo 6, $\mathcal{I}_6$ the group of fractional ideals of k generated by prime ideals which are prime to 6 and we put $\mathcal{S}_6 = \{(\xi) \mid \xi \in k^{\times},\, \xi \equiv 1 \pmod{6}\}$. Then we can compute

$$k(6) = k(\beta_0,\, \beta_1,\, \gamma) \quad \text{and} \quad G(k(6)/k) \cong \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/5\mathbb{Z}\,,$$

where

$$\beta_0 = \sqrt{\frac{-1+\sqrt{13}}{2}}\,,\ \beta_1 = \sqrt{\frac{-1-\sqrt{13}}{2}}$$

and γ is a root of the equation

$$X^5 - 40X^4 - 1220X^3 - 50800X^2 - 138460X - 1897012 = 0.$$

We extend the action of $\sigma \in G(\mathbb{Q}(\zeta)/\mathbb{Q})$ to $\mathbb{Q}(\zeta)k(6)$ by $\beta_0^{\sigma} = \beta_1$, $\beta_1^{\sigma} = \beta_0$ and $\gamma^{\sigma} = \gamma$.

After these preparation, we can now describe our main results.

Theorem 1. *Notations being as above, we put*

$$\varepsilon = \frac{1}{2^3}\Phi(z_1\,;\, \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \end{pmatrix}, \begin{pmatrix} \frac{2}{3} \\ 0 \end{pmatrix}\,;\, \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \end{pmatrix}, \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \end{pmatrix})^3.$$

Then ε is a Minkowski unit of the ray class field $k(6)$ of k modulo 6.

Theorem 2. *Notations being as in Theorem 1, let $K_0 = k(\sqrt{-3})$, $x_i = \log\theta_i$, where $\theta_i = N_{k(6)/K_0}(\varepsilon^{\sigma^i})$. Let χ be the non-trivial character of $\mathcal{I}_6/\mathcal{S}_6$ corresponding to K_0 and $L_k(s\,;\,\chi)$ the L-function of χ. Then we have*

$$10^2 L_k(1\,;\,\chi) = \frac{2^6\cdot 5\pi^2}{3^3\cdot 13\sqrt{13}}\big(x_0^2 + 2x_1^2 + x_2^2 + 2x_0x_1 + 2x_1x_2\big).$$

Theorem 3. *Notations being as in Theorem 2, we put $K_1 = k(\beta_0)$, $K_2 = \mathbb{Q}(\beta_0)$, χ the non-trivial character of K_1 over K_2 and $L_{K_2}(s\,;\,\chi)$ the L-function of χ. Then we have*

$$60L_{K_2}(1\,;\,\chi) = \frac{2^4\pi^3}{13^2\sqrt{3}}\log|(\varepsilon_0^2\varepsilon_1^{-4}\varepsilon_3^6)^{\sigma}|,$$

where $\varepsilon_i = N_{k(6)/K_1}(\varepsilon^{\sigma^i})$.

Theorem 4. *Notations being as in Theorem 3, let χ be the non-trivial character of K_1 over k and $L_k(s\,;\,\chi)$ the L-function of χ. Then we have*

$$\begin{aligned} 60^2 L_k(1;\chi) = \frac{2^6\pi^2}{3\cdot 13\sqrt{13}}\Big(&\log|\varepsilon_0\varepsilon_1^{-2}\varepsilon_3^3|\,\log|(\varepsilon_0^{12}\varepsilon_1\varepsilon_3)^{1+\sigma}| \\ &- \log|\varepsilon_0^2\varepsilon_1\varepsilon_3|\,\log|(\varepsilon_0\varepsilon_1^{-2}\varepsilon_3^3)^{1+\sigma}|\Big). \end{aligned}$$

3 Integrality of Special Values

We recall some properties of Siegel modular forms. For a positive integer N, we put $\mathit{\Gamma}_N = \{\, A \in GL_4(\mathbb{Z}) \mid {}^tAJA = J,\ A \equiv I_4 \pmod{NM_4(\mathbb{Z})} \,\}$. We let every element $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ act on $\mathfrak{S}_2$ by $A(z) = (A_{11}z + A_{12})(A_{21}z + A_{22})^{-1}$. For a positive integer r and a subring R of C, let $M_r(\mathit{\Gamma}_N, R)$ denote the vector space of all modular forms f on $\mathfrak{S}_2$ such that $f(A(z)) = \det(A_{21}z + A_{22})^r f(z)$ for all $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \in \mathit{\Gamma}_N$ and that $f(z) = \sum_\xi a(\xi) e(\mathrm{Tr}(\xi z)/N)$ with $a(\xi) \in R$, where ξ runs over all semi-integral semi-definite symmetric matrices of degree 2 (i.e. $\xi = \begin{pmatrix} a & b/2 \\ b/2 & d \end{pmatrix}$ with $a, b, d \in \mathbb{Z}$). If all $a(\xi)$ are contained in $\mathbb{Q}(\zeta_N)$, we let τ act on $f(z)$ by $f^\tau(z) = \sum_\xi a(\xi)^\tau e(\mathrm{Tr}(\xi z)/N)$. Then it is well known that $f^\tau(z) \in M_r(\mathit{\Gamma}_N, \mathbb{Q}(\zeta_N))$ for all $f(z) \in M_r(\mathit{\Gamma}_N, \mathbb{Q}(\zeta_N))$.

The following lemma is a refinement of Proposition 2 in [7].

Lemma 1. *Let $v_p : \mathbb{Q} \to \mathbb{Z}$ be the additive p-adic valuation. Let C be a curve of genus 2 defined by $y^2 = \sum_{i=0}^{6} u_i x^{6-i}$ with $u_i \in \mathbb{Z}$, $z_* \in \mathfrak{S}_2$ a point corresponding to the Jacobian variety of C by the standard normalization of its period matrix and J_ν the Igusa J-invariants of C (cf. p.177 in [4]). Let N be an positive integer and $r, s \in (\frac{1}{N}\mathbb{Z})^2$. If $v_p(J_{10}) \leq \frac{10}{\nu} v_p(J_\nu)$ for all prime number p and $\nu = 2, 4, 6, 8, 10$, then $\mathit{\Phi}(z_*\,;\, r, s\,;\, 0, 0)$ is an algebraic integer.*

Proof. (cf. p.322 in [7]). We put

$$X_{10}(z) = \prod_{r,s} \mathit{\Theta}(0, z\,;\, r, s)^2$$

for $z \in \mathfrak{S}_2$, where r, s runs over $(\frac{1}{2}\mathbb{Z})^2/\mathbb{Z}^2$ satisfying that $4({}^t rs)$ is even. Let T be the set of representations for $\mathit{\Gamma}_1/\mathit{\Gamma}_{2N^2}$ and $T = \{\, X_{10}(z)\mathit{\Phi}^\tau(Az\,;\, r, s) \mid A \in T,\ \tau \in G(\mathbb{Q}(\zeta_{N^2})/\mathbb{Q}) \,\}$. Let $f(z)$ be the fundamental symmetric polynomial of T of degree n. Then we have $f \in M_{10n}(\mathit{\Gamma}_1, \mathbb{Z})$ (cf. p.322 in [7]). By Lemma 14 in [4], there exists non-zero element $\mu \in \mathbb{C}$ and integers $a_{i_2 i_4 i_6 i_8 i_{10}} \in \mathbb{Z}$ such that

$$f(z_*) = \sum_{2i_2 + 4i_4 + 6i_6 + 8i_8 + 10i_{10} = 10n} a_{i_2 i_4 i_6 i_8 i_{10}} J_2^{i_2} J_4^{i_4} J_6^{i_6} J_8^{i_8} J_{10}^{i_{10}}$$

and $X_{10}(z_*) = \mu^{10} J_{10}$. Hence we have

$$\frac{f(z_*)}{X_{10}(z_*)^n} \in \mathbb{Z}$$

by

$$\sum_{\nu=1}^{5} i_{2\nu} v_p(J_{2\nu}) \geq \sum_{\nu=1}^{5} \frac{2\nu}{10} i_{2\nu} v_p(J_{10}) = n v_p(J_{10}).$$

This shows $\mathit{\Phi}(z_*\,;\, r, s\,;\, 0, 0)$ is an algebraic integer. □

Murabayashi, Umegaki and Wamelen have found equations of hyperelliptic curves whose Jacobian varieties have complex multiplications.

Especially, they showed that $k = \mathbb{Q}(\alpha)$ is the CM-field corresponding to the Jacobian variety of the curve

$$C\,:\,y^2 = x^5 - 156X^4 + 10816X^3 - 421824X^2 + 8998912X - 8042776$$

whose J-invariants are as follows:

$$J_2 = 2^9 13^2,\ J_4 = 2^{11}13^6 7,\ J_6 = 2^{16}3^9 11,\ J_8 = 2^{20}13^3 3,\ J_{10} = 2^{20}13^{15}.$$

Hence $\Phi(z_1\,;\,r,s\,;\,0,0)$ is an algebraic integer for $r,s \in (\frac{1}{N}\mathbb{Z})^2$ by Lemma 1. Furthermore, by an argument similar to that in [2], one can show $\Phi(z_1\,;\,r,s\,;\,0,0)^3$ is contained in $k(6)$ for $r,s \in (\frac{1}{3}\mathbb{Z})^2$.

Now we recall Shimura's reciprocity law which plays essential roles in the next section. Let v be a non-zero integer and A a matrix in $M_4(\mathbb{Z})$ with ${}^tAJA = vJ$. we suppose that the determinant of A is v^2 and that v is prime to $2N$.

Then it is well known that there exists a matrix B in Γ_1 with

$$A \equiv \begin{pmatrix} I_2 & 0 \\ 0 & vI_2 \end{pmatrix} B \pmod{2N^2}$$

We recall $\Phi(z\,;\,r,s\,;\,r_1,s_1)$ is a Siegel modular function of level $2N^2$ if r,s,r_1,s_1 are contained in $(\frac{1}{N}\mathbb{Z})^2$. We let A act on $\Phi(z\,;\,r,s\,;\,r_1,s_1)$ by

$$\Phi^A(z\,;\,r,s\,;\,r_1,s_1) = \Phi(B(z)\,;\,r,vs\,;\,r_1,vs_1).$$

We note that Φ^A is also a Siegel modular function of level $2N^2$. Let R be a regular representation of k with respect to the basis α, α^σ, $-(\alpha^{\sigma^2}+\alpha^{\sigma^3})$, $-(\alpha^{\sigma^2}+2\alpha^{\sigma^3})$ and ω an integer of k which is prime to $2N^2$. Then we have the following:

Lemma 2. *(Proposition 2.2 in [11]). Let $k(2N^2)$ be the ray class field of k modulo $2N^2$. Then we have $\Phi(z_1\,;\,r,s\,;\,r_1,s_1) \in k(2N^2)$ and*

$$\Phi(z_1\,;\,r,s\,;\,r_1,s_1)^{\left(\frac{k(2N^2)/k}{(\omega)}\right)} = \Phi^{R(\omega\omega^{\sigma^3})}(z_1\,;\,r,s\,;\,r_1,s_1).$$

4 Norm Computation

To avoid the complicated expressions, we write

$$\Psi(z\,;\,r_1,r_2,r_3,r_4\,;\,s_1,s_2,s_3,s_4) = \Phi(z\,;\,\begin{pmatrix} r_1/6 \\ r_2/6 \end{pmatrix}, \begin{pmatrix} r_3/6 \\ r_4/6 \end{pmatrix}\,;\,\begin{pmatrix} s_1/6 \\ s_2/6 \end{pmatrix}, \begin{pmatrix} s_3/6 \\ s_4/6 \end{pmatrix})$$

and put

$$\alpha_1 = \Psi(z_1\,;\,2,4,2,2\,;\,0,0,0,0)^3,\ \alpha_2 = \Psi(z_1\,;\,2,2,4,0\,;\,0,0,0,0)^3.$$

Then α_1 and α_2 are algebraic integers of $k(6)$ and $\varepsilon = \alpha_2/\alpha_1$, where ε is the element defined in Theorem 1. Since $N_{k(6)/\mathbb{Q}}(\alpha_i)$ is rational integer, one can determine exact value of $N_{k(6)/\mathbb{Q}}(\alpha_i)$ by approximation with some luck as follows.

It is easy to see that $(\alpha - \alpha^{\sigma^2})$ and $(\alpha - \alpha^{\sigma^2} + \alpha^{\sigma^3})$ are prime ideals of k lying above 13 and 29, respectively. We define

$$\tau_1 = \left(\frac{k(6)/k}{\alpha - \alpha^{\sigma^2}}\right) \quad \text{and} \quad \tau_2 = \left(\frac{k(6)/k}{\alpha - \alpha^{\sigma^2} + \alpha^{\sigma^3}}\right)$$

be Artin symbols. Then $G(k(6)/k) = \langle\, \tau_1,\, \tau_2 \,\rangle$ and $\tau_1^2 = \tau_2^{10} = 1$.

First we note that

$$N_{k(6)/\mathbb{Q}}(\alpha_2) = \left|N_{k(6)/k}(\alpha_2)\right|^2 \left|N_{k(6)/k}(\alpha_2^{\sigma})\right|^2.$$

The actions of $G(k(6)/k)$ for α_2 are explicitly given by Shimura's reciprocity law and easy to compute. On the other hand, there are no theories which are able to handle the action of σ. But it is known that

$$\alpha_2^{\alpha} = \Psi(z_1\,;\, r_1, r_2, r_3, r_4\,;\, s_1, s_2, s_3, s_4)^3 \zeta_6^m$$

for some $r_i \in \mathbb{Z}/6\mathbb{Z}$, $s_i \in \mathbb{Z}/2\mathbb{Z}$ and $m \in \mathbb{Z}/6\mathbb{Z}$. So we put

$$\beta_1 = \Psi(z_1\,;\, r_1, r_2, r_3, r_4\,;\, s_1, s_2, s_3, s_4)^3$$

and compute the approximate value of

$$\left|N_{k(6)/k}(\alpha_2)\right|^2 \left|N_{k(6)/k}(\beta_1)\right|^2 \tag{1}$$

for all r_i and s_i. Our calculation shows that the only possible integral value for (1) is 2^{96}. Hence we can conclude that $N_{k(6)/\mathbb{Q}}(\alpha_2) = 2^{96}$. At the same time, β_1 is a candidate of α_2^{σ}. Strictly speaking, we have

$$\alpha_2^{\sigma} = \beta_1^{\tau_1^{i_{11}} \tau_2^{i_{12}}} \zeta_6^{m_1} \quad \text{for some } 0 \le i_{11} \le 1,\ 0 \le i_{12} \le 9,\ 0 \le m_1 \le 5.$$

In a similar manner, we get candidates β_2, β_3 of $\alpha_2^{\sigma^2}$, $\alpha_2^{\sigma^3}$. Namely,

$$\alpha_2^{\sigma^2} = \beta_2^{\tau_1^{i_{21}} \tau_2^{i_{22}}} \zeta_6^{m_2}, \quad \alpha_2^{\sigma^3} = \beta_3^{\tau_1^{i_{31}} \tau_2^{i_{32}}} \zeta_6^{m_3}.$$

If $m_1, m_2, m_3, i_{11}, i_{12}, i_{21}, i_{22}, i_{31}, i_{32}$ are determined, then the action of σ for α_2 is explicitly known.

Now, noting that $G(k(6)/\mathbb{Q}) = \{\, \sigma^{i_0} \tau_1^{i_1} \tau_2^{i_2} \mid 0 \le i_0 \le 3,\ 0 \le i_1 \le 1,\ 0 \le i_2 \le 9 \,\}$, we choose m_1, m_2, m_3 so that all the coefficients of the monic polynomial with roots α_2^{ρ} $(\rho \in G(k(6)/\mathbb{Q})$ are close to rational integers. Next, using an integral basis $\{v_i \mid 0 \le i \le 79\}$ of $k(6)$ over $\mathbb{Z}$ which is explained in §6, we choose $i_{11}, i_{12}, i_{21}, i_{22}, i_{31}, i_{32}$ so that the simultaneous equations

$$\sum_{i=0}^{79} x_i v_i^{\rho} = \alpha_2^{\rho} \qquad (\rho \in G(k(6)/\mathbb{Q})) \tag{2}$$

have solutions which are close to rational integers. Our calculation again shows that there is only one possibility of $(m_1, m_2, m_3, i_{11}, i_{12}, i_{21}, i_{22}, i_{31}, i_{32})$. Hence we were luckily able to determine the action of σ. Namely we have

$$\begin{aligned}\alpha_2^{\sigma} &= \Psi(z_1\,;\,0,4,1,0,0,0,3,0)^3\zeta_6^2\,,\\ \alpha_2^{\sigma^2} &= \Psi(z_1\,;\,2,3,4,4,0,3,0,0)^3\,,\\ \alpha_2^{\sigma^3} &= \Psi(z_1\,;\,4,5,5,2,0,3,3,0)^3\zeta_6^5\,.\end{aligned}$$

and, at the same time, get the coefficients of α_2 with respect to $\{v_i \mid 0 \le i \le 79\}$.

Under these preparations, we can prove Theorem 1.

Proof of Theorem 1. The proof is computational. In the same way as for α_2, we have

$$N_{k(6)/\mathbb{Q}}(\alpha_1) = N_{k(6)/\mathbb{Q}}(\alpha_2) = 2^{96}.$$

On the other hand, we know the integral expressions of α_1 and α_2 with respect to an integral basis of $k(6)$. It is then straightforward to see that α_2/α_1 is an integer of $k(6)$. Hence $\varepsilon = \alpha_2/\alpha_1$ is a unit of $k(6)$.

It is also a routine work to see that the rank of the 80×80 matrix

$$(\log|\varepsilon^{\rho_1\rho_2}|)_{\rho_1,\,\rho_2 \in G(k(6)/\mathbb{Q})}$$

is 39 and hence ε is a Minkowski unit. □

5 The Quotient of Regulators

For an algebraic number field F, we denote by $\zeta_F(s)$ the Dedekind zeta function of F, by D_F the discriminant of F and R_F the regulator of F. Let χ be a non-trivial character of K_1 over K_2. Then we have

$$L_{K_2}(1\,;\,\chi) = \lim_{s\to 1}\frac{\zeta_{K_1}(s)}{\zeta_{K_2}(s)} = \sqrt{\frac{|D_{K_2}|}{|D_{K_1}|}}\frac{(2\pi)^4}{2^2(2\pi)}\frac{R_{K_1}}{R_{K_2}} = \frac{(2\pi)^3}{2^2 13^2\sqrt{3}}\frac{R_{K_1}}{R_{K_2}}\,,$$

where $D_{K_1} = 3^2 \cdot 13^6$ and $D_{K_2} = -3 \cdot 13^2$.

The computation in §6 shows that $\beta_0^{\tau_1} = -\beta_0$, $\beta_1^{\tau_1} = -\beta_1$, $\beta_0^{\tau_2} = \beta_0$ and $\beta_1^{\tau_2} = -\beta_1$. Moreover we recall $\beta_0^{\sigma} = \beta_1$ and $\beta_1^{\sigma} = \beta_0$. Hence the embeddings of K_2 into $\mathbb{R}$ are $\tau_1|_{K_2}$ and $id|_{K_2}$. The embedding of K_2 into $\mathbb{C}$ is $\sigma|_{K_2}$. Moreover the embeddings of K_1 into $\mathbb{R}$ are empty and those of K_1 into $\mathbb{C}$ are $id|_{K_1}$, $\tau_1|_{K_2}$, $\tau_2|_{K_2}$, $\sigma|_{K_2}$.

Let E_{K_i} be the unit group of K_i. Then the free ranks of E_{K_1} and E_{K_2} are three and two, respectively. Since no prime of K_2 lying above 2 ramifies in K_1 over K_2, there exist units η_0, η_1, ξ_1 of E_{K_1} with

$$E_{K_2} = \langle -1,\, \eta_0,\, \eta_1\, \rangle \quad \text{and} \quad E_{K_1} = \langle -1,\, \xi_1,\, \eta_0,\, \eta_1\, \rangle. \tag{3}$$

Hence we have

$$\pm R_{K_2} = \begin{vmatrix} \log|\eta_0| & \log|\eta_0^{\tau_1}| \\ \log|\eta_1| & \log|\eta_1^{\tau_1}| \end{vmatrix}$$

and

$$\begin{aligned}
\pm R_{K_1} &= \begin{vmatrix} 2\log|\eta_0| & 2\log|\eta_0^{\tau_1}| & 2\log|\eta_0^{\sigma}| \\ 2\log|\eta_1| & 2\log|\eta_1^{\tau_1}| & 2\log|\eta_1^{\sigma}| \\ 2\log|\xi_1| & 2\log|\xi_1^{\tau_1}| & 2\log|\xi_1^{\sigma}| \end{vmatrix} \\
&= 2^2\Bigg(\log|\xi_1|\begin{vmatrix} \log|\eta_0^{\tau_1}| & 2\log|\eta_0^{\sigma}| \\ \log|\eta_1^{\tau_1}| & 2\log|\eta_1^{\sigma}| \end{vmatrix} - \log|\xi_1^{\tau_1}|\begin{vmatrix} \log|\eta_0| & 2\log|\eta_0^{\sigma}| \\ \log|\eta_1| & 2\log|\eta_1^{\sigma}| \end{vmatrix} \\
&\quad +2\log|\xi_1^{\sigma}|\begin{vmatrix} \log|\eta_0| & \log|\eta_0^{\tau_1}| \\ \log|\eta_1| & \log|\eta_1^{\tau_1}| \end{vmatrix}\Bigg) \\
&= 2^2\Big(\log|\xi_1| + \log|\xi_1^{\tau_1}| + 2\log|\xi_1^{\sigma}|\Big)\begin{vmatrix} \log|\eta_0| & \log|\eta_0^{\tau_1}| \\ \log|\eta_1| & \log|\eta_1^{\tau_1}| \end{vmatrix}
\end{aligned}$$

by $2\log|\eta_i^{\sigma}| = -\log|\eta_i| - \log|\eta_i^{\tau_1}|$. This shows the following:

Lemma 3. *For the above unit ξ_1 of K_1, we have*

$$\frac{R_{K_1}}{R_{K_2}} = \pm\, 2^2 \log|\xi_1^{1+\tau_1+2\sigma}|.$$

Now we can present the proof of Theorem 3. The proofs of Theorems 2 and 4 are similar and omitted.

Proof of Theorem 3. We construct η_0, η_1 and ξ_1 explicitly. Recall that $\varepsilon = \alpha_2/\alpha_1$ and $\varepsilon_i = N_{k(6)/K_1}(\varepsilon^{\sigma^i})$. Put $\Xi_0 = \varepsilon_0^4\varepsilon_1^2\varepsilon_3^2$, $\Xi_1 = \varepsilon_0^2\varepsilon_1^{-4}\varepsilon_3^6$ and $\Xi_2 = \varepsilon_0^{12}\varepsilon_1\varepsilon_3$. Then it is shown computationally that one of the 60-th roots of each Ξ_i is contained in K_1. It is easy to see by PARI that (3) holds if we put $\eta_0 = \xi_0$, $\eta_1 = \xi_0^{-1}\xi_2$ and $\xi_i = \sqrt[60]{\Xi_i}$. It is shown again computationally that $|\xi_1^{1+\tau_1}| = 1$ and $\log|\xi_1^{\sigma}| > 0$. Hence we have Theorem 3. □

6 Computational Techniques

In this section, we explain some techniques which were needed for computations in previous sections.

6.1 Construction of $k(6)$

First we see that $k(6)$ is the splitting field of

$$f_0(X) = X^2 - \alpha - \alpha^{\sigma^2},$$

$$f_1(X) = X^2 + 1 + \alpha + \alpha^{\sigma^2} \text{ and}$$

$$f_2(X) = X^5 - 20X^3 - 80(1 + 2\alpha + 2\alpha^{\sigma^2})X^2 - 810X - 382(1 + 2\alpha + 2\alpha^{\sigma^2})$$

by KASH (cf. [1]). It is easy to see that $f_0(\beta_0) = f_1(\beta_1) = 0$. Next we note that $f_2(X)^{1+\sigma} = g(X)^2$, where $g(X) = X^5 - 40X^4 - 1220X^3 - 50800X^2 - 138460X - 1897012$. Since $g(X)$ is irreducible over $\mathbb{Q}$, we can conclude that $k(6) = k(\beta_0, \beta_1, \gamma)$ for any root γ of $g(X)$.

6.2 Actions of τ_1 and τ_2 for $\varPhi(z_1\,;\,r,s\,;\,r_1,s_1)$

The regular representations of $\omega_1 = (\alpha - \alpha^{\sigma^2})^{1+\sigma^3}$ and $\omega_2 = (\alpha - \alpha^{\sigma^2} + \alpha^{\sigma^3})^{1+\sigma^3}$ are decomposed as follows:

$$R(\omega_1) = \begin{pmatrix} -5 & -2 & 0 & 2 \\ 4 & 5 & -2 & 0 \\ 0 & 2 & -5 & 4 \\ -2 & 0 & -2 & 5 \end{pmatrix} \equiv \begin{pmatrix} I_2 & 0 \\ 0 & 13I_2 \end{pmatrix} B_1 \pmod{72},$$

$$B_1 = \begin{pmatrix} -5 & -2 & 0 & 2 \\ 4 & 5 & -2 & 0 \\ -576 & 410 & -377 & 532 \\ -410 & 288 & -266 & 377 \end{pmatrix} \in S_p(2,\mathbb{Z}),$$

$$R(\omega_2) = \begin{pmatrix} 0 & 0 & 11 & -6 \\ 0 & 0 & -3 & -1 \\ -1 & 3 & 6 & -6 \\ 6 & 11 & 3 & -9 \end{pmatrix} \equiv \begin{pmatrix} I_2 & 0 \\ 0 & 29I_2 \end{pmatrix} B_2 \pmod{72},$$

$$B_2 = \begin{pmatrix} -72 & 216 & 11 & -6 \\ 0 & 0 & -3 & -1 \\ 4171 & -12513 & -258 & 474 \\ 9102 & -27305 & -561 & 1035 \end{pmatrix} \in S_p(2,\mathbb{Z}).$$

Hence Shimura's reciprocity law implies

$$\varPhi(z_1\,;\,r,s\,;\,r_1,s_1)^{3\tau_1} = \varPhi(B_1(z_1)\,;\,r,13s\,;\,r_1,13s_1)^3, \tag{4}$$

$$\varPhi(z_1\,;\,r,s\,;\,r_1,s_1)^{3\tau_2} = \varPhi(B_2(z_1)\,;\,r,29s\,;\,r_1,29s_1)^3 \tag{5}$$

for $r, s, r_1, s_1 \in (\frac{1}{6}\mathbb{Z})^2$.

However the convergence of $\varPhi(B_i(z_1)\,;\,\ldots)$ is very slow. We must transform $\varPhi(B_i(z_1)\,;\,\ldots)$ to the form $\varPhi(z_1\,;\,\ldots)$ which converges faster in order to calculate the approximate values of (4) and (5) with high precision. This is done easily. Namely we know that if $r_1, s_1, r_2, s_2 \in (\frac{1}{6}\mathbb{Z})^2$, then

$$\varPhi(B_i(z_1)\,;\,r_1,s_1\,;\,r_2,s_2) = \varPhi(z_1\,;\,r_1',s_1'\,;\,r_2',s_2')\zeta_{72}^m \tag{6}$$

for some $r_1', s_1', r_2', s_2' \in (\frac{1}{6}\mathbb{Z})^2$ and $m \in \mathbb{Z}/72\mathbb{Z}$. The transformation formula for theta series determines r_1', s_1', r_2', s_2' explicitly. So we determine m by calculating both side of (6) with low precision and next calculate right hand side of (6) with high precision.

6.3 Integral Basis of $k(6)$

It seems non-realistic to compute an integral basis of $k(6)$ straightforward using algorithms implemented on several number theoretic packages because $k(6)$ is a field of degree 80 with huge discriminant $2^{64}3^{40}13^{60}$ (use Theorem (2.6) in [10] noting that $G(k(2)/k) \cong \mathbb{Z}/5\mathbb{Z}$, $G(k(3)/k) \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ and $k(6)/k$ is tamely ramified). So we constructed an integral basis of $k(6)$ by combining those of two subfields k_1 and k_2 of $k(6)$.

Let $k_1 = \mathbb{Q}(\alpha,\, \beta_0,\, \beta_1) = \mathbb{Q}(\alpha + \beta_0 + \beta_1)$. Then k_1 is a subfield of $k(6)$ and a Galois extension of $\mathbb{Q}$ with degree 16. Put

$$t_i = \alpha^{\sigma^{i_0}} + (-1)^{i_1}\beta_0 + (-1)^{i_2}\beta_1 \quad (i = 4i_0 + 2i_1 + i_2,\, 0 \le i_0 \le 3,\, 0 \le i_1, i_2 \le 1).$$

Then

$$\begin{aligned}\prod_{i=0}^{15}(X - t_i) = X^{16} &+ 4X^{15} + 22X^{14} + 40X^{13} + 167X^{12} + 280X^{11} + 768X^{10} \\ &+3640X^9 + 2141X^8 + 4832X^7 + 3780X^6 + 11204X^5 \\ &+128999X^4 + 126752X^3 + 155662X^2 + 8312X + 6397\end{aligned}$$

is the minimal polynomial of $t_0 = \alpha + \beta_0 + \beta_1$ over $\mathbb{Q}$. An integral basis of k_1 is easily found as polynomials of t_0 using PARI. If we determine the actions of τ_1, τ_2, σ for t_0, then we can explicitly determine the action of $G(k(6)/\mathbb{Q})$ for k_1. Now we obtain $h_i(X) \in \mathbb{Q}[X]$ such that $t_i = h_i(t_0)$ again by PARI and define $\nu_i \in G(k_1/\mathbb{Q})$ by $t_0^{\nu_i} = h_i(t_0)$. By expressing $\alpha = (t_0 + t_3)/2 = (t_0 + h_3(t_0))/2$ and α^{ν_i} as polynomials of t_0, we see that $G(k_1/k) = \{\nu_0, \nu_1, \nu_2, \nu_3\}$. Next, noting that

$$\begin{aligned}\alpha - \alpha^{\sigma^2} &= \frac{t_0 + h_3(t_0) - h_8(t_0) - h_{11}(t_0)}{2}, \\ \alpha - \alpha^{\sigma^2} + \alpha^{\sigma^3} &= \frac{t_0 + h_3(t_0) - h_8(t_0) - h_{11}(t_0) + h_{12}(t_0) + h_{15}(t_0)}{2},\end{aligned}$$

we check the properties of Frobenius automorphisms

$$\frac{x^{\tau_1} - x^{13}}{\alpha - \alpha^{\sigma^2}},\ \frac{x^{\tau_2} - x^{29}}{\alpha - \alpha^{\sigma^2} + \alpha^{\sigma^3}} \in \mathfrak{O}_{k_1}$$

for several $x \in \mathfrak{O}_{k_1}$ and conclude that $\tau_1 = \nu_3$ and $\tau_2 = \nu_1$. Similarly, using the expression

$$\alpha^\sigma = \frac{h_4(t_0) + h_7(t_0)}{2},$$

we see that possible extensions of $\sigma \in G(\mathbb{Q}(\zeta)/\mathbb{Q})$ to k_1 are ν_4, ν_5, ν_6 and ν_7. We define $\sigma = \nu_4$ noting that $\nu = \nu_0,\, \nu_4,\, \nu_8,\, \nu_{12}$ satisfy $(\beta_0 + \beta_1)^\nu = \beta_0 + \beta_1$. Then actions of τ_1, τ_2 and σ for β_i are as follows:

$$\begin{aligned}\sigma &: \beta_0 \mapsto \beta_1,\ \ \beta_1 \mapsto \beta_0, \\ \tau_1 &: \beta_0 \mapsto -\beta_0,\ \ \beta_1 \mapsto -\beta_1, \\ \tau_2 &: \beta_0 \mapsto \beta_0,\ \ \beta_1 \mapsto -\beta_1.\end{aligned}$$

Next we note the roots of $g(X)$ are

$$\gamma_0 = -13.486416826327889668... - \sqrt{-1} \cdot 22.304896245305038177...\,,$$

$$\begin{aligned}
\gamma_1 &= -13.486416826327889668... + \sqrt{-1}\cdot 22.304896245305038177...,\\
\gamma_2 &= -0.9684041344417469795... - \sqrt{-1}\cdot 6.2914405142719518538...,\\
\gamma_3 &= -0.9684041344417469795... + \sqrt{-1}\cdot 6.2914405142719518538...,\\
\gamma_4 &= 68.909641921539273296....
\end{aligned}$$

Then $k_2 = \mathbb{Q}(\gamma_0)$ is a subfield of $k(6)$ with $[k_2 : \mathbb{Q}] = 5$. By a similar but slightly complicated way, because k_2 is not a Galois extension of $\mathbb{Q}$, we have $\gamma_0^{\tau_1} = \gamma_0$ and $\gamma_0^{\tau_2} = \gamma_1$. We note that τ_2 is not the complex conjugation because $\gamma_1^{\tau_2} = \gamma_3$. Furthermore we let act σ on k_2 trivially. An integral basis of k_2 is also easily found by PARI.

Then we can construct a submodule $M = \sum_{i=0}^{79} \mathbb{Z}w_i$ of $\mathfrak{O}_{k(6)}$ by combining integral bases of k_1 and k_2. Since the discriminants of k_1 and k_2 are $3^8 13^{12}$ and $2^4 13^3$ respectively, M may not be equal to $\mathfrak{O}_{k(6)}$. In fact, computing the discriminant of M, we see that $(\mathfrak{O}_{k(6)} : M) = 13^{24}$. Since α_2 is contained in $\mathfrak{O}_{k(6)}$, $13^{24}\alpha_2$ has an integral expression with respect to $\{w_i\}$. Computing the coefficients by solving the simultaneous equations

$$\sum_{i=0}^{79} x_i w_i^\rho = 13^{24}\alpha_2^\rho \qquad (\rho \in G(k(6)/\mathbb{Q})),$$

we see that $13\alpha_2 \in M$ and $\alpha_2 \notin M$. Hence, if we put $M' = M + \mathbb{Z}\alpha_2$, then $(\mathfrak{O}_{k(6)} : M') = 13^{23}$. Fortunately, we reached $\mathfrak{O}_{k(6)}$ repeating this procedure by using 80 conjugates of α_2. In this way, we constructed an integral basis of $k(6)$. It is then easy and may be worthy to notice that $\mathfrak{O}_{k(6)}/M \cong (\mathbb{Z}/13\mathbb{Z})^{24}$. The simple structure of $\mathfrak{O}_{k(6)}/M$ assisted our computations.

The multi-precision calculations in this work were carried on TCP (Tiny C interpreter linked with PARI library), which is available from
`ftp://tnt.math.metro-u.ac.jp/pub/math-packs/tc/` .

References

1. M. Daberkow, C. Fieker, J. Klüners, M. Pohst, K. Roegner, M. Schönig and K. Wildanger, *KANT V4*, J. Symb. Comput. **24** (1997), 267–283.
2. T. Fukuda and K. Komatsu, *On a unit group generated by special values of Siegel modular functions*, Math. Comp., **69-231** (2000), 1207–1212.
3. T. Fukuda and K. Komatsu, *On Minkowski units constructed by special values of Siegel modular functions*, submitted to J. Th. Nombres Bordeaux
4. J. Igusa, *Modular forms and projective invariants*, Amer. J. Math., **89** (1967), 817–855.
5. J. Igusa, *On the ring of the modular forms of degree two over* $\mathbb{Z}$, Amer. J. Math., **101** (1979), 149–183.
6. K. Katayama, *Kronecker's limit formulas and their applications*, J. Fac. Sci. Univ. Tokyo Sect. I, **13** (1966), 1-44.
7. K. Komatsu, *Construction of a normal basis by special values of Siegel modular functions*, Proc. Amer. Math. Soc. **128** (2000), 315–323.

8. S. Konno, *On Kronecker's limit formula in a totally imaginary quadratic field over a totally real algebraic number field*, J. Math. Soc. Japan, **17** (1965), 411–424.
9. N. Murabayashi and A. Umegaki, *Determination of all* $\mathbb{Q}$*-rational CM-points in the moduli space of principally polarized abelian surfaces* J. Algebra, **235-1** (2001), 267-274.
10. J. Neukirch, *Algebraic Number Theory*, Grundlehren der mathematischen Wissenschaften vol. 322, Springer, 1999.
11. G. Shimura, *Theta functions with complex multiplication*, Duke Math. J., **43** (1976), 673–696.
12. P. van Wamelen, *Proving that a genus 2 curve has complex multiplication*, Math. Comp., **68-228** (1999), 1663–1677.

Computational Aspects of NUCOMP

Michael J. Jacobson, Jr.[1,⋆] and Alfred J. van der Poorten[2]

[1] Department of Computer Science, University of Manitoba,
Winnipeg, MB, R3T 2N2, Canada
jacobs@cs.umanitoba.ca

[2] ceNTRe for Number Theory Research, Macquarie University,
Sydney 2109, Australia
alf@math.mq.edu.au

1 Introduction

In 1989, Shanks introduced the NUCOMP algorithm [10] for computing the reduced composite of two positive definite binary quadratic forms of discriminant Δ. Essentially by applying reduction before composing the two forms, the intermediate operands are reduced from size $O(\Delta)$ to $O(\Delta^{1/2})$ in most cases and at worst to $O(\Delta^{3/4})$. Shanks made use of this to extend the capabilities of his hand-held calculator to computations involving forms with discriminants with as many as 20 decimal digits, even though his calculator had only some 10 digits precision. Improvements by Atkin (described in [3], [4]) have also made NUCOMP very effective for computations with forms of larger discriminant.

Although there is nothing in Shanks' original description which suggests that NUCOMP is only applicable to positive definite forms, for years there were no documented applications in any other setting. Recently, van der Poorten [13] has shown that, with very little extra effort, NUCOMP can also be applied to computations in the infrastructure involving indefinite binary quadratic forms. This opens the door to practical improvements in real quadratic field-based applications such as regulator computation and key exchange protocols in the infrastructure.

Until now NUCOMP has been applied exclusively to computations in number fields. However, Cantor's algorithm [2,12] for adding reduced divisors on hyperelliptic curves (equivalently ideal multiplication in function fields) is virtually identical to the composition and reduction algorithms for binary quadratic forms; the main difference being that coefficients of the binary quadratic forms are polynomials over a finite field rather than integers. Thus, there is no reason to believe that NUCOMP cannot also be applied in function fields. Intuitively, one would expect that applying NUCOMP will reduce the degrees of the intermediate operands from $O(2g)$ to at most $O(3g/2)$, where g is the genus of the hyperelliptic curve or function field. Furthermore, by combining van der Poorten's

[⋆] The first author acknowledges the assistance of an Australian Research Council International Research Exchange Grant, held by the second author, in allowing him to visit Macquarie University, Sydney, thus initiating the present work.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 120–133, 2002.

ideas [13] we can also apply NUCOMP to computations in the infrastructure of a real quadratic function field.

In this paper, we show that NUCOMP does in fact yield significant improvements in speed over ordinary composition with reduction in all of the above settings for certain sizes of discriminants. We begin with a description of NUCOMP as presented in [13] which incorporates the improvements described in [3], followed by versions which are suitable for implementation in function fields over any finite field, even or odd characteristic. We then present extensive computations in imaginary quadratic number fields, imaginary quadratic function fields, real quadratic fields, and real quadratic function fields, all of which clearly demonstrate the efficiency of NUCOMP.

2 Description of the Algorithm

2.1 Number Fields

Let $\varphi_1 = u_1X^2 + v_1XY + w_1Y^2 = (u_1, v_1, w_1)$ and $\varphi_2 = u_2X^2 + v_2XY + w_2Y^2 = (u_2, v_2, w_2)$ be two binary quadratic forms of discriminant $\Delta = v_1^2 - 4u_1w_1 = v_2^2 - 4u_2w_2$. Algorithm 1 is based on Algorithm 3 from [13]. The modifications in computing the near reduced composite (Step 6 and Step 7) are from [3]. The relative generator γ can be used for distance computations in real quadratic fields, or not computed at all when working in imaginary quadratic fields.

Algorithm 1 (NUCOMP). Given two quadratic forms $\varphi_1 = (u_1, v_1, w_1)$ and $\varphi_2 = (u_2, v_2, w_2)$ with the same discriminant Δ, compute $\varphi_3 = (u_3, v_3, w_3)$ and γ such that $\varphi_3 = (1/\gamma)\varphi_1\varphi_2$. Precompute $L = |\Delta|^{1/4}$.

```
1. If $w_1 < w_2$ swap $\varphi_1$ and $\varphi_2$. Set $s \leftarrow \frac{1}{2}(v_1 + v_2)$; then $m \leftarrow v_2 - s$.
2. Use Euclid's extended algorithm to compute $(b, c, F)$ such that $bu_2 + cu_1 = F = \gcd(u_1, u_2)$. If $F \mid s$, set $G \leftarrow F$, $A_x \leftarrow G$, $B_x \leftarrow mb$, $B_y \leftarrow u_1/G$, $C_y \leftarrow u_2/G$, $D_y \leftarrow s/G$, and go to Step 5.
3. If $F \nmid s$, use Euclid's extended algorithm again to compute $(y, G)$ so that $xF + ys = G = \gcd(F, s)$, and set $H \leftarrow F/G$. Also set $B_y \leftarrow u_1/G$, $C_y \leftarrow u_2/G$, $D_y \leftarrow s/G$.
4. Compute $l \leftarrow y(b(w_1 \bmod H) + c(w_2 \bmod H)) \bmod H$, $B_x \leftarrow b(m/H) + l(B_y/H)$.
5. Set $b_x \leftarrow B_x \bmod B_y$ and $b_y \leftarrow B_y$. Then execute a partial Euclidean algorithm on $b_x$, $b_y$ :
   (a) Set $x \leftarrow 1$, $y \leftarrow 0$, $z \leftarrow 0$.
   (b) If $|b_y| > L$ and $b_x \neq 0$ go to substep 5(c). Otherwise, if $z$ is odd set $b_y \leftarrow -b_y$, $y \leftarrow -y$. Then set $a_x \leftarrow Gx$, $a_y = Gy$. Go to Step 6.
   (c) Let $q \leftarrow \lfloor b_y/b_x \rfloor$ and simultaneously $t \leftarrow b_y \bmod b_x$. Now set $b_y \leftarrow b_x$ and $b_x \leftarrow t$. Then set $t \leftarrow y - qx$, followed by $y \leftarrow x$ and $x \leftarrow t$. Finally let $z \leftarrow z + 1$ and go to substep 5(b).
6. [$z = 0$] If $z \neq 0$ go to Step 7. Otherwise, compute the near reduced composite $\varphi_3 = (u_3, v_3, w_3)$ as follows:
   (a) $Q_1 \leftarrow C_y b_x$, $c_x \leftarrow (Q_1 - m)/B_y$
   (b) $d_x \leftarrow (b_x D_y - w_2)/B_y$
   (c) $u_3 \leftarrow b_y C_y$
   (d) $w_3 \leftarrow b_x c_x - Gdx$
   (e) $v_3 \leftarrow v_2 - 2Q_1$
   Go to Step 8.
7. [$z \neq 0$] Compute the near reduced composite $\varphi_3 = (u_3, v_3, w_3)$ as follows:
   (a) $c_x \leftarrow (C_y b_x - mx)/B_y$
   (b) $Q_1 \leftarrow b_y c_x$, $Q_2 \leftarrow Q_1 + m$
   (c) $d_x \leftarrow (D_y b_x - w_2 x)/B_y$
   (d) $Q_3 \leftarrow y d_x$, $Q_4 \leftarrow Q_3 + D_y$, $d_y \leftarrow Q_4/x$
   (e) If $b_x \neq 0$ set $c_y \leftarrow Q_2/b_x$; otherwise set $c_y \leftarrow (c_x d_y - w_1)/d_x$.
   (f) $u_3 \leftarrow b_y c_y - a_y d_y$
   (g) $w_3 \leftarrow b_x c_x - a_x d_x$
   (h) $v_3 \leftarrow G(Q_3 + Q_4) - Q_1 - Q_2$
8. Set $\gamma \leftarrow x + y(v_3 - \sqrt{\Delta})/(2Gu_3)$
```

Remark. As in the regular composition algorithm, it is important to compute only the required coefficients in Euclid's extended algorithm. If only one of the two multipliers is required, some gain in speed will be obtained by not computing the second.

There are two subtle differences between our presentation of NUCOMP here as opposed to that in [13]. First, we ensure that $w_2 < w_1$ by initially swapping φ_1 and φ_2 if necessary. The quantity w_2 is used to compute φ_3 in Steps 6 and 7, so this simple operation makes sure that it is the smaller of the two third coefficients. Second, we iterate the partial Euclidean algorithm until *both* values b_x and b_y are less than $L = |\Delta|^{1/4}$. According to computational experiments, taking these extra Euclidean steps resulted in a small improvement in the overall execution time.

Steps 6 and 7 incorporate the modifications from [3]. In the following we prove that these modifications are equivalent to the corresponding steps of Algorithm 3 of [13].

Proposition 1. *If $z = 0$ after Step 5 of Algorithm 1 (NUCOMP), then Step 6 correctly computes φ_3.*

Proof. Since $z = 0$, we have $c_y = C_y$ and $d_y = D_y$. From Algorithm 3 of [13] we get

$$\begin{aligned} v_3 &= (a_x d_y + a_y d_x) - (b_x c_y + b_y c_x) && \\ &= G d_y - b_x c_y - b_y c_x && (x = 1, y = 0) \\ &= G d_y - b_x c_y - c_y b_x + m && (c_x = (c_y b_x - m)/B_y \text{ and } b_y = B_y) \\ &= s + m - 2 b_x c_y && (d_y = s/G) \\ &= v_2 - 2Q_1 && (s + m = v_2)\ . \end{aligned}$$

Proposition 2. *If $z \neq 0$ after Step 5 of Algorithm 1 (NUCOMP), then Step 7 correctly computes φ_3.*

Proof. Clearly c_x, d_x, u_3, w_3 are as in Algorithm 3 of [13]. Now

$$c_y = Q_2/b_x = (Q_1 + m)/b_x = (b_y c_x + m)/b_x$$

$$d_y = Q_4/x = (Q_3 + D_y)/x = (y d_x + D_y)/x,$$

so c_y and d_y are also correct. From Algorithm 3 of [13] we get

$$\begin{aligned} v_3 &= (a_x d_y + a_y d_x) - (b_x c_y + b_y c_x) \\ &= G(x d_y + y d_x) - b_x c_y - b_y c_x && (a_x = Gx, a_y = Gy) \\ &= G(Q_3 + Q_4) - Q_1 - b_x c_y && (Q_4 = y d_x + D_y = x d_y) \\ &= G(Q_3 + Q_4) - Q_1 - Q_2 && (Q_2 = b_y c_x + m = b_x c_y) \ . \end{aligned}$$

The following algorithm, NUDUPL, corresponds to the special case of NUCOMP where $\varphi_1 = \varphi_2$, i.e., squaring a form. As with NUCOMP, a relative generator γ with respect to φ_1^2 is also produced. Algorithm 2 is based on Algorithm 4 from [13], with a few efficiency modification added. The modifications in computing the near reduced composite (Step 6 and Step 7) are from [3].

Algorithm 2 (NUDUPL). Given a quadratic form $\varphi = (u, v, w)$, compute $\varphi_3 = (u_3, v_3, w_3)$ and γ such that $\varphi_3 = (1/\gamma)\varphi_1^2$.

```
1. Use Euclid's extended algorithm to compute $(y, G)$ such that $xu + yv = G = \gcd(u, v)$ and set $A_x \leftarrow G$, $B_y \leftarrow u/G$, $D_y \leftarrow v/G$.
2. Compute $B_x \leftarrow yw \bmod B_y$.
3. Set $b_x \leftarrow B_x$ and $b_y \leftarrow B_y$. Then execute a partial Euclidean algorithm on $b_x$, $b_y$ :
   (a) Set $x \leftarrow 1$, $y \leftarrow 0$, $z \leftarrow 0$.
   (b) If $|b_y| > L$ and $b_x \neq 0$ go to substep 3(c). Otherwise, if $z$ is odd set $b_y \leftarrow -b_y$, $y \leftarrow -y$. Then set $a_x \leftarrow Gx$, $a_y = Gy$. Go to Step 4.
   (c) Let $q \leftarrow \lfloor b_y/b_x \rfloor$ and simultaneously $t \leftarrow b_y \bmod b_x$. Now set $b_y \leftarrow b_x$ and $b_x \leftarrow t$. Then set $t \leftarrow y - qx$, followed by $y \leftarrow x$ and $x \leftarrow t$. Finally let $z \leftarrow z + 1$ and go to substep 3(b).
4. [$z = 0$] If $z \neq 0$ go to Step 5. Otherwise, compute the near reduced composite $\varphi_3 = (u_3, v_3, w_3)$ as follows:
   (a) $d_x \leftarrow (b_x D_y - w)/B_y$
   (b) $u_3 \leftarrow b_y^2$, $w_3 \leftarrow b_x^2$
   (c) $v_3 \leftarrow v - (b_x + b_y)^2 + u_3 + w_3$
   (d) $w_3 \leftarrow w_3 - G d_x$
   Go to Step 6.
5. [$z \neq 0$] Compute the near reduced composite $\varphi_3 = (u_3, v_3, w_3)$ as follows:
   (a) $d_x \leftarrow (b_x D_y - wx)/B_y$
   (b) $Q_1 \leftarrow d_x y$, $d_y \leftarrow Q_1 + D_y$
   (c) $v_3 \leftarrow G(d_y + Q_1)$
   (d) $d_y \leftarrow d_y/x$
```

```
    (e) $u_3 \leftarrow b_y^2$, $w_3 \leftarrow b_x^2$
    (f) $v_3 \leftarrow v_3 - (b_x + b_y)^2 + u_3 + w_3$
    (g) $u_3 \leftarrow u_3 - a_y d_y$, $w_3 \leftarrow w_3 - a_x d_x$
6. Set $\gamma \leftarrow x + y(v_3 - \sqrt{\Delta})/(2Gu_3)$
```

Proposition 3. *If $z = 0$ after Step 3 of Algorithm 2 (NUDUPL), then Step 4 correctly computes φ_3.*

Proof. From Algorithm 4 of [13] we have

$$\begin{aligned} v_3 &= (a_x d_y + a_y d_x) - 2b_x b_y \\ &= (a_x + a_y)(d_x + d_y) - (b_x + b_y)^2 + u_3 + w_3 && (u_3 = b_y^2 - a_y d_y, \\ & && \ w_3 = b_x^2 - a_x d_x) \\ &= G(d_x + d_y) - (b_x + b_y)^2 + u_3 + w_3 && (x = 1, y = 0) \\ &= Gd_x + Gd_y - (b_x + b_y)^2 + u_3 + (b_x^2 - Gd_x) \\ &= v - (b_x + b_y)^2 + b_y^2 + b_x^2 && (D_y = b_1/G) \ . \end{aligned}$$

Proposition 4. *If $z \neq 0$ after Step 3 of Algorithm 2 (NUDUPL), then Step 5 correctly computes φ_3.*

Proof. Clearly d_x, d_y, u_3, w_3 are as in Algorithm 4 from [13]. From Algorithm 4 from [13] we have

$$\begin{aligned} v_3 &= (a_x d_y + a_y d_x) - 2b_x b_y \\ &= a_x d_y + a_y d_x - (b_x + b_y)^2 + b_y^2 + b_x^2 \\ &= Gx d_y + a_y d_x - (b_x + b_y)^2 + b_y^2 + b_x^2 && (a_x = Gx) \\ &= GD_y + 2a_y d_x - (b_x + b_y)^2 + b_y^2 + b_x^2 && (x d_y = d_x y + D_y) \\ &= G(D_y + 2Q_1) - (b_x + b_y)^2 + b_y^2 + b_x^2 \\ &= G((Q_1 + D_y) + Q_1) - (b_x + b_y)^2 + b_y^2 + b_x^2 \ . \end{aligned}$$

2.2 Function Fields — Odd Characteristic

Let $K = GF(q)$, $q = p^n$ for some odd prime p, be a finite field of odd characteristic. Given a square-free, monic polynomial Δ with coefficients over K, the quadratic congruence function field of discriminant Δ is formed by adjoining $\sqrt{\Delta}$ to the field of rational functions $K(X)$. The resulting field is very similar algebraically to a quadratic number field. In particular, one can study equivalence classes of ideals, infrastructure, and other properties of quadratic number fields.

Ideals in function fields are represented here almost exactly as in number fields; the two polynomials $u(X) = u$ and $v(X) = v$ represent the $K[X]$-module $uK[x] + (v + \sqrt{\Delta})K[x]$ of norm u, where $u \mid v^2 - \Delta$. If we set $w = (v^2 - \Delta)/u$, then we have a three coefficient representation $\varphi = (u, v, w)$ of the ideal. When viewed in this light, one realizes that the composition algorithms for binary quadratic forms, including NUCOMP and NUDUPL, generalize almost immediately to function fields. The main difference is that the formulas presented above will compute $\varphi_3 = (u_3, 2v_3, w_3)$ rather than (u_3, v_3, w_3), which is easily corrected as long as the ground field has odd characteristic. The modifications to Algorithm 1 (NUCOMP) for function fields over constant fields of odd characteristic are as follows:

- Step 1. $s \leftarrow v_1 + v_2$, $m \leftarrow v_2 - v_1$
- Step 6(e). $v_3 \leftarrow v_2 - Q_1$
- Step 7(h). $v_3 \leftarrow [G(Q_3 + Q_4) - Q_1 - Q_2]/2$
- Step 8. $\gamma \leftarrow x + y(v_3 - \sqrt{\Delta})/(Gu_3)$

The modifications for Algorithm 2 (NUDUPL) are the following:

- Step 1. $\ldots D_y \leftarrow 2v/G$
- Step 4(c). $v_3 \leftarrow [2v - (b_x + b_y)^2 + u_3 + w_3]/2$
- Step 5(f). $v_3 \leftarrow [v_3 - (b_x + b_y)^2 + u_3 + w_3]/2$
- Step 6. $\gamma \leftarrow x + y(v_3 - \sqrt{\Delta})/(Gu_3)$

In practice, the relative generator γ is not explicitly computed in function fields. Computing the degree of γ is sufficient, since it is more convenient to work with distances, i.e., the degrees of principal ideal generators and relative generators [9].

As in number fields, the main advantage of NUCOMP over composition in function fields is that the sizes of the intermediate operands remain small. In function fields, the size of the operands is measured by polynomial degree. Since reduced ideals in function fields satisfy $\deg(u) \leq g$, we want to use the partial Euclidean algorithm (Step 5 of NUCOMP and Step 3 of NUDUPL) to force $\deg(b_x), \deg(b_y) < L \approx g/2$, so that $\deg(u_3) \approx 2\deg(b_y) \approx g$ and φ_3 will be almost reduced. We found that taking $L = (g + 2)/2$ for imaginary quadratic function fields ($\deg(\Delta)$ is odd) and $L = (g + 1)/2$ for real quadratic function fields seemed to work the best.

2.3 Function Fields — Even Characteristic

Let $K = GF(q)$, $q = 2^n$, and let ρ be a root of the equation $y^2 + h(X)y = f(X)$ defined over $K[X]$. Adjoining ρ to the field of rational functions yields a quadratic congruence function field. As in the odd characteristic case, we can represent ideals in the function field by triples (u, v, w) where $w = (v^2 + h(X) + f(X))/u$. The composition and reduction algorithms are very similar, the main difference being that the conjugate ideal of (u, v, w) is given by $(u, v + h(X), w)$.

The modifications to Algorithm 1 (NUCOMP) for function fields over constant fields of even characteristic follow easily from Remark 5.4 of [13], and are described below. As above, ρ is a root of $y^2 + h(X)y = f(X)$ and we write h for $h(X)$.

- Step 1. $m \leftarrow v_1 + v_2$, $s \leftarrow m + h$
- Step 6(e). $v_3 \leftarrow v_2 + Q_1$
- Step 7(d). $Q_3 \leftarrow yd_x$, $d_y = (Q_3 + s)/x$
- Step 7(h). $v_3 \leftarrow Q_3 + Q_1 + v_1$
- Step 8. $\gamma \leftarrow x + y(v_3 + h + \rho)/(Gu_3)$

The modifications for Algorithm 2 (NUDUPL) are the following:

- Step 1. Use Euclid's extended algorithm to compute (y, G) such that $xu + yh = G = \gcd(u, h)$ and set $A_x \leftarrow G$, $B_y \leftarrow u/G$, and $D_y \leftarrow h/G$.
- Step 4(c). $v_3 \leftarrow v + b_x b_y$
- Step 5(b,c,d) $v_3 \leftarrow d_x y$, $d_y \leftarrow (v_3 + D_y)/x$
- Step 5(f). $v_3 \leftarrow v_3 + b_y b_x + v$
- Step 6. $\gamma \leftarrow x + y(v_3 + h + \rho)/(Gu_3)$

3 Performance in Practice

In the following, the algorithms for composition in all cases are the optimized ideal multiplication and squaring algorithms from [5, Chapter 2]. In our experience, composition can be performed more efficiently using ideals rather than binary quadratic forms. The NUCOMP and NUDUPL algorithms are implemented as described above, but the reduction algorithm is the optimized version from [5, Chapter 2]. Thus, we are using the most efficient ideal arithmetic and reduction using standard ideal multiplication known to us, as well as the most efficient NUCOMP and reduction with forms, allowing for as unbiased a comparison as possible. All runtimes are given in CPU seconds, and the computations are performed on an 800 MHz Pentium III processor running Linux. The algorithms were implemented using the NTL computer algebra library [11] with the GNU gmp multiprecision integer package installed as the integer arithmetic kernel, and compiled with the GNU g++ compiler version 2.91.66.

3.1 Imaginary Quadratic Fields

In order to compare the performance of NUCOMP and NUDUPL versus composition, we have implemented the Diffie-Hellmann key exchange protocol in the class group of an imaginary quadratic order [1]. For each discriminant size given in Table 1, we performed 5000 key exchanges with both NUCOMP and composition, using random discriminants of the given size and random exponents of the same bit-length as and bounded by $\sqrt{|\Delta|}$. Each communication partner performs two exponentiations per key exchange, so we expect each partner to perform about $\log_2 |\Delta|$ NUDUPL or ideal squaring operations and half as many NUCOMP or ideal multiplication operations per key exchange. The total time for all 5000 key exchanges per communication partner and the average time for a single key exchange per partner, using composition and NUCOMP, are given in the table, as well as the ratio of the total time for all key exchanges using NUCOMP over the total time using ideal multiplication. Our computations show that NUCOMP is already more efficient for discriminants of 64 bits, and becomes even more efficient as the discriminants grow in size.

Table 1: Imaginary quadratic field key exchange comparison.

	Comp. Time		NUCOMP Time		
$\lceil \log_2 \lvert\Delta\rvert \rceil$	Total	Avg.	Total	Avg.	NUCOMP/comp
32	7.82	0.00	8.98	0.00	1.1491
64	26.92	0.01	24.26	0.00	0.9012
128	102.95	0.02	77.83	0.02	0.7560
256	394.35	0.08	284.75	0.06	0.7221
512	1630.78	0.33	1057.69	0.21	0.6486
768	3848.80	0.77	2412.04	0.48	0.6267
1024	7291.36	1.46	4406.37	0.88	0.6043
2048	38390.05	7.68	20054.58	4.01	0.5224

3.2 Imaginary Quadratic Function Fields

We have also implemented the Diffie-Hellmann key exchange protocol in the class group of an imaginary quadratic congruence function field [7], where the ground field is any finite field of odd characteristic. The results in Table 2 were obtained using prime fields $\mathbb{F}_p$ as ground fields, where the prime was selected to be the smallest odd prime with the given number of bits. The results in Table 3 were obtained using various extensions of $\mathbb{F}_2$. In each of the tables, for each finite field and genus pair we performed a number of key exchanges using random function fields of the given genus and random exponents having the same bit-length as and bounded by q^g, where q is the cardinality of the finite field. For $g \leq 5$ we performed 4000 key exchanges using both NUCOMP and composition, for $5 < g \leq 10$ we performed 2000, for $10 < g \leq 15$ we performed 1000, and for $g > 15$ we performed 500. Here, we expect each communication partner to perform $2 \log_2 q^g$ NUDUPL or ideal squaring operations and half as many NUCOMP or ideal multiplication operations per key exchange. The ratio of the total time for all key exchanges using NUCOMP over the total time using ideal multiplication is given for each genus/field pair. In both tables we have not included computations for $g = 1$ (elliptic curves), since in this case simple direct formulas exist for group arithmetic.

Table 2: Imaginary function field over $\mathbb{F}_p$ key exchange — NUCOMP/composition.

	$\lceil \log_2 p \rceil$						
g	2	4	8	16	32	64	128
2	1.0778	1.2763	1.1848	1.1911	1.0979	1.0724	1.0371
3	1.2627	1.2492	1.3092	1.2922	1.1722	1.1562	1.1398
4	1.2450	1.2528	1.2698	1.2671	1.1225	1.1135	-
5	1.2365	1.1389	1.1426	1.1303	0.9997	0.9987	-
6	1.1331	1.1015	1.0792	1.0831	0.9717	0.9756	-
7	1.0987	1.0272	1.0089	1.0120	0.9111	0.9179	-

Table 2: (continued)

	$\lceil \log_2 p \rceil$						
g	2	4	8	16	32	64	128
8	1.0000	0.9931	0.9869	0.9950	0.8971	-	-
9	0.9903	0.9503	0.9292	0.9329	0.8411	-	-
10	0.9568	0.9218	0.9199	0.9187	0.8360	-	-
11	0.8991	0.8821	0.8732	0.8721	0.8061	-	-
12	0.8722	0.8634	0.8667	0.8641	0.8018	-	-
13	0.8552	0.8265	0.8205	0.8216	0.7681	-	-
14	0.8339	0.8206	0.8209	0.8212	0.7668	-	-
15	0.7995	0.7741	0.7751	0.7740	0.7480	-	-
20	0.7252	0.7204	0.7187	0.7217	-	-	-
25	0.6815	0.6808	0.6834	0.6848	-	-	-
30	0.6556	0.6561	0.6601	0.6631	-	-	-

Table 3: Imaginary function field over $GF(2^n)$ key exchange — NUCOMP/composition.

	n							
g	1	2	4	8	16	32	64	128
2	1.7143	1.2393	1.0495	1.0237	1.0145	0.9984	0.9893	0.9629
3	1.5154	1.2348	1.1231	1.1558	1.1291	1.1222	1.1115	1.0665
4	1.2981	1.1151	1.1528	1.1425	1.1182	1.0892	1.0749	-
5	1.3746	1.1348	1.0836	1.0668	1.0507	1.0230	1.0041	-
6	1.1875	1.0657	1.0770	1.0740	1.0503	1.0098	0.9818	-
7	1.2437	1.0145	1.0052	1.0025	0.9981	0.9519	0.9483	-
8	1.0511	1.0164	0.9977	0.9978	0.9956	0.9587	-	-
9	1.0796	0.9764	0.9430	0.9415	0.9447	0.9085	-	-
10	0.9820	0.9506	0.9333	0.9293	0.9346	0.9051	-	-
11	0.9813	0.9204	0.8967	0.8942	0.9018	0.8795	-	-
12	0.9443	0.8960	0.8885	0.8937	0.9011	0.8861	-	-
13	0.9063	0.8705	0.8590	0.8589	0.8674	0.8639	-	-
14	0.9114	0.8577	0.8526	0.8589	0.8691	0.8693	-	-
15	0.8805	0.8306	0.8249	0.8254	0.8416	0.8476	-	-
20	0.8086	0.7820	0.7874	0.7918	0.8051	-	-	-
25	0.7594	0.7397	0.7442	0.7545	0.7699	-	-	-
30	0.7222	0.7176	0.7270	0.7365	0.7528	-	-	-

According to our data, NUCOMP is more efficient than composition for function fields of fairly small genus, with the trade-off point lying between genus 5 and 10, depending on the ground field. In addition, NUCOMP becomes increasingly more efficient as both the genus and the size of the ground field increase (hence the discrepancies between the trade-off points for different ground fields). Both observations are explained by the fact that NUCOMP attempts to minimize the sizes of intermediate operands. In the case of function fields, we expect the degrees of the polynomial operands to be bounded by $O(3g/2)$ as opposed to $O(2g)$ for composition. As the genus increases, the difference between the degrees of the operands becomes greater, and the overall speed of NUCOMP as compared to composition also increases.

The fact that NUCOMP keeps the degrees of the intermediate operands small is also significant as the size of the ground field increases. If the cost of multiplying coefficients of the polynomials is expensive, then even small reductions in the polynomial degrees become beneficial. Thus, as the ground fields become larger, the trade-off points for which NUCOMP out-performs composition occur for smaller genus.

3.3 Real Quadratic Fields

In real quadratic fields, the corresponding Diffie-Hellmann key exchange protocol takes place in the principal ideal class [8,6]. The protocol essentially consists of each partner performing two binary exponentiations of principal ideals while keeping track of the principal ideal generator or its natural logarithm (distance). In practice, maintaining these distances to sufficient accuracy is somewhat problematic. We have used the approach of (f,p)-representations from [6] to keep track of the distances, using the same precision for the distance approximations for both composition and NUCOMP. Incorporating NUCOMP into the algorithms from [6] is fairly straightforward. Our implementation using NUCOMP always produced unique key ideals, even though the accuracy of the distance approximations is only guaranteed theoretically for regular composition [6].

For each discriminant size given in Table 4, we have performed 5000 key exchanges using random discriminants of the given size and random exponents of the same bit-length as and bounded by $\sqrt{\Delta}$. Each communication partner performs two exponentiations per key exchange, so we expect each partner to perform about $\log_2 \Delta$ NUDUPL or ideal squaring operations and half as many NUCOMP or ideal multiplication operations per key exchange. The total time for all 5000 key exchanges per communication partner and the average time for a single key exchange per partner, using regular composition and NUCOMP, are given in the table, as well as the ratio of the total time for all key exchanges using NUCOMP over the total time using ideal multiplication. Our computations show that NUCOMP is more efficient for discriminants of 32 bits or more, and as in the imaginary case, it becomes even more efficient as the discriminants grow in size.

Table 4: Real quadratic field key exchange comparison.

	Comp. Time		NUCOMP Time		
$\lceil \log_2 \Delta \rceil$	Total	Avg.	Total	Avg.	NUCOMP/comp
32	25.74	0.01	21.70	0.00	0.8430
64	99.47	0.02	70.19	0.01	0.7056

Table 4: (continued)

	Comp. Time		NUCOMP Time		
$\lceil \log_2 \varDelta \rceil$	Total	Avg.	Total	Avg.	NUCOMP/comp
128	408.71	0.08	262.38	0.05	0.6420
256	1825.36	0.37	1150.90	0.23	0.6305
512	7536.19	1.51	4535.24	0.91	0.6018
768	18371.47	3.67	10786.01	2.16	0.5871
1024	34749.08	6.95	20182.38	4.04	0.5808
2048	173514.36	34.70	96699.88	19.34	0.5573

Upon comparing the data for key exchange in real quadratic fields with that of imaginary quadratic fields, one finds that the benefits of using NUCOMP are somewhat more pronounced in the real case. The ideal multiplication part of the algorithms are the same in both cases, but reduction is more expensive using (f, p)-representations because fairly high precision distance approximations must be maintained. Since one benefit of NUCOMP is that a large portion of the reduction is done beforehand, it is to be expected that NUCOMP will yield a more substantial savings in the real case, since many of the expensive reduction steps involving the distance approximations are avoided.

One area in which NUCOMP and NUDUPL are especially effective is computations where one can take advantage of the relatively small operand sizes and use single precision arithmetic rather than multiprecision. Since NUCOMP requires intermediate operands of size $O(\varDelta^{3/4})$ [13], one can implement NUCOMP for fields with discriminant less than 10^{15} using almost exclusively single precision arithmetic (assuming 32-bit word size). For discriminants larger than 10^{10}, standard ideal arithmetic requires multiprecision arithmetic since the intermediate operands can be as large as $O(\varDelta)$.

To illustrate the effect of NUCOMP and NUDUPL in such settings, we have implemented a simple $O(\varDelta^{1/4+\epsilon})$ baby-step giant-step regulator computation routine. For each discriminant size given in Table 5, where we denote $\log_{10} |\varDelta|$ by size($\varDelta$), we have computed 10000 regulators using random discriminants of the given size. The total time for all 10000 regulator computations using both regular composition and NUCOMP are given in the table, as well as the ratio of the total time using NUCOMP over the total time using ideal multiplication.

Table 5. Quadratic field regulator comparison (single precision).

size($\varDelta$)	Regular composition	NUCOMP	NUCOMP/regular
7	144.46	79.48	0.55019
8	248.48	127.77	0.51421
9	431.27	209.19	0.48506
10	735.21	345.60	0.47007
11	1392.63	606.11	0.43523
12	2584.00	1053.50	0.40770

As expected, the effect of NUCOMP is rather dramatic in this case, cutting the total runtime in half.

3.4 Real Quadratic Function Fields

Unlike the case of real quadratic fields, maintaining distances in real quadratic function fields is easy, since they are integers (degrees of polynomials). The corresponding key exchange protocol in the principal class [9] is very similar to that in real quadratic number fields; each communication partner has to perform two binary exponentiations of principal ideals and maintain the corresponding distances. We have also implemented this protocol, and for each finite field and genus pair in Table 6 and Table 7, we have performed a number of key exchanges using random field discriminants of the given genus and random exponents bounded by q^g. As in the imaginary function field case, we expect each communication partner to perform $2\log_2 q^g$ NUDUPL or ideal squaring operations and half as many NUCOMP or ideal multiplication operations per key exchange. We performed 4000 key exchanges using both NUCOMP and composition for $g \leq 5$, 2000 for $5 < g \leq 10$, 1000 for $10 < g \leq 15$, and 500 for $g > 15$. The ratio of the total time for all key exchanges using NUCOMP over the total time using ideal multiplication is given for each genus/field pair. Again, we omit the data for $g = 1$ (elliptic curves), since the explicit formulas for the group law are more efficient than composition or NUCOMP.

Table 6: Real function field over $\mathbb{F}_p$ key exchange — NUCOMP/composition.

	$\lceil \log_2 p \rceil$						
g	2	4	8	16	32	64	128
2	1.1632	1.2673	1.2823	1.2719	1.2482	1.2647	1.2886
3	1.0928	1.2228	1.2651	1.2874	1.2223	1.2296	1.2338
4	1.2165	1.1511	1.1439	1.1447	1.0531	1.0693	-
5	1.1232	1.1393	1.1344	1.1363	1.0571	1.0656	-
6	1.0704	1.0563	1.0386	1.0449	0.9595	0.9769	-
7	1.0598	1.0491	1.0486	1.0485	0.9693	0.9782	-
8	1.0506	0.9835	0.9580	0.9603	0.8898	-	-
9	1.0026	0.9810	0.9722	0.9719	0.9013	-	-
10	0.9669	0.9261	0.9171	0.9216	0.8518	-	-
11	0.9643	0.9272	0.9240	0.9257	0.8641	-	-
12	0.9089	0.8842	0.8725	0.8724	0.8175	-	-
13	0.9038	0.8809	0.8749	0.8767	0.8291	-	-
14	0.8796	0.8509	0.8423	0.8457	0.7964	-	-
15	0.8709	0.8423	0.8351	0.8386	0.8090	-	-
20	0.7804	0.7692	0.7663	0.7703	-	-	-
25	0.7485	0.7475	0.7479	0.7513	-	-	-
30	0.7185	0.7146	0.7174	0.7195	-	-	-

Table 7: Real function field over $GF(2^n)$ key exchange — NUCOMP/composition.

	n							
g	1	2	4	8	16	32	64	128
2	0.8066	0.9766	1.0972	1.1664	1.1583	1.1793	1.1890	1.2176
3	0.8045	1.0597	1.1841	1.1910	1.1726	1.1839	1.1696	1.1595
4	0.8501	1.0464	1.0822	1.0741	1.0657	1.0662	1.0532	-
5	0.8989	1.0925	1.1082	1.1045	1.0940	1.0740	1.0484	-
6	0.9867	1.0351	1.0219	1.0108	1.0090	0.9911	0.9867	-
7	0.9488	1.0520	1.0258	1.0292	1.0291	0.9942	0.9945	-
8	1.0292	0.9834	0.9545	0.9579	0.9662	0.9462	-	-
9	1.0031	0.9837	0.9654	0.9710	0.9793	0.9549	-	-
10	1.0222	0.9360	0.9101	0.9141	0.9283	0.9148	-	-
11	0.9866	0.9358	0.9249	0.9289	0.9390	0.9266	-	-
12	0.9771	0.8956	0.8809	0.8851	0.8991	0.8966	-	-
13	0.9427	0.9012	0.8928	0.8988	0.9093	0.9097	-	-
14	0.9492	0.8676	0.8555	0.8605	0.8775	0.8871	-	-
15	0.9329	0.8697	0.8673	0.8744	0.8897	0.8909	-	-
20	0.8596	0.8103	0.8095	0.8123	0.8240	-	-	-
25	0.8176	0.7908	0.7954	0.8060	0.8242	-	-	-
30	0.7840	0.7630	0.7681	0.7769	0.7933	-	-	-

The same observations hold here as in the imaginary function field case. The performance of NUCOMP relative to composition improves as the genus increases and as the size of the ground field increases. However, unlike the number field case, NUCOMP does not seem to have as dramatic an effect in the real case as in the imaginary case when working in function fields. In function fields, the computational differences between the imaginary and real cases is not nearly as drastic as in number fields, since floating point approximations are not used to maintain distances. In particular, the reduction algorithms are almost identical, the only difference being that extra reduction steps are taken in the real case to ensure that the resulting composite has distance close to a given quantity. Thus, we expect that the absolute difference between the total runtimes using NUCOMP and composition to be roughly the same for the imaginary and real function field cases. This is exactly what we observed. The difference between the ratios of total NUCOMP time to total composition time between the two cases is accounted for by the fact that the amount of extra work required for the real case is the same for both NUCOMP and composition.

4 Further Work

One immediate extension of our work is a detailed complexity analysis of NUCOMP in function fields using the model of [12]. By comparing our results from this analysis with that of the usual composition and reduction algorithms from

[12], we will be able to precisely predict the trade-off points where NUCOMP out-performs composition. As a part of this analysis, we will determine bounds on the degrees of the intermediate operands. Preliminary experiments indicate that NUCOMP performs exceptionally well in function fields; the vast majority of near-reduced composites are in fact already reduced and the degrees of the intermediate operands do appear to be close to $3g/2$. We will conduct more refined experiments as a complement to our analysis.

Our computations with NUCOMP in real quadratic fields rely upon the (f, p)-representations of distances as described in [6]. As mentioned earlier, the floating-point precision required to guarantee unique keys in the key exchange protocol is only valid for composition. The analysis of the precision requirements needs to be extended if NUCOMP and NUDUPL are to be used with confidence in this key exchange protocol. This, also, is work in progress.

References

1. J. Buchmann and H.C. Williams, *A key-exchange system based on imaginary quadratic fields*, Journal of Cryptology **1** (1988), 107–118.
2. D.G. Cantor, *Computing in the Jacobian of a hyperelliptic curve*, Math. Comp. **48** (1987), no. 177, 95–101.
3. H. Cohen, *A course in computational algebraic number theory*, Springer-Verlag, Berlin, 1993.
4. S. Düllmann, *Ein Algorithmus zur Bestimmung der Klassengruppe positiv definiter binärer quadratischer Formen*, Ph.D. thesis, Universität des Saarlandes, Saarbrücken, Germany, 1991.
5. M.J. Jacobson, Jr., *Subexponential class group computation in quadratic orders*, Ph.D. thesis, Technische Universität Darmstadt, Darmstadt, Germany, 1999.
6. M.J. Jacobson, Jr., R. Scheidler, and H.C. Williams, *The efficiency and security of a real quadratic field based key exchange protocol*, Public-Key Cryptography and Computational Number Theory (Warsaw, Poland), de Gruyter, 2001.
7. N. Koblitz, *Hyperelliptic cryptosystems*, Journal of Cryptology **1** (1989), 139–150.
8. R. Scheidler, J. Buchmann, and H.C. Williams, *A key-exchange protocol using real quadratic fields*, Journal of Cryptology **7** (1994), 171–199.
9. R. Scheidler, A. Stein, and H.C. Williams, *Key-exchange in real quadratic congruence function fields*, Designs, Codes and Cryptography **7** (1996), 153–174.
10. D. Shanks, *On Gauss and composition I, II*, Proc. NATO ASI on Number Theory and Applications (R.A. Mollin, ed.), Kluwer Academic Press, 1989, pp. 163–179.
11. V. Shoup, *NTL: A library for doing number theory*, Software, 2001; see `http://www.shoup.net/ntl`.
12. A. Stein, *Sharp upper bounds for arithmetics in hyperelliptic function fields*, J. Ramanujan Math. Soc. **16** (2001), no. 2, 1–86.
13. A.J. van der Poorten, *A note on NUCOMP*, to appear in *Math. Comp.*

Efficient Computation of Class Numbers of Real Abelian Number Fields

Stéphane R. Louboutin

Institut de Mathématiques de Luminy, UPR 906
163, avenue de Luminy, Case 907
13288 Marseille Cedex 9, FRANCE
loubouti@iml.univ-mrs.fr

Abstract. Let $\{K_m\}$ be a parametrized family of real abelian number fields of known regulators, e.g. the simplest cubic fields associated with the $\mathbf{Q}$-irreducible cubic polynomials $P_m(x) = x^3 - mx^2 - (m+3)x - 1$. We develop two methods for computing the class numbers of these K_m's. As a byproduct of our computation, we found 32 cyclotomic fields $\mathbf{Q}(\zeta_p)$ of prime conductors $p < 10^{10}$ for which some prime $q \geq p$ divides the class numbers h_p^+ of their maximal real subfields $\mathbf{Q}(\zeta_p)^+$ (but we did not find any conterexample to Vandiver's conjecture!).

1 Introduction

This paper is an abridged version of [Lou5] in which the reader will find the proofs we omit here, and in which he will also find various supplementary examples (families of real cyclic quartic, sextic and octic fields). Our aim is to explain how one can generalize the technique developed in [Lou1] not only to compute efficiently class numbers of real abelian number fields of known regulators, but also to compute efficiently exact values of Gauss sums and roots numbers associated with primitive Dirichlet characters of large conductors.

In [Bye], [Lou4], [LP], [Sha] and [Wa], various authors dealt with the so called **simplest cubic fields**, the real cyclic cubic number fields K_m associated with the cubic polynomials

$$P_m(x) = x^3 - mx^2 - (m+3)x - 1$$

of discriminants $d_m = \Delta_m^2$, where $\Delta_m := m^2 + 3m + 9$, and roots θ_m, $\sigma(\theta_m) = -1/(\theta_m + 1)$ and $\sigma^2(\theta_m) = -(\theta_m + 1)/\theta_m$. Since $-x^3 P_m(1/x) = P_{-m-3}(x)$, we may assume that $m \geq -1$. In this paper, we assume that Δ_m is square-free. In that case, the conductor of K_m is equal to Δ_m, its discriminant is equal to Δ_m^2, the set $\{-1, \theta_m, \sigma(\theta_m)\}$ generates the full group of algebraic units of K_m and the regulator of K_m is

$$\mathrm{Reg}_{K_m} = \log^2 \theta_m - (\log \theta_m)(\log(1 + \theta_m)) + \log^2(1 + \theta_m),$$

with

$$\theta_m = \frac{1}{3}\Big(2\sqrt{\Delta_m}\cos\Big(\frac{1}{3}\arctan(\frac{\sqrt{27}}{2m+3})\Big) + m\Big).$$

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 134–147, 2002.

In [Jean] and [SW], S. Jeanin, R. Schoof and L. C. Washington dealt with the so called **simplest quintic fields**, the real cyclic quintic number fields K_m associated with the quintic polynomials

$$\begin{aligned}P_m(x) = x^5 + m^2x^4 - (2m^3 + 6m^2 + 10m + 10)x^3 \\ +(m^4 + 5m^3 + 11m^2 + 15m + 5)x^2 + (m^3 + 4m^2 + 10m + 10)x + 1\end{aligned}$$

of discriminants $d_m = (m^3 + 5m^2 + 10m + 7)^2\Delta_m^4$, $\Delta_m = m^4 + 5m^3 + 15m^2 + 25m+25$ and roots θ_m, $\sigma(\theta_m) = ((m+2)+m\theta_m - \theta_m^2)/(1+(m+2)\theta_m)$, $\sigma^2(\theta_m)$, $\sigma^3(\theta_m)$ and $\sigma^4(\theta_m)$. In this paper, we assume that Δ_m is square-free. In that case, the conductor of K_m is equal to Δ_m, its discriminant is equal to Δ_m^4, the set $\{-1, \theta_m, \sigma(\theta_m), \sigma^2(\theta_m), \sigma^3(\theta_m)\}$ generates the full group of algebraic units of K_m and the regulator of K_m is

$$\mathrm{Reg}_{K_m} = \frac{1}{5}\prod_{1\le i\le 4}\Big(\sum_{0\le j\le 4}\zeta_5^{ij}\log|\sigma^j(\theta_m)|\Big).$$

Since $P_m(m+1)P_m(m+2) = -(m^3+5m^2+10m+7)^2 < 0$ we can use Newton's method for computing efficiently as good as desired numerical approximations of a root $\theta_m \in (m+1, m+2)$ of $P_m(x)$. Then, the four other roots are computed inductively by the transformation $\theta \mapsto \sigma(\theta) := ((m+2)+m\theta-\theta^2)/(1+(m+2)\theta)$.

One of the motivation for computing class numbers of simplest cubic and quintic fields stems from Vandiver's conjecture according to which p never divides the class number h_p^+ of the maximal real subfield $\mathbf{Q}(\zeta_p)^+ = \mathbf{Q}(\cos(2\pi/p))$ of a cyclotomic field $\mathbf{Q}(\zeta_p)$ of prime conductor p. However, as the computation of h_p^+ is impossible to perform (except for very small values of p, say $p \le 67$ (see [Wa, Tables, pages 420-423])), the idea is to compute class numbers h_K of subfields K of $\mathbf{Q}(\zeta_p)^+$ of small degrees:

Theorem 1. (i). *Let $p \equiv 1 \pmod{12}$ be a prime and let h_2, h_3 and h_6 denote the class numbers of the real quadratic, cubic and sextic subfields of the cyclotomic field $\mathbf{Q}(\zeta_p)$. Then, h_2h_3 divides h_6 and h_6 divides the class number h_p^+ of the maximal real subfield $\mathbf{Q}(\zeta_p)^+ = \mathbf{Q}(\cos(2\pi/p))$ of $\mathbf{Q}(\zeta_p)$ (see* [CW, Lemmas 1 and 2]*). However, all the prime factors q of h_6 are less than p (see* [Mos]*).* (ii). *Let $p \equiv 1 \pmod{10}$ be a prime and let h_5 denote the class number of the real quintic subfield of the cyclotomic field $\mathbf{Q}(\zeta_p)$. Then, h_5 divides h_p^+.*

Since the simplest cubic and quintic fields have small regulators we might expect to find some of them of prime conductors and large class numbers. Therefore, by using simplest cubic fields we might expect to find examples of cyclotomic fields of prime conductors p for which $h_p^+ \ge p$ but for which, unfortunately, all the prime factors q of h_p^+ could be less than p. Up to now, only one such example had been found (see [CW] and [SWW]), and we will find three more examples (see Table 3). In the same way, by using simplest quintic fields we might expect to find examples of cyclotomic fields of prime conductors p for which some prime factor q of h_p^+ satisfies $q \ge p$. Up to now, only one such example had been found (see [SW] and [Jean]), and we will find 31 more examples (see Table 2).

2 First Method for Computing Class Numbers

Let K be a real abelian number field of degree $q > 1$, discriminant d_K and conductor f_K associated with a $\mathbf{Q}$-irreducible unitary polynomial $P_K(X) = X^q + a_{q-1}X^{q-1} + \cdots + a_0 \in \mathbf{Z}[X]$. Let X_K denote the group (of order q) of primitive even Dirichlet characters associated with K and let Reg_K denote the regulator of K. According to the analytic class number formula (see [Lan, Chapter XIII, section 3, Th. 2 page 259]), $s \mapsto F_K(s) = (d_K/\pi^q)^{s/2}\Gamma^q(s/2)\zeta_K(s)$ has a simple pole at $s = 0$ of residue

$$\mathrm{Res}_{s=0}(F_K(s)) = -2^{q-1}h_K\mathrm{Reg}_K = 2^q \lim_{s\to 0} s^{q-1}\zeta_K(s).$$

Since $\zeta_K(s) = \prod_{\chi\in X_K} L(s,\chi)$ and $L(0,\chi) = -1/2$ if $\chi = 1$ but $L(0,\chi) = 0$ for $1 \neq \chi \in X_K$, we obtain

$$h_K\mathrm{Reg}_K = \prod_{1\neq\chi\in X_K} L'(0,\chi). \tag{1}$$

Lemma 1. *(See* [Sta]*). If χ is a (non-necessarily primitive) non-trivial even Dirichlet character modulo $f > 1$, then $L'(0,\chi) = -\sum_{1\leq k<f/2} \chi(k)\log\sin(k\pi/f)$.*

From now on, to simplify, we assume that $q \geq 3$ is an odd prime. Then, $K/\mathbf{Q}$ is cyclic of degree q, the conductor f_K of K is odd, X_K is cylic of order q, the conductors f_χ of all the $1 \neq \chi \in X_K$ are equal to f_K, and the characters $1 \neq \chi \in X_K$ come in conjugate pairs $\{\chi, \bar{\chi}\}$. Hence, using (1) and letting χ_K denote any one of the $q-1$ generators of X_K, we obtain the simple formula

$$h_K\mathrm{Reg}_K = \prod_{1\leq l\leq(q-1)/2} \Big| \sum_{1\leq k\leq f_K/2} \chi_K^l(k)\log\sin(k\pi/f_K)\Big|^2. \tag{2}$$

To further simplify, we finally assume that f_K is square-free. Then, $f_K = \prod_{1\leq i\leq t} p_i$ is a product of $t \geq 1$ pairwise distinct odd primes $p_i \equiv 1 \pmod q$ and $\chi = \prod_{1\leq i\leq t} \chi_{p_i}$ where each χ_{p_i} is a character of order q modulo p_i. Now, for a given prime $p \equiv 1 \pmod q$, we set $g_p = \min\{g \geq 1;\ g^{(p-1)/q} \not\equiv 1 \pmod p\}$, $G_p = g_p^{(p-1)/q} \bmod p$ and we let ϕ_p be the character of order q mod p defined by $\phi_p(x) = \zeta_q^{k(x)}$ where $k(x) = \min\{k \in \{0,1,\cdots,q-1\};\ G_p^k \equiv x^{(p-1)/q} \pmod p\}$ (for $\gcd(x,p) = 1$). To each $n \in \{0,1,\cdots,(q-1)^{t-1}-1\}$ we associate its $(q-1)$-adic development $n = \sum_{2\leq i\leq t}(a_i-1)(q-1)^{i-2}$, $a_i \in \{1,2,\cdot,q-1\}$, and the primitive mod f_K character of order q

$$\psi_n = \phi_{p_1}\prod_{2\leq i\leq t} \phi_{p_i}^{a_i}.$$

There exists a unique $n_K \in \{0,1,\cdots,(q-1)^{t-1}-1\}$ such that the primitive character $\chi_K := \psi_{n_K}$ of order q generates the cyclic group X_K of primitive Dirichlet characters associated with K. The following algorithm provides us with an efficient technique for determining this unique n_K:

```
1. $n := 0$, $n' := (q-1)^{t-1} - 1$.
2. If $n = n'$ then go to step 9.
3. $p := 3$.
4. While $p$ divides $d_P$ do $p :=$next prime.
5. If $P_K(X)$ has no root modulo $p$ then do $p :=$next prime and go to step 4.
```

Now, since $P_K(X)$ has at least one root modulo p and since p does not divide the discriminant d_P of $P_K(X)$, it holds that p splits in K and we must have $\psi_n(p) = \chi(p) = +1$. Hence, we now do:

```
6. If $\psi_n(p) \neq 1$ then $\{n := n+1$; go to step 2$\}$.
7. If $\psi_{n'}(p) \neq 1$ then $\{n' := n'-1$; go to step 2$\}$.
8. $p :=$next prime and go to step 4
9. Return($n$).
```

Practically, this algorithm is fast for we only have to use Step 5 for small primes p. In fact, assume the Generalized Riemann Hypothesis. Then, for any distinct Dirichlet characters χ and χ' mod f there exists some prime $p \leq 3\log^2 f$ which does not divide f such that $\chi(p) \neq \chi'(p)$ (apply [Ba, Theorem 3] with $G = \ker(\chi\chi'^{-1})$). Hence, the primes p which arise in our algorithm satisfy $p \leq 3\log^2 \Delta_m$. For example, we used this method in the case of simplest quintic fields to compute the data given in Table 1 in 2h35mn. The computations were all carried out on a PC microcomputer with Pentium IV, 1Ghz, by using Pr. Y. Kida's UBASIC language which allows fast arbitrary precision calculation on PC's. (See also [Lou4] for another example of the use of this method in the case of simplest cubic fields).

3 A Faster Method for Computing Class Numbers

In this section we develop a more complicated but much more efficient technique for computing class numbers of real abelian number fields of a given degree $q > 1$ and known regulators (it will practically require only $O(f_K^{0.5+\epsilon})$ elementary operations to compute h_K, whereas our previous techniques based on (1) and (2) requires $O(f_K^{1+\epsilon})$ elementary operations to compute h_K). The idea is to generalize [WB, Section 3] to compute efficiently good enough numerical approximations to $L'(0,\chi)$ for $1 \neq \chi \in X_K$, and to use (1). Let χ be a primitive even Dirichlet character modulo $f > 1$. Set

$$\tau(\chi) = \sum_{1 \leq n \leq f} \chi(n)\exp(2n\pi i/f) \qquad \text{(Gauss sum)}, \tag{3}$$

$$W(\chi) = \tau(\chi)/\sqrt{f} \qquad \text{(root number)} \tag{4}$$

and

$$\theta(x,\chi) = \sum_{n \geq 1} \chi(n)e^{-\pi n^2 x/f} \qquad (x > 0). \tag{5}$$

Then, $|W(\chi)| = 1$ and using

$$\theta(1/x,\chi) = W(\chi)\sqrt{x}\theta(x,\bar{\chi}) \qquad (x > 0), \tag{6}$$

we obtain

$$\begin{aligned}&(f/\pi)^{s/2}\Gamma(s/2)L(s,\chi)\\ &=\int_0^\infty \theta(x,\chi)x^{s/2}\frac{dx}{x}=\int_1^\infty \theta(x,\chi)x^{s/2}\frac{dx}{x}+W(\chi)\int_1^\infty \theta(x,\bar\chi)x^{(1-s)/2}\frac{dx}{x},\end{aligned}$$

$L(0,\chi)=0$ and the following result which enables us to compute numerical approximations to $L'(0,\chi)$ to any prescribed accuracy:

Theorem 2. *Let χ be a primitive even Dirichlet character modulo $f>1$. Then,*

$$L'(0,\chi)=\frac{1}{2}\sum_{n\geq 1}\chi(n)\int_{\pi n^2/f}^\infty e^{-t}\frac{dt}{t}\ +W(\chi)\sqrt{\frac{f}{\pi}}\sum_{n\geq 1}\frac{\bar\chi(n)}{n}\int_{\sqrt{\pi n^2/f}}^\infty e^{-t^2}dt. \quad (7)$$

Hence, setting

$$\begin{aligned}E_1(x):=\int_x^\infty e^{-t}\frac{dt}{t}&=-\log x\ -\gamma\ -\sum_{k\geq 1}\frac{(-1)^k}{k\cdot k!}x^k\\ &=e^{-x}\Big(\frac{1}{z+}\,\frac{1}{1+}\,\frac{1}{z+}\,\frac{2}{1+}\,\frac{2}{z+}\,\frac{3}{1+}\,\frac{3}{z+}\cdots\Big)\end{aligned}$$

(where $\gamma=0.577\ 215\ 664\ 901\ 532\cdots$ denotes Euler's constant),

$$\begin{aligned}E_2(x):=\frac{2}{\sqrt\pi}\int_{\sqrt x}^\infty e^{-t^2}dt&=1-2\sqrt{\frac{x}{\pi}}\sum_{k\geq 0}\frac{(-1)^k}{(2k+1)\cdot k!}x^k\\ &=\frac{1}{\sqrt\pi}e^{-x}\Big(\frac{1}{z+}\,\frac{\frac12}{z+}\,\frac{\frac22}{z+}\,\frac{\frac32}{z+}\,\frac{\frac42}{z+}\,\frac{\frac52}{z+}\cdots\Big),\end{aligned}$$

and

$$L_N'(0,\chi)=\frac{1}{2}\sum_{1\leq n\leq N}\chi(n)E_1(\pi n^2/f)\ +\frac{W(\chi)\sqrt f}{2}\sum_{1\leq n\leq N}\frac{\bar\chi(n)}{n}E_2(\pi n^2/f)$$

($N\geq 1$ a positive integer), it holds that

$$|L'(0,\chi)-L_N'(0,\chi)|\leq\frac{1}{2M^t\sqrt{\pi t^3}}\frac{f^{\frac12-t}}{\log^{3/2}(Mf)}$$

for

$$N\geq B(t,f,M):=\sqrt{\frac{tf}{\pi}\log(Mf)}. \quad (8)$$

Corollary 1. *(See* [Lou3, Proof of Theorem 7]*). Let $q\geq 2$ be a given prime. Fix $t>(q-1)/2$ and $M>0$, and let K range over a family of real abelian numbers fields K of degree q for which all the root numbers $W(\chi)$, $1\neq\chi\in X_K$, are known. Then, as $f_K\longrightarrow\infty$ and for $N\geq B(t,f_K,M)$, the limit $|\frac{1}{\mathrm{Reg}_K}\prod_{1\neq\chi\in X_K}L'(0,\chi)-\frac{1}{\mathrm{Reg}_K}\prod_{1\neq\chi\in X_K}L_N'(0,\chi)|$ is equal to zero.*

4 Efficient Computation of Root Numbers

According to Corollary 1, (1) and Theorem 2 could be used to compute efficiently numerical approximations $L'_N(0,\chi)$ to $L'(0,\chi)$ for primitive even Dirichlet characters of order $n_\chi > 1$ and class numbers of real abelian number fields. However, as there is no known general formula for Gauss sums (see [BE]), we will now explain how to compute efficiently these root numbers $W(\chi)$ (notice that since the use of (4) to compute the exact value of $W(\chi)$ requires $\gg f_\chi$ elementary operations, it would be much simpler to use Theorem 1 than to use (4) and Theorem 2). We point out that we are going to end up with a method for computing class numbers of real abelian number fields which is more satisfactory than the one previously used (see [SW] and [SWW]): we compute exact values of root numbers, whereas in [SWW] they had three choices for $W(\chi)$ for simplest cubic fields of a given prime conductor and in [SW, Top of page 553] they had twenty choices for $W(\chi)$ for simplest quintic fields of a given prime conductor. In their computations they were lucky enough for in all cases considered only one of their possible choices gave rise to an approximation of the class number sufficiently close to an integer. To begin with, let us fix some notation. Throughout this fourth section, we let χ denote a primitive even Dirichlet character or order $n_\chi > 1$ and conductor $f_\chi > 1$. We set $\omega(\chi) = (\tau(\chi))^{n_\chi}$, $\zeta_\chi = \exp(2\pi i/n_\chi)$ and $\mathbf{Q}(\chi) = \mathbf{Q}(\zeta_\chi)$. We let $\phi_\chi = \phi(n_\chi)$ and $\mathbf{Z}[\chi] = \mathbf{Z}[\zeta_\chi]$ denote the degree and the ring of algebraic integers of the cyclotomic field $\mathbf{Q}(\chi)$, respectively. Finally, for any l relatively prime to n_χ, we let σ_l denote the $\mathbf{Q}$-automorphism of $\mathbf{Q}(\chi)$ which is defined by $\sigma_l(\zeta_\chi) = \zeta_\chi^l$. Notice that if $\gcd(l, n_\chi) = 1$, then χ^l is also a primitive Dirichlet character of order n_χ and conductor f_χ and that χ^l is even (respectively odd) if χ is even (respectively odd).

Theorem 3. *Let χ be a primitive Dirichlet character of conductor $f_\chi > 1$ and order $n_\chi > 1$. Then*

$$\omega(\chi) := (\tau(\chi))^{n_\chi} = f_\chi^{n_\chi/2}(W(\chi))^{n_\chi} \in \mathbf{Z}[\chi] \tag{9}$$

and $\sigma_l(\omega(\chi)) = \omega(\chi^l)$ for $\gcd(l, n_\chi) = 1$. Moreover, if n_χ is prime and f_χ is square-free, then $\omega(\chi) \in f_\chi\mathbf{Z}[\chi]$.

4.1 Exact Computation of $\omega(\chi)$

Fix a $\mathbf{Z}$-basis $\mathcal{B} = \{\epsilon_1, \cdots, \epsilon_{\phi_\chi}\}$ and write

$$\omega(\chi) = \sum_{1 \le k \le \phi_\chi} b(k,\chi)\epsilon_k \in \mathbf{Z}[\chi]. \tag{10}$$

with $b(k,\chi) \in \mathbf{Z}$, $1 \le k \le \phi_\chi$. Let $\mathcal{B}^\perp = \{\eta_1, \cdots, \eta_{\phi_\chi}\}$ be the dual basis of $\mathcal{B}$, relative to the trace form. Hence, $\mathrm{Tr}_{\mathbf{Q}(\chi)/\mathbf{Q}}(\epsilon_k\eta_k) = 1$ but $\mathrm{Tr}_{\mathbf{Q}(\chi)/\mathbf{Q}}(\epsilon_k\eta_l) = 0$ for $k \ne l$, and

$$b(k,\chi) = \mathrm{Tr}_{\mathbf{Q}(\chi)/\mathbf{Q}}(\eta_k\omega(\chi)) = f_\chi^{n_\chi/2} \sum_{\substack{1 \le l \le n_\chi \\ \gcd(l,n_\chi)=1}} \sigma_l(\eta_k)(W(\chi^l))^{n_\chi} \tag{11}$$

(for $\sigma_l(\omega(\chi)) = \omega(\chi^l)$), and these coordinates $b(k,\chi)$ are rational integers of reasonable size: $|b(k,\chi)| \leq M(\mathcal{B}^\perp)\phi_\chi f_\chi^{n_\chi/2}$ where

$$M(\mathcal{B}^\perp) = \max\{|\sigma_l(\eta_j)|;\ 1 \leq l \leq n_\chi,\ \gcd(l,n_\chi) = 1,\ 1 \leq j \leq \phi_\chi\}. \tag{12}$$

For example, if $n_\chi = q \geq 3$ is prime, then $\mathcal{B}^\perp = \{\eta_l := (\zeta_q^{-l} - 1)/q;\ 1 \leq l \leq q-1\}$ is the dual basis of the $\mathbf{Z}$-basis $\mathcal{B} = \{\zeta_q^k;\ 1 \leq k \leq q-1\}$ of the ring of algebraic integers $\mathbf{Z}[\zeta_q]$ of $\mathbf{Q}(\zeta_q)$, and $M(\mathcal{B}^\perp) \leq 2/q \leq 1$.
Now assuming that

$$\textbf{Hypothesis: } \theta(\chi^l) := \sum_{n \geq 1} \chi^l(n) e^{-\pi n^2/f_\chi} \neq 0 \tag{13}$$

for $1 \leq l \leq n_\chi$ and $\gcd(l,n_\chi) = 1$, we explain how one can compute efficiently as good as desired numerical approximations $b_N(k,\chi)$ to these coordinates $b(k,\chi) \in \mathbf{Z}$ of $\omega(\chi)$, hence how one can compute their exact values. The key point is that $\theta(\chi^l) \neq 0$ implies

$$W(\chi^l) = \theta(\chi^l)/\overline{\theta(\chi^l)},$$

by (6). According to **Section 4.4** below, this Hypothesis should always be satisfied. The following Lemma 2 will enable us to compute as good as desired numerical approximations $\theta_N(\chi^l)$ to $\theta(\chi^l)$. These approximations will then enable us to check the **Hypothesis** (13) prior to using Lemma 3 for computing as good as desired numerical approximations $b_N(k,\chi)$ to the rational integers $b(k,\chi)$ defined in (10), whose exact values can therefore be deduced.

Lemma 2. *Let χ be a Dirichlet character modulo $f > 1$. Set*

$$\theta_N(\chi) = \sum_{1 \leq n \leq N} \chi(n) e^{-\pi n^2/f}$$

($N \geq 1$ a positive integer). If $N \geq B(t,f,M)$ (as in (8)), then

$$|\theta(\chi) - \theta_N(\chi)| \leq \frac{1}{2M^t\sqrt{\pi t}} \frac{f^{\frac{1}{2}-t}}{\sqrt{\log(Mf)}}. \tag{14}$$

Lemma 3. *Let χ be a primitive even Dirichlet character of order $n_\chi > 1$ and conductor $f_\chi > 1$. Assume that $\theta_N(\chi^l) \neq 0$ for $\gcd(l,n_\chi) = 1$, set*

$$W_N(\chi^l) = \theta_N(\chi^l)/\overline{\theta_N(\chi^l)}$$

and

$$b_N(k,\chi) = f_\chi^{n_\chi/2} \sum_{\substack{1 \leq l \leq n_\chi \\ \gcd(l,n_\chi)=1}} \sigma_l(\eta_k)(W_N(\chi^l))^{n_\chi} \qquad (1 \leq k \leq \phi_\chi), \tag{15}$$

and fix ϵ such that $0 \leq \epsilon \leq 1$. Assume that $|\theta(\chi^l) - \theta_N(\chi^l)| \leq \epsilon|\theta_N(\chi^l)|/n_\chi$ for $1 \leq l \leq n_\chi$ and $\gcd(l,n_\chi) = 1$. Then,

$$|b_N(k,\chi) - b(k,\chi)| \leq \frac{27(e-1)}{4} M(\mathcal{B}^\perp)\phi_\chi f_\chi^{n_\chi/2}\epsilon.$$

Proof. Let us simplify the notation: we set $n = n_\chi$, $\theta = \theta(\chi^l)$, $\theta_N = \theta_N(\chi^l)$, $W = W(\chi^l) = \theta/\bar{\theta}$ (notice that $\theta_N \neq 0$ and $|\theta - \theta_N| \leq \epsilon|\theta_N|/n$ imply $\theta \neq 0$), $W_N = W_N(\chi^l) = \theta_N/\bar{\theta}_N$ and write $\theta = \theta_N + \epsilon_N\theta_N$ with $|\epsilon_N| \leq \epsilon/n$. Then,

$$|W^n - W_N^n| = \frac{|(1+\epsilon_N)^n - (1+\bar{\epsilon}_N)^n|}{|1+\bar{\epsilon}_N|^n} = \frac{|\sum_{k=1}^n \binom{n}{k}(\epsilon_N^k + \bar{\epsilon}_N^k)|}{|1-\epsilon_N|^n}$$

yields

$$|W^n - W_N^n| \leq 2\frac{\sum_{k=1}^n \binom{n}{k}|\epsilon_N|^k}{(1-|\epsilon_N|)^n} = 2\frac{(1+|\epsilon_N|)^n - 1}{(1-|\epsilon_N|)^n} \leq 2\frac{(1+\epsilon/n)^n - 1}{(1-1/n)^n}.$$

Since $(1-1/n)^n \geq (1-1/3)^3 = 8/27$ for $n \geq 3$ and since $(1+\epsilon/n)^n - 1 \leq e^\epsilon - 1 \leq (e-1)\epsilon$ for $0 \leq \epsilon \leq 1$, we obtain $|W(\chi^l)^n - W_N(\chi^l)^n| \leq 27(e-1)\epsilon/4$ for $1 \leq l \leq n_\chi$ and $\gcd(l, n_\chi) = 1$, and the desired results, by (11), (12) and (15).

4.2 Exact Computation of $W(\chi)$ and $\tau(\chi)$

Now that we know how to compute the exact value of $\omega(\chi) := (\tau(\chi))^{n_\chi}$, let us explain how one can determine which of its n_χth root is equal to $\tau(\chi)$:

Lemma 4. *Fix $\epsilon \in (0,1]$. Let χ be a primitive even Dirichlet character of order $n_\chi > 2$ and conductor $f_\chi > 1$. Assume that $\omega(\chi)$ is known and that N is such that $\theta_N(\chi) \neq 0$ and $|\theta(\chi) - \theta_N(\chi)| \leq \epsilon|\theta_N(\chi)|/n_\chi$. Fix W a n_χth root of $(W(\chi))^{n_\chi} = \omega(\chi)/f^{n_\chi/2}$. Then, $W(\chi) = \zeta_\chi^{k_0}W$ where k_0 in the unique integer $k \in \{0, 1, \cdots, n-1\}$ such that $|W_N(\chi) - \zeta_\chi^k W| < 2\epsilon/n_\chi$ (and it holds that $|W_N(\chi) - \zeta_\chi^k W| > (4-2\epsilon)/n_\chi \geq 2\epsilon/n_\chi$ for $k \neq k_0$).*

Proof. Since $|\theta - \theta_N| \leq \epsilon|\theta_N|/n_\chi$, we have $\theta \neq 0$, $\theta_N \neq 0$ and

$$|W(\chi) - W_N(\chi)| = |\frac{\theta}{\bar{\theta}} - \frac{\theta_N}{\bar{\theta}_N}| = \frac{|\theta(\bar{\theta}_N - \bar{\theta}) + \theta(\bar{\theta}_N - \bar{\theta})|}{|\bar{\theta}\bar{\theta}_N|} \leq 2\frac{|\theta - \theta_N|}{|\theta_N|} \leq \frac{2\epsilon}{n_\chi}.$$

There exists a unique $k_0 \in \{0, 1, \cdots, n-1\}$ such that $W(\chi) = \zeta_\chi^{k_0}W$. Since for $a \neq b$ we have $|\zeta_\chi^a W - \zeta_\chi^b W| = |\zeta_\chi^a - \zeta_\chi^a| \geq 2\sin(\pi/n_\chi) > 4/n_\chi$, we have $|W_N(\chi) - \zeta_\chi^{k_0}W| = |W_N(\chi) - W(\chi)| < 2\epsilon/n_\chi$ and $|W_N(\chi) - \zeta_n^k W| = |(W_N(\chi) - W(\chi)) + (\zeta_\chi^{k_0}W - \zeta_\chi^k W)| \geq 2\sin(\pi/n_\chi) - |W_N(\chi) - W(\chi)| > 4/n - 2\epsilon/n_\chi \geq 2\epsilon/n_\chi$ for $k \neq k_0$.

4.3 Computation of Class Numbers of Simplest Quintic Fields

First, we checked our present method by recomputing Table 1. Second, we used it to compute the class numbers of all the simplest quintic fields K_m's of conductors $\Delta_m = m^4 + 5m^3 + 15m^2 + 25m + 25 \leq 10^{10}$ a prime. We obtain the following consequence: there are 32 simplest quintic fields K_m of prime conductors $p \leq 10^{10}$ whose class numbers are divisible by some prime $q \geq p$ (see Table 2). Third, we used it to compute the class numbers of all the simplest cubic fields K_m's with $-1 \leq m \leq 554869$ and $\Delta_m = m^2+3m+9 \equiv 1 \pmod{12}$ a prime. We obtain the following consequence: in the range $-1 \leq m \leq 554869$, there are only 4 simplest cubic fields K_m of prime conductors $\Delta_m = m^2+3m+9 \equiv 1 \pmod{12}$ for which the product h_2h_3 of the class number h_2 of the real quadratic field $\mathbf{Q}(\sqrt{\Delta_m})$ and of the class number h_3 of the simplest cubic field K_m of conductor Δ_m is greater than or equal to Δ_m (see Table 3).

4.4 A Conjecture

According (i) to our numerical evidence (the computation of approximations to $\theta(\chi)$ for the 10582203 primitive even Dirichlet characters χ of prime conductors $p \leq 20000$ and for numerous examples of cubic, quartic and quintic primitive even Dirichlet characters of (non necessarily prime) large conductors associated with simplest cubic, quartic and quintic fields), and (ii) to the fact that as $p \geq 5$ ranges over the odd primes it holds that $\sum_{\chi\in X_p^+} |\theta(\chi)|^2$ is asymptotic to $p^{3/2}/(4\sqrt{2})$ and that $\theta(\chi) \neq 0$ for at least $\gg p/\log p$ of the $(p-3)/2$ characters $1 \neq \chi \in X_p^+$ (adapt the proof of [Lou2, Theorem 1]), we put forward the following conjecture:

Conjecture 1. (i) (See **Hypothesis** (13)). For any primitive even Dirichlet character of conductor $f_\chi > 1$ it holds that $\theta(\chi) \neq 0$. (ii) Let $p \geq 5$ denote an odd prime and let X_p^+ denote the set of order $(p-3)/2$ of the primitive even Dirichlet characters modulo p. For $a \geq 0$ real, the limit

$$g_+(a) = \lim_{p\to\infty} \frac{2}{p-3} \#\{\chi \in X_p^+;\ |\theta(\chi)| \leq ap^{1/4}\}$$

exits, $a \mapsto g_+(a)$ is continuous, strictly increasing, $g_+(0) = 0$ and $g_+(\infty) = 1$.

Now, fix $t_0 < 1/4$. Then, at least for real cyclic fields K of a given prime degree q and of large prime conductors f_K, we might expect to have $|\theta(\chi)| \geq f_K^{t_0}$ for all the $1 \neq \chi \in X_K$. In that case, for $t > 1/2 - t_0$ we can use (14) with $N \geq B(t, f_K, M)$ to check numerically that $\theta(\chi) \neq 0$ for all the $1 \neq \chi \in X_K$. Then, for $t > (q+1)/2 - t_0$ we can use Lemma 3 with $N \geq B(t, f_K, M)$ to compute the exact value of $\omega(\chi)$ for all the $1 \neq \chi \in X_K$. Finally, for $t > 1/2 - t_0$ we can use Lemma 4 to compute the exact value of $W(\chi)$ for all the $1 \neq \chi \in X_K$. Hence, according to Corollary 1, we might expect that our second method for computing class numbers of real abelian number fields K of a given degree and known regulators requires only $O(f_K^{0.5+\epsilon})$ elementary operations. In practice, it is indeed amazingly efficient and of the conjectured complexity.

5 Explicit Formulae for $\omega(\chi)$

Finally, we explain how we can dispense with Subsection 4.1 when dealing with simplest cubic and quintic fields: we know beforehand $\omega(\chi_{K_m})$ and we can use Lemma 4 for computing the root number $W(\chi_{K_m})$, making in these two cases our method for computing class numbers simpler and faster.

5.1 Simplest Cubic Fields

Set $\omega = (-1+i\sqrt{3})/2$. The units in $\mathbf{Z}[\omega]$ are $\{\pm 1, \pm\omega, \pm\omega^2\}$. An algebraic integer $\alpha = a + b\omega \in \mathbf{Z}[\omega]$ is *primary* if $\alpha \equiv -1 \pmod{3\mathbf{Z}[\omega]}$, i.e. if $a \equiv -1 \pmod 3$ and $b \equiv 0 \pmod 3$. The order of the multiplicative group $(\mathbf{Z}[\omega]/3\mathbf{Z}[\omega])^*$ is equal to 6 and the six units in $\mathbf{Z}[\omega]$ form a set of representatives of this group. Therefore, if $\alpha \in \mathbf{Z}[\omega]$ and 3 does not divide its norm $N(\alpha) = \alpha\bar{\alpha}$, then exactly one of its six associates is primary. If follows that if $0 \neq \alpha \in \mathbf{Z}[\omega]$ is a nonunit element such that $\alpha \equiv (-1)^t \pmod{3\mathbf{Z}[\omega]}$, where t denotes the number of irreducible factors of α (counted with multiplicity), then α can be written in a unique way as a product of primary irreducibles. Now, let $\pi \in \mathbf{Z}[\omega]$ be a primary irreducible element of norm a rational prime $p \equiv 1 \pmod 3$. For $\alpha \in \mathbf{Z}[\omega]$ coprime with π, let $\chi_\pi(\alpha) \in \{1, \omega, \omega^2\}$ be the cubic residue symbol defined by $\alpha^{(p-1)/3} \equiv \chi_\pi(\alpha) \pmod \pi$. Then, $\tau(\chi_\pi)^3 = p\pi$ (see [IR, Corollary page 115]). It follows:

Theorem 4. *Assume that $\Delta_m = m^2 + 3m + 9$ is square-free, write $\Delta_m = \prod_{k=1}^t p_k$ where the p_k's are distinct odd primes and set*

$$\delta_m := (-1)^t \left(\frac{m}{3}\right) \frac{2m+3+3i\sqrt{3}}{2} \equiv (-1)^t \pmod{3\mathbf{Z}[\omega]}.$$

Then, δ_m can be written in a unique way as a product $\delta_m = \prod_{1 \leq k \leq t} \pi_k$ of t primary irreducibles elements $\pi_k \in \mathbf{Z}[\omega]$ with $p_k = |\pi_k|^2$. Set $\chi_{\delta_m} = \prod_{1 \leq k \leq t} \chi_{\pi_k}$. Then, χ_{δ_m} is a primitive cubic character modulo Δ_m, and

$$\omega(\chi_{\delta_m}) := \tau(\chi_{\delta_m})^3 = \Delta_m \delta_m. \tag{16}$$

Hence, setting $\epsilon_m = (1 - (-1)^t \left(\frac{m}{3}\right))/2 \in \{0,1\}$, there exists $k_m \in \{0,1,2\}$ such that

$$\arg(W(\chi_{\delta_m})) \equiv \frac{1}{3}\arctan\left(\frac{3\sqrt{3}}{2m+3}\right) + \frac{2k_m + \epsilon_m}{3}\pi \pmod{2\pi}, \tag{17}$$

and, if $\theta(\chi_{\delta_m}) \neq 0$, then k_m can be efficiently computed by using **Lemma 4**. *Moreover, χ_{δ_m} is associated with the simplest cubic field K_m, i.e. the character χ_{K_m} associated with K_m obtained by using the technique developed in* **Section 2** *is equal either to χ_{δ_m} or to its conjugate character $\bar{\chi}_{\delta_m}$.*

Since $P_m(x) = x^3 - mx^2 - (m+3)x - 1$ has no root in the finite field with two elements, we have $1 \neq \chi_{K_m}(2) \in \{\omega, \omega^2\}$ where χ_{K_m} is the cubic character associated with K_m obtained by using the technique developed in Section 2. According to the law of cubic reciprocity (see [IR, Theorem 1, page 114]), we have

$$\chi_{\delta_m}(2) = \chi_2(\delta_m) \equiv \delta_m \equiv \begin{cases} \omega^2 \pmod{2\mathbf{Z}[\omega]} & \text{if } m \equiv 0 \pmod 2 \\ \omega \pmod{2\mathbf{Z}[\omega]} & \text{if } m \equiv 1 \pmod 2, \end{cases}$$

hence

$$\chi_{\delta_m}(2) = \begin{cases} \omega^2 & \text{if } m \equiv 0 \pmod 2 \\ \omega & \text{if } m \equiv 1 \pmod 2. \end{cases}$$

Hence, by computing $\chi_{K_m}(2)$ and by changing χ_{K_m} into its conjugate if necessary, we may assume that $\chi_{K_m}(2) = \chi_{\delta_m}(2)$, which implies $\chi_{K_m} = \chi_{\delta_m}$.

5.2 Simplest Quintic Fields

In the same way, we have:

Theorem 5. *Assume that* $\Delta_m = m^4 + 5m^3 + 15m^2 + 25m + 25$ *is square-free. Since* $P_m(x) = x^5 + m^2x^4 - (2m^3 + 6m^2 + 10m + 10)x^3 + (m^4 + 5m^3 + 11m^2 + 15m + 5)x^2 + (m^3 + 4m^2 + 10m + 10)x + 1$ *has no root in the finite field with two elements, we may assume that the quintic character* χ_{K_m} *associated with the simplest quintic field* K_m *(obtained by using the technique developed in* **Section 2***) satisfies*

$$\chi_{K_m}(2) = \begin{cases} \zeta_5^3 & \text{if } m \equiv 0 \pmod 2 \\ \zeta_5^4 & \text{if } m \equiv 1 \pmod 2. \end{cases}$$

In that case, it holds that $\omega(\chi_{K_m}) := \tau(\chi_{K_m})^5 = (-1)^{t-1}(\frac{m}{5})\Delta_m\delta_m$ *where*

$$\begin{aligned}\delta_m = {} & (m^6 + 5m^5 + 5m^4 + 25m^2 + 125m + 125)\zeta_5 \\ & +(m^6 + 5m^5 - 5m^4 - 75m^3 - 175m^2 - 125m)\zeta_5^2 \\ & +(m^6 + 10m^5 + 25m^4 - 100m^2 - 125m)\zeta_5^3 \\ & +(m^6 + 10m^5 + 40m^4 + 75m^3 + 50m^2)\zeta_5^4 \in \mathbf{Z}[\zeta_5].\end{aligned}$$

6 Simplest Cubic Fields and Class Numbers of the Maximal Real Subfields of Some Cyclotomic Fields

Assume that $\Delta_m = m^2 + 3m + 9 \equiv 1 \pmod{12}$ is square-free. Let $L_m = K_m(\sqrt{\Delta_m})$ denote the compositum of the real quadratic field $k_m = \mathbf{Q}(\sqrt{\Delta_m})$ and of the simplest cubic field K_m, both of conductor Δ_m. Then, L_m is a so-called **simplest sextic field** of conductor Δ_m associated with the sextic polynomial

$$P_m(x) = x^6 - 2mx^5 - 5(m+3)x^4 - 20x^3 + 5mx^2 + 2(m+3)x + 1,$$

and a subgroup of finite index Q_{L_m} (dividing 12) of the group of algebraic units of L_m is known (see [Gra]). Using this subgroup, and following the proof of [Lou4, Theorem 4], we obtain:

Theorem 6. *(See* [Lou5]*). Assume that* $\Delta_m = m^2 + 3m + 9 \equiv 1 \pmod{12}$ *is square-free. Then,*

$$h_{L_m} \geq \frac{\Delta_m^2}{15e\log^6(4\Delta_m)}. \tag{18}$$

In particular, for $m \geq 10^5$ *it holds that* $h_{L_m} > \Delta_m$.

Notice that, in the special case that $\Delta_m = m^2 + 3m + 9 \equiv 1 \pmod{12}$ is prime, we have $h_{k_m} = h_2$, $h_{K_m} = h_3$ and $h_{L_m} = h_6$, with the notation of Theorem 1. With this notation, we have seen that the simplest cubic fields K_m for which $\Delta_m = m^2 + 3m + 9 \equiv 1 \pmod{12}$ is prime and such that $h_2h_3 \geq \Delta_m$ are few and far between (see Table 3). However, according to the previous lower bound for h_{L_m}, as soon as m is large enough we have $h_6 \geq \Delta_m$. Moreover, using this lower bound for h_{L_m} and following the proof of [CW, Theorem 2] (see [CW, Page 269]), we obtain:

Corollary 2. *(See* [CW, Theorem 2] *for a worse and non-effective result). Let* $\epsilon > 0$ *be given. Set* $c = \frac{1}{3}\prod_{p\equiv 1 \pmod 3}(1-2p^{-2}) = 0.311\cdots$. *For at least* $(c+o(1))x^{1/2}$ *positive odd square-free integers* $f \leq x$ *is holds that the class numbers* h_f^+ *of the maximal real subfields* $\mathbf{Q}(\zeta_f)^+$ *of the cyclotomic fields* $\mathbf{Q}(\zeta_f)$ *satisfy* $h_f^+ \gg_\epsilon f^{2-\epsilon}$, *and the constants involved in these* $o(1)$ *and* $\gg_\epsilon$ *are explicit.*

Proof. Let f range over the positive square-free integers of the form $f = \Delta_m := m^2+3m+6 \equiv 1 \pmod{12}$.

7 Tables

Table 1. class numbers h_{K_m} of the simplest quintic fields K_m of square-free conductors $\Delta_m < 10^7$

m	Δ_m	h_{K_m}	m	Δ_m	h_{K_m}	m	Δ_m	h_{K_m}
$-1, -2$	11	1	-21	154291	108691	-39	2038711	4521505
-3	31	1	19	170531	44605	37	2148911	27105755
1	71	1	-22	187751	76901	38	2382131	6728105
-4	101	1	21	247951	308605	-41	2505371	6340275
2	191	11	-24	270721	153005	39	2633851	7503505
3	451	5	22	295331	478775	-42	2766691	20599841
-6	631	11	23	349211	186091	-43	3047951	24153305
4	941	16	-26	378611	189305	41	3196631	8088176
-7	1271	55	24	410161	591775	-44	3350141	6495280
-8	2321	305	-27	443311	289025	42	3509671	61395955
6	3091	80	-28	515981	2372005	43	3845171	17264525
-9	3931	256	26	555671	721151	-46	4021391	21321025
7	5051	1451	-29	597251	540905	-47	4392551	12722855
8	7841	421	27	641491	1566401	-48	4788841	49860400
-11	9951	541	28	736901	1764400	46	4997051	42769375
9	11671	655	-31	788231	1217821	-49	5211371	56285605
-12	13981	1375	29	842591	760055	47	5433131	88151275
-13	19811	4705	-32	899321	798256	48	5897161	17478875
11	23411	2000	-33	1021771	4680055	-51	6139711	21966025
-14	27311	7255	31	1087691	1386275	49	6390311	74338555
12	31861	9680	-34	1156331	1402000	-53	7186931	155197205
13	42431	9455	32	1228601	4822625	51	7468771	28850896
-17	62891	9455	33	1382791	2148080	-54	7758151	37142851
-18	80251	37631	-36	1464901	4628591	52	8056541	118690480
16	90281	19301	34	1551071	2160455	53	8678351	44646025
-19	100991	203305	-37	1640531	1636721	-56	9002081	106111555
17	112871	83275	-38	1831511	11812625	54	9335491	54898055
18	139471	32605	36	1933261	3869525	-57	9677371	73297775

Table 2. the simplest quintic fields K_m of prime conductors $\Delta_m = p < 10^{10}$ for which some prime $q \geq p$ (in bold face letters) divides their class numbers h_{K_m}

m	Δ_m	h_{K_m}
27	641491	$1566401 = \mathbf{1566401}$
-61	12765251	$66431941 = \mathbf{66431941}$
66	20479231	$182277211 = 61 \cdot \mathbf{2988151}$
73	30425111	$335434451 = 11 \cdot \mathbf{30494041}$
77	37526591	$3233114891 = \mathbf{3233114891}$
-84	46927381	$2068985771 = \mathbf{2068985771}$
-88	56676161	$5912208301 = \mathbf{5912208301}$
-99	91352671	$3144379001 = \mathbf{3144379001}$
-102	103090711	$3626779141 = \mathbf{3626779141}$
-121	205717691	$11420513591 = \mathbf{11420513591}$
122	230839031	$60390377311 = 11 \cdot \mathbf{5490034301}$
128	279170201	$24178878281 = \mathbf{24178878281}$
129	287909191	$32215474121 = \mathbf{32215474121}$
139	387022451	$42590939281 = 11 \cdot \mathbf{3871903571}$
-147	451386751	$155312785456 = 2^4 \cdot \mathbf{9707049091}$
-163	684652511	$785372557471 = 41 \cdot \mathbf{19155428231}$
162	710402911	$421336924016 = 2^4 \cdot \mathbf{26333557751}$
178	1032554351	$320881058831 = \mathbf{320881058831}$
-187	1190654831	$259187494511 = 11 \cdot \mathbf{23562499501}$
-237	3089232931	$1634411025661 = \mathbf{1634411025661}$
238	3276804731	$3314877124271 = 71 \cdot \mathbf{46688410201}$
242	3501489071	$4793050096976 = 2^4 \cdot \mathbf{299565631061}$
-249	3767856571	$2253716261071 = 11 \cdot \mathbf{204883296461}$
-263	4694424311	$9653048507861 = 11 \cdot \mathbf{887549864351}$
-264	4766572561	$3419567237581 = 11^2 \cdot \mathbf{2860886261}$
268	5256015221	$4240933367591 = 151 \cdot \mathbf{28085651441}$
271	5494201451	$6532834598131 = \mathbf{6532834598131}$
282	6437395351	$18156246542621 = 11 \cdot \mathbf{1650567867511}$
291	7295360131	$5988407760191 = \mathbf{5988407760191}$
293	7497114671	$10748665628261 = 11^2 \cdot \mathbf{88831947341}$
-303	8291171431	$25938285252521 = 2311 \cdot \mathbf{11223836111}$
-312	9325450081	$15721799752591 = 41 \cdot \mathbf{383458530551}$

Table 3. least values of $m \geq -1$ for which Δ_m is prime and $h_2 h_3 \geq \Delta_m$

m	Δ_m	$\lvert\theta(\chi_{\delta_m})\rvert$	$\arg W(\chi_{\delta_m})$	h_2	h_3	$h_2 h_3/\Delta_m$
102496	10505737513	$20.268\cdots$	$\frac{1}{3}\arctan(\frac{3\sqrt{3}}{2m+3}) + \frac{\pi}{3}$	891	13152913	$1.115\cdots$
106253	11290018777	$34.364\cdots$	$\frac{1}{3}\arctan(\frac{3\sqrt{3}}{2m+3})$	2685	6209212	$1.476\cdots$
319760	102247416889	$202.162\cdots$	$\frac{1}{3}\arctan(\frac{3\sqrt{3}}{2m+3})$	1887	57772549	$1.066\cdots$
554869	307881271777	$88.861\cdots$	$\frac{1}{3}\arctan(\frac{3\sqrt{3}}{2m+3}) + \frac{\pi}{3}$	7983	93739324	$2.430\cdots$

References

Ba. E. Bach. Explicit bounds for primality testing and related problems. *Math. Comp.* **55** (1990), 355–380.

BE. B.C. Berndt and R.J. Evans. The determination of Gauss sums. *Bull. Amer. Math. Soc.* **5** (2) (1981), 107–129. Corrigendum im **7** (2) (1982), 411.

Bye. D. Byeon. Class number 3 problem for the simplest cubic fields. *Proc. Amer. Math. Soc.* **128** (2000), 1319–1323.

CW. G. Cornell and L. C. Washington. Class numbers of cyclotomic fields. *J. Number Theory* **21** (1985), 260–274.

Gra. M. N. Gras. Special units in real cyclic sextic fields. *Math. Comp.* **48** (1988), 543–556.

IR. K. Ireland and M. Rosen. *A classical introduction to modern number theory. Second edition.* Graduate Texts in Mathematics, **84**. Springer-Verlag, New York, 1990.

Jean. S. Jeannin. Nombre de classes et unités des corps de nombres cycliques quintiques d'E. Lehmer. J. Théor. Nombres Bordeaux **8** (1996), no. 1, 75–92.

Lan. S. Lang. *Algebraic Number Theory. Second edition.* Graduate Texts in Mathematics, **110**. Springer-Verlag, New York, 1994.

Lou1. S. Louboutin. Computation of relative class numbers of imaginary abelian number fields. *Experimental Math.* **7** (1998), 293–303.

Lou2. S. Louboutin. Sur le calcul numérique des constantes des équation fonctionnelles des fonctions L associées aux caractères impairs. *C. R. Acad. Sci. Paris* **329** (1999), 347–350.

Lou3. S. Louboutin. Computation of relative class numbers of CM-fields by using Hecke L-functions. *Math. Comp.* **69** (2000), 371–393.

Lou4. S. Louboutin. The exponent three class group problem for some real cyclic cubic number fields. *Proc. Amer. Math. Soc.* **130** (2002), 353–361.

Lou5. Efficient computation of root numbers, Gauss sums, and class numbers of real abelian number fields. *In preparation.*

LP. F. Lemmermeyer and A. Pethö. Simplest cubic fields. *Manuscripta Math.* **88** (1995), 53–58.

Mos. C. Moser. Nombre de classes d'une extension cyclique réelle de **Q** de degré 4 ou 6 et de conducteur premier. *Math. Nachr.* **102** (1981), 45-52.

Sha. D. Shanks. The simplest cubic fields. *Math. Comp.* **28** (1974), 1137–1152.

Sta. H. M. Stark. Dirichlet's class-number formula revisited. *Contemp. Math.* **143** (1993), 571–577.

SW. R. Schoof and L.C. Washington. Quintic polynomials and real cyclotomic fields with large class numbers. *Math. Comp.* **50** (1987), 179–182.

SWW. E. Seah, L.C. Washington and H.C. Williams. The calculation of a large cubic class number with an application to real cyclotomic fields. *Math. Comp.* **41** (1983), 303–305.

Wa. L. C. Washington. Class numbers of the simplest cubic fields. *Math. Comp.* **48** (1987), 371–384.

WB. H. C. Williams and J. Broere. A computational technique for evaluating $L(1,\chi)$ and the class number of a real quadratic field. *Math. Comp.* **30** (1976), 887-893.

An Accelerated Buchmann Algorithm for Regulator Computation in Real Quadratic Fields

Ulrich Vollmer[*]

Technische Universität Darmstadt, Fachbereich Informatik
Fachgebiet Kryptographie und Computeralgebra
Alexanderstr. 10, 64283 Darmstadt

Abstract. We present a probabilistic algorithm for computing the regulator R of a real quadratic order of discriminant Δ running in time $L(\frac{1}{2}, 3/\sqrt{8} + o(1))$.

1 Introduction

In his paper [Buc90], Buchmann proposed a generalization of Hafner and McCurley's subexponential algorithm for class group computation in imaginary quadratic fields [HM89] to the computation of class group and regulator of arbitrary number fields. While his algorithm depends on an as yet unproven "smoothness assumption for reduced ideals" for fields of degree exceeding two, it does extend unconditionally Hafner and McCurley's algorithm to real quadratic fields.

In this paper we present two modifications of Buchmann's algorithm for the real quadratic case. Their goal is to improve the asymptotics of the expected run time. Correctness, and running time bounds for both algorithms depend on a Generalized Riemann Hypothesis (GRH).

The expected run time needed by Buchmann's original algorithm in order to compute class group and regulator of a number field with discriminant Δ and fixed degree was bounded by $L_\Delta(\frac{1}{2}, 1.7)$ where

$$L_\Delta(a, b) = \exp(b(\log|\Delta|)^a(\log\log|\Delta|)^{1-a}).$$

Our first algorithm, RqClR, computes class group and regulator of a real quadratic order with discriminant Δ in time $L_\Delta(\frac{1}{2}, \sqrt{2})$. It confirms the correctness of its result by computing an approximation to the special value of the L-function of the field at 1.

The second proposed algorithm, RqR, computes only the regulator in time $L_\Delta(\frac{1}{2}, 3/\sqrt{8})$. It produces with probability given a priori the correct result. However, it does not verify the correctness of the result.

The results of this paper are collected in the following theorem.

Theorem 1. (GRH) *For any positive real number $p \leq 1$, and $\epsilon > 0$, there is some $\Delta_0 = \Delta_0(\epsilon)$, and a probabilistic algorithm that has the following property:*

[*] research supported by the DFG

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 148–162, 2002.

Given the positive discriminant $\Delta > \Delta_0$ of the quadratic order $\mathcal{O}$, the algorithm computes an integer R that differs from some positive multiple $m \cdot R_\Delta$ of the regulator R_Δ of $\mathcal{O}$ by less than one. Independent of Δ, the probability that $m = 1$ taken over all random input of the algorithm is at least p.

The expected run time of the algorithm is bound by $L_\Delta(\frac{1}{2}, c)$ where

a. $c = 3/\sqrt{8} + \epsilon$ *if* $p < 1$*;*
b. $c = \sqrt{2} + \epsilon$ *if* $p = 1$.

In case b, the algorithm also computes the class number, and the elementary divisors of the class group of $\mathcal{O}$.

2 Previous Work

The details of Buchmann's algorithm for the quadratic case were spelled out in Abel's thesis [Abe94]. Her algorithm is applicable to arbitrary quadratic orders, not only maximal ones. Abel was able to prove on the basis of some Generalized Riemann Hypothesis (GRH) that her variant of the algorithm runs in time bound by $L_\Delta(\frac{1}{2}, 5/6\sqrt{3} + o(1))$.

In [Vol00], the author indicated briefly that three sub-algorithms used by Abel can be substituted by faster ones. Mentioned were:

- Replacement of the factorization algorithm used in the process of generating relations. This suggestion was already made in [HM89].
- Computation of an approximation of the regulator from logarithms of units that form a generating set of the unit group with the help of an algorithm proposed by Maurer in his thesis [Mau00];
- Use of the fast algorithm for computation of the determinant of the relation lattice proposed in [Vol00] itself.

This paper takes up the suggestions of [Vol00], incorporating them into a complete algorithm.

Practical implementation. The focus of this paper is in presenting an algorithm whose complexity can be rigorously proved (assuming GRH), although some of the ideas might also lead to practical improvements.

For advice on the practical implementation of Buchmann's algorithm for quadratic fields, we refer the reader to [Coh93], and [Jac99]. [Coh93] gives a detailed description of the algorithm as implemented in the well-known PARI package. (Please refer to the fourth printing, and the author's web site for the corrected text of the relevant passage.)

[Jac99] shows how the Multiple Quadratic Polynomial Sieve can be employed for rapid generation of relations in the quadratic case. The resulting algorithm is implemented in the LiDIA package of Buchmann et al.

To the best of the author's knowledge there are no published rigorous, or heuristic analyses of the expected run times of the algorithms proposed in the cited works. It is, however, to be expected that they share the asymptotic behavior of RQR, or RQCLR, depending on the linear algebra algorithms employed.

3 Overview

Let $\mathcal{O}$ be a real quadratic order, and K its fractional field which we assume to be embedded into $\mathbb{R}$. For simplicity we will assume in this overview that $\mathcal{O}$ is maximal.

We denote the discriminant of $\mathcal{O}$ by $\varDelta$, the group of invertible $\mathcal{O}$-ideals by $\mathcal{I}_\varDelta$, its subgroup of principle ideals by $\mathcal{P}_\varDelta$, the class group $\mathcal{I}_\varDelta/\mathcal{P}_\varDelta$ of $\mathcal{O}$ by $Cl_\varDelta$, the class number by $h_\varDelta$, the regulator by $R_\varDelta$, and the non-trivial automorphism of K by σ.

We assume in the following that $R_\varDelta \gg \log \varDelta$, since otherwise there are deterministic algorithms that are more efficient than the probabilistic ones proposed here.

Buchmann's algorithm uses the fact that we can compute "small" representatives of each ideal class, called reduced ideals, in polynomial time. For background on reduced ideals, the properties of the reduction operator, and cycles of reduced ideals we refer the reader to [Len82]. Here, we will just give the definition.

Definition 1. *An integral ideal* $\mathfrak{a} \in \mathcal{I}_\varDelta$ *is called* primitive *if* $\mathfrak{a} \subseteq q\mathbb{Z}$ *implies* $q = 1$. *It is called* reduced *if it is primitive, and* $q = \min(\mathfrak{a} \cap \mathbb{N})$ *is a minimum of* $\mathfrak{a}$*, i.e.* $|\alpha|, |\alpha^\sigma| < q$ *imply* $\alpha = 0$ *for any* $\alpha \in \mathfrak{a}$.

This definition coincides with the classical one introduced by Gauss in the language of binary forms.

In [Buc90], Buchmann introduced, generalizing ideas by Seysen [Sey87], and Hafner/McCurley [HM89], lattices $L^{(m)} \subset \mathbb{Z}^m \oplus \mathbb{R}$ with determinant $h_\varDelta R_\varDelta$, and showed how to produce a generating set for $L^{(m)}$ for suitably chosen $m \gg 0$.

We recall the definition of $L^{(m)}$. Roughly spoken, it is the lattice of "relations" over a large set of prime ideals.

We define the relevant set of prime ideals of $\mathcal{O}$. For $b \in \mathbb{R}$, let $\mathcal{F}_b = \{\mathfrak{p} \in \mathcal{I}_\varDelta \mid \mathrm{N}\,\mathfrak{p} = p \text{ prime}, \gcd(\varDelta, p) = 1, p < b\}$. Set $c = L_\varDelta(\frac{1}{2}, z)$, where z is later chosen such that $\mathcal{F}_c$ is large enough for random reduced ideals to factor over $\mathcal{F}_c$ with sufficiently high probability. The cardinality of $\mathcal{F}_c$ will be denoted by m.

Let $\mathcal{I}_{\varDelta,c}$ denote the free subgroup of $\mathcal{I}_\varDelta$ generated by $\mathcal{F}_c$. Enumerate the elements of $\mathcal{F}_c$ such that $\mathcal{F}_c = \{\mathfrak{p}_i \mid 1 \le i \le m\}$ and $N\mathfrak{p}_i \le \mathrm{N}\,\mathfrak{p}_j$ whenever $i < j$, and use this enumeration to identify $\mathcal{I}_{\varDelta,c}$ with $\mathbb{Z}^m$ in the natural way. The algorithms presuppose that the restriction of the projection $\psi : \mathcal{I}_\varDelta \longrightarrow Cl_\varDelta$ to $\mathcal{I}_{\varDelta,c}$ is surjective. Due to a well known result of Bach, cf. [Bac90] this is certainly the case if $c > g = 12 \log^2 \varDelta$ which we will henceforth assume throughout. Denote $\mathcal{F}_g$ by $\mathcal{G}$, and $\operatorname{card} \mathcal{G}$ by l.

Let

$$\phi : K^* \longrightarrow \mathcal{P}_\varDelta : \alpha \longmapsto (\alpha),$$

$$\operatorname{Log} : K^* \longrightarrow \mathbb{R} : \alpha \longmapsto \frac{1}{2} \log \left| \frac{\alpha}{\alpha^\sigma} \right|,$$

and $\mathcal{O}_c = \phi^{-1}(\mathcal{I}_{\varDelta,c} \cap \mathcal{P}_\varDelta)$.

We define the lattice $L^{(m)}$ to be the image of $\mathcal{O}_c$ under (ϕ, Log). We will call its elements *relations*. The pre-image of a relation under (ϕ, Log) is called its generator. From the diagram

$$\begin{tikzcd}
1 \arrow[r] & \pm 1 \arrow[r] \arrow[d] & \mathcal{O}_c \arrow[r, "(\phi,\mathrm{Log})"] \arrow[d, equal] & \mathbb{Z}^m \oplus \mathbb{R} \arrow[d, "\pi"'] & & \\
1 \arrow[r] & \mathcal{O}^* \arrow[r] & \mathcal{O}_c \arrow[r, "\phi"] & \mathbb{Z}^m \arrow[r, "\psi"] & Cl_\Delta \arrow[r] & 1
\end{tikzcd}$$

we see that $\pi | L^{(m)}$ has kernel $(0, R_\Delta \mathbb{Z})$ and the sequence

$$0 \longrightarrow \mathbb{R}/R_\Delta\mathbb{Z} \longrightarrow (\mathbb{Z}^m \oplus \mathbb{R})/L^{(m)} \longrightarrow Cl_\Delta \longrightarrow 1$$

is exact.

For any $v = (\mathbf{v}, \mathrm{Log}\,\alpha) \in L^{(m)}$, we call $\mathbf{v} = \pi(v)$ its integral part. For any sub-lattice $M \subseteq L^{(m)}$, we will denote $\pi(M)$ also simply by M'.

Both RqR, and RqClR compute R_Δ by producing couples of elements of $L^{(m)}$ with the same image under π. To achieve this they proceed roughly in the following manner:

1. Construct the elements of the factor base $\mathcal{F}_c$.
2. Choose some $n \in \mathbb{N}$. For each j with $1 \leq j \leq n$ generate a random relation $v_j \in L^{(m)}$ and enter its coefficients into a matrix A. (Instead of the value of Log, we record its argument in compact representation.)
3. RqClR *only:* Compute the determinant $\tilde{h}$ of the column space of A.
4. Choose randomly two relations w_s with generators γ_s, $s = 1, 2$. Express each $\pi(w_s)$ as an integral linear combination of $\pi(v_j)$. Each found expression yields an element E_s of the kernel of π.
5. Compute the real GCD $\tilde{R}$ of E_1 and E_2 using e.g. algorithm `rgcd_cfrac` in [Mau00].
6. RqClR *only:* Calculate bounds for the product of class number and regulator using the L function of field K. If $\tilde{h}\tilde{R}$ does not lie within these bounds, start over.
7. Output $R' = \tilde{R}$, and, if we are in RqClR, also $h = \tilde{h}$.
8. RqClR *only:* Compute the Smith Normal Form of A, and extract the class group structure.

The algorithms differ in the relation generation in step 2. In RqR we choose n large and compute many relations with few non-zero entries, which we will call “sparse”. In RqClR we compute fewer relations, but the relations may have non-zero entries at each place. The reason for the different asymptotic behavior of RqR, and RqClR lies in this difference. *NB:* in practical implementations one chooses n only slightly larger than m, and generates only sparse relations. It is still unclear why this succeeds.

We outline the rest of the paper. In the sections 4 through 6 we will treat those aspects of the proposed algorithms that are specific to our approach. For the general framework, and results not listed here see [Sey87,HM89,Abe94,Mau00].

The generation of random reduced ideals in an ideal class is treated in section 4. Section 5 establishes how many random relations need to be generated in order to find two integral linear dependencies among them. Section 6 deals with the extraction of a generating set of units from these dependencies.

In section 7 we will give listings for RqR and RqClR, and conclude the proof of Theorem 1.

Throughout the following sections we will drop subscripting the symbols R_Δ, h_Δ, and $L_\Delta(1/2, z)$, the dependency upon Δ being understood. Δ is no longer assumed to be fundamental. We will let $\mathcal{I}$, and $\mathcal{P}$ denote the subgroups of $\mathcal{I}_\Delta$, and $\mathcal{P}_\Delta$ containing only ideals prime to the conductor f of Δ. Recall that we have unique factorization in $\mathcal{I}$, and $\mathcal{I}/\mathcal{P} \simeq \mathcal{I}_\Delta/\mathcal{P}_\Delta$. Obviously, $\mathcal{I}_{\Delta,c} \subset \mathcal{I}$, and $\phi(\mathcal{O}_c) \subset \mathcal{P}$.

Functions o_i with $i = 1, 2, \ldots$ will denote effectively computable auxiliary functions that depend on Δ only, and tend to 0 with growing Δ.

4 Random Relations

In [Buc89], Buchmann has given, and analyzed a method for the construction of a generating system for the lattice $L^{(m)}$ in the case of an arbitrary number field. This method relies in the real-quadratic case on the following proposition which can be proved in analogy to Proposition 4.4 of [Sey87] giving the same result for the imaginary quadratic case.

Proposition 1. (GRH) *The number N_c of reduced $\mathcal{O}$-ideals that factor completely over the the set $\mathcal{F}_c$ of ideals with prime norm smaller than $c = L(\frac{1}{2}, z)$ and co-prime with f is at least $hR \cdot L(\frac{1}{2}, -1/(4z))$.*

Buchmann proceeds by taking power products over $\mathcal{F}_c$ with exponents up to Δ, and choosing—by a method called PV—a random reduced ideal in the resulting class. For ease of reference, we will describe a simple variant of PV for the real-quadratic case which we will call RandomReduced that enjoys—with minor modifications—the same properties as the more general algorithm.

Another, slightly more elaborate variant of PV for the quadratic case was given by Abel in her thesis [Abe94].

Let $\mathfrak{a} \in \mathcal{I}$ be some invertible $\mathcal{O}$-ideal. For any $d \in \mathbb{N}$ we define the set

$$S_d = S_d(\mathfrak{a}) = \{(\mathfrak{b}, \alpha) \mid \mathfrak{b} \text{ is reduced}, \mathfrak{b} = \alpha\mathfrak{a}, d \leq \operatorname{Log}\alpha/\log\Delta < (d+1)\}.$$

Let $D > R$ be given. RandomReduced proceeds as follows.

1. Choose some random $d \in [1, D)$.
2. Enumerate all elements in S_d.
3. Choose randomly among them.

The following lemmata are needed to show that RandomReduced has the desired properties. For any $\mathfrak{a} \in \mathcal{I}$ and $D > 0$ we denote by T_D the range of values of RandomReduced

$$T_D = T_D(\mathfrak{a}) = \bigcup_{d=1}^{D} S_d(\mathfrak{a}).$$

Lemma 1. *Fix* $\mathfrak{a} \in \mathcal{I}$*. Let* $d \geq 1$*.*
Then $2 \leq \operatorname{card} S_d(\mathfrak{a}) \leq 2\log_2 \Delta$*, and* $2D \leq \operatorname{card} T_D(\mathfrak{a}) \leq 2D\log_2 \Delta$*.*

This is a trivial consequence of the properties of the reduction operator ρ proved in [Len82].

Lemma 2. *Given* $\mathfrak{a}, \mathfrak{b} \in \mathcal{I}$*, where* $\mathfrak{a} \sim \mathfrak{b}$*, and* $\mathfrak{b}$ *is reduced. Then* $\operatorname{card}\{d \mid 1 \leq d < D, \exists\alpha \textit{ such that} (\mathfrak{b}, \alpha) \in S_d(\mathfrak{a})\} = D\log\Delta/R + \delta$ *with* $-2 \leq \delta \leq 1$*.*

Proof. Let $\mathfrak{b} = \alpha\mathfrak{a}$, where α is chosen such that $0 \leq \operatorname{Log}\alpha < R$. Then $(\mathfrak{b}, \alpha') \in S_d(\mathfrak{a})$ for some α' if and only if $d\log\Delta \leq \operatorname{Log}\alpha + tR < (d+1)\log\Delta$ for some $t \in \mathbb{Z}$. Since we assumed that $R \gg \log\Delta$ the claim follows.

Let $D \leq \Delta$, $\mathfrak{a} \in \mathcal{I}$ and $d \in [1, D)$. We show that it is possible to enumerate all elements in S_d in polynomial time. For this to be possible, the field elements need to be given in compact representation. The following lemma follows immediately from results in [BTW95].

Lemma 3. *Given* $\alpha \in K$ *in compact representation,* $\mathfrak{a} \in \mathcal{I}$*, and* $f \in \mathbb{N}$*, it is possible to compute the compact representation of* α^f*, and* $\alpha\mathfrak{a}$ *in time polynomial in the size of* $\mathfrak{a}$*,* α*, and* $\log n$*.*

Thus we may proceed as follows.

1. compute some α_0 with $\operatorname{Log}\alpha_0 \in [\frac{1}{2}\log\Delta, \log\Delta)$;
2. compute $l_0 = \operatorname{Log}\alpha_0$ with precision $\log_2 \Delta$;
3. set $f = \lfloor d \cdot (\log\Delta/l_0) \rfloor$, and compute $\alpha_1 = \alpha^f$;
4. compute β_0 such that $\mathfrak{b} = \beta_0\alpha_1\mathfrak{a} = \rho_0(\alpha_1\mathfrak{a})$ is reduced;
5. compute $\operatorname{Log}(\beta_0\alpha_1)$, and—through successive reduction—all β_i such that $\beta_i\mathfrak{b}$ is reduced, and $\operatorname{Log}\beta_i + \operatorname{Log}(\beta_0\alpha_1) \in [d\log\Delta, (d+1)\log\Delta)$.

Note that we can assure that all reduced ideals in S_d get enumerated, but due to the imprecise computation of logarithms in this enumeration process, the enumeration may inadvertently contain ideals with relative generators from a slightly larger interval. Since at most 2 ideals are thus erroneously listed, this will not affect the probability estimates that follow, and is, hence, ignored.

Note further that the ideal $\mathfrak{a}$ might already be given as the product of a principle ideal (with generator γ in compact representation) with a reduced ideal $\mathfrak{c}$. In this case we start from $\mathfrak{c}$, and adjust f in step 3 accordingly.

We summarize the properties of RandomReduced in the following proposition.

Proposition 2. *Let* $\mathfrak{a}$ *be a given invertible* $\mathcal{O}$*-ideal.* RandomReduced *computes randomly in polynomial time some* $\mathcal{O}$*-ideal* $\mathfrak{b}$*, and* $\alpha \in K$ *in compact representation such that* $\mathfrak{b} = \alpha \cdot \mathfrak{a}$ *is reduced.*

For any reduced $\mathfrak{b}$ *equivalent to* $\mathfrak{a}$ *the probability that* RandomReduced *outputs* $\mathfrak{b}$ *on input* $\mathfrak{a}$ *is contained in the interval* $(\log(2)/R - 1/D, \log\Delta/(2R) +$

$1/D$). Moreover, the probability that the second component α of the output of RandomReduced *fulfills $\operatorname{Log}\alpha \in [tR, (t+1)R)$ conditional on the event that the first component is some fixed $\mathfrak{b}$ is bounded from below by $1/N$ with $N = \lceil D \log \Delta / R \rceil$ if $t < (D \log \Delta - R)/R$.*

Proof. All but the last claim follow in a straightforward manner from the preceding lemmata. We turn to the latter.

Fix some reduced $\mathfrak{b}$ in the ideal class of $\mathfrak{a}$. Let α be a generator of $\mathfrak{b}$ relative to $\mathfrak{a}$ with $0 \leq \operatorname{Log}\alpha < R$. Let further $B = \{d \mid 1 \leq d < D, \exists t \,\text{such that}\, d \leq (\operatorname{Log}\alpha_0 + tR)/\log\Delta < d+1\}$. Then the sought conditional probability is certainly bounded from below by $1/N$ where $N = \operatorname{card} B$.

Now, $1 \leq d < D$, and $d \leq (\operatorname{Log}\alpha + tR)/\log\Delta < d+1$ imply $0 \leq t < D\log\Delta/R$. The claim follows.

We are now in the position to show how to generate random relations. The procedure will be called RandomRelation.

Fix some $\mathcal{H}$ with $\mathcal{G} \subseteq \mathcal{H} \subseteq \mathcal{F}_c$ that parameterizes RandomRelation in the sense that it determines whether we generate "sparse" ($\mathcal{H} = \mathcal{G}$), or "dense" ($\mathcal{H} = \mathcal{F}_c$) relations. Sparse relations have $O(\log^2 \Delta)$ non-zero entries in their integral parts. For dense relations there is no such restriction. An $n \times m$ relation matrix is sparse in the usual sense if all contained relations are sparse since $\log^2 \Delta = m^{o(1)}$.

Let $\mathfrak{q} \in \mathcal{I}$ be some ideal which will later be chosen to be some power of an element of $\mathcal{F}_c$ that "offsets" the relation at one place.

```
1. For each $\mathfrak{p} \in \mathcal{H}$ choose $e_\mathfrak{p}$ with $0 \leq e_\mathfrak{p} < \Delta$. Set $e_\mathfrak{p} = 0$ for $\mathfrak{p} \in \mathcal{F} \setminus \mathcal{H}$.
2. Compute $\mathfrak{a} = \mathfrak{q} \cdot \prod_{\mathfrak{p}\in\mathcal{H}} \mathfrak{p}^{e_\mathfrak{p}}$.
3. Compute $(\mathfrak{b}, \alpha) = \text{RandomReduced}(\mathfrak{a})$ with $D = \Delta$.
4. if $\mathfrak{b} \notin \mathcal{I}_{\Delta,c}$ then return Failure.
5. Compute $b_\mathfrak{p}$ such that $\mathfrak{b} = \prod_{\mathfrak{p}\in\mathcal{F}_c} \mathfrak{p}^{f_\mathfrak{p}}$
6. return $((e_\mathfrak{p} - f_\mathfrak{p})_{\mathfrak{p}\in\mathcal{F}_c}, \alpha)$.
```

In step 2, each computation of an ideal product is followed by reduction. Hence, the ideal $\mathfrak{a}$ computed in step 2 is computed and stored as the product of some $\alpha_0 \in K^*$ (in compact representation) and a reduced ideal.

For steps 4, and 5 we factor the norm of $\mathfrak{b}$ with the elliptic curve method, cf. Algorithm 7.2 of [LP92].

Lemma 4. *For any class $C \in Cl_\Delta$, the probability that $\mathfrak{a}$ computed in step 2 belongs to C is contained in an interval $((1-o_1)/h, (1+o_1)/h)$ with $o_1(\Delta) = o(1)$.*

Proof. This lemma follows from lemma 4.5 of [Sey87].

Lemma 5. *For the probability p that a given reduced ideal is computed in step 3 we have $hR \cdot p \in (\log(2) - o_2, \log\Delta + o_2)$ for some $o_2 = o(1)$ provided $R = o(D)$.*

Proof. This follows from Proposition 2, and Lemma 4.

Corollary 1. *The probability that the ideal* $\mathfrak{b}$ *computed in step 3 lies in* $\mathcal{I}_{\Delta,c}$ *is bounded from below by* $(\log(2) - o_3)L(\frac{1}{2}, -1/(4z))$.

Proof. Consequence of Proposition 1, and 2, and Lemma 5.

The repeated call to the procedure above with identical parameters until it returns successfully yielding some relation $(\mathbf{v}, \alpha)$ will be called RandomRelation.

5 Relation Lattices

In this section we study sub-lattices of $L^{(m)}$, subsequently simply denoted by L, and of $L' = \pi(L^{(m)})$. Our goal is to estimate the number of relations which need to be generated to achieve one of the following two goals:

1. the lattice generated by the integer parts of the obtained relations equals L';
2. the likelihood that the integer part of a randomly chosen relation is contained in the lattice generated by the integer parts of the other relations exceeds some a priori given bound.

Both algorithms, RqR and RqClR, start out by generating m relations whose integral parts form a square diagonally dominant matrix, as originally proposed by Seysen.

```
1. for $i = 1$ to $m$
2.     $(\mathbf{v}_i, \alpha_i) \leftarrow$ RandomRelation$(\mathcal{G}, \mathfrak{p}_i^{2m\Delta})$
```

Let A_0 denote the matrix containing the integral parts $\mathbf{v}_i$ of the relations v_i generated this way, and L_0 the lattice generated by $\{v_i\}$. Then $\log_2 \det A_0 = \log_2[L' : \pi(L_0)] < (1 + o_4)m \log_2 \Delta$ where $o_4 = o(1)$ can be explicitly given.

Lemma 6. *Let* $(w_i), i = 1, \ldots, k$ *be a sequence of relations* $w_i \in L$. *Let further for any* $j = 1, \ldots, k$ *the sub-lattice* $L_j \subseteq L$ *be generated by* L_0, *and all* w_i *with* $i \leq j$. *Then we have* $\pi(w_{j+1}) \in \pi(L_j)$ *for at least* $n - (1 + o_4)m \log_2 \Delta$ *values of* j.

Proof. This follows from the fact that any chain of sub-lattices $M_i \subset L$ with $L_0 \subset M_1 \subset \cdots \subset M_e \subset L$ has length e smaller than $(1 + o_4)m \log_2 \Delta$.

Thus we only need to produce $k = (1+o_4)m \log_2 \Delta/(1-p)$ additional relations w_i with RandomRelation in order to ensure that with probability p a relation randomly chosen from among them is contained in the lattice generated by the rest.

Lemma 7. *Given some* $\mathbf{v} = (a_i) \in L'$ *with* $0 \leq a_i \leq \Delta - \log \Delta$, *the probability that a call to* RandomRelation$(\mathcal{F}_c, (1))$ *yields a* v *with* $\pi(v) = \mathbf{v}$ *is at least* $(1 - o_5)h/(2\Delta^m \log_2 \Delta)$.

Proof. Let $\mathfrak{c}$ correspond to $\mathbf{v}$, and let $\mathfrak{b}$ run through the set of all c-smooth reduced ideals. RandomRelation arrives at some v with $\pi(v) = \mathbf{v}$ if it chooses in step 2 the ideal $\mathfrak{c} \cdot \mathfrak{b}$, which has exponents smaller than Δ at each place by assumption, and $\mathfrak{b}$ in step 3. For the second choice there are N_c possibilities differing in probability by a factor of $2 \log_2 \Delta$. Each such choice can follow the selection of $(1 + o_1)\Delta^m/h$ different power products in steps 1 and 2, every one of which occurs with the same probability. The claim follows.

Next we prove an estimate for the number of lattice points of L that are not in some sub-lattice. Define $B(d) = \{(a_i) \in \mathbb{Z}^m \mid 0 \leq a_i \leq d\}$.

Lemma 8. *Let M' be some proper sub-lattice of L'. If $D \gg h$, then $L' \setminus M'$ contains at least $(D - 2h)^m/(2h)$ elements in $B(D)$.*

Proof. We know that there is a basis of L' with positive coefficients smaller than or equal to h. Let $\mathbf{w}$ be an element of that basis that is not in M'. Then we can assign to each $\mathbf{v} \in M' \cap B(D - h)$ the lattice point $\mathbf{v} + \mathbf{w} \in B(D)$ which is obviously in $L' \setminus M'$.

Now $L' \cap B(D - h)$ contains at least $(D - 2h)^m/h$ elements. Thus we have either $\operatorname{card}((L' \setminus M') \cap B(D-h)) \geq (D-2h)^m/(2h)$, in which case we are done, or $\operatorname{card}(M' \cap B(D-h)) \geq (D-2h)^m/(2h)$. Using the assignment from the previous paragraph we find again the desired number of elements in $(L' \setminus M') \cap B(D)$.

Setting $D = \Delta - h$ in the preceding lemma, and applying lemma 7 we obtain an estimate for the probability that a call to RandomRelation enlarges the relation lattice.

Proposition 3. *Let M' be some proper full rank sub-latticce of L'. Then the probability that a call to* RandomRelation$(\mathcal{F}_c, (1))$ *results in a vector $w = (\mathbf{w}, \alpha)$ with $\mathbf{w} \in L' \setminus M'$ is bounded from below by $(1 - o_6)/(4 \log_2 \Delta)$.*

If $L'_n = L'$ then we call the corresponding $m \times n$ matrix A a *full* relation matrix. The last proposition yields finally the desired conclusion about the number of relations we need to compute in order to arrive at a full relation matrix.

Corollary 2. *There is an effectively computable function $o_7 = o(1)$ such that for $k = L(\frac{1}{2}, z + o_7)$ the probability that $L'_k = L'$ is bounded from below by 1/2.*

6 Extracting a Generating Set of Units

In this section we assume that we are given the following data:

- Some $m \times n$ relation matrix $A = (a_{\mathfrak{p},j})$ with vector of generators α_j. We have $\prod_{\mathfrak{p} \in \mathcal{F}_c} \mathfrak{p}^{a_{\mathfrak{p},j}} = (\alpha_j)$.
- Two sparse relations $(\mathbf{w}_s, \gamma_s)$, $s = 1, 2$, obtained through a call to RandomRelation$(\mathcal{G}, (1))$.
- Two vectors $\mathbf{x}_s = (x_{j,s})$ with $A\mathbf{x}_s = \mathbf{w}_s$.

We have seen in Lemma 6 how to find a $\mathbf{w}$ which lies in the column space of a sparse relation matrix. If, on the other hand, we choose to compute dense relations, then Corollary 2 assures us that we can quickly compute a full relation matrix. Two more calls to RandomRelation yield the desired dependent relations.

The vectors $\mathbf{x}_s$ are computed with the algorithm DiophantineSolver proposed in [MS99]. This algorithm finds a solution to the Diophantine system $A\mathbf{x} = \mathbf{w}$ with size restricted by

$$\log||\mathbf{x}|| = O(m \log(m||A||) + \log||\mathbf{w}||) \tag{1}$$

On the basis of the above data, we can assign a unique unit to each relation vector: $\epsilon_s = \gamma_s / \prod \alpha_j^{x_{j,s}}$ is a unit of $\mathcal{O}$, since γ_s, and $\prod \alpha_j^{x_{j,s}}$ generate the same ideal. We denote ϵ_s by $\textsc{Unit}(\gamma_s, A, \mathbf{x}_s)$.

We will show that for two independently, and randomly chosen sparse relations with generators γ_s, $s = 1, 2$ the units $\pm\textsc{Unit}(\gamma_s, A, \mathbf{x}_s)$ generate the full unit group with probability $(1 - o(1))/2$.

Let $\operatorname{Log} \textsc{Unit}(\gamma_s, A, \mathbf{x}_s) = t_s R$. Then $\langle \pm\textsc{Unit}(\gamma_s, A, \mathbf{x}_s)\rangle = \mathcal{O}^*$ is equivalent to $\gcd(t_1, t_2) = 1$. We will first give size limits for the t_i, and then estimate the probability that the two t_s are co-prime.

Lemma 9. *If* $\operatorname{Log} \textsc{Unit}(\gamma_s, A, \mathbf{x}_s) = t_s R$, *then* $\log t_s < (1 + o(1)) m \log \Delta$.

Proof. This is a consequence of (1) and $\operatorname{Log} \alpha_j < \Delta \log \Delta$ which holds by construction.

Lemma 10. *Let* $U, V, D \in \mathbb{Z}$ *with* $0 < \log|U - V| < D/100$. *Consider the set* $S = \{(x, y) \in \mathbb{Z}^2 \mid U \le x < U + D, V \le y < V + D\}$. *If* $0 \ll D$ *then there are more than* $D^2/2$ *pairs* $(x, y) \in S$ *with* $\gcd(x, y) = 1$.

Proof. We define the following subsets of S:

$$\begin{aligned} T &= \{(x, y) \in S \mid \gcd(x, y) \ne 1\}, \\ T_p &= \{(x, y) \in S \mid p | \gcd(x, y)\} \end{aligned}$$

where p denotes some prime number. We need to show that $\operatorname{card} T < D^2/2$. We will show instead that

$$\sum_{p \le D} \operatorname{card} T_p + \operatorname{card} \bigcup_{p > D} T_p < D^2/2$$

which is certainly sufficient. Note that for any two $p, q > D$ the sets T_p and T_q are disjoint.

Let $p \le D$. Then a simple counting argument shows that $\operatorname{card} T_p < (1 + \lfloor D/p \rfloor)^2$. Thus

$$\begin{aligned} \sum_{p \le D} \operatorname{card} T_p &< \sum_{p \le D} (1 + D/p)^2 \\ &< D(\log\log D + O(1)) + D^2 P(2), \end{aligned}$$

where P is the prime zeta function, and $P(2) = 0.452....$

Let $p > D$. Then $\operatorname{card} T_p \leq 1$. For any $d \in \mathbb{Z}$ we define yet another set $U_d = \{(x, y) \in S \mid x-y = d\}$. If $T_p \cap U_d \neq \emptyset$ then $p|d$. Thus since $|d| < |U-V|+D$

$$\operatorname{card}(U_d \cap \bigcup_{p>D} T_p) < \log(|U-V| + D).$$

From this we deduce

$$\begin{aligned}\operatorname{card} \bigcup_{p>D} T_p &= \sum_{d=U-V-D}^{U-V+D} \operatorname{card}(U_d \cap \bigcup_{p>D} T_p) \\ &< 2D(\log(|U-V| + D)) < D^2/50 + D \log D.\end{aligned}$$

Adding the two estimates we obtain the desired result for sufficiently large D.

Corollary 3. *Let $A = (a_{\mathfrak{p},j})$ be an $m \times n$ relation matrix with vector of generators α_j, so that we have $\prod_{\mathfrak{p} \in \mathcal{F}_c} \mathfrak{p}^{a_{\mathfrak{p},j}} = (\alpha_j)$. Let $(\mathbf{w}_s, \gamma_s)$ for $s = 1, 2$ be the output of two independent calls to* RandomRelation$(\mathcal{G}, (1))$ *for which there exist $\mathbf{x}_s$ such that $\mathbf{w}_s = A\mathbf{x}_s$. Let $\mathbf{x} = \mathbf{x}(A, \mathbf{w})$ be some random variable taking values in the solution space of the Diophantine equation $A\mathbf{x} = \mathbf{w}$. Let* $\operatorname{Log} \textsc{Unit}(\gamma_s, A, \mathbf{x}(A, \mathbf{w}_s)) = t_s R$. *Then the probability that* $\gcd(t_1, t_2) = 1$, *taken over all random input of* RandomRelation *and* $\mathbf{x}$, *exceeds* $(1 - o_8)/2$.

Proof. Keep the notation from the corollary. For $s = 1, 2$, we fix two exponent vectors $\mathbf{e}_s = (e_{\mathfrak{p},s})$, and two c-smooth reduced ideals $\mathfrak{b}_s = \prod \mathfrak{p}^{f_{\mathfrak{p},s}}$ in the ideal classes represented by the power products $\mathfrak{a}_s = \prod \mathfrak{p}^{e_{\mathfrak{p},s}}$. Let $\mathbf{f}_s = (f_{\mathfrak{p},s})$, and $\mathbf{w}_s = \mathbf{e}_s - \mathbf{f}_s$.

Fix further $\mathbf{x}_s = (x_{j,s})$ with $\mathbf{w}_s = A\mathbf{x}_s$ which we assume to exist. Then $\textsc{Unit}(\gamma_s, A, \mathbf{x}_s) = \gamma_s/\beta_s$ where $\beta_s = \prod \alpha_j^{x_{j,s}}$ and γ_s is a generator of $\mathfrak{a}_s/\mathfrak{b}_s$.

It suffices to show that the probability that $\gcd(t_1, t_2) = 1$ conditional on the event that 1) during the calls to RandomRelation those exponent vectors and ideals were chosen, and 2) $\mathbf{x}$ took value $\mathbf{x}_s$ exceeds $1/2$.

If $(\mathbf{w}, \gamma)$ is one of the possible values of RandomRelation under the set condition then any other can be written as $(\mathbf{w}, \gamma')$ with $\gamma' = \gamma\epsilon^u$ where ϵ is the fundamental unit of $\mathcal{O}$, and u varies in an interval of width $\Delta \log \Delta / R$. Thus $t_s = U_s + u_s$ with fixed U_s, and $0 \leq u_s < \Delta$.

Lemma 9 implies $U_s < m \log \Delta(1 + o(1))$. Since $\log m \ll \Delta$ we can apply Lemma 10. We conclude that half the pairs (t_1, t_2) yield $\gcd(t_1, t_2) = \gcd(U_1 + u_1, U_2 + u_2) = 1$.

Now, Proposition 2 gives a lower bound for the conditional probability that a particular u_s is chosen. This bound implies the claim.

7 Conclusion

In this section we give listings of RqR and RqClR, and conclude the proof of Theorem 1. We will refer to the steps of RqR, and RqClR using the numbering in the listings. Algorithm DetEss used in RqClR was introduced in [Vol00].

Algorithm 1. Probabilistic invariant computation

Input: Discriminant Δ

Output: Class number h_Δ,
elementary divisors d_i of Cl_Δ, regulator approximation R'

```
RqClR($\Delta$)
1.  Validation interval Find $a, b$ such that $a < h_\Delta R_\Delta < b < 2a$ through
    approximation of $\sqrt{\Delta} L(1, \chi_\Delta)$.
2.  Parameters Let $z \leftarrow 1/\sqrt{8}$, $c \leftarrow L_\Delta(\frac{1}{2}, z)$, and $k \leftarrow L_\Delta(\frac{1}{2}, z + o_2(\Delta))$.
3.  Factor base Compute and store all prime ideals in $\mathcal{F}_c$. Let $m \leftarrow \operatorname{card} \mathcal{F}_c$.
4.  Generating set Let $g \leftarrow 6 \log^2 \Delta$, $\mathcal{G} = \mathcal{F}_g$, and $l \leftarrow \operatorname{card} \mathcal{G}$.
5.  Full rank relation lattice for $i = 1$ to $m$
6.      $(\mathbf{v}_i, \alpha_i) \leftarrow$ RandomRelation($\mathcal{G}, \mathfrak{p}_i^{2m\Delta}$)
7.  Full relation lattice for $j = 1$ to $k$
8.      $(\mathbf{v}_{m+j}, \alpha_{m+j}) \leftarrow$ RandomRelation($\mathcal{F}_c, (1)$)
9.  $A \leftarrow (\mathbf{v}_j)_{j=1}^{k+m}$.
10. Class number Compute $\tilde{h} \leftarrow$ DetEss($A$).
11. HNF Compute with Hafner and McCurley's algorithm $H \leftarrow \mathrm{HNF}(A, \tilde{h})$.
12. Units Call RandomRelation($\mathcal{G}, (1)$) twice. Let $(\mathbf{w}_s, \gamma_s)$ be the resulting
    relations.
13.     $\mathbf{x}_s \leftarrow$ DiophantineSolver($A, \mathbf{w}_s$).
14.     Compute $\epsilon_s =$ Unit($\gamma_s, A, \mathbf{x}_s$).
15.     Compute the real GCD $\tilde{R}$ of $(\operatorname{Log} \epsilon_1, \operatorname{Log} \epsilon_2)$
        using algorithm rgcd_cfrac in [Mau00].
16. Verification if $\tilde{h}\tilde{R} \notin (a, b)$ then return Failure
17. $h \leftarrow \tilde{h}, R' \leftarrow \tilde{R}$.
18. Class group Compute the Smith Normal Form of $H$ which yields the
    elementary divisors $d_i$ of $Cl_\Delta$.
19. return $(h, R', (d_i)_{i=1}^{l})$.
```

We analyze the probability with which RqClR produces correct output. Corollary 2 assures that steps 5 through 9 produce a matrix A whose column space equals L' with probability exceeding 1/2. We obtain an approximation to R_Δ in steps 12 through 15 with probability exceeding 1/4 according to Corollary 3.

Next, we assure ourselves that RqClR never returns incorrect results. $\tilde{h}$ computed in step 10 is always a multiple of the class number even when the previous steps yielded an A which is not a full relation matrix.

Likewise, ϵ_1, ϵ_2 computed in step 14 are always units since they are quotients of two generators of the same ideal. So $\tilde{R} = \gcd(\operatorname{Log} \epsilon_1, \operatorname{Log} \epsilon_2)$ computed approximately in step 15 is close to a multiple of R_Δ. Thus, step 16 assures that $\tilde{h} = h_\Delta$, and $\tilde{R} \approx R_\Delta$, and the precision is ensured by Maurer's algorithm.

The same argument implies that $\tilde{R}$ obtained by RqR in each round is an approximation to a multiple of the regulator.

Algorithm 2. Probabilistic regulator computation

Description: Monte-Carlo algorithm for the computation of the regulator of a real-quadratic field

Input: Discriminant Δ, error probability p

Output: regulator approximation R' with $|R' - R_\Delta| < 1$

```
RqR($\Delta$)
*Parameters* Let $z \leftarrow 1/\sqrt{8}$, $c \leftarrow L_\Delta(\frac{1}{2}, z)$, and $k \leftarrow 2L_\Delta(\frac{1}{2}, z)\log_2 \Delta(1 + o_1(\Delta))$.
*Factor base* Compute and store all prime ideals in $\mathcal{F}_c$. Let $m \leftarrow \operatorname{card} \mathcal{F}_c$.
*Generating set* Let $g \leftarrow 6\log^2 \Delta$, $\mathcal{G} = \mathcal{F}_g$, and $l \leftarrow \operatorname{card} \mathcal{G}$.
*Full rank relation lattice* **for** $i = 1$ **to** $m$
    $(\mathbf{v}_i, \alpha_i) \leftarrow$ RandomRelation($\mathcal{G}, \mathfrak{p}_i^{2m\Delta}$)
*Relation sequence* **for** $j = 1$ **to** $k$
    $(\mathbf{v}_{m+j}, \alpha_{m+j}) \leftarrow$ RandomRelation($\mathcal{G}, (1)$)
Set $r \leftarrow 0$ and **repeat**
    Set $r \leftarrow r + 1$. Choose randomly $m < j_1, j_2 \leq k + m$.
    Let $\mathbf{w}_s \leftarrow \mathbf{v}_{j_s}$ for $s = 1, 2$
    Let $A = (\mathbf{v}_j \mid j \neq j_1, j_2)$.
    $\mathbf{x}_s \leftarrow$ DiophantineSolver($A, \mathbf{w}_s$) for $s = 1, 2$
    Compute $\epsilon_s =$ Unit($\gamma_s, A, \mathbf{x}_s$) for $s = 1, 2$.
    Compute the real GCD $\tilde{R}$ of ($\operatorname{Log} \epsilon_1, \operatorname{Log} \epsilon_2$)
    using algorithm rgcd_cfrac in [Mau00].
    $R' \leftarrow \min(R', \tilde{R})$.
**until** $(3/4)^{r-1} < p$
**return** $R'$.
```

By Lemma 6, and Corollary 3 this multiple is the regulator itself with probability exceeding 1/4. Hence, after the execution of $O(\log(1/(1-p)))$ rounds, the minimum of all $\tilde{R}$ computed will be an approximation to R_Δ with probability p.

Finally, we verify the time, and space complexity bound of Theorem 1. Due to Lemma 6, we need to call RandomRelation in RqR $m + 2m\log_2 \Delta(1 + o_4) = L_\Delta(\frac{1}{2}, z + o_9)$ times. Each call takes estimated time bounded by $L_\Delta(\frac{1}{2}, 1/(4z) + o_{10})$. In RqClR we need $L_\Delta(\frac{1}{2}, z + o_5)$ relations, but this time each call to RandomRelation costs time $L_\Delta(\frac{1}{2}, z + 1/(4z))$ due to the longer time needed to compute the random power product. Note that the estimated time needed for the factorizations in RandomRelation can be subsumed into the $o(1)$ term, cf. [LP92].

The solution of the two Diophantine systems to obtain the two integral linear dependencies takes time $L_\Delta(\frac{1}{2}, 3z + o(1))$. The remaining steps needed to arrive at the regulator multiple take only time $L_\Delta(\frac{1}{2}, 2z + o(1))$ due to Lemma 9, and Theorem 12.1.5 of [Mau00].

The optimum run time of both algorithms will be achieved with $z = 1/\sqrt{8}$ which yields the run time bounds of Theorem 1, and concludes the proof of the theorem.

8 Corrigendum

The run-time analysis of the algorithms given in [Vol00] ignored the cost involved in the generation of what we call in this article "dense" relations. The algorithms presented in the cited paper run in time bounded by $L_\Delta(1/2, \sqrt{2})$ instead of $L_\Delta(1/2, 3/\sqrt{8})$ as was claimed. We will present in a forthcoming paper a modification that reinstates the run-time bound given in [Vol00]. Moreover, it will also allow for the computation of the class number of a real-quadratic order within the smaller time bound thus improving upon RqClR given here.

The modification rests on the following strengthening of proposition 3. (We keep the notation used throughout this paper.)

Proposition 4. *Let $M \subset \pi(L^{(m)}) = L'$ be a sub-lattice that does not contain some vector $v \in L'$ with the following properties*

1. $0 \le v_i \le h_\Delta$ *for all* $0 \le i \le l$*;*
2. $v_i = 0$ *for all* $i > l$*.*

Then RandomRelation$(\mathcal{G}, (1))$ *produces an element* $w = (\mathbf{w}, \gamma) \in L$ *such that* $\mathbf{w} \in L' \setminus M$ *with probability bounded from below by a positive inverse linear function in* $\log \Delta$*.*

Thus we get again an effectively computable function $o_{11} = o(1)$ such that for $n = L_\Delta(1/2, z + o_{11})$ the probability that the lattice $M = L'_n$ contains all elements in $\pi(L^{(m)})$ with the properties specified in the proposition is bounded from below by 1/2. Now, the methods of [Vol00] allow us to extract the class number h_Δ, and the primary invariants of Cl_Δ from M even though it is not a full relation lattice.

Likewise we can produce relation lattices that contain with probability given a priori a sought DL relation.

Acknowledgments

The author thanks the German Science Foundation and Prof. Johannes Buchmann for their support, and the referee for many helpful and detailed comments.

References

Abe94. Christine Abel. *Ein Algorithmus zur Berechnung der Klassenzahl und des Regulators reellquadratischer Ordnungen.* PhD thesis, Universität des Saarlandes, Saarbrücken, Germany, 1994. German.

Bac90. Eric Bach. Explicit bounds for primality testing and related problems. *Math. Comp.*, 55(191):355–380, 1990.

BTW95. Johannes Buchmann, Christoph Thiel, and Hugh C. Williams. Short representation of quadratic integers. In Wieb Bosma and Alf J. van der Poorten, editors, *Computational Algebra and Number Theory, Sydney 1992*, volume 325 of *Mathematics and its Applications*, pages 159–185. Kluwer Academic Publishers, 1995.

Buc90. Johannes Buchmann. A subexponential algorithm for the determination of class groups and regulators of algebraic number fields. In Catherine Goldstein, editor, *Séminaire de Théorie des Nombres, Paris 1988–1989*, volume 91 of *Progress in Mathematics*, pages 27–41. Birkhäuser, 1990.

Buc89. J. Buchmann. A subexponential algorithm for the determination of class groups and regulators of algebraic number fields. In *Séminaire de Théorie des Nombres*, pages 27–41, Paris, 1988-89.

Coh93. H. Cohen. *A course in computational algebraic number theory*. Springer, Heidelberg, 1993.

HM89. James L. Hafner and Kevin S. McCurley. A rigorous subexponential algorithm for computation of class groups. *J. Am. Math. Soc.*, 2(4):837–850, 1989.

Jac99. Michael J. Jacobson, Jr. Applying sieving to the computation of quadratic class groups. *Mathematics of Computation*, 68(226):859–867, 1999.

Len82. Hendrik W. Lenstra, Jr. On the calculation of regulators and class numbers of quadratic fields. In J. V. Armitage, editor, *Journees Arithmetiques, Exeter 1980*, volume 56 of *London Mathematical Society Lecture Notes Series*, pages 123–150. Cambridge University Press, 1982.

LP92. H.W. Lenstra Jr. and C. Pomerance. A rigorous time bound for factoring integers. *J. Amer. Math. Soc.*, 5:483–516, 1992.

Mau00. Markus Maurer. *Regulator approximation and fundamental unit computation for real quadratic orders*. PhD thesis, Technische Universität Darmstadt, Fachbereich Informatik, Darmstadt, Germany, 2000.

MS99. Thom Mulders and Arne Storjohann. Diophantine linear system solving. ACM Press, 1999.

Sey87. Martin Seysen. A probablistic factorization algorithm with quadratic forms of negative discriminant. *Mathematics of Computation*, 48:757–780, 1987.

Vol00. Ulrich Vollmer. Asymptotically fast discrete logarithms in quadratic number fields. In Wieb Bosma, editor, *Algorithmic Number Theory Symposium IV*, volume 1838 of *Lecture Notes in Computer Science*, pages 581–594. Springer-Verlag, 2000.

Some Genus 3 Curves with Many Points[⋆]

Roland Auer[1] and Jaap Top[2]

[1] Department of Mathematics and Statistics, University of Saskatchewan,
106 Wiggins Road, Saskatoon, S7N 5E6, Canada
auer@snoopy.usask.ca
[2] IWI, Rijksuniversiteit Groningen,
Postbus 800, NL-9700 AV Groningen, The Netherlands
top@math.rug.nl

Abstract. We explain a naive approach towards the problem of finding genus 3 curves C over any given finite field $\mathbb{F}_q$ of odd characteristic, with a number of rational points close to the Hasse-Weil-Serre upper bound $q+1+3[2\sqrt{q}]$. The method turns out to be successful at least in characteristic 3.

1 Introduction

1.1 Curves of Genus ≤ 3 over Finite Fields

The maximal number of rational points that a (smooth, geometrically irreducible) curve of genus g over a finite field $\mathbb{F}_q$ can have, is denoted by $N_q(g)$. One has the estimate (see [Se1])

$$N_q(g) \leq q+1+g[2\sqrt{q}]$$

in which the notation $[r]$ for $r \in \mathbb{R}$ means the largest integer $\leq r$. The upper bound here is called the Hasse-Weil-Serre bound.

For $g=1$, it is a classical result of Deuring [De], [Wa] that $N_q(1) = q+1+[2\sqrt{q}]$, except when $q=p^n$ with p prime and $n \geq 3$ odd and p divides $[2\sqrt{q}]$, in which case $N_q(1) = q+[2\sqrt{q}]$. For $g=2$ an explicit formula is due to J-P. Serre. He stated and proved the result during a course [Se3] he gave at Harvard university in 1985; a nice survey including some modifications of the original proof can be found in Chapter 5 of the thesis [Sh]. The final result is that if q is a square and $q \neq 4,9$ then $N_q(2) = q+1+2[2\sqrt{q}]$. Moreover $N_9(2) = 20 = 9+1+2[2\sqrt{9}]-2$ and $N_4(2) = 10 = 4+1+2[2\sqrt{4}]-3$. In case q is not a square, then also $N_q(2) = q+1+2[2\sqrt{q}]$ except when either $\gcd(q,[2\sqrt{q}]) > 1$ or q can be written in one of the forms n^2+1, n^2+n+1 or n^2+n+2. In these remaining cases, one has that if $2\sqrt{q}-[2\sqrt{q}] \geq \frac{\sqrt{5}-1}{2}$ then $N_q(2) = q+2[2\sqrt{q}]$ and if $2\sqrt{q}-[2\sqrt{q}] < \frac{\sqrt{5}-1}{2}$ then $N_q(2) = q+2[2\sqrt{q}]-1$.

[⋆] It is a pleasure to thank Hendrik Lenstra for his interest in this work, and for his remarks which led to Section 2 of this paper.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 163–171, 2002.

For $g \geq 3$ no such result is known. The best known lower bounds in case $g \leq 50$ and q a power of 2 or 3 which is ≤ 128 can be found in [G-V]. In [Se2, § 4] J-P. Serre gives values of $N_q(3)$ for $q \leq 19$ and for $q = 25$. Moreover he shows in [Se3, p. 64-69] that $N_{23} = 48$. Hence we have the following table.

q	2	3	4	5	7	8	9	11	13	16	17	19	23	25	27	29	31	32	37	41
$N_q(3)$	7	10	14	16	20	24	28	28	32	38	40	44	48	56	56	60	≥ 56	64	≥ 68	≥ 72

The entries for $q = 29, 31, 37$ are obtained using the technique from the current paper; its main goal is to give lower bounds for $N_q(3)$ by restricting ourselves to one specific family of curves of genus 3.

1.2 Plane Quartics with 24 Automorphisms

Let k be a field of characteristic different from 2. The plane quartic C_λ given by $x^4+y^4+z^4 = (\lambda+1)(x^2y^2+y^2z^2+z^2x^2)$ is for $\lambda \in k$ with $\lambda \neq -3, 1, 0$ a geometrically irreducible, smooth curve of genus 3. The degree 4 polynomials given here are fixed by the subgroup $G < \mathrm{PGL}(3, k)$ generated by $\sigma : (x, y, z) \mapsto (y, z, x)$ and $\tau : (x, y, z) \mapsto (y, -x, z)$. The group G is isomorphic to S_4, the symmetric group on 4 elements. Hence G is contained in the group of automorphisms of C_λ. For general λ the automorphism group of C_λ in fact equals G.

These curves occur in the classification of non-hyperelliptic genus 3 curves with nontrivial automorphism group, as given in [He, p. 2.88] and in [Ve, Table 5.6, pp. 63-64].

Suppose $\lambda \neq 0, 1$. By E_λ we denote the elliptic curve given by the equation $y^2 = x(x-1)(x-\lambda)$. If moreover $\lambda \neq -3$ then we write $E_\lambda^{(\lambda+3)}$ for the elliptic curve with equation $(\lambda+3)y^2 = x(x-1)(x-\lambda)$. The relation with the curves C_λ is as follows.

Lemma 1.1. *Suppose k is a field of characteristic different from 2 and $\lambda \in k \setminus \{0, 1, -3\}$. Then the jacobian of the curve C_λ given by $x^4 + y^4 + z^4 = (\lambda + 1)(x^2y^2+y^2z^2+z^2x^2)$ is over k isogenous to the product $E_\lambda^{(\lambda+3)} \times E_\lambda^{(\lambda+3)} \times E_\lambda^{(\lambda+3)}$, where $E_\lambda^{(\lambda+3)}$ denotes the elliptic curve with equation $(\lambda+3)y^2 = x(x-1)(x-\lambda)$.*

Proof. Most of this is shown in [To, pp. 40-41]; one takes the quotient of C_λ by the involution $(x, y, z) \mapsto (-x, y, z)$. The resulting curve has genus 1 and it admits an involution without any fixed points. Taking the quotient again results in an elliptic curve, given by $y^2 = x^3 + 2(\lambda+1)(\lambda+3)x^2 + (\lambda-1)(\lambda+3)x$. The 2-isogeny with kernel generated by $(0, 0)$ maps this curve onto $E_\lambda^{(\lambda+3)}$ (compare the formulas for 2-isogenies as given in [Si-T, III § 4]). Write $\pi : C_\lambda \to E_\lambda^{(\lambda+3)}$ for the composition of all maps described here. Then

$$\rho = (\pi, \pi\sigma, \pi\sigma^2) : \ C_\lambda \to E_\lambda^{(\lambda+3)} \times E_\lambda^{(\lambda+3)} \times E_\lambda^{(\lambda+3)}$$

where $\sigma : (x, y, z) \mapsto (y, z, x)$ is one of the automorphisms of C_λ. The fact that ρ induces an isomorphism between the spaces of regular 1-forms implies that ρ induces an isogeny between $\mathrm{Jac}(C_\lambda)$ and the triple product of $E_\lambda^{(\lambda+3)}$. □

Corollary 1.2. *With notations as above, one finds for $\lambda \in \mathbb{F}_q$ with q odd and $\lambda \neq 0, 1, -3$ that*

$$\#C_\lambda(\mathbb{F}_q) = 3\#E_\lambda^{(\lambda+3)}(\mathbb{F}_q) - 2q - 2.$$

Proof. It is a well known fact that $\#C_\lambda(\mathbb{F}_q)$ equals $q+1-t$, where t is the trace of Frobenius acting on a Tate module of $\mathrm{Jac}(C_\lambda)$. Lemma 1.1 implies that this Tate module is isomorphic to a direct sum of three copies of the Tate module of $E_\lambda^{(\lambda+3)}$. Hence $t = 3t'$ where t' is the trace of Frobenius on the Tate module of $E_\lambda^{(\lambda+3)}$. Since this trace equals $q + 1 - \#E_\lambda^{(\lambda+3)}(\mathbb{F}_q)$, the result follows. □

1.3 Results

Our strategy for finding a curve of genus 3 over a finite field $\mathbb{F}_q$ with odd characteristic should now be clear: find λ such that $\#E_\lambda^{(\lambda+3)}(\mathbb{F}_q)$ is as large as possible and use Corollary 1.2. This works quite well for small q, using a direct search. In fact, as will be explained in Section 4 below, it is not even necessary here to calculate $\#E_\lambda^{(\lambda+3)}(\mathbb{F}_q)$ for many values $\lambda \in \mathbb{F}_q$.

We obtain a general result when the characteristic of $\mathbb{F}_q$ equals 3, because in that case we deal with a curve $E_\lambda^{(\lambda)}$ which is isomorphic to the curve E_μ with $\mu = 1/\lambda$. Since it is precisely known which values $\#E_\mu(\mathbb{F}_q)$ attains (see [A-T] and also Section 2 below), one obtains a nice explicit lower bound for $N_{3^n}(3)$. In fact, the result implies that the difference between $N_{3^n}(3)$ and the Hasse-Weil-Serre bound is bounded independently of n:

Proposition 1.3. *For every $n \geq 1$ the inequality*

$$3^n + 1 + 3[2\sqrt{3^n}] - N_{3^n}(3) \leq \begin{cases} 0 & \text{if } n \equiv 2 \bmod 4; \\ 12 & \text{if } n \equiv 0 \bmod 4; \\ 21 & \text{if } n \equiv 1 \bmod 2 \end{cases}$$

holds.

For the proof we refer to Section 3. Note that this proves a special case of a conjecture of J-P. Serre [Se3, p. 71], which says that for *all* q the difference $q + 1 + 3[2\sqrt{q}] - N_q(3)$ should be bounded independently of q.

In characteristic at least 5 we have not been able to obtain a general result such as given in Proposition 1.3. However, the fact that a curve $E_\lambda^{(\lambda+3)}$ is either isomorphic to E_λ or it is a quadratic twist of E_λ, implies (again using [A-T]) that for every finite field $\mathbb{F}_q$ of odd characteristic, a curve C_λ as above exists for which $\#C(\mathbb{F}_q)$ is at most 21 off from either the Hasse-Weil-Serre upper bound $q+1+3[2\sqrt{q}]$, or from the analogous lower bound $q+1-3[2\sqrt{q}]$. This is proven in Section 4. We note that a sharper result of the same kind (with 21 replaced by 3) was obtained by Kristin Lauter [Lau], [Lau-Se] using an entirely different method.

As Everett Howe pointed out to us, it is in fact possible to improve our result slightly by replacing the product $E_\lambda^{(\lambda+3)} \times E_\lambda^{(\lambda+3)} \times E_\lambda^{(\lambda+3)}$ we use, by a product

$E \times E \times E_\lambda$ in which $E/\mathbb{F}_q$ is an elliptic curve with a rational point of order 2 and $\#E(\mathbb{F}_q)$ maximal under that condition, and $E_\lambda/\mathbb{F}_q$ is a Legendre elliptic curve over $\mathbb{F}_q$ with as many rational points as possible. The result of Everett Howe, Franck Leprévost and Bjorn Poonen [H-L-P, Prop. 15] in this case implies that either this product or its standard quadratic twist is isogenous over $\mathbb{F}_q$ to the jacobian of a smooth genus 3 curve over $\mathbb{F}_q$. It may be noted that the estimate obtained in this way is in general still weaker than Lauter's result (it replaces our 21 by 9 instead of by 3).

2 A Characterization of Legendre Elliptic Curves

Suppose K is a field of characteristic $\neq 2$, and E/K is an elliptic curve. We will say that E/K is a Legendre elliptic curve over K if there is a $\lambda \neq 0, 1$ in K such that E is over K isomorphic to E_λ given by $y^2 = x(x-1)(x-\lambda)$. A necessary but in general not sufficient condition for an elliptic curve E/K to be a Legendre elliptic curve over K is that all points of order 2 on E are K-rational. An intrinsic description of Legendre elliptic curves is given as follows. Take a separable closure K^{sep} of K and write $G_K = \mathrm{Gal}(K^{\mathrm{sep}}/K)$ for its Galois group.

Lemma 2.1. *The statements*

1. *E is a Legendre elliptic curve over K;*
2. *E can be given by an equation $y^2 = (x-a)(x-b)(x-c)$ in which at least one of $\pm(a-b), \pm(b-c), \pm(c-a)$ is a square in K^*;*
3. *E has all its points of order 2 rational over K, and there exists a point $P \in E(K^{sep})[4]$ such that $-P$ is not in the G_K-orbit of P.*

are equivalent.

Proof. The equivalence of (1) and (2) is easy. To verify that (2) and (3) are equivalent, suppose (after possibly permuting a, b, c) that $a-b$ is a square and that E is given by $y^2 = (x-a)(x-b)(x-c)$. The point $T_b = (b, 0)$ in $E(K)$ has order 2, and the quotient $E' := E/\langle T_b \rangle$ admits an isogeny of degree 2: $\varphi : E' \to E$ defined over K (the dual isogeny of the quotient map). A very well known property (compare [Si-T, III § 5]) of φ is that the image $\varphi(E'(K)) \subset E(K)$ equals the kernel of the homomorphism $E(K) \to K^*/K^{*2}$ defined by $T_b \mapsto (b-a)(b-c)$ and $(x,y) \mapsto x-b$ for all $(x,y) \in E(K)$ with $(x,y) \neq T_b$. Hence the condition that $a-b$ be a square is equivalent with the property that the point $T_a := (a, 0) \in E(K)$ is in the image of $E'(K)$. This means precisely that a pair of points $\{P, P+T_b\} \subset E$ exists which is G_K-stable, and $2P = T_a$. Hence P is a point of order 4 on E, and for all $\sigma \in G_K$ we have $\sigma(P) - P \in \{O, T_b\}$. In particular $\sigma(P) - P \neq 2P$, which means $\sigma(P) \neq -P$ for all $\sigma \in G_K$.

Vice versa, suppose given a point P of order 4 with the property $\sigma(P) \neq -P$ for all $\sigma \in G_K$. Since all 2-torsion of E is K-rational, we have that $\sigma(P) - P \in E(K)[2]$ and moreover the condition $\sigma(P) \neq -P$ implies that $\sigma(P) - P$ is in a cyclic subgroup of $E(K)[2]$ which is independent of σ. Hence we have points T and $2P$ of order 2, where $\{P, P+T\}$ is G_K-stable. As we have seen, this implies the statements (1) and (2). □

Corollary 2.2. *Suppose q is a power of an odd prime and $E/\mathbb{F}_q$ is an elliptic curve. Let $\pi \in \mathrm{End}(E)$ be the Frobenius endomorphism (raising coordinates to the power q). Then E is a Legendre elliptic curve over $\mathbb{F}_q$ if and only if $\pi + 1 \in 2\,\mathrm{End}(E)$ but $\pi + 1 \notin 4\,\mathrm{End}(E)$.*

Proof. The Galois group $G_{\mathbb{F}_q}$ is topologically generated by the q-th power map, and this generator acts on E via the endomorphism π. The condition $\pi + 1 \in 2\,\mathrm{End}(E)$ is equivalent with the statement that E has all its points of order 2 rational over $\mathbb{F}_q$. In the same manner, the condition $\pi + 1 \notin 4\,\mathrm{End}(E)$ precisely means that a point P of order 4 exists, with the property $\pi(P) \neq -P$. Since the Galois group $G_{\mathbb{F}_q}$ acts on $E(\overline{\mathbb{F}_q})[4]$ via a (cyclic) subgroup of the kernel of $\mathrm{GL}_2(\mathbb{Z}/4\mathbb{Z}) \xrightarrow{\bmod 2} \mathrm{GL}_2(\mathbb{Z}/2\mathbb{Z})$, it follows that $\sigma(P) \neq -P$ for all $\sigma \in G_{\mathbb{F}_q}$. Hence Lemma 2.1 implies that E is a Legendre elliptic curve over $\mathbb{F}_q$.

Vice versa, if E is a Legendre elliptic curve over $\mathbb{F}_q$, then by Lemma 2.1 we know that $P \in E(\overline{\mathbb{F}_q})[4]$ exists with $\pi(P) \neq -P$, which implies that $\pi + 1$ is not divisible by 4 in $\mathrm{End}(E)$. We have that $\pi + 1 \in 2\,\mathrm{End}(E)$ since π acts trivially on all points of order 2.

This proves the corollary. □

Proposition 2.3. *An elliptic curve $E/\mathbb{F}_q$ (with q odd) for which $\#E(\mathbb{F}_q) \in 4\mathbb{Z}$ is isogenous to a Legendre elliptic curve over $\mathbb{F}_q$, except in the following case: $q = r^2$ with $r \in 1 + 4\mathbb{Z}$, and $\#E(\mathbb{F}_q) = q + 1 + 2r$.*

Proof. (This result was first presented in [A-T], however, with a somewhat different proof. The present proof is more conceptual, but it gives less information concerning the possible values of Legendre parameters λ in the supersingular case.)

Let $\pi \in \mathrm{End}(E)$ be the (q-th power) Frobenius. The proof considers two cases.

First, suppose $\pi = r \in \mathbb{Z}$. Then $q = \deg(\pi) = r^2$ and $\#E(\mathbb{F}_q) = (r-1)^2$. Any curve E' isogenous to E then also satisfies $\#E'(\mathbb{F}_q) = (r-1)^2$ and Frobenius in $\mathrm{End}(E')$ is equal to r. By Corollary 2.2, one (and equivalently, all of them) such curve E' is a Legendre elliptic curve over $\mathbb{F}_q$ precisely when $r + 1$ is even, but not divisible by 4. The latter condition is equivalent with $r \equiv 1 \bmod 4$. This proves the statement in the case $\pi \in \mathbb{Z}$.

If $\pi \notin \mathbb{Z}$ then $\mathbb{Z}[\pi] \subset \mathrm{End}(E)$ is an order in the ring of integers of an imaginary quadratic field K. We have that $\#E(\mathbb{F}_q) = (1-\pi)(\overline{1-\pi})$ where the bar denotes complex conjugation in K. The condition $\#E(\mathbb{F}_q) \equiv 0 \bmod 4$ implies that $(1-\pi)/2$ is integral. Now consider the order $A := \mathbb{Z}[(1+\pi)/2]$. By construction, $\pi \in A$ satisfies $\pi + 1 \in 2A$ and $\pi + 1 \notin 4A$. It is a result of Waterhouse [Wa, Thm. 4.5] (compare [Sch, p. 194] where a mistake in the original result is corrected), that a curve $E'/\mathbb{F}_q$ exists with an isomorphism $\mathrm{End}(E') \cong A$ such that under this isomorphism Frobenius on E' corresponds to $\pi \in A$. This implies in particular that $\#E'(\mathbb{F}_q) = \#E(\mathbb{F}_q)$ and hence E' and E are isogenous. Moreover, using Corollary 2.2 we know that E' is a Legendre elliptic curve over $\mathbb{F}_q$. This proves the proposition. □

3 Characteristic 3

We will now prove Proposition 1.3. Take $n \geq 1$ and write $q := 3^n$, $m := [2\sqrt{q}]$ and $q + 1 + m = N + r$ with $N \in 4\mathbb{Z}$ and $0 \leq r \leq 3$. As explained in (1.3), we will examine how close to the upper bound $q + 1 + m$ the number of $\mathbb{F}_q$-points on a Legendre elliptic curve $E_{1/\lambda} \cong E_\lambda^{(\lambda)}$ can be, for $\lambda \in \mathbb{F}_q$.

If n is odd and moreover $N \equiv 1 \bmod 3$ (the smallest n where this is the case, is $n = 11$ which gives $m = 841$ and $N = 3^{11} + 1 + 840$), then we replace N by $N - 4$. The resulting number N satisfies $q + 1 - m \leq N \leq q + 1 + m$, and moreover we know from [De] that $E/\mathbb{F}_q$ exists with $\#E(\mathbb{F}_q) = N$. If n is odd, then Proposition 2.3 implies the existence of $\lambda \in \mathbb{F}_q$ with $\#E_\lambda^{(\lambda)}(\mathbb{F}_q) = N$. Hence Corollary 1.2 yields a genus 3 curve C_λ with $\#C_\lambda(\mathbb{F}_q) = 3N - 2q - 2$. In particular, this shows that

$$q + 1 + 3m - N_q(3) \leq q + 1 + 3m - 3N + 2q + 2 = 3r + 12 \leq 21$$

for odd n (in fact, even $\leq 3r \leq 9$ unless m is divisible by 3).

If n is even, then $m = 2 \cdot 3^{n/2}$ and (again using Deuring's results [De]) an elliptic curve $E/\mathbb{F}_q$ exists with $\#E(\mathbb{F}_q) = q + 1 + m$. By Proposition 2.3, this number of points occurs for a Legendre elliptic curve only in case $m/2 \equiv 3 \bmod 4$, i.e., when $n \equiv 2 \bmod 4$. Hence under this condition we obtain a curve C_λ whose number of points attains the Hasse-Weil-Serre bound.

In the remaining case we have $n \equiv 0 \bmod 4$. Here the number $q + 1 + m$ does not occur as $\#E_\lambda^(\lambda)(\mathbb{F}_q)$, for any $\lambda \in \mathbb{F}_q$. Hence we take the largest smaller possibility, which is $q+1+m-4$. Proposition 2.3 implies that a Legendre elliptic curve with this number of points over $\mathbb{F}_q$ indeed occurs. It follows that a genus 3 curve $C_\lambda/\mathbb{F}_q$ exists with $\#C_\lambda(\mathbb{F}_q) = 3(q+1+m-4) - 2q - 2 = q + 1 + 3m - 12$. This implies the inequality given in Proposition 1.3. □

4 Examples in Characteristic > 3

The problem which arises when one attempts to adapt the argument presented in Section 3 to finite fields of characteristic > 3, can already be seen in the following result.

Proposition 4.1. *Suppose q is a power of a prime $p > 3$, and $m := [2\sqrt{q}]$. Over $\mathbb{F}_q$, a curve C_λ of genus 3 exists such that either $\#C_\lambda(\mathbb{F}_q) \geq q + 1 + 3m - 21$ or $\#C_\lambda(\mathbb{F}_q) \leq q + 1 - 3m + 21$.*

As we mentioned in the introduction, a somewhat stronger result has been obtained by Kristin Lauter [Lau], [Lau-Se] using quite different techniques. Moreover a variant of our proof may be obtained by using a result of Everett Howe, Franck Leprévost and Bjorn Poonen [H-L-P, Prop. 15].

Proof. Write $q + 1 + m = N + r$ with $N \in 4\mathbb{Z}$ and $0 \leq r \leq 3$. Then one of $N, 2q + 2 - N$ occurs as the number of points on a Legendre elliptic curve

$E_\lambda/\mathbb{F}_q$, except possibly when $r > 0$ and p divides $m - r$. In that case, we replace N by $N' := N - 4$ and we obtain a number of points which does occur.

This gives us an elliptic curve E_λ. The corresponding curve $E_\lambda^{(\lambda+3)}$ has either N or N' points, or in case $\lambda + 3$ is not a square in $\mathbb{F}_q$ this number is $2q + 2 - N$ or $2q + 2 - N'$. Since this number is at distance at most 7 from one of $q + 1 \pm m$, Corollary 1.2 implies the result. □

Proposition 4.2. *Suppose $p \equiv 3 \bmod 4$ is a prime number, $n \geq 1$ is an odd integer and $q = p^{2n}$. Then $N_q(3) = q + 1 + 6p^n$ equals the Hasse-Weil-Serre bound.*

Proof. Take $\lambda = -1 \in \mathbb{F}_p$. Since $p \equiv 3 \bmod 4$, the elliptic curve $E_\lambda/\mathbb{F}_p$ is supersingular. This implies $\#E_\lambda(\mathbb{F}_p) = p + 1$ (in case $p = 3$, this follows from the fact that the number of points is a multiple of 4, and also of course from a direct calculation). One concludes that $\#E_\lambda(\mathbb{F}_q) = q + 1 + 2p^n$. Since $\lambda + 3 \neq 0$ as an element of $\mathbb{F}_p$ is a square in $\mathbb{F}_q$, the two curves E_λ and $E_\lambda^{(\lambda)}$ are isomorphic over $\mathbb{F}_q$. Corollary 1.2 therefore yields that the genus 3 curve C_λ attains the Hasse-Weil-Serre bound over $\mathbb{F}_q$. □

Note that the genus 3 curve used in the above proposition is in fact the famous Fermat quartic. Hence the result is probably well known.

4.1 Legendre Curves with Prescribed Order

In practice, a fairly efficient method to find $\lambda \in \mathbb{F}_q$ for which $\#E_\lambda^{(\lambda+3)}(\mathbb{F}_q)$ equals a given number $N \equiv 0 \bmod 4$ can be given in case $q = p$ a prime or $q = p^2$ the square of a prime. This works as follows. Write $N = q + 1 - t$.

We first treat the case $q = p^2$ and $t = \pm 2p$. Exactly one of the two numbers $p^2 + 1 \pm 2p$ occurs as a number of points of a Legendre elliptic curve over $\mathbb{F}_{p^2}$, and this number is attained in our family precisely for the supersingular $\lambda \neq -3$ such that $\lambda + 3$ is a square in $\mathbb{F}_{p^2}$; the number with the opposite choice of sign occurs for the ones such that $\lambda + 3$ is a nonsquare.

In the remaining cases, the Hasse inequality tells us $|t| < 2\sqrt{q}$. Hence we find t exactly if we know $t \bmod 4p$. Now $t \bmod 4$ is already known, hence it suffices to find a λ such that $\#E_\lambda^{(\lambda+3)}(\mathbb{F}_q) \equiv q + 1 - t \bmod p$. If we write $\chi : \mathbb{F}_q^* \to \pm 1$ for the nontrivial character with kernel $\mathbb{F}_q^{*2}$, this means we look for $\lambda \neq -3$ such that $\#E_\lambda(\mathbb{F}_q) \equiv 1 - \chi(\lambda + 3)t$. It is well known [Si, V § 4] that $\#E_\lambda(\mathbb{F}_q) \equiv 1 - \left((-1)^{(p-1)/2} H_p(\lambda)\right)^e$, with $e = 1$ if $q = p$ and $e = 1 + p$ if $q = p^2$. Here $H_p(\lambda) = \sum_{i=0}^{(p-1)/2} \binom{(p-1)/2}{i}^2 \lambda^i$ is the so-called Hasse polynomial, whose coefficients can be computed using an easy recursion. Hence in case $q = p$ we have to solve $H_p(\lambda) = \pm t$ for $\lambda \in \mathbb{F}_p$, and then check whether $\chi(\lambda + 3)$ has the correct value. Similarly, when $q = p^2$ we look for solutions in $\mathbb{F}_q$ of $H_p(\lambda)H_p(\lambda^p) = \pm t$. This works reasonably efficient for -say- $q < 10^7$.

4.2 Numerical Results

Using the package KANT, we tested which values $\#E(\mathbb{F}_q) \equiv 0 \bmod 4$ occur for the curves $E_\lambda^{(\lambda+3)}/\mathbb{F}_q$, for all odd $q < 100000$. It turns out that for most q, all values are attained. In the table below, we list all $q < 100000$ where this is *not* the case, and for each of them the missing value(s) $\#E(\mathbb{F}_q)$. As can be seen from the data, usually there is only one such missing value, which moreover is always one of the minimal or the maximal possible number $\equiv 0 \bmod 4$ in the interval $\left[q+1-[2\sqrt{q}], q+1+[2\sqrt{q}]\right]$. We list a sign ± indicating which of these possibilities occurs for a missing value. There are exactly two exceptional cases for $q < 100000$. The first one is $q = 7^4$: here two values don't occur, namely the maximum $2500 = 7^4 + 1 + 2 \cdot 7^2$ and also $2396 = 7^4 + 1 - 6$. The other one is $q = 5^6$. The two values missing here are the minimal one $5^6 + 1 - 2 \cdot 5^3$ and also $15380 = q + 1 - 2\sqrt{q} + 4$. The following table gives all other $q < 100000$ with a missing value.

q	$\pm$	q	$\pm$	q	$\pm$	q	$\pm$	q	$\pm$	q	$\pm$	q	$\pm$	q	$\pm$
5	+	7	−	3^2	−	13	−	19	−	5^2	−	7^2	−	67	1
3^4	+	5^3	−	13^2	−	173	+	293	+	7^3	−	487	−	23^2	−
5^4	+	3^6	−	733	−	787	+	907	+	2503	+	3253	+	4493	−
4903	−	5333	+	5479	−	5779	−	3^8	+	7573	−	9413	+	10639	−
11239	−	11243	+	12547	−	11^4	+	14887	−	17959	+	18773	+	23719	+
24967	−	25603	−	27893	−	13^4	+	31687	−	33287	+	33493	−	37253	+
42853	−	46663	−	51991	+	52903	−	58567	+	3^{10}	−	64013	+	65539	+
67607	+	71293	−	76733	+	17^4	+	85853	+	92419	+	94253	−	99859	−

The table shows that for all but 30 values $q < 100000$, the maximal value of $\#C_\lambda(\mathbb{F}_q)$ equals $q + 1 + 3t$, where $q + 1 + t \equiv 0 \bmod 4$ is the maximal number of points of an elliptic curve over $\mathbb{F}_q$ with all its points of order 2 rational.

Whenever the Hasse-Weil-Serre bound is divisible by 4, we may be in the lucky circumstance that it is reached using our family of curves. This happens quite frequently, for instance when q equals any of the primes 19, 29, 53, 67, 71, 89, 103, 107, 151,.... In the case $q = 173$ the bound $q + 1 + 3m$ is a multiple of 4, but as the table above shows, our curves do not attain it.

The data seems to indicate that for much more than 50% of the prime powers q, all possible values $N_q \equiv 0 \bmod 4$ are attained by the family $E_\lambda^{(\lambda+3)}$. Moreover, the only occurrences of a q for which more than one value is missing, happened at 'high' even powers of a prime number. We have no theoretical explanation for this. A numerical test over all $q = p^{2n} < 10^7$ revealed exactly one more case where two values are missing, namely at $q = 7^6$. We have also not been able to explain why in all cases where we found that exactly one value is missing, this missing value is one of the maximal or minimal number of points.

References

A-T. R. Auer and J. Top, *Legendre elliptic curves over finite fields,* accepted for publ. in J. Number Theory, 2001.

De. M. Deuring, *Die Typen der Multiplicatorenringe elliptischer Funktionenkörper*, Abh. Math. Sem. Univ. Hamburg **14** (1941), 197–272.

G-V. G. van der Geer and M. van der Vlugt, *Tables for the function $N_q(g)$,* available from `http://www.science.uva.nl/~geer/`

He. P. Henn, *Die Automorphismengruppen der algebraischen Funktionenkörper vom Geschlecht* 3. Inaugural dissertation, Heidelberg, 1976.

H-L-P. E. W. Howe, F. Leprévost and B. Poonen, *Large torsion subgroups of split Jacobians of curves of genus two or three,* Forum Math. **12** (2000), 315–364.

Lau. K. Lauter, *Genus three curves over finite fields*, Notes of a lecture, available from `http://msri.org/publications/ln/msri/2000/crypto/lauter/1/`

Lau-Se. K. Lauter, *The maximum or minimum number of rational points on curves of genus three over finite fields,* (with an appendix by J-P. Serre), available from `http://arXiv.org/abs/math.AG/0104086`, accepted for publication in Compos. Math., 2002.

Sch. R. Schoof, *Nonsingular plane cubic curves over finite fields,* J. Combin. Theory Ser. A **46** (1987), 183–211.

Se1. J-P. Serre, *Sur le nombre des points rationnels d'une courbe algébrique sur un corps fini*, C. R. Acad. Sci. Paris Sér. I **296** (1983), 397–402.

Se2. J-P. Serre, *Résumé des cours de 1983–1984*, Annuaire du Collège de France (1984), 79–83. Reprinted in Vol. 3 of Jean-Pierre Serre, Œvres, Collected Papers. New York, etc.: Springer-Verlag, 1985.

Se3. J-P. Serre, Rational points on curves over finite fields, Lectures given at Harvard University, 1985. Notes by F. Q. Gouvêa.

Sh. V. Shabat, *Curves with many points.* Ph.D. thesis, Amsterdam, 2001.

Si. J. H. Silverman, The arithmetic of elliptic curves. New York, etc.: Springer-Verlag, GTM 106, 1986.

Si-T. J. H. Silverman and J. Tate, Rational points on elliptic curves. New York, etc.: Springer-Verlag, 1992.

To. J. Top, *Hecke L-series related with algebraic cycles or with Siegel modular forms.* Ph.D. thesis, Utrecht, 1989.

Ve. A. M. Vermeulen, *Weierstrass points of weight two on curves of genus three.* Ph.D. thesis, Amsterdam, 1983.

Wa. W. C. Waterhouse, *Abelian varieties over finite fields*, Ann. Sci. École Norm. Sup. **2** (1969), 521–560.

Trinomials $ax^7 + bx + c$ and $ax^8 + bx + c$ with Galois Groups of Order 168 and $8 \cdot 168$

Nils Bruin[1] and Noam D. Elkies[2*]

[1] Pacific Institute for Mathematical Sciences (SFU, UBC).
Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6 Canada
bruin@members.ams.org

[2] Department of Mathematics, Harvard University, Cambridge, MA 02138 USA
elkies@math.harvard.edu

Abstract. We obtain the curves of genus 2 parametrizing trinomials $ax^7 + bx + c$ whose Galois group is contained in the simple group G_{168} of order 168, and trinomials $ax^8 + bx + c$ whose Galois group is contained in $G_{1344} = (\mathbf{Z}/2)^3 \rtimes G_{168}$. In the degree-7 case, we find rational points of small height on this curve over $\mathbf{Q}$ and recover four inequivalent trinomials: the known $x^7 - 7x + 3$ (Trinks-Matzat) and $x^7 - 154x + 99$ (Erbach-Fischer-McKay), and two new examples,

$$37^2x^7 - 28x + 9 \qquad \text{and} \qquad 499^2x^7 - 23956x + 3^4 113.$$

We prove that there are no further rational points, and thus that every trinomial $ax^7 + bx + c$ with Galois group $\subseteq G_{168}$ over $\mathbf{Q}$ is equivalent to one of those four examples. In the degree-8 case, we again find some rational points of small height and compute the associated trinomials. This time all our examples are new:

$$x^8 + 16x + 28, \quad x^8 + 576x + 1008, \quad \text{and} \quad 19^4 53\, x^8 + 19x + 2,$$

each with Galois group G_{1344}; and

$$x^8 + 324x + 567,$$

with Galois group G_{168} acting transitively on the eight roots. We conjecture, but do not prove, that there are no further rational points, and thus that every trinomial $ax^8 + bx + c$ with Galois group $\subseteq G_{1344}$ over $\mathbf{Q}$ is equivalent to one of those four examples.

1 Introduction

1.1 Old and New Trinomials

Let G_{168} be the non-abelian simple group of second smallest order 168, isomorphic with both $\mathrm{PSL}_2(\mathbf{Z}/7\mathbf{Z})$ and $\mathrm{GL}_3(\mathbf{Z}/2\mathbf{Z})$ (also PGL_3, SL_3, PSL_3). The latter isomorphism yields actions of G_{168} on the 7 points and 7 lines of the projective

[*] Supported in part by the Packard Foundation.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 172–188, 2002.

plane of order 2 (Fano plane), either of which realizes G_{168} as a subgroup of index 15 in the alternating group A_7. In 1968 W.Trinks [T] showed that the trinomial $x^7 - 7x + 3$, of degree 7 and discriminant $3^8 7^8$, has Galois group G_{168}.[1] Trinks' unpublished manuscript [T] was cited a decade later by Erbach, Fischer, and McKay in a paper [EFM] that exhibits a new trinomial $x^7 - 154x + 99$, not equivalent with Trinks', which they show also has Galois group G_{168}. We find two further examples,

$$37^2 x^7 - 28x + 9 \qquad \text{and} \qquad 499^2 x^7 - 23956x + 3^4 113, \tag{1}$$

not equivalent with each other or with the Trinks-Matzat and Erbach-Fischer-McKay trinomials, and prove that every trinomial $ax^7 + bx + c$ over $\mathbf{Q}$ with Galois group contained in G_{168} is equivalent to one of the four such trinomials exhibited above.

Likewise let G_{1344} be the semidirect product $(\mathbf{Z}/2)^3 \rtimes G_{168}$. This is the automorphism group of the $(3,4,8)$ Steiner system, and the group $\mathrm{AGL}_3(\mathbf{Z}/2\mathbf{Z})$ of invertible affine linear transformations of a three-dimensional space over $\mathbf{Z}/2\mathbf{Z}$. (The Steiner system consists of the affine planes in this space.) Then G_{1344} is a subgroup of index 15 in A_8. We find four equivalence classes of trinomials $ax^8 + bx + c$ of degree 8 whose Galois group over $\mathbf{Q}$ is contained in G_{1344}, all new to our knowledge. Three of these, represented by

$$x^8 + 16x + 28, \quad x^8 + 576x + 1008, \quad \text{and} \quad 19^4 53 x^8 + 19x + 2, \tag{2}$$

have Galois group G_{1344}. The fourth, represented by

$$x^8 + 324x + 567, \tag{3}$$

has Galois group $G_{168} \cong \mathrm{PSL}_2(\mathbf{Z}/7\mathbf{Z})$, transitively permuting the eight roots as it does the points of the projective line over $\mathbf{Z}/7\mathbf{Z}$. We conjecture, but do not prove, that every trinomial $ax^8 + bx + c$ over $\mathbf{Q}$ with Galois group contained in G_{1344} is equivalent to one of our four trinomials exhibited above.

In each case we find a curve of genus 2 parametrizing trinomials satisfying the Galois condition. A direct search yields points of small height from which we recover our trinomials. To prove that these are the only ones, we must show that each curve has no further rational points. For the trinomials of degree 7, we are able to extend the methods of [B1,B2,B3] to obtain a proof. The curve parametrizing trinomials of degree 8 with Galois group contained in G_{1344} is too hard for us to treat in the same way with our present computational power.

1.2 Trinomials and Curves

Consider more generally trinomials $ax^n + bx + c$ of any degree $n \geq 2$ over a field K of characteristic zero. The equivalence class of such a trinomial consists of the

[1] A referee familiar with [T] reports that it contains "a hint that this polynomial has been found by Matzat via a computer search"; presumably Matzat surmised the Galois group from the degrees of the irreducible factors of $x^7 - 7x + 3$ modulo primes other than $3, 7$.

trinomials of the form $\alpha(a(mx)^n + b(mx) + c)$ for some $\alpha, m \in K^*$. Equivalent trinomials have the same Galois group because their roots are proportional. Define the *invariant* of a trinomial (other than the degenerate ax^n or c) by

$$T := b^n/ac^{n-1} \in \mathbf{P}^1(K). \tag{4}$$

We readily see that two trinomials are equivalent if and only if they have the same invariant. Trinomials with $ac = 0$ have $T = \infty$; those with $b = 0$ have $T = 0$; and each $T \in K^*$ is attained by the trinomial

$$P_T := x^n + Tx + T \quad (T \in K). \tag{5}$$

The polynomial P_T is separable unless T vanishes or equals

$$\gamma_n := (-n)^n/(n-1)^{n-1}. \tag{6}$$

We may regard (5) as a degree-n map from the x-line to the T-line, branched only at $T = 0$, $T = \infty$, and $T = \gamma_n$ with ramification indices (n), $(n-1, 1)$, and $(2, 1^{n-2})$. The corresponding extension $K(x)/K(T)$ of function fields has Galois group S_n [M, III, Satz 1]. Let B_n be the curve whose function field $K(B_n)$ is the Galois closure. This is the curve that parametrizes trinomials with a factorization

$$P_T(x) = \prod_{i=1}^{n}(x - r_i) \tag{7}$$

into linear factors; it is a normal cover of the T-line with Galois group S_n, ramified only above $T = 0$, $T = \infty$, and $T = \gamma_n$. We can also realize B_n as the smooth complete intersection of hypersurfaces $\sigma_j = 0$ of degree $j = 2, 3, \ldots, n-2$ in $\mathbf{P}^{n-2}$. Namely, σ_j is the elementary symmetric function of degree j in n variables $r_1, \ldots, r_n$ whose sum (the elementary symmetric function σ_1) vanishes. The genus of B_n is $(n^2 - 5n + 2)(n-2)!/4 + 1$; it can be calculated by applying either the adjunction formula to that complete intersection or the Riemann-Hurwitz formula to the degree-$(n!)$ map from B_n to the T-line.

For example, B_2 and B_3 are rational curves with actions of S_2 and S_3 by fractional linear transformations; B_4 is isomorphic with the conic $s_1^2 + s_2^2 + s_3^2 = 0$, with $s_j = r_1 \pm r_2 \pm r_3 \pm r_4$ (two minus signs) and S_4 acting by signed permutations of the s_i; and B_5 is the Bring curve of genus 4, whose automorphism group S_5 is the largest of any curve of its genus. For each prime p other than the primes dividing n or $n-1$ (that is, at which γ_n coincides with either 0 or ∞), the curve B_n and the function T on B_n have good reduction at p.

Now let G be any subgroup of S_n, and let $B_n(G)$ be the quotient of B_n by G, corresponding to the subfield of $K(B_n)$ fixed by G. This is the curve parametrizing trinomials $ax^n + bx + c$ whose Galois group is contained in G. The rational function T of degree $[S_n : G]$ on $B_n(G)$ realizes this parametrization: its value at each point of $B_n(G)$ is the invariant of the trinomials this point parametrizes. For instance, $B_n(S_n)$ is the T-line itself; $B_n(\{1\})$ is just B_n; if G is the point stabilizer S_{n-1} then $B_n(G)$ is the x-line, with $T = -x^n/(x+1)$

by (5). Considering T geometrically as a map from $B_n(G)$ to the T-line, we see that it is unramified away from $T = 0, \infty, \gamma_n$. This map can be used to calculate the genus of $B_n(G)$ via the Riemann-Hurwitz formula.[2] When the genus exceeds 1, the curve has finitely many K-rational points for any number field K by Faltings [Fa1,Fa2]. Hence there are finitely many equivalence classes of trinomials $ax^n + bx + c$ with Galois group contained in G. We can then ask for a provably complete list of such equivalence classes for some given K, notably $\mathbf{Q}$. In particular, this happens for the curves $B_7(G_{168})$ and $B_8(G_{1344})$, which turn out to have genus 2.

We can now state our results in the following equivalent form:

Theorem 1. *The curve* $B_7(G_{168})$ *has the hyperelliptic model*

$$Y^2 = X(81X^5 + 396X^4 + 738X^3 + 660X^2 + 269X + 48). \tag{8}$$

Theorem 2. *The* $\mathbf{Q}$*-rational points of the hyperelliptic curve (8) are the Weierstrass point* $(X, Y) = (0, 0)$*, the two points at infinity, and the two point pairs* $(-3, \pm 84)$ *and* $(1/9, \pm 28/9)$*.*

Theorem 3. *The curve* $B_8(G_{1344})$ *has the hyperelliptic model*

$$Y^2 = 2X^6 + 28X^5 + 196X^4 + 784X^3 + 1715X^2 + 2058X + 2401. \tag{9}$$

Conjecture 1. The only $\mathbf{Q}$-rational points of the hyperelliptic curve (9) are the four pairs $(X, Y) = (0, \pm 49)$, $(-1, \pm 38)$, $(-3, \pm 32)$, and $(-7, \pm 196)$.

In the next section we prove Theorem 1 and recover the degree-7 trinomials with Galois group G_{168} from four of the rational points listed in Theorem 2. (The other points do not yield trinomials because they are zeros or poles of T.) We then outline the proof of Theorem 3, and recover our degree-8 trinomials with Galois group $\subseteq G_{1344}$ from the rational points listed in Conjecture 1, including the curious reappearance of G_{168} for the trinomial (3). In the final section we prove Theorem 2, and indicate our difficulty in proving Conjecture 1.

1.3 Using Distinct Residue Characteristics in Chabauty Arguments

In order to prove Theorem 2 we make use of covering techniques and the method of Chabauty. In Chabauty's method, one takes an embedding of a curve C in an abelian variety A defined over a number field K. By considering C as a subvariety of A, we have $C(K) \subset A(K)$. Let p be a finite prime of K. Then $A(K)$ is a finitely generated subgroup of the p-adic Lie group $A(K_p)$. The topological completion $\overline{A(K)} \subset A(K_p)$ is a sub Lie group. Similarly, $C(K_p)$ is a p-adic submanifold of $A(K_p)$. Naturally, $C(K) \subset C(K_p) \cap A(K_p)$. The latter is an intersection of a p-adic curve and, provided certain nontrivial technical conditions are met, a

[2] This computation, while elementary, can be tricky to perform accurately, as witness the claim [M, p.95] that $B_7(G_{168})$ has genus 3.

submanifold of positive codimension in $A(K_p)$. One would expect this to be 0-dimensional and finite. Its size can be determined by p-adic analytic means and provides an upper bound for $\#C(K)$.

Naturally, $C(K_p) \cap A(K_p)$ may contain points that do not correspond to points in $C(K)$. In fact, for larger p one would expect such points. Here we describe how one can exhibit such points using information obtained from other residue characteristics.

Suppose $A(K) = \langle P_1, \ldots, P_r \rangle$. Suppose that p is a prime of good reduction of C. Let k be the residue field of K at p. We fix a reduction map $A(K_p) \to A(k)$ Let $P = n_1 P_1 + \cdots + n_r P_r \in A(K)$. If $P \in C(K)$, then certainly $P \bmod p \in C(k)$. This gives certain congruences on $n_1, \ldots, n_r \in \mathbf{Z}$ by considering

$$C(k) \cap \langle P_1, \ldots, P_r \rangle \bmod p \subset A(k).$$

These congruences are modulo the kernel of reduction modulo p. These need not be independent from the congruences obtained from another prime q and may be used to sharpen the bound on $\#C(K)$.

The above observation removes the need for choosing a particularly small prime for the Chabauty argument (which might not be available). In Section 3, we give an example how this idea can be used in conjunction with Chabauty techniques as described in [B2].

2 Hyperelliptic Models for $B_7(G_{168})$ and $B_8(G_{1344})$

2.1 Computing a Hyperelliptic Model for $B_7(G_{168})$

We obtain our hyperelliptic model for the curve $B_7(G_{168})$ by finding low-degree rational functions X, Y on the curve, proving that they generate the curve's function field, and computing the polynomial relation (8) satisfied by X, Y. Our strategy for finding X, Y is as follows.

Let $(r_1 : \cdots : r_7)$ be homogeneous coordinates on $\mathbf{P}^6$. In this projective space we have the hyperplane on which $\sigma_1 = \sum_{i=1}^{7} r_i$ vanishes, and further hypersurfaces $\sigma_j = 0$ $(j = 2, 3, 4, 5)$ whose complete intersection with the hyperplane $\sigma_1 = 0$ is the curve $B_7 = B_7(\{1\})$. The rational coordinate T on the line $B_7(S_7) = B_7/S_7$ is the quotient σ_6^7/σ_7^6 of two homogeneous polynomials of the same degree, both invariant under S_7. Likewise the quotient of any two homogeneous polynomials of the same degree in the r_i, both invariant under G_{168}, is a rational function on $B_7(G_{168}) = B_7/G_{168}$. We exhibit homogeneous polynomials p_j of degree $j = 3, 4, 5, 6$ in the r_i that are invariant under G_{168} but not under S_7, and obtain polynomial relations between them and the σ_j. Setting $\sigma_j = 0$ for $j \leq 5$, we obtain simpler relations involving only p_j, σ_6, σ_7 that hold on B_7 and its quotient $B_7(G_{168})$. We eliminate r_6, σ_6, σ_7 to obtain a relation between r_3, r_4, r_5, which we write as a polynomial in

$$X := p_4^2/p_3 p_5 \quad \text{and} \quad Z := 7 p_4 p_5 / p_3^3. \tag{10}$$

This polynomial has degree 2 in Z, and thus defines a hyperelliptic curve. We show that the curve has genus 2 and recover its hyperelliptic model (8) by computing the discriminant with respect to Z and factoring it as a square times a polynomial of degree 6 in X.

This curve is then the image of $B_7(G_{168})$ under a nonconstant rational map; we must also show that the map is an isomorphism, and obtain T as a rational function on the curve. We do both by writing σ_6, σ_7 as rational functions of r_3, r_4, r_5. This yields $T = \sigma_6^7/\sigma_7^6$ as a rational function of X and Z. The function field $K(X, Z)$ is then known to contain $K(T)$ and to be contained in its Galois extension $K(B_7)$; we can then use Galois theory to identify $K(X, Z)$ with the function field of $B_7(G_{168})$, completing the proof of Theorem 1. Alternatively the last step can be done by showing independently that $B_7(G_{168})$ has genus 2 and quoting the fact that a nonconstant rational map between curves of the same genus $g > 1$ must be an isomorphism.

Let $\mathbf{r} = \{r_1, r_2, \ldots, r_7\}$, and let Φ be a collection of seven 3-element subsets of $\mathbf{r}$ that are the lines of a Fano plane. For instance, we may take

$$\Phi = \{\{r_i, r_{i+1}, r_{i+3}\} \mid i \in \mathbf{Z}/7\mathbf{Z}\}. \tag{11}$$

Let $\overline{\Phi}$ be the collection of 4-element subsets of $\mathbf{r}$ complementary to those in Φ. Thus if we choose Φ by (11) then

$$\overline{\Phi} = \{\{r_{i+2}, r_{i+4}, r_{i+5}, r_{i+6}\} \mid i \in \mathbf{Z}/7\mathbf{Z}\}. \tag{12}$$

Now G_{168} is the group of permutations of $\mathbf{r}$ that fix Φ. Our G_{168}-invariant polynomials p_j are defined by

$$p_3 = \sum_{l \in \Phi} r r' r'', \qquad p_4 = \sum_{\bar{l} \in \overline{\Phi}} r r' r'' r''',$$

$$p_5 = \sum_{l \in \Phi} r r' r'' (r^2 + r'^2 + r''^2), \qquad p_6 = \sum_{\bar{l} \in \overline{\Phi}} r r' r'' r''' (r^2 + r'^2 + r''^2 + r'''^2),$$

where $l = \{r, r', r''\}$ and $\bar{l} = \{r, r', r'', r'''\}$. Each p_j is a new element of the space of G_{168}-invariant polynomials of degree j in $\mathbf{r}$: it is not contained in the polynomials of degree j in the $p_{j'}$ $(j' < j)$ and σ_j. But the p_j must satisfy algebraic dependences with the σ_j. We find the first few such relations in degrees 10, 11 (one each), and 12 (two relations):

$$21 p_3 \sigma_7 = 5 p_4 \sigma_6 + p_4 p_6 + p_5^2 + 2 p_3^2 p_4, \tag{13}$$

$$49 p_4 \sigma_7 = -7 p_5 \sigma_6 + 7 p_5 p_6 + 2 p_3^2 p_5 + 3 p_3 p_4^2, \tag{14}$$

$$49 p_5 \sigma_7 = 42 p_6 \sigma_6 + 9 p_3^2 \sigma_6 - 35 p_6^2 - 36 p_3^2 p_6 + 10 p_3 p_4 p_5 + 4 p_4^3 - 9 p_3^4, \tag{15}$$

$$49 p_5 \sigma_7 = 84 \sigma_6^2 - 3 p_3^2 \sigma_6 + 7 p_6^2 + 12 p_3^2 p_6 + 4 p_3 p_4 p_5 - 2 p_4^3 + 3 p_3^4. \tag{16}$$

We next reduce these to a single polynomial relation in p_3, p_4, p_5 by eliminating p_6, σ_6, σ_7. Choose any two, say p_6 and σ_6, We may regard (13,14) as simultaneous linear equations in p_6 and σ_6. We solve them, substitute into (15,16), and clear denominators to obtain two polynomials in p_3, p_4, p_5, σ_7 which are quadratic in σ_7. Their resultant with respect to σ_7 is then a polynomial relation satisfied by p_3, p_4, p_5. Switching the roles of σ_6, σ_7 we find another such polynomial. The gcd of these polynomials has two irreducible factors, of degrees 20 and 36. The first of these is spurious. Expressed in terms of X, Z (see (10)), it is the curve

$$Z^2 - (X^3 - 5X^2 + 12X)Z + 18X^2 - 27X = 0 \tag{17}$$

of genus 1. This factor can be ruled out in various ways, such as using degree-13 relations in p_j, σ_6, σ_7, or even calculating that it has bad reduction at the primes $11, 17$ while $B_7(168)$ must have good reduction away from the primes $2, 3, 7$ dividing $7(7-1)$. We are left with the curve

$$\begin{aligned}(11X^2 + 13X + 4)Z^2 - (81X^5 + 315X^4 + 467X^3 + 335X^2 + 90X)Z \\ = 37X^4 + 171X^3 + 216X^2 + 108X.\end{aligned} \tag{18}$$

(Like (17), this equation has smaller coefficients because of the factor of 7 introduced into Z in (10).) The curve (18) is a quadratic cover of the X-line. Its discriminant as a polynomial in Z is

$$(9X^2 + 13X + 6)^2 X(81X^5 + 396X^4 + 738X^3 + 660X^2 + 269X + 48). \tag{19}$$

Therefore (18) yields our hyperelliptic model (8) for $B_7(G_{168})$, with

$$Y = \frac{2(11X^2 + 13X + 4)Z - X(81X^4 + 315X^3 + 467X^2 + 335X + 90)}{9X^2 + 13X + 6}. \tag{20}$$

It is a welcome sanity check that this curve has bad reduction only at the primes $2, 3, 7$ dividing $7(7-1)$: the sextic in (19) has discriminant $2^{24}3^{11}7^8$.

2.2 Theorem 1, and Septic Trinomials over Q

We next show that p_6, σ_6, and σ_7 are rational functions of p_3, p_4, p_5. The elimination of p_6, σ_6 left two equations, both quadratic in σ_7. In the process of eliminating σ_7 we obtained a linear combination of these two equations that is linear in σ_7, with coefficients in $\mathbf{Q}(p_3, p_4, p_5)$. Solving it yields $\sigma_7 \in \mathbf{Q}(p_3, p_4, p_5)$. We can either repeat this argument for p_6 and σ_6 to show that $p_6, \sigma_6 \in \mathbf{Q}(p_3, p_4, p_5)$; alternatively, since we already know that $p_6, \sigma_6 \in \mathbf{Q}(p_3, p_4, p_5, \sigma_7)$ by solving (13,14) for (p_6, σ_6), we can deduce $p_6, \sigma_6 \in \mathbf{Q}(p_3, p_4, p_5)$ from $\sigma_7 \in \mathbf{Q}(p_3, p_4, p_5)$.

In particular, $T = \sigma_6^7/\sigma_7^6 \in \mathbf{Q}(p_3, p_4, p_5)$; since T is homogeneous of degree 0, it is thus contained in the function field $\mathbf{Q}(X, Z)$ of our hyperelliptic curve (8).

We can now conclude the proof of Theorem 1. We just showed that $\mathbf{Q}(X, Z)$ contains $\mathbf{Q}(T)$. On the other hand, X and Z are homogeneous functions of the r_i, so $\mathbf{Q}(X, Z)$ is contained in the function field of B_7. But this function field $\mathbf{Q}(B_7)$ is a normal extension of $\mathbf{Q}(T)$ with Galois group S_7. Hence the intermediate field $\mathbf{Q}(X, Z)$ is $\mathbf{Q}(B_7)/G$ for some group $G \subseteq S_7$, namely the stabilizer of X and Z in S_7. Clearly then $G \supseteq G_{168}$. But G can be no larger than G_{168}. We can show this directly, by checking that no element of $S_7 - G_{168}$ fixes X. Alternatively we may recall that the only subgroups of S_7 properly containing G_{168} are S_7 itself and A_7, and noting that $\mathbf{Q}(B_7)/S_7 = \mathbf{Q}(T)$ and $\mathbf{Q}(B_7)/A_7 = \mathbf{Q}(\sqrt{\gamma_7 - T}\,)$ are both rational function fields and thus not isomorphic with $\mathbf{Q}(X, Z)$. Either way, it follows that $G = G_{168}$ and that $\mathbf{Q}(X, Z)$ is the function field of $B_7(G_{168})$, and we are done.

[We could also have completed the proof by showing independently that $B_7(G_{168})$ has genus 2 and thus that the map from $B_7(G_{168})$ to $\mathbf{Q}(X, Z)$ must be an isomorphism. We can compute the genus of $B_7(G_{168})$ by applying the Riemann-Hurwitz formula to the map of degree 30 (or 15) from $B_7(G_{168})$ to $\mathbf{Q}(T) = B_7(S_7)$ (or $B_7(A_7)$). Alternatively, we can count holomorphic differentials on $B_7(G_{168})$. By the adjunction formula, the holomorphic differentials on B_7 are the sections of $O(8)$, that is, homogeneous polynomials of degree 8 in $r_1, \ldots, r_7$ modulo $(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5)$. Such a section descends to a holomorphic differential on $B_7(G_{168})$ if and only if it is invariant under G_{168}. We find that the space of invariant sections is two-dimensional, generated by p_4^2 and $p_3 p_5$. This confirms that $B_7(G_{168})$ has genus 2, and also that $X = p_4^2/p_3 p_5$ gives the degree-2 map from $B_7(G_{168})$ to $\mathbf{P}^1$.]

We can also compute septic trinomials over $\mathbf{Q}$ with Galois group contained in G_{168}. A search for rational points reveals the seven points listed in the statement of Theorem 2. Of these, three yield degenerate septics: the two points with $X = -3$, and the point $(X, Y, Z) = (1/9, -28/9, -7/9)$, are zeros of σ_7 but not of σ_6, and thus poles of T. The Weierstrass point $(0, 0, 0)$ yields $T = -7^7/3^6$, the invariant of the Trinks-Matzat trinomial $x^7 - 7x + 3$. The Erbach-Fischer-McKay trinomial $x^7 - 154x + 99$, with invariant $-14^7 11/3^{12}$, arises from the point at infinity $(X, Y, Z) = (\infty, \infty^3, \infty^3)$. The remaining two rational points yield our new septics exhibited in (1): the other point at infinity, $(X, Y, Z) = (\infty, \infty^3, 0)$, yields $T = -28^7/3^{12}37^2$ and the trinomial $37^2x^7 - 28x + 99$; and $(X, Y, Z) = (1/9, -28/9, 3493/1017)$, the hyperelliptic conjugate of the pole $(1/9, -28/9, -7/9)$, has $T = 2^{14}5^3 7^7 113/3^{24}499^2$, the invariant of the trinomial $499^2x^7 - 23956x + 3^4 113$. We verified with GP, and again with Magma [BCP], that each of these trinomials has Galois group exactly G_{168}.

2.3 The Hyperelliptic Model for $B_8(G_{1344})$, and Octic Trinomials

We sketch the proof of Theorem 3, that is, the computation of the hyperelliptic model (9) for the curve $B_8(G_{1344})$. We proceed much as we did with $B_7(G_{1344})$. Here we need G_{1344}-invariant polynomials in eight variables $r_1, \ldots, r_8$ satisfying $\sigma_j = 0$ for $1 \le j \le 6$. Let Ψ be a $(3, 4, 8)$ Steiner system of fourteen 4-element subsets b of $\{r_1, \ldots, r_8\}$. For instance, we may take for Ψ the disjoint union of $\overline{\Phi}$ and the set obtained from Φ by extending each line by r_8. This time we need five new invariants, in degrees $4, 6, 7, 8, 9$. and find relations in degrees $13, 14$ (one each) and $15, 16$ (two each). Specifically, we take

$$p_{j+4} = \sum_{b \in \Psi} r r' r'' r''' (r^j + r'^j + r''^j + r'''^j) \qquad (j = 0, 2, 3, 4, 5)$$

where $b = \{r, r', r'', r'''\}$, and find

$$\begin{gathered}
2p_4p_9 + p_6p_7 + 11p_6\sigma_7 = 0,\\
4p_4^2p_6 + 10p_6p_8 - 144p_6\sigma_8 + 3p_7^2 - 21p_7\sigma_7 + 294\sigma_7^2 = 0,\\
3p_4^2p_7 + 5p_4^2\sigma_7 + 3p_6p_9 - 3p_7p_8 - 24p_7\sigma_8 + 14\sigma_7p_8 = 0,\\
-p_4^2p_7 - 3p_4^2\sigma_7 + 2p_7p_8 + 12p_7\sigma_8 + 84\sigma_7\sigma_8 = 0,\\
288\sigma_8^2 - 20p_8\sigma_8 - 8p_4^2\sigma_8 - 3p_7p_9 - 2p_8^2 + p_4^2p_8 = 0,\\
1512p_8\sigma_8 + 432p_4^2\sigma_8 + 441p_9\sigma_7 + 63p_7p_9 + 35p_8^2 - 50p_4^2p_8 - 27p_4p_6^2 - 4p_4^4 = 0.
\end{gathered}$$

We then solve simultaneous linear equations to write σ_8, p_8, p_9 as rational functions of p_4, p_6, p_7, σ_7, and use resultants to eliminate p_6. This leaves a polynomial in x_4, x_7, σ_7 that is quadratic in in x_4^7. Its discriminant with respect to x_4^7 is a homogeneous polynomial of degree 22 in p_7, σ_7 which contains a square factor of degree $2 \cdot 8$. Eliminating this factor we obtain the sextic in the right-hand side of (9), where $X = p_7/\sigma_7$. (Thus X is the quotient $p_6p_7/p_6\sigma_7$ of degree-13 polynomials that generate the space of holomorphic differentials on $B_8(G_{1344})$ by the adjunction formula and the degree-13 relation.) The discriminant of the sextic is $-2^{24}7^{18}$, again confirming good reduction away from the prime factors of $n(n-1)$. Curiously the sextics for $B_7(G_{168})$ and $B_8(G_{1344})$ both have Galois group isomorphic with S_5, but with different permutation representations.

A search for rational points reveals the eight points listed in the statement of Conjecture 1. Three of these, one of the $X = -1$ points and both $X = -3$ points, are zeros of T; the remaining five yield genuine trinomials. The three trinomials listed in (2) come from points with $X = -7, 0, -1$ respectively. The remaining two points, with $X = -7, 0$, both yield $T = 18^4/7^7$, the invariant of the octic trinomial $x^8 + 324x + 567$ of (3). Again using GP (in a version with `polgalois` extended to maximal degree 11) and checking with Magma, we confirmed that each of the three trinomials in (2) has Galois group exactly G_{1344}. On the other hand, $x^8 + 324x + 567$ has Galois group G_{168}, acting transitively on the eight roots. It appears twice because there are two embeddings of G_{168} into G_{1344} not equivalent by conjugation in G_{1344}. We may identify $\{r_1, \ldots, r_8\}$ with the projective line over $\mathbf{Z}/7\mathbf{Z}$ by taking r_8 to ∞ and r_j $(j \le 7)$ to $j \bmod 7$. Then $\mathrm{PSL}_2(\mathbf{Z}/7\mathbf{Z}) \cong G_{168}$ preserves Ψ, and is thus contained in $G_{1344} = \mathrm{Aut}(\Psi)$. But $\mathrm{PGL}_2(\mathbf{Z}/7\mathbf{Z}) \cong \mathrm{Aut}(G_{168})$ does not preserve Ψ, so conjugation by an element of $\mathrm{PGL}_2(\mathbf{Z}/7\mathbf{Z}) - \mathrm{PSL}_2(\mathbf{Z}/7\mathbf{Z})$ yields an inequivalent embedding of G_{168} into G_{1344}. This is why $x^8 + 324x + 567$, and indeed any irreducible trinomial $ax^8 + bx + c$ with Galois group G_{168} over some field K of characteristic zero, must come from two K-rational points of $B_8(G_{1344})$ with the same value of T.

3 Determining Rational Points on Curves

3.1 Proof of Theorem 2

First we introduce a model for $B_7(G_{168})$ that is slightly better suited for computation. We define

$$C : y^2 = 48x^5 + 29x^4 + 64x^3 - 108x^2 + 64x - 16,$$

which is isomorphic to the model in Theorem 1 via

$$(X, Y) = \left(\frac{1}{x-1}, \frac{y}{(x-1)^3}\right).$$

We use the techniques from [B3] to determine $C(\mathbf{Q})$. See [B4] for a more elaborate exposition on how these techniques apply in practice. Here, we leave extensive computational details to an electronic reference [BE]. We will concentrate on a technical difficulty that typically arises in examples other than extremely small ones.

First, we establish that the techniques from [Fl] do not apply. We write J for the Jacobian variety of C.

Lemma 1. $\mathrm{rk} J(\mathbf{Q}) = 2$

Proof. Using a 2-descent one obtains an upper bound of 2 on the rank of $J(\mathbf{Q})$. For our purposes a lower bound is more relevant. We get one by checking that the divisor classes represented by $[(1,9) - \infty]$ and $[(2/3, 28/9) - \infty]$ are independent. One can do so either by checking that their images generate the 2-Selmer group or by checking that their height-pairing matrix is nonsingular. These computations can be performed by Stoll's implementation [St] in Magma [BCP]. See [BE] for details.

We see that the rank of $J(\mathbf{Q})$ is equal to the geometric dimension of J. This rules out an application of the method of Chabauty-Coleman directly to C as a subvariety of J.

We take the approach of [B3]. We determine the rational points on a set of twists of an unramified cover of C. We take the cover that is obtained by pulling back an embedding of C in J along the multiplication-by-2 map $J \xrightarrow{2} J$. See [BF] for a description of this cover. Rather than working with this cover directly, which would be a curve of genus 17, we use the many subcovers that this curve has. The following lemmas show that, in order to determine $C(\mathbf{Q})$, it suffices to find the points of certain curves of genus 1 over a number field K that satisfy certain additional arithmetic properties.

Consider the number field $K = \mathbf{Q}(\alpha)$ defined by the relation $\alpha^5 - 2\alpha^4 + 3\alpha^3 - 4\alpha^2 + 5\alpha - 6 = 0$. Over K, we have the following factorization of F.

$$\begin{aligned}
F(x) &= Q(x)R(x), \text{ where}\\
Q(x) &= (-2\alpha^3 + 5\alpha^2 - 7\alpha + 6)x - \alpha^4 + 2\alpha^3 - 3\alpha^2 + 4\alpha - 1\\
R(x) &= (95\alpha^4 - 52\alpha^3 + 213\alpha^2 - 62\alpha + 391)x^4 +\\
&\quad (108\alpha^4 - 56\alpha^3 + 233\alpha^2 - 79\alpha + 422)x^3 +\\
&\quad (172\alpha^4 - 82\alpha^3 + 392\alpha^2 - 98\alpha + 696)x^2 +\\
&\quad (-124\alpha^4 + 60\alpha^3 - 280\alpha^2 + 72\alpha - 496)x +\\
&\quad 64\alpha^4 - 32\alpha^3 + 144\alpha^2 - 40\alpha + 256
\end{aligned}$$

The following lemma links $C(\mathbf{Q})$ to the rational points on certain curves of genus 1 over K.

Lemma 2. *Let*

$$\begin{aligned}\delta_1 &= -5\alpha^4 + 14\alpha^3 - 18\alpha^2 + 9\alpha + 5,\\ \delta_2 &= 123\alpha^4 - 262\alpha^3 + 188\alpha^2 + 151\alpha - 383,\\ \delta_3 &= 6\alpha^4 - 3\alpha^3 + 11\alpha^2 + 19.\end{aligned}$$

If $(x, y) \in C(\mathbf{Q})$*, then there is an* $i \in \{1, 2, 3\}$ *and* $y_1 \in K$ *such that*

$$\delta_i y_1^2 = R(x).$$

Proof. First note that $F(x)$ has no rational roots. Therefore, if $x \in \mathbf{Q}$, then $Q(x), R(x) \in K^*$. Let $\mathfrak{p}$ be a prime of K so that resultant$(Q(x), R(x))$ is a unit at $\mathfrak{p}$. It is straightforward to verify that if $Q(x)R(x)$ has even valuation at $\mathfrak{p}$, then so do $Q(x)$ and $R(x)$ individually. Thus, if $Q(x)R(x) = y^2$, then there exists $\delta \in K$, representing an element from the subgroup of K^*/K^{*2} of elements that have an even valuation at all primes outside the primes above $\{2, 3, 7\}$ so that there are $y_1, y_2 \in K^*$ satisfying

$$\begin{aligned}\delta y_1^2 &= R(x),\\ \delta y_2^2 &= Q(x),\\ y_1 y_2 &= y.\end{aligned}$$

Following [Si, Theorem X.1.1], we write $K(\{2, 3, 7\}, 2)$ for this group. The group $K(\{2, 3, 7\}, 2)$ is finite. In fact, since K has class number 1, the group is represented by the square-free elements of the $\{2, 3, 7\}$-unit group.

We employ local arguments to show that we only need classes represented by the three elements given in the lemma. Let $\mathfrak{p}$ be a prime of K and let p be the rational prime below K. For each class in $K(\{2, 3, 7\}, 2)$, we choose a representing element $\delta \in K^*$. We test if the equations $\delta y_1^2 = R(x), \delta y_2^2 = Q(x)$ can be simultaneously satisfied for $x \in \mathbf{Q}_p$, $y_1, y_2 \in K_{\mathfrak{p}}$. For the finite primes, this procedure is completely automatic in [B5] built on top of [K]. See [B4] for an example and for a transcript.

Besides the values mentioned in the lemma, we also find $\delta_4 = -1 - 2\alpha + 4\alpha^2 - 3\alpha^3 + \alpha^4$. To rule out this value, note that K has only one real place. If we embed K in $\mathbf{R}$ by $\alpha \mapsto 1.4918\ldots$, we find that $\delta_4 R(x)$ is definite negative for $x \in \mathbf{R}$ and thus is never a square.

This leaves us the three values mentioned in the lemma. Note that the choice of representatives is arbitrary and thus that the same procedure executed twice may return different but equivalent values.

Next we show that there is a good reason why the three values for δ in the lemma above occur. Each of the curves $\delta_i y^2 = R(x)$ actually has a rational point with $y \in K$ and $x \in \mathbf{Q}$. Consequently, these curves are isomorphic to their Jacobians, i.e., are elliptic curves. We compute Weierstrass-models of these.

Lemma 3. *For* $i = 1, 2, 3$*, the curve* $\delta_i y^2 = R(x)$ *has a* K*-rational point with a* $\mathbf{Q}$*-rational* x*-coordinate* x_i*. The curve is isomorphic to the Weierstrass-model*

E_i, where the relevant data is given in the following table.

i	x_i	E_i
1	1	$Y^2 = X^3 + (-21\alpha^4 - 282\alpha^3 - 138\alpha^2 + 324\alpha + 864)X - 1330\alpha^4 + 4338\alpha^3 - 360\alpha^2 + 5080\alpha - 14592$
2	2/3	$Y^2 = X^3 + (-924\alpha^4 - 318\alpha^3 - 90\alpha^2 + 2274\alpha + 1629)X + 11312\alpha^4 + 21394\alpha^3 - 7230\alpha^2 - 25520\alpha - 71778$
3	∞	$Y^2 = X^3 + (795\alpha^4 - 1584\alpha^3 + 738\alpha^2 - 2562\alpha + 3501)X + 14068\alpha^4 - 6586\alpha^3 + 3894\alpha^2 - 39856\alpha + 2982$

Proof. The proof is completely standard. See [Ca] for a nice recipe for finding a Weierstrass-model for a curve of the form $Y^2 =$ quartic in x with a rational point. In [B5], this procedure is implemented. See [BE] for more information.

Determining the Mordell-Weil group of E_i, or rather a subgroup of finite index in $E(K)$, is the most difficult step. First we bound the rank by a 2-descent and then we hope that we can find sufficient independent points. In the process of the 2-descent we need a field extension A of K over which E_i acquires a 2-torsion point. This is the same, cubic extension of K for all the E_i. This should come as no surprise, since the models $\delta_i y^2 = R(x)$ already indicate that the E_i are quadratic twists and thus have isomorphic 2-torsion.

We need full classgroup and unit information of A, which is a degree 15 extension of $\mathbf{Q}$. As it turns out, this is in fact doable. Using an implementation of the relation methods (see [H], [Co]) in MAGMA [BCP], we find that A has trivial class number. The same method gives information about the group of $\{2,3,7\}$-units in A.

As is often the case, finding the class group information is much easier than proving that the information is correct. To find the information, one only needs consider prime ideals above rational primes up to 300. The computation takes a few seconds on PentiumIII 600Mhz laptop running Linux. To verify that the obtained results are correct assuming GRH involves checking the primes up to 34225 and takes about 3 minutes. To verify the results unconditionally, one needs to check all primes up to 5028282, which takes about 12 hours on a Sun Ultra5. See [BE] for a transcript.

Lemma 4. *For $i = 1, 2, 3$, the group $E_i(K)$ is torsion-free and of indicated rank. A subgroup of finite, odd index is generated by the point $P_{i,j}$, where $X(P_{i,j})$ is given in the table below.*

i	rk$E_i(K)$	$X(P_{i,j})$
1	3	$(4\alpha^4 - 25\alpha^3 + 36\alpha^2 - 69\alpha + 86)/8$ $8\alpha^4 - 26\alpha^3 + 7\alpha^2 - 14\alpha + 52$ $-2\alpha^4 + 4\alpha^3 - 2\alpha^2 - 10\alpha + 16$
2	2	$-22\alpha^4 + 34\alpha^3 - 18\alpha^2 + 64\alpha - 77$ $(-45\alpha^4 - 74\alpha^3 - 9\alpha^2 + 68\alpha + 192)/12$
3	3	$(-47\alpha^4 + 30\alpha^3 + 2\alpha^2 + 156\alpha - 104)/4$ $(261\alpha^4 - 1374\alpha^3 + 601\alpha^2 - 1026\alpha + 3462)/25$ $(-13\alpha^4 + 24\alpha^3 + 34\alpha - 66)/9$

Proof. These facts can be verified using [B5]. See [BE] or [B4] for more information.

The curves $\delta_i y^2 = R(x)$ and E_i are isomorphic over K. We will interpret x as a degree 2 morphism $E_i \to \mathbf{P}^1$, not to be confused with X, the X-coordinate of the Weierstrass-model. Lemma 2 assures that by determining $x(E_i(K)) \cap \mathbf{P}^1(\mathbf{Q})$ for $i = 1, 2, 3$, we obtain a set that contains the x-coordinates of $C(\mathbf{Q})$. If this set is finite, then it is only a finite amount of work to obtain $C(\mathbf{Q})$ from it.

To lighten notation, fix i and write $E = E_i$, $r = \mathrm{rk} E(K)$ and $P_1, \ldots, P_r = P_{i,1}, \ldots, P_{i,r}$. From Lemma 4 we know that $E_i(K) \simeq \mathbf{Z}^r$ and that the index $I = \#(E(K)/\langle P_1, \ldots, P_r\rangle)$ is finite. In our situation, we expect that $I = 1$, but do not need to prove it. For any $P \in E(K)$ we have $n_1, \ldots, n_r \in \mathbf{Z}$ so that $IP = n_1 P_1 + \cdots + n_r P_r$.

Let p be a rational prime so that E has good reduction at all the primes $\mathfrak{p}_1, \ldots, \mathfrak{p}_s$ of K above p. We define $\Lambda_p \subset \langle P_1, \ldots, P_r\rangle$ to be the intersection of the kernels of reduction modulo $\mathfrak{p}_1, \ldots, \mathfrak{p}_s$.

$$\Lambda_p = \{P \in \langle P_1, \ldots, P_r\rangle : \text{for all } i \in \{1, \ldots, s\} \text{ we have } P \bmod \mathfrak{p}_i = O\}$$

If $P \in E(K)$ has $x(P) \in \mathbf{P}^1(\mathbf{Q})$, then for any i, j we have $x(P) \bmod \mathfrak{p}_i \in \mathbf{P}^1(\mathbf{F}_p)$ and $x(P) \bmod \mathfrak{p}_i = x(P) \bmod \mathfrak{p}_j$. We define

$$\begin{aligned} V_p = \{P \in \langle P_1, \ldots, P_r\rangle :\ & \text{for all } i, j \in \{1, \ldots, s\} \text{ we have} \\ & x(P) \bmod \mathfrak{p}_i \in \mathbf{P}^1(\mathbf{F}_p) \text{ and } x(P) \bmod \mathfrak{p}_i = x(P) \bmod \mathfrak{p}_j\}. \end{aligned}$$

Assume that for each i, we have that I is coprime with the index of

$$\langle P_1, \ldots, P_r\rangle \bmod \mathfrak{p}_i \subset (E \bmod \mathfrak{p}_i)(\mathbf{F}_{N\mathfrak{p}_i}).$$

Then $\langle P_1, \ldots, P_r\rangle \bmod \mathfrak{p}_i = E(K) \bmod \mathfrak{p}_i$. It follows that if $P \in E(K)$ with $x(P) \in \mathbf{P}^1(\mathbf{Q})$, then there is a $Q \in V_p$ so that $P \bmod \mathfrak{p}_i = Q \bmod \mathfrak{p}_i$ for $i = 1, \ldots, s$. In other words, $(P - Q) \bmod \mathfrak{p}_i = O$.

In order to bound the number of $P \in E(K)$ with $x(P) \in \mathbf{P}^1(\mathbf{Q})$ that reduce to a fixed $Q \in V_p$, we use that the group structure on the kernel of reduction $E^1(K_{\mathfrak{p}_i})$ is given by a formal group. Again, we do not need that $I = 1$. We only need that $\Lambda_p \otimes \mathbf{Z}_p$ is equal to the intersection of the kernels of reduction $E(K) \cap E^1(K_{\mathfrak{p}_i})$. Since any prime $q \neq p$ is a unit in $\mathbf{Z}_p$, this follows if $I \bmod p \neq 0$. For details, we refer the reader to [B2].

Here we will concentrate on ways to reduce the number of $Q \in V_p$ that need further consideration. Let q be another rational prime satisfying the necessary assumptions and assume that $\Lambda_p + \Lambda_q$ does *not* equal the entire $\langle P_1, \ldots, P_r\rangle$. We consider

$$V_{p,q} = \{P \in V_p : (P + \Lambda_p) \cap V_q \neq \emptyset\}$$

Obviously, if $P \in E(K)$ with $x(P) \in \mathbf{P}^1(\mathbf{Q})$, then there is a $Q \in V_{p,q}$ so that $P \bmod \mathfrak{p}_i = Q \bmod \mathfrak{p}_i$ for $i = 1, \ldots, s$. However, $V_{p,q}$ may be a strict subset of V_p. In this way, we can get extra information by combining data at distinct residue characteristics.

Note that this argument can be used cumulatively. Furthermore, $\{P \in V_p : (P + \Lambda_p) \cap V_{q,q'}\}$ may be a proper subset of $V_{p,q} \cap V_{p,q'}$. Of course, if we combine information modulo Λ_p and Λ_q, the resulting information is most naturally expressed modulo $\Lambda_p \cap \Lambda_q$. This information tends to be much more bulky, though. In practice it seems to be preferable to just keep the information modulo Λ_p.

Lemma 5. *For $i = 1, 2, 3$, we have that the only solutions of $\delta_i y^2 = R(x)$ with $y \in K$ and $x \in \mathbf{Q}$, have x as indicated in the table below.*

i	*x-coordinates*
1	$1, 10$
2	$2/3$
3	∞

Proof. Again, the computations involved are automated in [B5]. For E_1, the desired result can be obtained by a Chabauty argument using the primes above 1439 augmented with congruence information at 947. For E_2, a straightforward argument at 5 suffices. For E_3, we had trouble finding one rational prime that yields enough information. Here we combined a Chabauty argument with congruence information at 1439 and 947. This also involves first combining the information at 1439 and 947 before combining it with the information at 71. Since this procedure is not fully automated in [B5], we give some detail here on how to proceed. Since the output format of the routines is rather bulky, the following output is edited for brevity. If the reader is interested in the full details, he or she is referred to [BE].

First we define the cover. Note that we apply $x \mapsto 1/x$, so that $x = \infty$ corresponds to $x = 0$ in this session.

```
kash> O:=OrderMaximal(x^5 - 2*x^4 + 3*x^3 - 4*x^2 + 5*x - 6);;
kash> ec:=Ell(1,0,0,0,Elt(O,[3501, -2562, 738, -1584, 795]),
>   Elt(O,[2982, -39856, 3894, -6586, 14068]));;
kash> P1:=EllXtoPnt(ec,Elt(O,[-104, 156, 2, 30, -47] / 4));;
kash> P2:=EllXtoPnt(ec,Elt(O,[3462,-1026,601,-1374,261])/25);;
kash> P3:=EllXtoPnt(ec,Elt(O,[-66, 34, 0, 24, -13] / 9));;
kash> EllGenInit([P1,P2,P3],3);;
kash> cov:=QuarCov(HypEllRev(deltas[3]*Rpol),0,ec);;
kash> Unbind(cov.IsEllDblCov);
```

Next we compute V_{71}, V_{947} and V_{1439}.

```
kash> L1:=EllCovFibStrict(cov,PlaceSupport(71*O));;
Warning. Results only valid if 2 is prime to index in MW-group.
Result of FibStrict:
 [71, [ 6, 9, 18, 19, 22, 24, 28, 29, 34, 35, 37, 38, 40, 42, 44,
        46, 47, 57, 60, 70, 0 ] ]
kash> L2:=EllCovFibStrict(cov,PlaceSupport(947*O));;
Warning. Results only valid if 2 is prime to index in MW-group.
Result of FibStrict:
[947, [ 14, 37, 50, 149, 151, 162, 218, 225, 250, 274, 288, 333,
    357, 369, 373, 395, 397, 450, 466, 480, 612, 625, 636, 652,
    656, 692, 767, 776, 812, 825, 826, 838, 844, 857, 944, 0 ] ]
kash> L3:=EllCovFibStrict(cov,PlaceSupport(1439*O));;
Warning. Results only valid if 4 is prime to index in MW-group.
Result of FibStrict:
[1439, [ 24, 55, 79, 98, 112, 181, 183, 265, 289, 368, 369, 413,
    471, 527, 540, 570, 589, 611, 635, 695, 726, 731, 787, 848,
    865, 910, 944, 973, 978, 987, 1049, 1077, 1097, 1134, 1226,
    1261, 1271, 1337, 1359, 1377, 0 ] ]
```

The printed information gives $x(P) \bmod p$ for $P \in V_p$. Internally, more information is stored in L1,L2,L3, but that information is too bulky to print. The program notes that $[E(K) : \langle P_1, \ldots, P_r\rangle]$ should not be divisible by 2. Lemma 4 ensures this.

Next, we determine $V_{947,1439}$ and combine that information with V_{71}.

```
kash> L23:=EllCovFibSect(L2,L3);;
Fiber intersection yields:[ 947, [ 450, 0 ] ]
kash> L123:=EllCovFibSect(L1,L23);;
Fiber intersection yields:[ 71, [ 0 ] ]
```

Note that $V_{947,1439}$ indicates only 2 possible residue classes for $x(P) \bmod 947$, while V_{947} indicates 36 possible residue classes. This information combined with V_{71} leaves only one residue class for $x(P) \bmod 71$. Here we check using a power series argument that there are no points 71-adically close to $O \in E(K)$ that have a rational image under x.

```
kash> EllCovThetaTest(cov,PlaceSupport(71*O),EllZero(ec));
Computing Theta^G for G=( 0: 1: 0 )...
G is only point in fiber if the following matrix has maximal rank mod 71
[41 31 10]
[42 70 29]
[24 59 33]
[67  1 52]
true
```

We see that any $P \in C(\mathbf{Q})$ has $x(P) \in \{1, 10, 2/3, \infty\}$. Theorem 2 follows.

3.2 A Line of Attack for Conjecture 1

An isomorphic model for $B_8(G_{1344})$ is

$$C : y^2 = x^6 - 3x^5 + 25x^4/4 - 6x^3 + 20x^2 + 4.$$

To decide Conjecture 1, we need to determine $C(\mathbf{Q})$. Similar to $B_7(G_{168})$, the Mordell-Weil group of the Jacobian of C has rank 2. Therefore, a direct Chabauty argument will not work. In principle, we can apply the method from the previous section. To factor $x^6 - 3x^5 + 25x^4/4 - 6x^3 + 20x^2 + 4$ into a quadratic factor $Q(x)$ and a quartic factor $R(x)$, we need a degree 15 extension K.

To get the analogue of Lemma 2, one could use the information on the 2-Selmer group of the Jacobian of C, together with local arguments.

For the analogue of Lemma 4, one would in general need a further degree 3 extension in order to perform a 2-descent. This would lead to a degree 45 extension. Classgroup information is probably not feasibly computable for such a field. For C, things are not that grim, though. The Galois-group of $x^6 - 3x^5 + 25x^4/4 - 6x^3 + 20x^2 + 4$ is S_5 acting transitively on the 6 roots. As a consequence, the Jacobians of the curves $\delta y^2 = R(x)$ have a 2-torsion point over K. This enables us to do a 2-isogeny descent. We only need classgroup-information of K. Surprisingly, the Minkowski-bound for K is only 196195. The classgroup information of K is unconditionally computable (K has class number 1).

Thus we can get upper bounds for the ranks of the elliptic curves involved. Actually finding the Mordell-Weil groups, however, involves finding rational points on elliptic curves over a degree 15 extension of $\mathbf{Q}$. Also, the rank bounds obtained by a 2-(isogeny-)descent are not necessarily sharp. With present techniques, solving this equation by the above method would involve an inordinate amount of luck. We did not have the courage to test ours.

References

BCP. Wieb Bosma, John Cannon, and Catherine Playoust: The Magma algebra system. I. The user language. *J. Symbolic Comput.*, 24(3-4):235–265, 1997. Computational algebra and number theory (London, 1993).

B1. Nils Bruin: *Chabauty Methods and Covering Techniques applied to Generalised Fermat Equations.* PhD thesis, Universiteit Leiden, 1999.

B2. Nils Bruin: Chabauty methods using elliptic curves. Technical Report W99–14, Leiden, 1999.

B3. Nils Bruin: Chabauty methods using covers on curves of genus 2. Technical Report W99–15, Leiden, 1999.

B4. Nils Bruin: On powers as sums of two cubes in Wieb Bosma (ed), *Algorithmic Number Theory* 4th International Symposium ANTS-IV Leiden, The Netherlands, July 2-7, 2000 Proceedings. Springer LNCS 1838.

B5. Nils Bruin: Algae, a program for 2-Selmer groups of elliptic curves over number fields. see `http://www.cecm.sfu.ca/~bruin/ell.shar`.

BE. Nils Bruin and Noam Elkies: Transcript of computations. available from `http://www.math.harvard.edu/~elkies/trinomials_bruin.g`, 2002.

BF. Nils Bruin and E. Victor Flynn: Towers of 2-covers of hyperelliptic curves. PIMS-01-12, `http://www.pims.math.ca/publications/#preprints`, 2001.

Ca. J. W. S. Cassels: *Lectures on Elliptic Curves.* LMS-ST 24. University Press, Cambridge, 1991.

Co. Henri Cohen: *A Course in Computational Algebraic Number Theory*, GTM 138 Springer, Berlin–Heidelberg–New York, 1993.

EFM. Erbach, D.W., Fischer J., and McKay, J.: Polynomials with Galois group PSL(2,7), *J. Number Theory* **11** (1979), 69–75.

Fa1. Faltings, G.: Endlichkeitssätze für abelsche Varietäten über Zahlkörpern, *Invent. Math.* **73** (1983), 349–366.

Fa2. Faltings, G.: Diophantine approximation on Abelian varieties, *Annals of Math.* (2) **133** (1991) #3, 549–576.

Fl. E.V. Flynn: A flexible method for applying Chabauty's theorem. *Compositio Mathematica*, 105:79–94, 1997.

H. Florian Heß: Zur Klassengruppenberechnung in algebraischen Zahlkörpern. Diplomarbeit, Technische Universität Berlin, 1996. `http://www.math.tu-berlin.de/~kant/publications/diplom/hess.ps.gz`.

K. M. Daberkow, C. Fieker, J. Klüners, M. Pohst, K. Roegner, M. Schörnig, K. Wildanger: KANT V4, *J. of Symbolic Comput.*, 3-4:267–283, 1997.

M. Matzat, B.H.: *Konstruktive Galoistheorie.*, Springer Lect. Notes Math. **1284**, 1987.

Si. Joseph H. Silverman: *The Arithmetic of Elliptic Curves.* GTM 106. Springer-Verlag, 1986.

St. Michael Stoll: Implementing 2-descent for Jacobians of hyperelliptic curves. *Acta Arith.*, 98(3):245–277, 2001.

T. Trinks, W.: Ein Beispiel eines Zahlkörpers mit der Galoisgruppe $\mathrm{PSL}(3,2)$ über $\mathbf{Q}$, manuscript, Univ. Karlsruhe, Karlsruhe, 1968.

Computations on Modular Jacobian Surfaces

Enrique González-Jiménez[1,⋆], Josep González[2,⋆⋆], and Jordi Guàrdia[2,⋆⋆⋆]

[1] Department de Matemàtiques, Universitat Autònoma de Barcelona,
E-08193 Bellaterra, Barcelona, Spain
enrikegj@mat.uab.es
http://mat.uab.es/enrikegj/

[2] Escola Universitària Politècnica de Vilanova i la Geltrú,
Av. Víctor Balaguer s/n, E-08800 Vilanova i la Geltrú, Spain
{josepg, guardia}@mat.upc.es

Abstract. We give a method for finding rational equations of genus 2 curves whose jacobians are abelian varieties A_f attached by Shimura to normalized newforms $f \in S_2(\Gamma_0(N))$. We present all the curves corresponding to principally polarized surfaces A_f for $N \leq 500$.

1 Introduction

Given a normalized newform $f = \sum_{n>0} a_n q^n \in S_2(\Gamma_0(N))$, Shimura [5]-[6] attaches to it an abelian variety A_f defined over $\mathbb{Q}$ of dimension equal to the degree of the number field $E_f = \mathbb{Q}(\{a_n\})$. The Eichler-Shimura congruence makes it possible to compute at every prime $p \nmid N$ the characteristic polynomial of the Frobenius endomorphism acting on the Tate module of $A_f/\mathbb{F}_p$ from the coefficient a_p and its Galois conjugates. In consequence, when A_f is $\mathbb{Q}$-isogenous to the jacobian of a curve C defined over $\mathbb{Q}$, the number of points of the reduction of this curve mod a prime p of good reduction can be obtained from the characteristic polynomial of the Hecke operator T_p acting on $H^0(A_f, \Omega^1)$. Among these *jacobian-modular curves*, those which are hyperelliptic of low genus are especially interesting for public key cryptography.

As an optimal quotient of the jacobian of $X_0(N)$, $J_0(N)$, the abelian variety A_f has a natural polarization induced from $J_0(N)$. We will focus our attention on polarized surfaces A_f which are $\mathbb{Q}$-isomorphic to jacobians of genus 2 curves. Wang [7] gave a first step in the determinations of such curves. More precisely, using modular symbols he computed the periods of f and its Galois conjugate and presented A_f as a complex torus with an explicit polarization. For those principally polarized A_f, Wang computed numerically Igusa invariants by means of even Thetanullwerte and built an hyperelliptic curv e $C/\mathbb{Q}$ such that $\operatorname{Jac} C \simeq A_f$ over $\overline{\mathbb{Q}}$. The curves C obtained with this procedure have two drawbacks: they have huge coefficients, and, moreover, we only know that their jacobians

[⋆] The first author was supported in part by DGI Grant BHA2000-0180.
[⋆⋆] The second author was supported in part by DGI Grant BFM2000-0794-C02-02.
[⋆⋆⋆] The third author was supported in part by DGICYT Grant BFM2000-0627.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 189–197, 2002.

are $\overline{\mathbb{Q}}$-isomorphic to the corresponding abelian varieties A_f, but we don't know whether they are $\mathbb{Q}$-isomorphic, or even $\mathbb{Q}$-isogenous. Frey and Muller [2] looked for a curve $C'/\mathbb{Q}$ among the twisted curves of C such that the local factors of the L-series of $\operatorname{Jac} C'$ and A_f agree for all primes less than a large enough bound.

In this paper we want to go one step further in the determination of these jacobian modular surfaces. We describe a more arithmetical and efficient method, based on odd Thetanullwerte, which solves the problem up to numerical approximations. Our method provides equations $C_F : y^2 = F(x)$ with $F(x) \in \mathbb{Q}[x]$ such that $\operatorname{Jac} C_F$ or $\operatorname{Jac} C_{-F}$ is A_f. The sign is chosen using the Eichler-Shimura congruence.

We have implemented a program in Magma to determine modular jacobian surfaces and equations for the corresponding curves. We have found all the modular jacobian surfaces of level $N \leq 500$. The equations obtained for the corresponding curves are presented at the end of the paper. It is remarkable that almost all of them are minimal equations over $\mathbb{Z}[1/2]$.

2 Theoretical Foundations

A polarized abelian variety (A, Θ) of dimension g defined over $\mathbb{C}$ can be realized as a complex torus $T_A = \mathbb{C}^g/\Lambda$, where Λ is the period lattice of A with respect to a basis of $H^0(A, \Omega^1)$, with a nondegenerate Riemann form defined on Λ. We choose a symplectic basis for Λ, and write it as a $2g \times g$ matrix $\Omega = (\Omega_1|\Omega_2)$. The normalized period matrix $Z = \Omega_1^{-1}\Omega_2$ satisfies the Riemann conditions $Z = {}^tZ$, $Y = \operatorname{Im} Z$ is positive definite and the Riemann theta function:

$$\theta(z) := \theta(z; Z) := \sum_{n \in \mathbb{Z}^g} \exp(\pi i {}^t n.Z.n + 2\pi i {}^t n.z)$$

is holomorphic in $\mathbb{C}^g$. The values of the Riemann theta function at 2-torsion points are called Thetanullwerte. Historically, only the even Thetanullwerte, i.e., the values of the theta function at even 2-torsion points have been studied, since the values at odd 2-torsion points are always zero. Anyway, the values of the derivatives of the theta function at the odd 2-torsion points have nice properties, and also do provide useful geometrical information ([4]).

We now give the theoretical results which allow one to recognize when a principally polarized abelian surface is the jacobian of a genus 2 curve.

Proposition 1. *Let (A, Θ) be an irreducible principally polarized abelian surface defined over a number field K. There exists a hyperelliptic curve C of genus 2 defined over K such that $A = \operatorname{Jac} C$.*

Proof: It is well known that the irreducibility of A implies that $A = \operatorname{Jac} C$ for a certain hyperelliptic curve C defined over $\mathbb{C}$. But for genus 2 curves, the Abel-Jacobi map in degree 1 is an isomorphism between the curve C and the Θ divisor in $\operatorname{Jac} C = A$. Hence, we can assume that $C = \Theta$, which is defined over K. □

Proposition 2. *A principally polarized abelian surface (A, Θ) is not irreducible if and only if there is an even 2-torsion point P such that the corresponding even Thetanullwerte vanishes.*

Proof: If (A, Θ) is irreducible principally polarized, then it is isomorphic to the jacobian of a hyperelliptic genus 2 curve, and hence every even Thetanullwerte is non-zero.

Conversely, assume that (A, Θ) is the product of two elliptic curves E_1, E_2. This means that the theta function θ_A associated to the pair (A, Θ) is equal to $\theta_1\theta_2$, where we denote by θ_i the theta function associated to the elliptic curve E_i. Let O_i be the zero point in E_i, which is the unique odd 2-torsion point in E_i. The pair $O = (O_1, O_2) \in E_1 \times E_2$ gives an even two torsion point in A, which satisfies $\theta_A(O) = 0$. □

Once we know that a principally polarized abelian surface A is a jacobian, we want a method to find a curve C such that $A \simeq \operatorname{Jac} C$. We would like to be careful enough to assure that, when A is defined over a number field K, the curve C and the isomorphism $A \simeq \operatorname{Jac} C$ are also defined over K. The following result, which can be found in [4], will be basic for our purpose.

Theorem 1. *Let $F(X) = a_6X^6 + a_5X^5 + \ldots + a_1X + a_0 \in \mathbb{C}[X]$ be a separable polynomial of degree 5 or 6. Let $\Omega = (\Omega_1|\Omega_2)$ be the period matrix of the hyperelliptic curve $C_F : y^2 = F(x)$ with respect to the basis $\omega_1 = \frac{dx}{y}$, $\omega_2 = \frac{xdx}{y}$ of $H^0(C_F, \Omega^1)$ and any symplectic basis of $H_1(C_F, \mathbb{Z})$, and take $Z_F = \Omega_1^{-1}\Omega_2$.*

a) The roots α_k of the polynomial F are the ratios $\frac{x_{k,2}}{x_{k,1}}$, given by the solutions $(x_{k,1}, x_{k,2})$ of the six homogeneous linear equations

$$\left(\frac{\partial\theta}{\partial z_1}(w_k)\ \frac{\partial\theta}{\partial z_2}(w_k) \right) \Omega_1^{-1} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = 0,$$

where $w_1, \ldots, w_6$ are the six odd 2-torsion points of $J(C_F)$, given by

$$w_1 = \tfrac{1}{2}Z_F\begin{pmatrix}0\\1\end{pmatrix} + \tfrac{1}{2}\begin{pmatrix}0\\1\end{pmatrix},\quad w_2 = \tfrac{1}{2}Z_F\begin{pmatrix}0\\1\end{pmatrix} + \tfrac{1}{2}\begin{pmatrix}1\\1\end{pmatrix},$$
$$w_3 = \tfrac{1}{2}Z_F\begin{pmatrix}1\\0\end{pmatrix} + \tfrac{1}{2}\begin{pmatrix}1\\0\end{pmatrix}, w_4 = \tfrac{1}{2}Z_F\begin{pmatrix}1\\0\end{pmatrix} + \tfrac{1}{2}\begin{pmatrix}1\\1\end{pmatrix},$$
$$w_5 = \tfrac{1}{2}Z_F\begin{pmatrix}1\\1\end{pmatrix} + \tfrac{1}{2}\begin{pmatrix}0\\1\end{pmatrix}, w_6 = \tfrac{1}{2}Z_F\begin{pmatrix}1\\1\end{pmatrix} + \tfrac{1}{2}\begin{pmatrix}1\\0\end{pmatrix}.$$

When $\deg F = 5$, one of these ratios is infinity and we discard it.

b) Let $W_j = (\alpha_j, 0)$ be the Weierstrass point corresponding to w_j. Denote by $H[W_j]$ the hyperplane of $\mathbb{P}^1$ given by the equation

$$H[W_j](X_1, X_2) := \left(\frac{\partial\theta}{\partial z_1}(w_j)\ \frac{\partial\theta}{\partial z_2}(w_j) \right) \Omega_1^{-1} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

The discriminant $\Delta_{alg}(C_F)$ of the polynomial F satisfies the relation

$$\Delta_{alg}(C_F)^7 = 2^{120} a_6^{10} \pi^{60} \det \Omega_1^{-30} \prod_{j<k} H[W_j](1,\alpha_k)^2 \text{ if } \deg(F) = 6;$$

$$\Delta_{alg}(C_F)^5 = 2^{80} a_5^{10} \pi^{80} \det \Omega_1^{-20} \prod_{j<k} H[W_j](1,\alpha_k)^2 \text{ if } \deg(F) = 5.$$

3 Determination of Hyperelliptic Equations

We explain here how one can, given an irreducible abelian surface (A, Θ) defined over K, look for a hyperelliptic curve $C_F : Y^2 = F(X)$ such that A is K-isomorphic to $\operatorname{Jac} C_F$. We have divided our method into four steps.

Step 1: Period matrix. The first step consists in choosing a suitable period matrix Ω for A. We have to fix a symplectic basis of $H_1(A, \mathbb{Z})$, a convenient basis of $H^0(A, \Omega^1_{A/K})$ and compute the corresponding period matrix. The following result assures us that the basis of regular differentials can be chosen arbitrarily.

Proposition 3. *([3]). Let C/K be a genus 2 curve. For every linearly independent pair of regular differentials $\omega_1, \omega_2 \in H^0(C, \Omega^1_{C/K})$, there exists a polynomial $F(X) \in K[X]$ of degree 5 or 6 without double roots such that the functions on C given by*

$$x = \frac{\omega_1}{\omega_2}, \quad y = \frac{dx}{\omega_2}$$

satisfy the equation $y^2 = F(x)$.

Step 2: Weierstrass points. In this step, we compute the roots α_k of the polynomial F given by the first part of the theorem 1, and we take the monic polynomial $F_0(X) = \prod_k (X - \alpha_k) \in K[X]$.

Step 3: Leading coefficient. With the formulas given for the discriminant in part *b)* of theorem 1, we obtain $a_6^{10} \in K$ (or $a_5^{10} \in K$ if $\deg F_0 = 5$). We choose one of the tenth roots $a_6' \in K$ of this value and take the polynomial $F_1(X) = a_6' F_0(X) \in K[X]$.

Step 4: Hyperelliptic equation. At this point, it only remains to find the tenth root of unity ζ such that $F = \zeta F_1$. Since the curves C_F and $C_{\lambda^2 F}$ with $\lambda \in K^*$ are K-isomorphic, it suffices to consider only the cases $\zeta = 1$ and $\zeta = -1$, when $-1 \notin K^2$. First we check whether C_F and C_{-F} are K-isomorphic. If they are not, then we look if $\operatorname{Jac} C_F$ and $\operatorname{Jac} C_{-F}$ are not K-isogenous. In this case, by Faltings Theorem, only one of their L-series will agree with the L-series of A and this will give the right sign for $F = \pm F_1$. In fact, it will suffice to find a prime $\mathfrak{p}$ in K of good reduction for the curves C_F and C_{-F} such that their reductions mod $\mathfrak{p}$ have a different number of points.

In the case that C_F and C_F are not K-isomorphic and $\operatorname{Jac} C_F$ and $\operatorname{Jac} C_{-F}$ are K-isogenous, we cannot determine the right sign. Anyway, we know that both jacobians $\operatorname{Jac} C_F$ and $\operatorname{Jac} C_{-F}$ are $K(\sqrt{-1})$-isomorphic to A_f, and one of them is K-isomorphic.

4 Modular Computations

We apply the method described in the previous section to present the irreducible principally polarized two-dimensional factors of $J_0(N)^{\text{new}}$ as jacobians of curves, for $N \leq 500$.

In order to do this, we begin looking for the normalized newforms $f = \sum a_n q^n \in S_2(\Gamma_0(N))$ such that the number field $E_f = \mathbb{Q}(\{a_n\})$ is quadratic. For each of these newforms, we take an integral basis of the $\mathbb{C}$-vector space generated by f and its Galois conjugate ${}^\sigma f$. We also determine a symplectic basis of $H_1(A_f, \mathbb{Z})$. If A_f is principally polarized, we compute the period matrix with respect to these bases, using the package on modular symbols written by W. Stein in `Magma`.

Next, we check the irreducibility of A_f by means of proposition 2. We remark that all the A_f studied are irreducible.

We now apply the method of section 3. We follow the steps described there, to find the corresponding curves $C_F : Y^2 = F(X)$. Since we are working over $\mathbb{Q}$, we can change the polynomial $F(X)$ in order to obtain an integral equation. We multiply $F(X)$ by $d = t/b$, where $t \in \mathbb{Z}$ is the square of the l.c.m. of the denominators of the coefficients of F, and $b \in \mathbb{Z}$ is the g.c.d. of their numerators divided by its maximum square-free factor. It is worth remarking that the equations obtained have very small coefficients, even before finding the integral model.

The only case in which we have found a curve C_F such that $\operatorname{Jac} C_F$ and $\operatorname{Jac} C_{-F}$ are $\mathbb{Q}$-isogenous occurs for $N = 256$, but in fact both curves are already $\mathbb{Q}$-isomorphic, because the corresponding polynomial $F(X)$ is odd.

We have used three tests to check the correctness of our equations. First, we have computed the absolute Igusa invariants of the curves C_F in two different ways: algebraically from the coefficients of our equations, and numerically from the even Thetanullwerte of the period matrix. They have agreed to high accuracy in all cases. Second, we have compared the local factors of the L-series of $\operatorname{Jac} C_F$ and A_f for all primes $p < 100$ not dividing $\Delta_{alg}(C_F)$. Finally, we have computed the odd part of the conductor of C_F using the program `genus2reduction` by Q. Liu. In all cases, this odd part agreed with the odd part of the square of the level of the newform f, as it should by [1]. It is worth noting that in almost all cases our equations are minimal over $\mathbb{Z}[1/2]$.

We illustrate our computations with an example. The first level for which $J_0(N)^{\text{new}}$ has a proper two-dimensional factor is $N = 63$. Using `Magma` we identify the corresponding normalized newform f:

$$f = q + \sqrt{3}q^2 + q^4 - 2\sqrt{3}q^5 + q^7 - \sqrt{3}q^8 - 6q^{10} + 2\sqrt{3}q^{11} + 2q^{13} + \dots$$

An integral basis of the space $\langle f, {}^\sigma f \rangle$ is

$$f_1 = q + q^4 + q^7 - 6q^{10} + 2q^{13} + \dots, \qquad f_2 = q^2 - 2q^5 - q^8 + 2q^{11} + \dots$$

A basis for $H_1(A_f, \mathbb{Z})$ in terms of modular symbols is given by

$$\gamma_1 = \{-\tfrac{1}{24}, 0\} - \{-\tfrac{1}{28}, 0\} + \{-\tfrac{1}{30}, 0\} - \{-\tfrac{1}{51}, 0\} - \{-\tfrac{1}{3}, -\tfrac{2}{7}\},$$

$$\gamma_2 = \{-\tfrac{1}{24}, 0\} - \{-\tfrac{1}{28}, 0\} + \{-\tfrac{1}{39}, 0\} - \{-\tfrac{1}{57}, 0\} - \{-\tfrac{1}{6}, -\tfrac{1}{7}\},$$

$$\gamma_3 = \{-\tfrac{1}{24}, 0\} + \{-\tfrac{1}{39}, 0\} - \{-\tfrac{1}{45}, 0\} - \{-\tfrac{1}{60}, 0\} - \{-\tfrac{1}{3}, -\tfrac{2}{7}\} - \{\tfrac{3}{7}, \tfrac{4}{9}\},$$

$$\gamma_4 = \{-\tfrac{1}{36}, 0\} - \{-\tfrac{1}{49}, 0\} + \{-\tfrac{1}{51}, 0\} - \{-\tfrac{1}{54}, 0\} + \{-\tfrac{1}{57}, 0\}$$
$$-\{-\tfrac{1}{60}, 0\} - \{-\tfrac{1}{3}, -\tfrac{2}{7}\}.$$

Computing the intersection matrix of these paths we see that A_f is principally polarized. We find a symplectic basis for $H_1(A_f, \mathbb{Z})$, and compute the periods of f_1, f_2 with respect to these bases. We obtain as period matrix $\varOmega = (\varOmega_1 \mid \varOmega_2)$ for A_f:

$$\varOmega_1 = \begin{pmatrix} 0.3590439\ldots + i * 0.6218823\ldots & -2.2150442\ldots + i * 1.2788564\ldots \\ -2.2150442\ldots + i * 3.8365691\ldots & 1.0771318\ldots + i * 0.6218823\ldots \end{pmatrix},$$

$$\varOmega_2 = \begin{pmatrix} -1.4969563\ldots + i * 1.2788564\ldots & -1.8560003\ldots - i * 0.6569740\ldots \\ -3.3529566\ldots + i * 0.6218823\ldots & -1.1379124\ldots + i * 3.2146868\ldots \end{pmatrix}.$$

We apply the method described in section 3, to obtain the monic polynomial

$$F_0(x) = x^6 - 54x^3 - 27.$$

The coefficient a_6 is $1/12$, so that $F_1(x) = 1/12 F_0(x)$. The first prime for which the local factors of C_{F_1} and C_{-F_1} are different is $p = 67$. Comparing with the polynomial

$$x^2(x + p/x - a_p)(x + p/x -^\sigma a_p),$$

we see that the right sign is -1. We multiply $-F_1(x)$ by 6^2 to obtain an integral equation. We can finally assert that A_f is the jacobian of the curve

$$y^2 = -3x^6 + 162x^3 + 81.$$

The Igusa invariants of this curve are

$$i_1 = \frac{2^3 \cdot 37^5}{3 \cdot 7^3}, \qquad i_2 = -\frac{3 \cdot 37^3 \cdot 103}{2 \cdot 7^3} \qquad i_3 = -\frac{5 \cdot 37^2 \cdot 881}{2^3 \cdot 7^3}.$$

We have also computed these Igusa invariants from the even Thetanullwerte associated to the period matrix Z, obtaining, of course, the same result.

Using Q. Liu's program, we find a minimal equation for the curve C:

$$Y^2 = X^6 + 54X^3 - 27,$$

which is obtained from our equation through the change $x = 3/X$, $y = 9Y/X^3$, which corresponds essentially to a different ordering of the modular forms f_1, f_2 as basis of $\langle f, {}^\sigma f \rangle$.

5 Tables

We present the equations that we have obtained in the following table. We have labelled the irreducible principally polarized two-dimensional factors A_f of $J_0(N)^{\mathrm{new}}$ as S_{NX}. We have ordered the two-dimensional factors of $J_0(N)^{\mathrm{new}}$ following the output of the `Magma` function `SortDecomposition`. The letter X denotes the position of A_f with respect to this ordering. The third column indicates when we know that the given equation is minimal over $\mathbb{Z}[1/2]$.

A_f	$C_F : y^2 \,=\, F(x), \quad \mathrm{Jac}\, C_F \simeq A_f$	minimal?
S_{23A}	$y^2 = x^6 - 8x^5 + 2x^4 + 2x^3 - 11x^2 + 10x - 7$	yes
S_{29A}	$y^2 = x^6 - 4x^5 - 12x^4 + 2x^3 + 8x^2 + 8x - 7$	yes
S_{31A}	$y^2 = x^6 - 8x^5 + 6x^4 + 18x^3 - 11x^2 - 14x - 3$	yes
S_{63B}	$y^2 = -3x^6 + 162x^3 + 81$	
S_{65B}	$y^2 = -x^6 - 4x^5 + 3x^4 + 28x^3 - 7x^2 - 62x + 42$	yes
S_{65C}	$y^2 = -15x^6 + 36x^4 - 30x^3 + 72x^2 - 39$	yes
S_{67B}	$y^2 = x^6 + 2x^5 + x^4 - 2x^3 + 2x^2 - 4x + 1$	yes
S_{73B}	$y^2 = x^6 - 4x^5 + 2x^4 + 6x^3 + x^2 + 2x + 1$	yes
S_{87A}	$y^2 = x^6 - 2x^4 - 6x^3 - 11x^2 - 6x - 3$	yes
S_{93A}	$y^2 = x^6 + 2x^4 - 6x^3 + 5x^2 + 6x + 1$	yes
S_{103A}	$y^2 = x^6 + 2x^4 + 2x^3 + 5x^2 + 6x + 1$	yes
S_{107A}	$y^2 = x^6 + 2x^5 + 5x^4 + 2x^3 - 2x^2 - 4x - 3$	yes
S_{115B}	$y^2 = x^6 + 2x^4 + 10x^3 + 5x^2 + 6x + 1$	yes
S_{117B}	$y^2 = x^6 - 10x^3 - 27$	yes
S_{117C}	$y^2 = -3x^6 - 12x^4 - 18x^3 - 48x^2 - 36x - 27$	yes
S_{125A}	$y^2 = x^6 + 2x^5 + 5x^4 + 10x^3 + 10x^2 + 8x + 1$	yes
S_{125B}	$y^2 = 5x^6 - 10x^5 + 25x^4 - 50x^3 + 50x^2 - 40x + 5$	yes
S_{133A}	$y^2 = x^6 - 2x^5 + 5x^4 - 6x^3 + 10x^2 - 8x + 1$	yes
S_{133B}	$y^2 = -3x^6 - 22x^5 - 35x^4 + 50x^3 + 74x^2 - 100x + 29$	yes
S_{135D}	$y^2 = x^6 + 6x^4 - 10x^3 + 9x^2 - 30x - 11$	yes
S_{147D}	$y^2 = x^6 - 4x^4 + 2x^3 + 8x^2 - 12x + 9$	yes
S_{161B}	$y^2 = x^6 + 6x^5 + 17x^4 + 22x^3 + 26x^2 + 12x + 1$	yes
S_{167A}	$y^2 = x^6 - 4x^5 + 2x^4 - 2x^3 - 3x^2 + 2x - 3$	yes
S_{175E}	$y^2 = x^6 + 2x^5 - 3x^4 + 6x^3 - 14x^2 + 8x - 3$	yes
S_{177A}	$y^2 = x^6 + 2x^4 - 6x^3 + 5x^2 - 6x + 1$	yes
S_{177B}	$y^2 = -15x^6 - 120x^5 - 530x^4 - 710x^3 - 515x^2 - 30x + 45$	
S_{188B}	$y^2 = x^5 - x^4 + x^3 + x^2 - 2x + 1$	yes
S_{189E}	$y^2 = x^6 - 2x^3 - 27$	yes

A_f	$C_F : y^2 \,=\, F(x), \quad \mathrm{Jac}\, C_F \simeq A_f$	minimal?
S_{191A}	$y^2 \,= x^6 + 2x^4 + 2x^3 + 5x^2 - 6x + 1$	yes
S_{205D}	$y^2 \,= x^6 + 2x^4 + 10x^3 + 5x^2 - 6x + 1$	yes
S_{209B}	$y^2 \,= x^6 - 4x^5 + 8x^4 - 8x^3 + 8x^2 + 4x + 4$	yes
S_{213B}	$y^2 \,= x^6 + 2x^4 + 2x^3 - 7x^2 + 6x - 3$	yes
S_{221C}	$y^2 \,= x^6 - 2x^5 + x^4 + 6x^3 + 2x^2 + 4x + 1$	yes
S_{224C}	$y^2 \,= -2x^6 - 8x^5 - 34x^4 - 48x^3 - 118x^2 + 56x + 154$	yes
S_{224D}	$y^2 \,= 2x^6 - 8x^5 + 34x^4 - 48x^3 + 118x^2 + 56x - 154$	yes
S_{243C}	$y^2 \,= x^6 + 6x^3 - 27$	yes
S_{250D}	$y^2 \,=\, 20\,x^6 - 140\,x^5 + 325\,x^4 + 1050\,x^3 + 425\,x^2 + 160\,x + 80$	
S_{256E}	$y^2 \,=\, 2\,x^5 - 128\,x$	yes
S_{261A}	$y^2 \,= x^6 - 6x^4 + 10x^3 + 21x^2 - 30x + 9$	yes
S_{261B}	$y^2 \,= -3x^6 + 18x^4 + 30x^3 - 63x^2 - 90x - 27$	yes
S_{261D}	$y^2 \,= -3x^6 + 6x^4 - 18x^3 + 33x^2 - 18x + 9$	yes
S_{262C}	$y^2 \,= -8x^5 + 56x^4 - 82x^3 - 312x^2 - 264x - 64$	yes
S_{266B}	$y^2 \,=\, 8\,x^6 + 16\,x^5 + 13\,x^4 + 6\,x^3 - 19\,x^2 - 8\,x - 16$	yes
S_{268C}	$y^2 \,= x^6 - 2x^5 + x^4 - 4x^3 + 2x^2 + 4x + 1$	yes
S_{275G}	$y^2 \,= -3x^6 - 2x^5 + x^4 - 14x^3 + 2x^2 - 8x + 1$	yes
S_{279A}	$y^2 \,=\, -3\,x^6 - 6\,x^4 - 18\,x^3 - 15\,x^2 + 18\,x - 3$	yes
S_{279B}	$y^2 \,=\, -3\,x^6 + 6\,x^5 - 3\,x^4 - 6\,x^3 + 18\,x^2 - 12\,x + 9$	yes
S_{287A}	$y^2 \,= x^6 + 2x^5 - 3x^4 - 6x^3 - 10x^2 - 4x - 3$	yes
S_{292A}	$y^2 \,= -x^6 - 2x^5 - 4x^4 - 4x^3 - 3x^2 - 2x + 1$	yes
S_{297E}	$y^2 \,=\, x^6 - 12\,x^4 - 8\,x^3 + 12\,x^2 - 12\,x + 4$	yes
S_{297F}	$y^2 \,=\, -3\,x^6 + 36\,x^4 - 24\,x^3 - 36\,x^2 - 36\,x - 12$	yes
S_{299A}	$y^2 \,= -3x^6 - 10x^5 - 7x^4 + 6x^3 + 6x^2 - 4x + 1$	yes
S_{325H}	$y^2 \,=\, -75\,x^6 + 180\,x^4 + 150\,x^3 + 360\,x^2 - 195$	yes
S_{335B}	$y^2 \,= x^6 - 4x^5 - 48x^2 - 20x - 4$	yes
S_{345G}	$y^2 \,=\, x^6 - 12\,x^5 + 32\,x^4 + 24\,x^3 + 8\,x^2 - 12\,x + 4$	yes
S_{351A}	$y^2 \,= x^6 - 6x^4 + 18x^3 + 9x^2 - 18x + 5$	yes
S_{351C}	$y^2 \,=\, -3\,x^6 + 18\,x^4 + 54\,x^3 - 27\,x^2 - 54\,x - 15$	yes
S_{351D}	$y^2 \,=\, 21\,x^6 - 210\,x^5 + 525\,x^4 - 602\,x^3 + 714\,x^2 + 336\,x + 665$	
S_{357E}	$y^2 \,= x^6 + 8x^4 - 8x^3 + 20x^2 - 12x + 12$	yes
S_{375C}	$y^2 \,=\, 105\,x^6 + 240\,x^5 + 550\,x^4 + 450\,x^3 + 325\,x^2 + 90\,x - 155$	yes
S_{376A}	$y^2 \,=\, -x^5 - x^4 + 3\,x^3 + 3\,x^2 - 4\,x + 1$	yes
S_{376B}	$y^2 \,= x^5 - x^3 + 2x^2 - 2x + 1$	yes
S_{380D}	$y^2 \,= x^5 - 7x^3 - 4x^2 + 5x + 5$	yes

A_f	$C_F : y^2 = F(x), \quad \mathrm{Jac}\, C_F \simeq A_f$	minimal?
S_{387F}	$y^2 = -12x^6 + 162x^3 + 324$	
S_{389B}	$y^2 = x^6 + 10x^5 + 23x^4 - 20x^3 - 45x^2 + 46x - 11$	yes
S_{391A}	$y^2 = x^6 + 10x^4 - 6x^3 - 11x^2 + 18x - 7$	yes
S_{424A}	$y^2 = x^6 - 2x^5 + 6x^4 - 8x^3 + 10x^2 - 8x + 5$	yes
S_{440E}	$y^2 = x^5 + 2x^3 - 11x^2 - 8x - 24$	yes
S_{440G}	$y^2 = x^5 - 2x^3 - 7x^2 - 8x + 8$	yes
S_{441I}	$y^2 = -3\,x^6 + 12\,x^4 + 6\,x^3 - 24\,x^2 - 36\,x - 27$	yes
S_{464I}	$y^2 = -x^6 - 2x^5 - 7x^4 - 6x^3 - 13x^2 - 4x - 8$	yes
S_{476B}	$y^2 = x^5 + 2x^4 + 3x^3 + 6x^2 + 4x + 1$	yes
S_{476D}	$y^2 = x^5 - 2x^4 + 3x^3 - 6x^2 - 7$	yes
S_{483C}	$y^2 = x^6 + 12\,x^5 + 26\,x^4 - 34\,x^3 - 67\,x^2 + 90\,x - 27$	yes
S_{488A}	$y^2 = -3x^6 + 18x^5 - 27x^4 - 12x^3 - 27x^2 - 36x - 24$	yes

References

1. H. Carayol, *Sur les représentations l-adiques associées aux formes modulaires de Hilbert*, Ann. Sci. École Norm. Sup. (4) **19** (1986), no. 3, 409–468.
2. G. Frey and M. Müller, *Arithmetic of modular curves and applications*, Algorithmic algebra and number theory (Heidelberg, 1997), Springer, Berlin, 1999, pp. 11–48.
3. E. González-Jiménez and J. González, *Modular curves of genus* 2, to appear in Math. Comp.
4. J. Guàrdia, *Jacobian nullwerte and algebraic equations*, to appear in Journal of Algebra.
5. G. Shimura, *Introduction to the arithmetic theory of automorphic functions*, Publications of the Mathematical Society of Japan, No. 11. Iwanami Shoten, Publishers, Tokyo, 1971, Kanô Memorial Lectures, No. 1.
6. G. Shimura, *On the factors of the jacobian variety of a modular function field*, J. Math. Soc. Japan **25** (1973), 523–544.
7. X. D. Wang, 2-*dimensional simple factors of* $J_0(N)$, Manuscripta Math. **87** (1995), no. 2, 179–197.

Integral Points on Punctured Abelian Surfaces

Andrew Kresch and Yuri Tschinkel

[1] Department of Mathematics,
University of Pennsylvania,
Philadelphia, PA 19104,
kresch@math.upenn.edu
[2] Department of Mathematics,
Princeton University,
Princeton, NJ 08544,
ytschink@math.princeton.edu

Abstract. We study the density of integral points on punctured abelian surfaces. Linear growth rates are observed experimentally.

1 Introduction

Let V be a smooth projective algebraic variety over a number field K. We now ask whether there exists a finite extension K' of K such that K'-rational points are Zariski dense. This property is called potential density of rational points, and is known to hold, e.g., for abelian varieties, certain classes of Fano varieties, and certain K3 surfaces (see [6], [1] and the references therein). Potential density is conjecturally related to global geometric invariants of V, such as the Kodaira dimension [10].

An analogous question can be asked about integral points. Let (V, Z) be a projective variety and a proper subvariety, both defined over K. Choose models $(\mathcal{V}, \mathcal{Z})$ over the ring of integers $\mathfrak{o}_K$. Let S be a finite set of non-archimedean places of K. A rational point Q on V determines a section s_Q of the structure map from $\mathcal{V}$ to $\mathrm{Spec}(\mathfrak{o}_K)$. We say that the point Q is S-integral (with respect to $\mathcal{Z}$) if the section s_Q does not meet $\mathcal{Z}$ outside S. We say that integral points are potentially dense for the pair (V, Z) if there exists a finite extension K' of K, a finite set S' of non-archimedean places of K', and models $(\mathcal{V}', \mathcal{Z}')$ over $\mathrm{Spec}(\mathfrak{o}_{K'})$ of the base-changed (V', Z') such that S'-integral points on $(\mathcal{V}', \mathcal{Z}')$ are Zariski dense in V'. Concretely, this means that after a finite extension of the base field, and allowing for a finite set of bad places, a given system of integral equations for V has a Zariski dense set of integral solutions such that their reductions, outside the fixed bad places, are away from the reduction of Z (given also by integral equations).

Conjecture 1 ([7]). Let V be a smooth algebraic variety whose rational points are potentially dense. Then integral points are potentially dense with respect to any codimension ≥ 2 subvariety $Z \subset V$.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 198–204, 2002.

This conjecture holds, e.g., for toric varieties and Del Pezzo surfaces [7]. Conversely, knowing potential density of integral points for certain varieties, we may deduce potential density of rational points in many new cases. For instance, Conjecture 1 implies potential density for rational points on general K3 surfaces (see [7]). An important test of the above conjecture is the case of punctured abelian varieties (that is, pairs (J, Z), where J is an abelian variety and $Z \subset J$ a codimension ≥ 2 subvariety).

For punctured abelian surfaces potential density is only known when the abelian surface is special (e.g., isogenous to products of elliptic curves, or admitting extra endomorphisms, see [7]). Here we study the case of simple abelian surfaces J over $\mathbb{Q}$, punctured at one rational point (which we may as well take to be the identity) and having a point $Q \in J(\mathbb{Q})$ of infinite order. We carry out a simple numerical experiment which strongly suggests that integral points on punctured abelian surfaces are not only Zariski dense, but moreover constitute a positive proportion of the multiples of Q. It would be interesting to have a conceptual interpretation of the proportionality constant.

Acknowledgments

The authors would like to thank Fedor Bogomolov, Bredan Hassett, Barry Mazur and Joe Shalika for inspiring discussions. We especially thank Michael Stoll for a careful reading of the paper and for many useful suggestions. We have used the packages `magma` and `maple`. The first author was supported in part by a National Science Foundation Postdoctoral Research Fellowship, and the second author was supported in part by the Clay Foundation and the National Science Foundation.

2 Divison Polynomials in Genus 2

Let $f \in \mathbb{Z}[X]$ be a polynomial of degree $2g + 1$ with no multiple factors and C the hyperelliptic curve (over $\mathbb{Z}$), defined by the equation

$$Y^2 = f(X).$$

Let (x, y) be a $\mathbb{Q}$-rational point on C, with $y \neq 0$, and let $Q := [(x, y) - \infty]$ be the corresponding point on the Jacobian $J = J(C)$. Denote by $\Theta = \Theta(J)$ the Θ-divisor. Cantor [2] has described a convenient algorithm for generating division polynomials $\psi_r(x)$ which vanish if and only if $r \cdot Q \in \Theta$. Moreover, $r \cdot Q = 0$ in J if and only if $\psi_{r'}(x) = 0$ for all r' with $|r' - r| \leq g - 1$. These polynomials give an efficient means of testing at which primes a given multiple of Q reduces to the identity in (the reduction modulo some prime of) the Jacobian.

Before stating basic facts about division polynomials, let us recall how to represent a point on a Jacobian. From now on we specialize to the case $g = 2$. Every point on J is expressible in the form $D - 2 \cdot \infty$ for an effective degree 2 cycle D on C, and D is unique except in the case of the zero element of J. The point $r \cdot Q$ can be put into this form by solving for polynomials $A(X)$ and $B(X)$

such that $A(X) - B(X)y$ vanishes to order r at Q, subject to degree bounds $\deg A \leq \lfloor (r+2)/2 \rfloor$ and $\deg B \leq \lfloor (r-3)/2 \rfloor$. Then $r \cdot Q \in \Theta$ is equivalent to the vanishing of the leading coefficient of A in the case r is even, or of B in the case r is odd. Cantor shows that one can produce universal polynomials A and B, whose coefficients are integer polynomials in the coefficients of f and in x (and y).

Concretely, let us continue to assume that f has coefficients in $\mathbb{Z}$. Cantor's algorithm generates polynomials $P_r(x)$ and $\psi_r(x)$ such that:

(i) $P_r(x) = 0$ if and only if $r \cdot Q \in \Theta$ (for all x in the algebraic closure $\overline{\mathbb{Q}}$ of $\mathbb{Q}$), $\deg P_r = r^2 - 4$ when r is even, and $\deg P_r = r^2 - 9$ when r is odd (this specifies P_r uniquely, up to a scalar multiple).

(ii) Define $\psi_r(x)$ to be proportional to $P_r(x)$ when r is even and to $f(x)P_r(x)$ when r is odd, and to have leading coefficient $\binom{r+1}{3}$; then $\psi_r(x)$ is an integer-coefficient polynomial of degree $r^2 - 4$.

(iii) The ψ_r satisfy the following recurrence relation:

$$\psi_r \psi_s \psi_{s+r} \psi_{s-r} = \det \begin{pmatrix} \psi_{s-2}\psi_r & \psi_{s-1}\psi_{r+1} & \psi_s \psi_{r+2} \\ \psi_{s-1}\psi_{r-1} & \psi_s \psi_r & \psi_{s+1}\psi_{r+1} \\ \psi_s \psi_{r-2} & \psi_{s+1}\psi_{r-1} & \psi_{s+2}\psi_r \end{pmatrix} \tag{1}$$

for any $s \geq r$.

The recurrence (1) determines ψ_r for all $r \geq 8$, given $\psi_1 = 0$, $\psi_2 = 1$, ..., ψ_7. One can effectively determine the universal polynomials ψ_3, ..., ψ_7 by solving for the coefficients of the polynomials $A(X)$ and $B(X)$ mentioned previously, for each $r \leq 7$. This is achieved economically by introducing a new variable v given by $vf(x) = x - X$. Then $\sqrt{f(X)/f(x)}$ is a power series in v which is easily computed (for reason of convention, the branch $-1 + \cdots$ of the square root is chosen for $g = 2$). Then one is reduced to solving

$$v^r \mid a(v) - b(v)\sqrt{f(X)/f(x)} \tag{2}$$

for polynomials $a(v)$ and $b(v)$ satisfying the same degree bounds as above (a differs from A by the change of variable, and b differs from B by the change of variable and multiplication by y). In particular, $a(0)+b(0) = 0$. We have $a(0) = 0$ for given $x \in \overline{\mathbb{Q}}$ if and only if $P_{r-1}(x) = 0$, and we can take $-a(0) = b(0) = P_{r-1}$. This means that for $r \leq 6$, (2) reduces to solving at most one equation for one unknown coefficient, and this is easily solved. For instance, ψ_4 is displayed in Table 1. For $r = 7$, the two unknown coefficients of the quadratic polynomial $b(v)$ must be solved for.

3 Results

We performed the following numerical experiment. Start with a curve C of genus 2 defined by $Y^2 = f(X)$, where $f(X)$ is a monic degree-5 polynomial with integral coefficients. Assume that the Jacobian J is simple, has Mordell-Weil rank

Table 1. The universal $\psi_4(x)$

$$
\begin{aligned}
f(X) &= X^5 + \alpha X^4 + \beta X^3 + \gamma X^2 + \delta X + \varepsilon,\\
\psi_4(x) &= 10x^{12} + 24\alpha x^{11} + (26\beta + 16\alpha^2)x^{10} + 20(2\alpha\beta + \gamma)x^9\\
&\quad + 10(4\alpha\gamma + 3\beta^2 - \delta)x^8 + 80(\beta\gamma - \varepsilon)x^7\\
&\quad + (-112\alpha\varepsilon + 68\beta\delta + 64\gamma^2 + 8\alpha\beta\gamma - 2\beta^3 - 16\alpha^2\delta)x^6\\
&\quad + (-4\beta^2\gamma - 8\beta\varepsilon - 64\alpha^2\varepsilon - 8\alpha\beta\delta + 16\alpha\gamma^2 + 152\gamma\delta)x^5\\
&\quad + 10(-8\alpha\beta\varepsilon + 4\alpha\gamma\delta + 11\delta^2 + 12\gamma\varepsilon - \beta^2\delta)x^4\\
&\quad + 40(\alpha\delta^2 - \beta^2\varepsilon + 6\delta\varepsilon)x^3 + 10(\beta\delta^2 + 16\varepsilon^2 - 4\beta\gamma\varepsilon + 8\alpha\delta\varepsilon)x^2\\
&\quad + (8\beta\delta\varepsilon - 16\gamma^2\varepsilon + 64\alpha\varepsilon^2 + 4\gamma\delta^2)x + 16\beta\varepsilon^2 - 8\delta\gamma\varepsilon + 2\delta^3.
\end{aligned}
$$

1 (over $\mathbb{Q}$), and that there is an integral point (x, y) such that $Q = [(x, y) - \infty]$ has infinite order in J.

Let S be the set of prime divisors of $2\,y\,\mathrm{disc}(f)$. Now the curve reduces well modulo all primes not in S, and we have an integral model for J over $\mathrm{Spec}(\mathbb{Z}) \smallsetminus S$, with an S-integral point Q disjoint from the zero section. We count positive integers r such that $r \cdot Q$ is as well disjoint from the zero section (again, over the complement of S); such r will be called *good*. For $r \cdot Q$ to be disjoint from zero outside S is equivalent to $\psi_{r-1}(x)$, $\psi_r(x)$, and $\psi_{r+1}(x)$ having no common prime factors outside S. A table is made of the density of the good integers r. Amazingly, we observe linear growth.

Remark 1. The significance of any sort of growth is that the set of good integers being infinite implies Zariski density of S-integral points on the punctured J (here we use the fact that J is simple).

We describe the procedure in detail for one curve, and then present tables giving the data from several curves.

The curve C_1 given by

$$y^2 = x^5 - 14x^4 + 65x^3 - 112x^2 + 60x$$

has rational point $(3, 6)$, and its Jacobian J_1 satisfies $J_1(\mathbb{Q}) = \mathbb{Z} \oplus (\mathbb{Z}/2\mathbb{Z})^4$ (see [4]). Here $S = \{2, 3, 5\}$. Then we have (at $x = 3$)

$$\psi_3 = 144, \qquad \psi_4 = -41472, \qquad \psi_5 = 585252864,$$
$$\psi_6 = -35588725014528, \qquad \psi_7 = 5004999490025816064.$$

Notice that 7 is a common factor of ψ_5, ψ_6, and ψ_7, so that $6 \cdot Q$ is not S-integral on the punctured J_1. Hence 6 and all its multiples are not good. The next integer, besides multiples of 6, which fails to be good is 22. The third is 38:

$$\gcd(\psi_{37}, \psi_{38}, \psi_{39}) = 2^{854} \cdot 3^{344} \cdot 17.$$

Table 2. Densities of S-integral points on J_i

range of r	density(J_1)	density(J_2)	density(J_3)	density(J_4)
1– 100	0.77	0.62	0.74	0.67
101– 200	0.69	0.63	0.70	0.67
201– 300	0.71	0.61	0.74	0.66
301– 400	0.74	0.62	0.69	0.70
401– 500	0.72	0.62	0.69	0.68
501– 600	0.72	0.63	0.74	0.67
601– 700	0.73	0.60	0.70	0.64
701– 800	0.70	0.64	0.72	0.70
801– 900	0.72	0.59	0.73	0.68
901–1000	0.72	0.63	0.69	0.67

The first two columns of Table 2 show integer ranges (1–100, ..., 901–1000) and the density of good r in each range.

We performed a similar experiment with the following curves:

$$\begin{aligned}
&C_2 : f(X) = X^5 + 9X^4 + 14X^3 - 18X^2 - 15X + 9, && (x, y) = (0, 3),\\
&C_3 : f(X) = X^5 + 2X^4 - 3X^3 - 2X^2 + 2X, && (x, y) = (2, 6),\\
&C_4 : f(X) = X^5 + 11X^4 + 7X^3 - 89X^2 + 2X + 88, && (x, y) = (-7, 54).
\end{aligned}$$

By a computation in [3], these are curves having Jacobians of Mordell-Weil rank 1 over $\mathbb{Q}$. It is easy to see that the Jacobians we are considering are simple over $\mathbb{Q}$ (e.g., by factoring the number of $\mathbb{F}_p$-points for suitable p). The corresponding columns of Table 2 indicate the experimentally observed densities for these Jacobians.

4 Heuristics

Let J be an abelian variety over $\mathbb{Q}$, and let Γ be the Mordell-Weil group $J(\mathbb{Q})$. Fix an integral model of J, and let S be the set of primes of bad reduction. Then, for p a prime not in S, let us denote by g_p the order of the subgroup of $J(\mathbb{F}_p)$ generated by Γ. The quantity

$$\rho(J) = \prod_{p \notin S} (1 - 1/g_p). \tag{3}$$

is a lower bound for the density of S-integral points on the punctured Jacobian. We do not know whether this product converges.

Conjecture 2. If J is simple of dimension ≥ 2 and has positive Mordell-Weil rank, then the product (3) converges.

Remark 2. Replacing Γ by a finite-index subgroup does not change the convergence of (3). Also, note that the conclusion of Conjecture 2 may fail if J is isogenous to a product of elliptic curves.

We computed the Euler products using the first 400 primes of good reduction, for the Jacobians J considered above. In our computation we used the subgroup generated by our point Q in place of the full Mordell-Weil group to obtain a quantity $\tilde{\rho}(J)$ for each Jacobian J. Numerically we observe convergence. The results are presented in Table 3.

Table 3. Values of Euler products for J_i

	J_1	J_2	J_3	J_4
$\tilde{\rho}(J)$	0.576	0.404	0.538	0.516

Remark 3. For J of dimension 2, a positive answer to Conjecture 2 would imply the density of integral points.

One can ask, for some abelian variety, how often the reduction of the cyclic group generated by a given point is the full group $J(\mathbb{F}_p)$; for elliptic curves, this question was raised by Lang and Trotter in [8]. Assuming the Generalized Riemann Hypothesis (GRH), Serre showed that for elliptic curves E, the number of primes $p \leq B$ such that $E(\mathbb{Z}/p\mathbb{Z})$ is cyclic is $\sim cB/log(B)$ (as $B \to \infty$ and for some c). Again, under GRH, the density is

$$\sum_{n \geq 1} \mu(n)/[K_n : Q],$$

where $\mu(n)$ is the Möbius function and K_n is the field generated by n-torsion points on E (see [9]). An unconditional lower bound $\gg B/log(B)^2$ (for elliptic curves with no rational 2-torsion points) has been proved by Gupta and Murty [5].

References

1. Bogomolov, F.A., Tschinkel, Yu.: Density of rational points on elliptic $K3$ surfaces. Asian J. Math. **4** (2000) 351–368
2. Cantor, D.G.: On the analogue of the division polynomials for hyperelliptic curves. J. Reine Angew. Math. **447** (1994) 91–145
3. Flynn, E.V.: Descent via isogeny in dimension 2. Acta Arith. **66** (1994) 23–43
4. Gordon, D.M., Grant, D.: Computing the Mordell-Weil rank of Jacobians of curves of genus two. Trans. Amer. Math. Soc. **337** (1993) 807–824
5. Gupta, R., Murty, M.R.: Cyclicity and generation of points mod p on elliptic curves. Invent. Math. **101** (1990) 225–235
6. Harris, J., Tschinkel, Yu.: Rational points on quartics. Duke Math. J. **104** (2000) 477–500
7. Hassett, B., Tschinkel, Yu.: Density of integral points on algebraic varieties. In: Peyre, E., Tschinkel, Yu. (eds.): Rational Points on Algebraic Varieties. Progr. Math., Vol. 199. Birkhäuser, Basel (2001) 169–197

8. Lang, S., Trotter, H.: Primitive points on elliptic curves. Bull. Amer. Math. Soc. **83** (1977) 289–292
9. Serre, J.-P.: Propriétés galoisiennes des points d'ordre fini des courbes elliptiques. Invent. Math. **15** (1972) 259–331
10. Vojta, P.: Diophantine approximation and value distribution theory. Lecture Notes in Math., Vol. 1239. Springer-Verlag, Berlin (1987)

Genus 2 Curves with $(3,3)$-Split Jacobian and Large Automorphism Group

Tony Shaska

University of California at Irvine, Irvine, CA 92697.
tshaska@math.uci.edu
http://www.math.uci.edu/~tshaska

Abstract. Let $\mathcal{C}$ be a genus 2 curve defined over k, $char(k) = 0$. If $\mathcal{C}$ has a $(3,3)$-split Jacobian then we show that the automorphism group $Aut(\mathcal{C})$ is isomorphic to one of the following: $\mathbb{Z}_2, V_4, D_8$, or D_{12}. There are exactly six $\mathbb{C}$-isomorphism classes of genus two curves $\mathcal{C}$ with $Aut(\mathcal{C})$ isomorphic to D_8 (resp., D_{12}) and with $(3,3)$-split Jacobian. We show that exactly four (resp., three) of these classes with group D_8 (resp., D_{12}) have representatives defined over $\mathbb{Q}$. We discuss some of these curves in detail and find their rational points.

1 Introduction

Let $\mathcal{C}$ be a genus 2 curve defined over an algebraically closed field k, of characteristic zero. We denote by $K := k(\mathcal{C})$ its function field and by $Aut(\mathcal{C}) := Aut(K/k)$ the automorphism group of $\mathcal{C}$. Let $\psi : \mathcal{C} \to \mathcal{E}$ be a degree n maximal covering (i.e. does not factor through an isogeny) to an elliptic curve $\mathcal{E}$ defined over k. We say that $\mathcal{C}$ has a *degree n elliptic subcover*. Degree n elliptic subcovers occur in pairs. Let $(\mathcal{E}, \mathcal{E}')$ be such a pair. It is well known that there is an isogeny of degree n^2 between the Jacobian $J_\mathcal{C}$ of $\mathcal{C}$ and the product $\mathcal{E} \times \mathcal{E}'$. We say that $\mathcal{C}$ has **(n,n)-split Jacobian**. The locus of such $\mathcal{C}$ (denoted by $\mathcal{L}_n$) is an algebraic subvariety of the moduli space $\mathcal{M}_2$ of genus two curves. For the equation of $\mathcal{L}_2$ in terms of Igusa invariants, see [18]. Computation of the equation of $\mathcal{L}_3$ was the main focus of [17]. For $n > 3$, equations of $\mathcal{L}_n$ have not yet been computed.

Equivalence classes of degree 2 coverings $\psi : \mathcal{C} \to \mathcal{E}$ are in 1-1 correspondence with conjugacy classes of non-hyperelliptic involutions in $Aut(\mathcal{C})$. In any characteristic different from 2, the automorphism group $Aut(\mathcal{C})$ is isomorphic to one of the following: $\mathbb{Z}_2$, $\mathbb{Z}_{10}$, V_4, D_8, D_{12}, $\mathbb{Z}_3 \rtimes D_8$, $GL_2(3)$, or 2^+S_5; see [18]. Here V_4 is the Klein 4-group, D_8 (resp., D_{12}) denotes the dihedral group of order 8 (resp., 12), and $\mathbb{Z}_2, \mathbb{Z}_{10}$ are cyclic groups of order 2 and 10. For a description of other groups, see [18]. If $Aut(\mathcal{C}) \cong \mathbb{Z}_{10}$ then $\mathcal{C}$ is isomorphic to $Y^2 = X^6 - X$. Thus, if $\mathcal{C}$ has extra automorphisms and it is not isomorphic to $Y^2 = X^6 - X$ then $\mathcal{C} \in \mathcal{L}_2$. We say that a genus 2 curve $\mathcal{C}$ has **large automorphism group** if the order of $Aut(\mathcal{C})$ is bigger then 4.

In section 2, we describe the loci for genus 2 curves with $Aut(\mathcal{C})$ isomorphic to D_8 or D_{12} in terms of Igusa invariants. From these invariants we are able to

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 205–218, 2002.

determine the field of definition of a curve $\mathcal{C}$ with $Aut(\mathcal{C}) \cong D_8$ or D_{12}. Further, we find the equation for this $\mathcal{C}$ and j-invariants of degree 2 elliptic subcovers in terms of i_1, i_2, i_3 (cf. section 2). This determines the fields of definition for these elliptic subcovers.

Let $\mathcal{C}$ be a genus 2 curve with $(3, 3)$-split Jacobian. In section 3 we give a short description of the space $\mathcal{L}_3$. Results described in section 3 follow from [17], even though sometimes nontrivially. We find equations of degree 3 elliptic subcovers in terms of the coefficients of $\mathcal{C}$. In section 4, we show that $Aut(\mathcal{C})$ is one of the following: $\mathbb{Z}_2$, V_4, D_8, or D_{12}. Moreover, we show that there are exactly six $\mathbb{C}$-isomorphism classes of genus two curves $\mathcal{C} \in \mathcal{L}_3$ with automorphism group D_8 (resp., D_{12}). We explicitly find the absolute invariants i_1, i_2, i_3 which determine these classes. For each such class we give the equation of a representative genus 2 curve $\mathcal{C}$. We notice that there are four cases (resp., three) such that the triple of invariants $(i_1, i_2, i_3) \in \mathbb{Q}^3$ when $Aut(C) \cong D_8$ (resp., $Aut(C) \cong D_{12}$). Using results from section 2, we determine that there are exactly four (resp., three) genus 2 curves $\mathcal{C} \in \mathcal{L}_3$ (up to $\bar{\mathbb{Q}}$-isomorphism) with group D_8 (resp., D_{12}) defined over $\mathbb{Q}$ and list their equations in Table 1. We discuss these curves and their degree 2 and 3 elliptic subcovers in more detail in section 5. Our focus is on the cases which have elliptic subcovers defined over $\mathbb{Q}$. In some of these cases we are able to use these elliptic subcovers to find the rational points of the genus 2 curve. This technique has been used before by Flynn and Wetherell [5] for the degree 2 elliptic subcovers.

Curves of genus 2 with degree 2 elliptic subcovers (or with elliptic involutions) were first studied by Legendre and Jacobi. The genus 2 curve with the largest known number of rational points has automorphism group isomorphic to D_{12}; thus it has degree 2 elliptic subcovers. It was found by Keller and Kulesz and it is known to have at least 588 rational points; see [10]. Using degree 2 elliptic subcovers Howe, Leprevost, and Poonen [8] were able to construct a family of genus 2 curves whose Jacobians each have large rational torsion subgroups. Similar techniques probably could be applied using degree 3 elliptic subcovers. Curves of genus 2 with degree 3 elliptic subcovers have already occurred in the work of Clebsch, Hermite, Goursat, Burkhardt, Brioschi, and Bolza in the context of elliptic integrals. For a history of this topic see Krazer [11] (p. 479). For more recent work see Kuhn [12] and [17]. More generally, degree n elliptic subfields of genus 2 fields have been studied by Frey [6], Frey and Kani [7], Kuhn [12], and this author [16].

Acknowledgements: The author wants to thank Professor Fried for his continuous support.

2 Genus Two Curves with Extra Automorphisms and the Moduli Space $\mathcal{M}_2$

Let k be an algebraically closed field of characteristic zero and $\mathcal{C}$ a genus 2 curve defined over k. Then $\mathcal{C}$ can be described as a double cover of $\mathbb{P}^1(k)$ ramified in 6 places $w_1, \dots, w_6$. This sets up a bijection between isomorphism classes of genus 2 curves and unordered distinct 6-tuples $w_1, \dots, w_6 \in \mathbb{P}^1(k)$ modulo automorphisms of $\mathbb{P}^1(k)$. An unordered 6-tuple $\{w_i\}_{i=1}^6$ can be described by a binary sextic (i.e. a homogenous equation $f(X, Z)$ of degree 6). Let $\mathcal{M}_2$ denote the moduli space of genus 2 curves; see [15]. To describe $\mathcal{M}_2$ we need to find polynomial functions of the coefficients of a binary sextic $f(X, Z)$ invariant under linear substitutions in X, Z of determinant one. These invariants were worked out by Clebsch and Bolza in the case of zero characteristic and generalized by Igusa for any characteristic different from 2; see [1], [9].

Consider a binary sextic, i.e. a homogeneous polynomial $f(X, Z)$ in $k[X, Z]$ of degree 6:

$$f(X, Z) = a_6X^6 + a_5X^5Z + \cdots + a_0Z^6.$$

Igusa J-invariants $\{J_{2i}\}$ of $f(X, Z)$ are homogeneous polynomials of degree $2i$ in $k[a_0, \dots, a_6]$, for $i = 1, 2, 3, 5$; see [9], [18] for their definitions. Here J_{10} is simply the discriminant of $f(X, Z)$. It vanishes if and only if the binary sextic has a multiple linear factor. These J_{2i} are invariant under the natural action of $SL_2(k)$ on sextics. Dividing such an invariant by another one of the same degree gives an invariant under $GL_2(k)$ action.

Remark 1. There many sets of $SL_2(k)$ invariants of binary sextics. The J_{2i} invariants that we use were first defined by Igusa [9] and work in all characteristics. One can download a Maple package that computes J_{2i} from author's web site. For more information on other sets of invariants the reader can see the *Igusa Invariants* package in Magma written by E. Howe.

Two genus 2 fields K (resp., curves) in the standard form $Y^2 = f(X, 1)$ are isomorphic if and only if the corresponding sextics are $GL_2(k)$ conjugate. Thus if I is a $GL_2(k)$ invariant (resp., homogeneous $SL_2(k)$ invariant), then the expression $I(K)$ (resp., the condition $I(K) = 0$) is well defined. Thus the $GL_2(k)$ invariants are functions on the moduli space $\mathcal{M}_2$ of genus 2 curves. This $\mathcal{M}_2$ is an affine variety with coordinate ring $k[\mathcal{M}_2] = k[a_0, \dots, a_6, J_{10}^{-1}]^{GL_2(k)}$ which is the subring of degree 0 elements in $k[J_2, \dots, J_{10}, J_{10}^{-1}]$; see Igusa [9]. The *absolute invariants*

$$i_1 := 144\frac{J_4}{J_2^2}, \quad i_2 := -1728\frac{J_2J_4 - 3J_6}{J_2^3}, \quad i_3 := 486\frac{J_{10}}{J_2^5} \tag{1}$$

are even $GL_2(k)$-invariants. Two genus 2 curves with $J_2 \neq 0$ are isomorphic if and only if they have the same absolute invariants. If $J_2 = 0$ then we can define new invariants as in [17]. For the rest of this paper if we say "there is a genus 2 curve $\mathcal{C}$ defined over k" we will mean the k-isomorphism class of $\mathcal{C}$.

One can define $GL_2(k)$ invariants with J_{10} in the denominator which will be defined everywhere. However, this is not efficient in doing computations since the degrees of these rational functions in terms of the coefficients of $\mathcal{C}$ will be multiples of 10 and therefore higher then degrees of i_1, i_2, i_3. For the purposes of this paper defining i_1, i_2, i_3 as above is not a restriction as it will be seen in the proof of Theorem 1. For the proofs of the following two lemmas, see [18].

Lemma 1. *The automorphism group G of a genus 2 curve $\mathcal{C}$ in characteristic $\neq 2$ is isomorphic to $\mathbb{Z}_2$, $\mathbb{Z}_{10}$, V_4, D_8, D_{12}, $\mathbb{Z}_3 \rtimes D_8$, $GL_2(3)$, or 2^+S_5. The case when $G \cong 2^+S_5$ occurs only in characteristic 5. If $G \cong \mathbb{Z}_3 \rtimes D_8$ (resp., $GL_2(3)$) then $\mathcal{C}$ has equation $Y^2 = X^6 - 1$ (resp., $Y^2 = X(X^4 - 1)$). If $G \cong \mathbb{Z}_{10}$ then $\mathcal{C}$ has equation $Y^2 = X^6 - X$.*

Remark 2. For the analogue of the above lemma for $g > 2$ in characteristic zero see [13] where sophisticated methods of computational group theory are used.

For the rest of this paper we assume that $char(k) = 0$.

Lemma 2. *i) The locus $\mathcal{L}_2$ of genus 2 curves $\mathcal{C}$ which have a degree 2 elliptic subcover is a closed subvariety of $\mathcal{M}_2$. The equation of $\mathcal{L}_2$ is given by equation (17) in [18].*

ii) The locus of genus 2 curves $\mathcal{C}$ with $Aut(\mathcal{C}) \cong D_8$ is given by the equation of $\mathcal{L}_2$ and

$$1706 J_4^2 J_2^2 + 2560 J_4^3 + 27 J_4 J_2^4 - 81 J_2^3 J_6 - 14880 J_2 J_4 J_6 + 28800 J_6^2 = 0 \tag{2}$$

iii) The locus of genus 2 curves $\mathcal{C}$ with $Aut(\mathcal{C}) \cong D_{12}$ is

$$\begin{aligned} -J_4 J_2^4 + 12 J_2^3 J_6 - 52 J_4^2 J_2^2 + 80 J_4^3 + 960 J_2 J_4 J_6 - 3600 J_6^2 = 0 \\ 864 J_{10} J_2^5 + 3456000 J_{10} J_4^2 J_2 - 43200 J_{10} J_4 J_2^3 - 2332800000 J_{10}^2 - J_4^2 J_2^6 \\ -768 J_4^4 J_2^2 + 48 J_4^3 J_2^4 + 4096 J_4^5 = 0 \end{aligned} \tag{3}$$

We will refer to the locus of genus 2 curves $\mathcal{C}$ with $Aut(\mathcal{C}) \cong D_{12}$ (resp., $Aut(\mathcal{C}) \cong D_8$) as the D_8-locus (resp., D_{12}-locus).

Each genus 2 curve $\mathcal{C} \in \mathcal{L}_2$ has a non-hyperelliptic involution $v_0 \in Aut(\mathcal{C})$. There is another non-hyperelliptic involution $v_0' := v_0 \, w$, where w is the hyperelliptic involution. Thus, degree 2 elliptic subcovers come in pairs. We denote the pair of degree 2 elliptic subcovers by (E_0, E_0'). If $Aut(\mathcal{C}) \cong D_8$ then $E_0 \cong E_0'$ or E_0 and E_0' are 2-isogenous. If $Aut(\mathcal{C}) \cong D_{12}$, then E_0 and E_0' are isogenous of degree 3. See [18] for details. The parameterizations of the following lemma were pointed out by G. Cardona.

Lemma 3. *Let $\mathcal{C}$ be a genus 2 curve defined over k. Then,*

i) $Aut(\mathcal{C}) \cong D_8$ if and only if $\mathcal{C}$ is isomorphic to

$$Y^2 = X^5 + X^3 + tX \tag{4}$$

for some $t \in k \setminus \{0, \frac{1}{4}, \frac{9}{100}\}$.

ii) $Aut(\mathcal{C}) \cong D_{12}$ if and only if $\mathcal{C}$ is isomorphic to

$$Y^2 = X^6 + X^3 + t \tag{5}$$

for some $t \in k \setminus \{0, \frac{1}{4}, -\frac{1}{50}\}$.

Proof. i) $Aut(\mathcal{C}) \cong D_8$: Then $\mathcal{C}$ is isomorphic to

$$Y^2 = (X^2 - 1)(X^4 - \lambda X^2 + 1)$$

for $\lambda \neq \pm 2$; see [18]. Denote $\tau := \sqrt{-2\frac{\lambda+6}{\lambda-2}}$. The transformation

$$\phi : (X, Y) \to (\frac{\tau x - 1}{\tau x + 1}, \frac{4\tau}{(\tau x + 1)^3} \cdot \frac{(\lambda + 6)^2}{\lambda - 2})$$

gives

$$Y^2 = X^5 + X^3 + tX$$

where $t = \frac{1}{4}(\frac{\lambda-2}{\lambda+6})^2$ and $t \neq 0, \frac{1}{4}$. If $t = \frac{9}{100}$ then $Aut(\mathcal{C})$ has order 24.

Conversely, the absolute invariants i_1, i_2, i_3 of a genus 2 curve $\mathcal{C}$ isomorphic to

$$Y^2 = X^5 + X^3 + tX$$

satisfy the locus as described in Lemma 2, part ii). Thus, $Aut(\mathcal{C}) \cong D_8$.

ii) $Aut(\mathcal{C}) \cong D_{12}$: In [18] it is shown that $\mathcal{C}$ is isomorphic to

$$Y^2 = (X^3 - 1)(X^3 - \lambda)$$

for $\lambda \neq 0, 1$ and $\lambda^2 - 38\lambda + 1 \neq 0$. Then,

$$\phi : (X, Y) \to ((\lambda + 1)^{\frac{1}{3}} X, (\lambda + 1) Y)$$

transforms $\mathcal{C}$ to the curve with equation

$$Y^2 = X^6 + X^3 + t$$

where $t = \frac{\lambda}{(\lambda+1)^2}$ and $t \neq 0, \frac{1}{4}$. If $t = -\frac{1}{50}$ then $Aut(\mathcal{C})$ has order 48.

The absolute invariants i_1, i_2, i_3 of a genus 2 curve $\mathcal{C}$ isomorphic to

$$Y^2 = X^6 + X^3 + t$$

satisfy the locus as described in Lemma 2, part iii). Thus, $Aut(\mathcal{C}) \cong D_{12}$. This completes the proof.

□

The following lemma determines a genus 2 curve for each point in the D_8 or D_{12} locus.

Lemma 4. *Let* $\mathfrak{p} := (J_2, J_4, J_6, J_{10})$ *be a point in* $\mathcal{L}_2$ *such that* $J_2 \neq 0$ *and* (i_1, i_2, i_3) *the corresponding absolute invariants.*

i) If $\mathfrak{p}$ *is in the* D_8*-locus, then the genus two curve* $\mathcal{C}$ *corresponding to* $\mathfrak{p}$ *is given by:*

$$Y^2 = X^5 + X^3 - \frac{3}{4} \cdot \frac{345 i_1^2 + 50 i_1 i_2 - 90 i_2 - 1296 i_1}{2925 i_1^2 + 250 i_1 i_2 - 9450 i_2 - 54000 i_1 + 139968} X.$$

ii) If $\mathfrak{p}$ *is in the* D_{12}*-locus, then the genus two curve* $\mathcal{C}$ *corresponding to* $\mathfrak{p}$ *is given by:*

$$Y^2 = X^6 + X^3 + \frac{1}{4} \cdot \frac{540 i_1^2 + 100 i_1 i_2 - 1728 i_1 + 45 i_2}{2700 i_1^2 + 1000 i_1 i_2 + 204525 i_1 + 40950 i_2 - 708588}.$$

Proof. i) By the previous lemma every genus 2 curve $\mathcal{C}$ with automorphism group D_8 is isomorphic to $Y^2 = X^5 + X^3 + tX$. Since $J_2 \neq 0$ then $t \neq -\frac{3}{20}$ and the absolute invariants are:

$$i_1 = -144\, t\, \frac{(20t-9)}{(20t+3)^2}, \quad i_2 = 3456\, t^2\, \frac{(140t-27)}{(20t+3)^3}, \quad i_3 = 243\, t^3\, \frac{(4t-1)^2}{(20t+3)^5} \tag{6}$$

From the above system we have

$$t = -\frac{3}{4}\, \frac{345 i_1^2 + 50 i_1 i_2 - 90 i_2 - 1296 i_1}{2925 i_1^2 + 250 i_1 i_2 - 9450 i_2 - 54000 i_1 + 139968}.$$

ii) By the previous lemma every genus 2 curve $\mathcal{C}$ with automorphism group D_{12} is isomorphic to $Y^2 = X^6 + X^3 + t$. The absolute invariants are:

$$i_1 = 1296 \frac{t(5t+1)}{(40t-1)^2}, \quad i_2 = -11664 \frac{t(20t^2+26t-1)}{(40t-1)^3}, \quad i_3 = \frac{729}{16}\, \frac{t^2(4t-1)^3}{(40t-1)^5}. \tag{7}$$

From the above system we have

$$t = \frac{1}{4}\, \frac{540 i_1^2 + 100 i_1 i_2 - 1728 i_1 + 45 i_2}{2700 i_1^2 + 1000 i_1 i_2 + 204525 i_1 + 40950 i_2 - 708588}.$$

This completes the proof.

□

Note: If $J_2 = 0$ then there is exactly one isomorphism class of genus 2 curves with automorphism group D_8 (resp., D_{12}) given by $Y^2 = X^5 + X^3 - \frac{3}{20}X$ (resp., $Y^2 = X^6 + X^3 - \frac{1}{40}$).

Remark 3. If the invariants $i_1, i_2, i_3 \in \mathbb{Q}$ then from the lemma above there is a $\mathcal{C}$ corresponding to these invariants defined over $\mathbb{Q}$. If a genus 2 curve does not have extra automorphisms (i.e. $Aut(\mathcal{C}) \cong \mathbb{Z}_2$), then an algorithm of Mestre determines if the curve is defined over $\mathbb{Q}$.

If the order of the automorphism group $Aut(C)$ is divisible by 4, then $\mathcal{C}$ has degree 2 elliptic subcovers. These elliptic subcovers are determined explicitly in [18]. Do these elliptic subcovers of $\mathcal{C}$ have the same field of definition as $\mathcal{C}$? In general the answer is negative. The following lemma determines the field of definition of these elliptic subcovers when $Aut(\mathcal{C})$ is isomorphic to D_8 or D_{12}.

Lemma 5. *Let $\mathcal{C}$ be a genus 2 curve defined over k, $char(k) = 0$.*

i) If $\mathcal{C}$ has equation

$$Y^2 = X^5 + X^3 + tX,$$

where $t \in k \setminus \{\frac{1}{4}, \frac{9}{100}\}$, then its degree 2 elliptic subfields have j-invariants given by

$$j^2 - 128 \frac{2000t^2 + 1440t + 27}{(4t-1)^2} j + 4096 \frac{(100t-9)^3}{(4t-1)^3} = 0.$$

ii) If $\mathcal{C}$ has equation

$$Y^2 = X^6 + X^3 + t,$$

where $t \in k \setminus \{\frac{1}{4}, -\frac{1}{50}\}$, then its degree 2 elliptic subfields have j-invariants given by

$$j^2 - 13824\, t\, \frac{500t^2 + 965t + 27}{(4t-1)^3} j + 47775744\, t\, \frac{(25t-4)^3}{(4t-1)^4} = 0.$$

Proof. The proof is elementary and follows from [18]. □

3 Curves of Genus 2 with Degree 3 Elliptic Subcovers

In this section we will give a brief description of the spaces $\mathcal{L}_2$ and $\mathcal{L}_3$. In the case $J_2 \neq 0$ we take these spaces as equations in terms of i_1, i_2, i_3, otherwise as homogeneous equations in terms of J_2, J_4, J_6, J_{10}. By a point $\mathfrak{p} \in \mathcal{L}_3$ we will mean a tuple (J_2, J_4, J_6, J_{10}) which satisfies the equation of $\mathcal{L}_3$. When it is clear that $J_2 \neq 0$ then $\mathfrak{p} \in \mathcal{L}_3$ would mean a triple $(i_1, i_2, i_3) \in \mathcal{L}_3$. As before k is an algebraically closed field of characteristic zero.

Definition 1. *A* **non-degenerate pair** *(resp.,* **degenerate pair***) is a pair $(\mathcal{C}, \mathcal{E})$ such that $\mathcal{C}$ is a genus 2 curve with a degree 3 elliptic subcover $\mathcal{E}$ where $\psi : \mathcal{C} \to \mathcal{E}$ is ramified in two (resp., one) places. Two such pairs $(\mathcal{C}, \mathcal{E})$ and $(\mathcal{C}', \mathcal{E}')$ are called isomorphic if there is a k-isomorphism $\mathcal{C} \to \mathcal{C}'$ mapping $\mathcal{E} \to \mathcal{E}'$.*

If $(\mathcal{C}, \mathcal{E})$ is a non-degenerate pair, then $\mathcal{C}$ can be parameterized as follows

$$Y^2 = (\mathfrak{v}^2 X^3 + \mathfrak{u}\mathfrak{v} X^2 + \mathfrak{v} X + 1)\,(4\mathfrak{v}^2 X^3 + \mathfrak{v}^2 X^2 + 2\mathfrak{v} X + 1), \tag{8}$$

where $\mathfrak{u}, \mathfrak{v} \in k$ and the discriminant

$$\Delta = -16\, \mathfrak{v}^{17}\, (\mathfrak{v} - 27)\, (27\mathfrak{v} + 4\mathfrak{v}^2 - \mathfrak{u}^2\mathfrak{v} + 4\mathfrak{u}^3 - 18\mathfrak{u}\mathfrak{v})^3$$

of the sextic is nonzero. We let $R := (27\mathfrak{v} + 4\mathfrak{v}^2 - \mathfrak{u}^2\mathfrak{v} + 4\mathfrak{u}^3 - 18\mathfrak{u}\mathfrak{v}) \neq 0$. For $4\mathfrak{u} - \mathfrak{v} - 9 \neq 0$ the degree 3 coverings are given by $\phi_1(X, Y) \to (U_1, V_1)$ and $\phi_2(X, Y) \to (U_2, V_2)$ where

$$\begin{aligned}
U_1 &= \frac{\mathfrak{v} X^2}{\mathfrak{v}^2 X^3 + \mathfrak{u}\mathfrak{v} X^2 + \mathfrak{v} X + 1}, \quad U_2 = \frac{(\mathfrak{v} X + 3)^2\, (\mathfrak{v}(4\mathfrak{u} - \mathfrak{v} - 9) X + 3\mathfrak{u} - \mathfrak{v})}{\mathfrak{v}\,(4\mathfrak{u} - \mathfrak{v} - 9)(4\mathfrak{v}^2 X^3 + \mathfrak{v}^2 X^2 + 2\mathfrak{v} X + 1)}, \\
V_1 &= Y\, \frac{\mathfrak{v}^2 X^3 - \mathfrak{v} X - 2}{\mathfrak{v}^2 X^3 + \mathfrak{u}\mathfrak{v} X^2 + \mathfrak{v} X + 1}, \\
V_2 &= (27 - \mathfrak{v})^{\frac{3}{2}}\, Y \frac{\mathfrak{v}^2(\mathfrak{v} - 4\mathfrak{u} + 8) X^3 + \mathfrak{v}(\mathfrak{v} - 4\mathfrak{u}) X^2 - \mathfrak{v} X + 1}{(4\mathfrak{v}^2 X^3 + \mathfrak{v}^2 X^2 + 2\mathfrak{v} X + 1)^2}
\end{aligned} \tag{9}$$

and the elliptic curves have equations:

$$\begin{aligned}
\mathcal{E} :&\quad V_1^2 = R U_1^3 - (12\mathfrak{u}^2 - 2\mathfrak{u}\mathfrak{v} - 18\mathfrak{v}) U_1^2 + (12\mathfrak{u} - \mathfrak{v}) U_1 - 4 \\
\mathcal{E}' :&\quad V_2^2 = c_3 U_2^3 + c_2 U_2^2 + c_1 U_2 + c_0
\end{aligned} \tag{10}$$

where

$$\begin{aligned}
c_0 &= -(9\mathfrak{u} - 2\mathfrak{v} - 27)^3 \\
c_1 &= (4\mathfrak{u} - \mathfrak{v} - 9)\,(729\mathfrak{u}^2 + 54\mathfrak{u}^2\mathfrak{v} - 972\mathfrak{u}\mathfrak{v} - 18\mathfrak{u}\mathfrak{v}^2 + 189\mathfrak{v}^2 + 729\mathfrak{v} + \mathfrak{v}^3) \\
c_2 &= -\mathfrak{v}\,(4\mathfrak{u} - \mathfrak{v} - 9)^2\,(54\mathfrak{u} + \mathfrak{u}\mathfrak{v} - 27\mathfrak{v}) \\
c_3 &= \mathfrak{v}^2\,(4\mathfrak{u} - \mathfrak{v} - 9)^3
\end{aligned} \tag{11}$$

The above facts can be deduced from Lemma 1 of [17]. The case $4\mathfrak{u} - \mathfrak{v} - 9 = 0$ is treated separately in [17]. There is an automorphism $\beta \in Gal_{k(\mathfrak{u},\mathfrak{v})/k(i_1,i_2,i_3)}$ given by

$$\begin{aligned} \beta(\mathfrak{u}) &= \frac{(\mathfrak{v}-3\mathfrak{u})(324\mathfrak{u}^2+15\mathfrak{u}^2\mathfrak{v}-378\mathfrak{u}\mathfrak{v}-4\mathfrak{u}\mathfrak{v}^2+243\mathfrak{v}+72\mathfrak{v}^2)}{(\mathfrak{v}-27)(4\mathfrak{u}^3+27\mathfrak{v}-18\mathfrak{u}\mathfrak{v}-\mathfrak{u}^2\mathfrak{v}+4\mathfrak{v}^2)} \\ \beta(\mathfrak{v}) &= -\frac{4(\mathfrak{v}-3\mathfrak{u})^3}{4\mathfrak{u}^3+27\mathfrak{v}-18\mathfrak{u}\mathfrak{v}-\mathfrak{u}^2\mathfrak{v}+4\mathfrak{v}^2} \end{aligned} \tag{12}$$

which permutes the j-invariants of $\mathcal{E}$ and $\mathcal{E}'$. The map

$$\theta : (\mathfrak{u}, \mathfrak{v}) \to (i_1, i_2, i_3)$$

defined when $J_2 \neq 0$ and $\Delta \neq 0$ has degree 2. Denote by J_θ the Jacobian matrix of θ. Then $det(J_\theta) = 0$ consist of the (non-singular) curve $\mathfrak{X}$ given by

$$\mathfrak{X} : \quad 8\mathfrak{v}^3 + 27\mathfrak{v}^2 - 54\mathfrak{u}\mathfrak{v}^2 - \mathfrak{u}^2\mathfrak{v}^2 + 108\mathfrak{u}^2\mathfrak{v} + 4\mathfrak{u}^3\mathfrak{v} - 108\mathfrak{u}^3 = 0 \tag{13}$$

and 6 isolated $(\mathfrak{u}, \mathfrak{v})$ solutions. These solutions correspond to the following values for (i_1, i_2, i_3):

$$(-\frac{8019}{20}, -\frac{1240029}{200}, -\frac{531441}{100000}),\ (\frac{729}{2116}, \frac{1240029}{97336}, \frac{531441}{13181630464}),\ (81, -\frac{5103}{25}, -\frac{729}{12500}) \tag{14}$$

We denote the image of $\mathfrak{X}$ in the $\mathcal{L}_3$ locus by $\mathfrak{Y}$. The map θ restricted to $\mathfrak{X}$ is unirational. The curve $\mathfrak{Y}$ can be computed as an affine curve in terms of i_1, i_2. For each point $\mathfrak{p} \in \mathfrak{Y}$ the degree 3 elliptic subcovers are isomorphic. If $\mathfrak{p}$ is an ordinary point in $\mathfrak{Y}$ and $\mathfrak{p} \neq \mathfrak{p}_6$ (cf. Table 1) then the corresponding curve $\mathcal{C}_\mathfrak{p}$ has automorphism group V_4.

If $(\mathcal{C}, \mathcal{E})$ is a degenerate pair then $\mathcal{C}$ can be parameterized as follows

$$Y^2 = (3X^2 + 4)(X^3 + X + c)$$

for some c such that $c^2 \neq -\frac{4}{27}$; see [17]. We define $\mathfrak{w} := c^2$. The map

$$\mathfrak{w} \to (i_1, i_2, i_3)$$

is injective as was shown in [17].

Definition 2. *Let $\mathfrak{p}$ be a point in $\mathcal{L}_3$. We say $\mathfrak{p}$ is a* **generic point** *in $\mathcal{L}_3$ if the corresponding $(\mathcal{C}_\mathfrak{p}, \mathcal{E})$ is a non-degenerate pair. We define*

$$e_3(\mathfrak{p}) := \begin{cases} |\theta^{-1}(\mathfrak{p})|, & if\ \mathfrak{p}\ is\ a\ generic\ point \\ 1 & otherwise \end{cases}$$

In [17] it is shown that the pairs $(\mathfrak{u}, \mathfrak{v})$ with $\Delta(\mathfrak{u}, \mathfrak{v}) \neq 0$ bijectively parameterize the isomorphism classes of non-degenerate pairs $(\mathcal{C}, \mathcal{E})$. Those $\mathfrak{w}$ with $\mathfrak{w} \neq -\frac{4}{27}$ bijectively parameterize the isomorphism classes of degenerate pairs $(\mathcal{C}, \mathcal{E})$. Thus, the number $e_3(\mathfrak{p})$ is the number of isomorphism classes of such pairs $(\mathcal{C}, \mathcal{E})$. In [17] it is shown that $e_3(\mathfrak{p}) = 0, 1, 2,$ or 4. The following lemma describes the locus $\mathcal{L}_3$. For details see [17].

Lemma 6. *The locus $\mathcal{L}_3$ of genus 2 curves with degree 3 elliptic subcovers is the closed subvariety of $\mathcal{M}_2$ defined by the equation*

$$C_8 J_{10}^8 + \cdots + C_1 J_{10} + C_0 = 0 \tag{15}$$

where coefficients $C_0, \ldots, C_8 \in k[J_2, J_6, J_{10}]$ are displayed in [17].

As noted above, with the assumption $J_2 \neq 0$ equation (15) can be written in terms of i_1, i_2, i_3.

4 Automorphism Groups of Genus 2 Curves with Degree 3 Elliptic Subcovers

Let $\mathcal{C} \in \mathcal{L}_3$ be a genus 2 curve defined over an algebraically closed field k, $char(k) = 0$. The following theorem determines the automorphism group of $\mathcal{C}$.

Theorem 1. *Let $\mathcal{C}$ be a genus two curve which has a degree 3 elliptic subcover. Then the automorphism group of $\mathcal{C}$ is one of the following: $\mathbb{Z}_2, V_4$, D_8, or D_{12}. Moreover, there are exactly six curves $\mathcal{C} \in \mathcal{L}_3$ with automorphism group D_8 and six curves $\mathcal{C} \in \mathcal{L}_3$ with automorphism group D_{12}.*

Proof. We denote by $G := Aut(\mathcal{C})$. None of the curves $Y^2 = X^6 - X$, $Y^2 = X^6 - 1$, $Y^2 = X^5 - X$ have degree 3 elliptic subcovers since their J_2, J_4, J_6, J_{10} invariants don't satisfy equation (15). From Lemma 1 we have the following cases:

i) If $G \cong D_8$, then $\mathcal{C}$ is isomorphic to

$$Y^2 = X^5 + X^3 + t\,X$$

as in Lemma 3. Igusa invariants are:

$$J_2 = 40t + 6,\ J_4 = 4t(9 - 20t),\ J_6 = 8t(22t + 9 - 40t^2),\ J_{10} = 16t^3(4t - 1)^2.$$

Substituting into the equation (15) we have the following equation:

$$(196t - 81)^4 (49t - 12)(5t - 1)^4 (700t + 81)^4 (490000\,t^2 - 136200\,t + 2401)^2 = 0 \tag{16}$$

For

$$t = \frac{81}{196}, \frac{12}{49}, \frac{1}{5}, -\frac{81}{700}$$

the triple (i_1, i_2, i_3) has the following values respectively:

$$(\frac{729}{2116}, \frac{1240029}{97336}, \frac{531441}{13181630464}), \quad (\frac{4288}{1849}, \frac{243712}{79507}, \frac{64}{1323075987}),$$

$$(\frac{144}{49}, \frac{3456}{8575}, \frac{243}{52521875}), \quad (-\frac{8019}{20}, -\frac{1240029}{200}, -\frac{531441}{10000})$$

If

$$490000\,t^2 - 136200\,t + 2401 = 0$$

then we have two distinct triples (i_1, i_2, i_3) which are in $\mathbb{Q}(\sqrt{2})$. Thus, there are exactly 6 genus 2 curves $\mathcal{C} \in \mathcal{L}_3$ with automorphism group D_8 and only four of them have rational invariants.

ii) If $G \cong D_{12}$ then $\mathcal{C}$ is isomorphic to a genus 2 curve in the form

$$Y^2 = X^6 + X^3 + t$$

as in Lemma 3. Then, $J_2 = -6(40t - 1)$ and

$$J_4 = 324t(5t + 1), \; J_6 = -162t(740t^2 + 62t - 1), \; J_{10} = -729t^2(4t - 1)$$

Then the equation of $\mathcal{L}_3$ becomes:

$$(25t-4)\,(11t+4)^3\,(20t-1)^6\,(111320000t^3 - 60075600t^2 + 13037748t + 15625)^3 = 0 \quad (17)$$

For

$$t = \frac{4}{25}, -\frac{4}{11}, \frac{1}{20}$$

the corresponding values for (i_1, i_2, i_3) are respectively:

$$(\frac{64}{5}, \frac{1088}{25}, \frac{1}{84375}), \quad (\frac{576}{361}, \frac{60480}{6859}, \frac{243}{2476099}), \quad (81, -\frac{5103}{25}, -\frac{729}{12500})$$

If

$$111320000t^3 - 60075600t^2 + 13037748t + 15625 = 0$$

then there are three distinct triples (i_1, i_2, i_3) none of which is rational. Hence, there are exactly 6 classes of genus 2 curves $\mathcal{C} \in \mathcal{L}_3$ with $Aut(\mathcal{C}) \cong D_{12}$ of which three have rational invariants.

iii) $G \cong V_4$. There is a 1-dimensional family of genus 2 curves with a degree 3 elliptic subcover and automorphism group V_4 given by $\mathfrak{Y}$.

iv) Generically genus 2 curves $\mathcal{C}$ have $Aut(\mathcal{C}) \cong \mathbb{Z}_2$. For example, every point $\mathfrak{p} \in \mathcal{L}_3 \backslash \mathcal{L}_2$ correspond to a class of genus 2 curves with degree 3 elliptic subcovers and automorphism group isomorphic to $\mathbb{Z}_2$. This completes the proof.

□

The theorem determines that there are exactly 12 genus 2 curves $\mathcal{C} \in \mathcal{L}_3$ with automorphism group D_8 or D_{12}. Only seven of them have rational invariants. From Lemma 4, we have the following:

Corollary 1. *There are exactly four (resp., three) genus 2 curves $\mathcal{C}$ defined over $\mathbb{Q}$ (up to $\bar{\mathbb{Q}}$-isomorphism) with a degree 3 elliptic subcover which have automorphism group D_8 (resp., D_{12}). They are listed in Table 1.*

Remark 4. All points $\mathfrak{p}$ in Table 1 are in the locus $det(J_\theta) = 0$. We have already seen cases $\mathfrak{p}_1, \mathfrak{p}_4$, and $\mathfrak{p}_7$ as the exceptional points of $det(J_\theta) = 0$; see equation (14). The class $\mathfrak{p}_3$ is a singular point of order 2 of $\mathfrak{Y}$, $\mathfrak{p}_2$ is the only point which belong to the degenerate case, and $\mathfrak{p}_6$ is the only ordinary point in $\mathfrak{Y}$ such that the order of $Aut(\mathfrak{p})$ is greater then 4.

Table 1. Rational points $\mathfrak{p} \in \mathcal{L}_3$ with $|Aut(\mathfrak{p})| > 4$

	$\mathcal{C}$	$\mathfrak{p} = (i_1, i_2, i_3)$	$e_3(\mathfrak{p})$	$Aut(\mathcal{C})$
$\mathfrak{p}_1$	$196X^5 + 196X^3 + 81X$	$i_1 = \frac{729}{2116}, i_2 = \frac{1240029}{97336}, i_3 = \frac{531441}{13181630464}$	2	D_8
$\mathfrak{p}_2$	$49X^5 + 49X^3 + 12X$	$i_1 = \frac{4288}{1849}, i_2 = \frac{243712}{79507}, i_3 = \frac{64}{1323075987}$	1	D_8
$\mathfrak{p}_3$	$5X^5 + 5X^3 + X$	$i_1 = \frac{144}{49}, i_2 = \frac{3456}{8575}, i_3 = \frac{243}{52521875}$	2	D_8
$\mathfrak{p}_4$	$700X^5 + 700X^3 - 81X$	$i_1 = -\frac{8019}{20}, i_2 = -\frac{1240029}{200}, i_3 = -\frac{531441}{10000}$	2	D_8
$\mathfrak{p}_5$	$25X^6 + 25X^3 + 4$	$i_1 = \frac{64}{5}, i_2 = -\frac{1088}{25}, i_3 = -\frac{1}{84375}$	1	D_{12}
$\mathfrak{p}_6$	$11X^6 + 11X^3 - 4$	$i_1 = \frac{576}{361}, i_2 = \frac{60480}{6859}, i_3 = \frac{243}{2476099}$	1	D_{12}
$\mathfrak{p}_7$	$20X^6 + 20X^3 + 1$	$i_1 = 81, i_2 = -\frac{5103}{25}, i_3 = -\frac{729}{12500}$	2	D_{12}

5 Computing Elliptic Subcovers

Next we will consider all points $\mathfrak{p}$ in Table 1 and compute j-invariants of their degree 2 and 3 elliptic subcovers. To compute j-invariants of degree 2 elliptic subcovers we use lemma 5 and the values of t from the proof of theorem 1. We recall that for $\mathfrak{p}_1, \dots, \mathfrak{p}_4$ there are four degree 2 elliptic subcovers which are two and two isomorphic. We list the j-invariant of each isomorphic class. They are 2-isogenous as mentioned before. For $\mathfrak{p}_5, \mathfrak{p}_6, \mathfrak{p}_7$ there are two degree 2 elliptic subcovers which are 3-isogenous to each other. To compute degree 3 elliptic subcovers for each $\mathfrak{p}$ we find the pairs $(\mathfrak{u}, \mathfrak{v})$ in the fiber $\theta^{-1}(\mathfrak{p})$ and then use equations (9). We focus on cases which have elliptic subcovers defined over $\mathbb{Q}$. There are techniques for computing rational points of genus two curves which have degree 2 subcovers defined over $\mathbb{Q}$ as in Flynn and Wetherell [5]. Sometimes the degree 3 elliptic subcovers are defined over $\mathbb{Q}$ even though the degree 2 elliptic subcovers are not; see Examples 2 and 6. These degree 3 subcovers help determine rational points of genus 2 curves as illustrated in examples 2, 4, 5, and 6.

Example 1. $\mathfrak{p} = \mathfrak{p}_1$: The j-invariants of degree 3 elliptic subcovers are $j = j' = 66^3$. A genus 2 curve $\mathcal{C}$ corresponding to $\mathfrak{p}$ is

$$\mathcal{C}: \ Y^2 = X^6 + 3X^4 - 6X^2 - 8.$$

Claim: The equation above has no rational affine solutions.

Indeed, two of the degree 2 elliptic subcovers (isomorphic to each other) have equations

$$\mathcal{E}_1: \ Y^2 = x^3 + 3x^2 - 6x - 8$$

$$\mathcal{E}_2: \ Y^2 = -8x^3 - 6x^2 + 3x + 1$$

where $x = X^2$ (i.e. $\phi : \mathcal{C} \to \mathcal{E}_1$ of degree 2 such that $\phi(X, Y) = (X^2, Y)$). The elliptic curve $\mathcal{E}_1$ has rank 0. Thus, the rational points of $\mathcal{C}$ are the preimages of the torsion points of $\mathcal{E}_1$. The torsion group of $\mathcal{E}_1$ has order 4 and is given by

$$Tor(\mathcal{E}_1) = \{\infty, (-1,0), (2,0), (-4,0)\}$$

None of the preimages is rational. Thus, $\mathcal{C}$ has no rational points except the point at infinity.

Example 2. $\mathfrak{p} = \mathfrak{p}_2$: The j-invariants of the degree 2 elliptic subcovers are

$$76771008 \pm 44330496\sqrt{3}.$$

The point $\mathfrak{p}_2$ belongs to the degenerate locus with $\mathfrak{w} = 0$. Thus, the equation of the genus 2 curve $\mathcal{C}$ corresponding to $\mathfrak{p}$ is

$$\mathcal{C}: \quad Y^2 = (3\,X^2 + 4)\,(X^3 + X).$$

Indeed, this curve has both pairs $(\mathcal{C}, \mathcal{E})$ and $(\mathcal{C}, \mathcal{E}')$ as degenerate pairs. It is the only such genus 2 curve defined over $\mathbb{Q}$. This fact was noted in [12] and [16]. Both authors failed to identify the automorphism group. The degree 3 coverings are

$$(U_1, V_1) = (X^3 + X, Y(3X^2 + 1)), \quad (U_2, V_2) = (\frac{X^3}{3X^2 + 4}, Y X^2\,[\frac{X^2 + 4}{(3X^2 + 4)^2}]^2)$$

and the elliptic curves have equations:

$$\mathcal{E}: \; V_1^2 = 27U_1^3 + 4U_1, \quad \textit{and} \quad \mathcal{E}': \; V_2^2 = U_2^3 + U_2.$$

$\mathcal{E}$ and $\mathcal{E}'$ are isomorphic with j-invariant 1728. They have rank 0 and rational torsion group of order 2, $Tor(\mathcal{E}) = \{\infty, (0,0)\}$. Thus, the only rational points of $\mathcal{C}$ are in the fibers $\phi_1^{-1}(0)$ and $\phi_2^{-1}(\infty)$. Hence, $\mathcal{C}(\mathbb{Q}) = \{(0,0), \infty\}$.

Example 3. $\mathfrak{p} = \mathfrak{p}_3$: All degree 2 and 3 elliptic subcovers are defined over $\mathbb{Q}(\sqrt{5})$.

Example 4. $\mathfrak{p} = \mathfrak{p}_4$: The degree 2 elliptic subcovers have j-invariants

$$\frac{1728000}{2809} \pm \frac{17496000}{2809}\sqrt{I}$$

where $I^2 = -1$. Thus, we can't recover any information from the degree 2 subcovers. One corresponding value for $(\mathfrak{u}, \mathfrak{v})$ is $(\frac{25}{2}, \frac{250}{9})$. Then $\mathcal{C}$ is

$$\mathcal{C}: \quad 3^8 \cdot Y^2 = (100X + 9)(2500X^2 + 400X + 9)\,(25X + 9)(2500X^2 + 225X + 9).$$

The degree 3 elliptic subcovers have equations

$$\begin{aligned} \mathcal{E}: \quad V_1^2 &= -\frac{1}{81}(10U_1 - 3)(8575U_1^2 - 2940U_1 + 108) \\ \mathcal{E}': \quad V_2^2 &= -\frac{686}{59049}\,(1700U_2 - 441)(1445000U_2^2 - 696150U_2 + 83853) \end{aligned} \tag{18}$$

where U_1, V_1, U_2, V_2 are given by formulas in (9).

Example 5. $\mathfrak{p} = \mathfrak{p}_5$: The degree 2 j-invariants are $j_1 = 0$ and $j_2 = -1228800$ and the degree 3 j-invariants as shown below are $j = j' = 0$. Let $\mathcal{C}$ be the genus 2 curve with equation

$$\mathcal{C}: \quad Y^2 = (X^3+1)(4X^3+1)$$

corresponding to $\mathfrak{p}$. The case is treated separately in [17]. The degree 3 elliptic subcovers have equations

$$\mathcal{E}: \; V_1^2 = -27U_1^3 + 4, \quad \mathcal{E}': \; V_2^2 = -16(27U_2^3 - 1)$$

where

$$(U_1, V_1) = (\frac{X^2}{X^3+1}, Y\frac{X^3-2}{(X+1)^2}), \quad (U_2, V_2) = (\frac{X}{4X^3+1}, Y\frac{8X^3-1}{(4X^3+1)^2}).$$

The rank of both $\mathcal{E}$ and $\mathcal{E}'$ is zero. Thus, the rational points of $\mathcal{C}$ are the preimages of the rational torsion points of $\mathcal{E}$ and $\mathcal{E}'$. The torsion points of $\mathcal{E}$ are $Tor(\mathcal{E}) = \{\infty, (0,2), (0,-2)\}$. Then $\phi_1^{-1}(0) = \{0, \infty\}$ and $\phi_1^{-1}(\infty) = \{-1, \frac{1}{2} \pm \frac{\sqrt{-3}}{2}\}$. Thus,

$$\mathcal{C}(\mathbb{Q}) = \{(0,1), (0,-1), (-1,0)\}$$

Example 6. $\mathfrak{p} = \mathfrak{p}_6$: This point is in $\mathfrak{Y}$ and it is not a singular point of $\mathfrak{Y}$. It has isomorphic degree 3 elliptic subcovers; see [17]. The corresponding $(\mathfrak{u}, \mathfrak{v})$ pair is $(\mathfrak{u}, \mathfrak{v}) = (20, 16)$ and $e_3(\mathfrak{p}) = 1$. Then the genus 2 curve has equation:

$$\mathcal{C}: \quad Y^2 = (256X^3 + 320X^2 + 16X + 1)\,(1024X^3 + 256X^2 + 32X + 1)$$

The degree 3 elliptic subcovers have j-invariants $j = j' = -32768$ and equations

$$\begin{aligned} \mathcal{E}: &\quad V_1^2 = 4(-5324U_1^3 + 968U_1^2 - 56U_1^2 + 1) \\ \mathcal{E}': &\quad V_2^2 = 11^3(-32000\,U_2^3 + 35200\,U_2^2 - 12320\,U_2 + 11^3) \end{aligned} \tag{19}$$

where U_1, V_1, U_2, V_2 are given by formulas in (9).

Both elliptic curves have trivial torsion but rank $r = 1$. One can try to adapt more sophisticated techniques in this case as Flynn and Wetherell have done for the degree 2 subcovers. This is the only genus 2 curve (up to $\mathbb{C}$-isomorphism) with automorphism group D_{12} and isomorphic degree 2 elliptic subcovers. Indeed all the degree 2 and 3 elliptic subcovers are $\mathbb{C}$-isomorphic with j-invariants $j = -32768$. The degree 2 elliptic subcovers also have rank 1 which does not provide any quick information about rational points of $\mathcal{C}$.

Example 7. $\mathfrak{p} = \mathfrak{p}_7$: All the degree 2 and 3 elliptic subcovers are defined over $\mathbb{Q}(\sqrt{5})$.

Throughout this paper we have made use of several computer algebra packages as Apecs, Maple, and GAP. The interested reader can check [18] and [17] for more details on loci $\mathcal{L}_2$ and $\mathcal{L}_3$. The equations for these spaces, j-invariants of elliptic subcovers of the degree 2 and 3, and other computational aspects of genus 2 curves can be downloaded from author's web site.

References

1. O. Bolza, On binary sextics with linear transformations into themselves. *Amer. J. Math.* **10**, 47-70.
2. J. W. S. Cassels and E. V. Flynn, Prolegomena to a Middlebrow Arithmetic of Curves of Genus Two, LMS, 230, 1996.
3. A. Clebsch, Theorie der Binären Algebraischen Formen, Verlag von B.G. Teubner, Leipzig, (1872).
4. T. Ekedahl and J. P. Serre, Exemples de courbes algébriques á jacobienne complétement décomposable. *C. R. Acad. Sci. Paris Sér. I Math.*, 317 (1993), no. 5, 509–513.
5. E. V. Flynn and J. Wetherell, Finding rational points on bielliptic genus 2 curves, *Manuscripta Math.* 100, 519-533 (1999).
6. G. Frey, On elliptic curves with isomorphic torsion structures and corresponding curves of genus 2. *Elliptic curves, modular forms, and Fermat's last theorem (Hong Kong, 1993)*, 79-98, Ser. Number Theory, I, *Internat. Press, Cambridge, MA*, (1995).
7. G. Frey and E. Kani, Curves of genus 2 covering elliptic curves and an arithmetic application. *Arithmetic algebraic geometry (Texel, 1989)*, 153-176, *Progr. Math.*, 89, *Birkhäuser Boston, Boston, MA, (1991).*
8. E. Howe, F. Leprévost, and B. Poonen, Large torsion subgroups of split Jacobians of curves of genus two or three. *Forum. Math*, **12** (2000), no. 3, 315-364.
9. J. Igusa, Arithmetic Variety of Moduli for genus 2. *Ann. of Math.* (2), 72, 612-649, (1960).
10. W. Keller, L. Kulesz, Courbes algébriques de genre 2 et 3 possédant de nombreux points rationnels. C. R. Acad. Sci. Paris Sér. I Math. 321 (1995), no. 11, 1469–1472.
11. A. Krazer, Lehrbuch der Thetafunctionen, Chelsea, New York, (1970).
12. M. R. Kuhn, Curves of genus 2 with split Jacobian. *Trans. Amer. Math. Soc* **307** (1988), 41-49
13. K. Magaard, T. Shaska, S. Shpectorov, and H. Völklein, The locus of curves with prescribed automorphism group, *RIMS Kyoto Technical Report Series*, Communications in Arithmetic Fundamental Groups and Galois Theory, 2001, edited by H. Nakamura.
14. P. Mestre, Construction de courbes de genre 2 á partir de leurs modules. In T. Mora and C. Traverso, editors, *Effective methods in algebraic geometry*, volume 94. *Prog. Math.* , 313-334. Birkhäuser, 1991. Proc. Congress in Livorno, Italy, April 17-21, (1990).
15. D. Mumford, The Red Book of Varieties and Schemes, Springer, 1999.
16. T. Shaska, Genus 2 curves with (n,n)-decomposable Jacobians, *Jour. Symb. Comp.*, Vol 31, **no. 5**, pg. 603-617, 2001.
17. T. Shaska, Genus 2 fields with degree 3 elliptic subfields, (submited for publication).
18. T. Shaska and H. Völklein, Elliptic Subfields and automorphisms of genus 2 function fields. *Proceeding of the Conference on Algebra and Algebraic Geometry with Applications: The celebration of the seventieth birthday of Professor S.S. Abhyankar*, Springer-Verlag, 2001.

Transportable Modular Symbols and the Intersection Pairing

Helena A. Verrill

Institut für Mathematik
Universität Hannover,Postfach 6009
D 30060 Hannover, Germany
verrill@math.uni-hannover.de
http://hverrill.net/

Abstract. Transportable modular symbols were originally introduced in order to compute periods of modular forms [18]. Here we use them to give an algorithm to compute the intersection pairing for modular symbols of weight $k \geq 2$. This generalizes the algorithm given by Merel [13] for computing the intersection pairing for modular symbols of weight 2. We also define a certain subspace of the space of transportable modular symbols, and give numerical evidence to support a conjecture that this space should replace the usual space of cuspidal modular symbols.

1 Introduction

In this paper $S_k(\Gamma_0(N), \mathbf{C})$ denotes the space of cuspidal modular forms of weight k and level N, and $\overline{S_k(\Gamma_0(N), \mathbf{C})}$ denotes the space of antiholomorphic cuspidal modular forms of weight k and level N. We will look at the following lattices, which are all equal when $k = 2$.

$$\begin{tikzcd}
\boldsymbol{\mathcal{S}}_k(\Gamma_0(N), \mathbf{Z}) \arrow[dr, dash] & & H_{k-1}(\mathcal{W}, \mathbf{Z}) \arrow[dl, dash, dashed] \arrow[ddl, dash] \\
& \boldsymbol{\mathcal{T}}_k(\Gamma_0(N), \mathbf{Z}) \arrow[d, dash] & \\
& \boldsymbol{\mathcal{U}}_k(\Gamma_0(N), \mathbf{Z}) &
\end{tikzcd}$$

Here $\boldsymbol{\mathcal{S}}_k(\Gamma_0(N), \mathbf{Z})$ is the space of integral weight k cuspidal modular symbols for $\Gamma_0(N)$, $\boldsymbol{\mathcal{T}}_k(\Gamma_0(N), \mathbf{Z})$ is the space of integral transportable modular symbols, and $\boldsymbol{\mathcal{U}}_k(\Gamma_0(N), \mathbf{Z})$ is a certain sublattice of $\boldsymbol{\mathcal{T}}_k(\Gamma_0(N), \mathbf{Z})$. The latter two spaces are Hecke submodules of finite index in $\boldsymbol{\mathcal{S}}_k(\Gamma_0(N), \mathbf{Z})$. The variety $\mathcal{W}$ is the Kuga-Sato variety, which is a smooth projective variety obtained from the $k-2$ fold fibre product of the universal family of elliptic curves over $X_0(N)$, as described by Deligne, [4] Lemme 5.4. The space of transportable symbols $\boldsymbol{\mathcal{T}}_k(\Gamma_0(N), \mathbf{Z})$ will be defined below. Note that usually one replaces

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 219–233, 2002.

$H_{k-1}(\mathcal{W},\mathbf{Z})$ in this diagram by its subspace given by the symmetric product $S^{k-2}H_1(\mathcal{E},\mathbf{Z})$, where $\mathcal{E}$ is the universal family of elliptic curves over $X_0(N)$. Elements of $\boldsymbol{\mathcal{T}}_k(\Gamma_0(N),\mathbf{Z})$ can be interpreted as elements of $H_{k-1}(\mathcal{W},\mathbf{Z})$, as described in § 4. However, there is not a unique embedding of $\boldsymbol{\mathcal{T}}_k(\Gamma_0(N),\mathbf{Z})$ in $H_{k-1}(\mathcal{W},\mathbf{Z})$. When we restrict to the part of $H_{k-1}(\mathcal{W},\mathbf{Z})$ corresponding to $S^{k-2}H_1(\mathcal{E},\mathbf{Z})$, we can obtain a unique embedding, over $\mathbf{Q}$. To have an inclusion of $\mathbf{Z}$-modules, we must pass to $\boldsymbol{\mathcal{U}}_k(\Gamma_0(N),\mathbf{Z})$, described in § 5. The interpretation of transportable modular symbols as elements of $H_{k-1}(\mathcal{W},\mathbf{Z})$ it is used to give a description of the intersection pairing on the space of cuspidal modular symbols.

Modular symbols first appear in papers of Birch[1], Mazur[11] and Swinnerton-Dyer[12]. Higher weight generalizations were carried out by Manin and Shokurov (e.g., in [10], [15]). Modular symbols give a concrete way to compute with modular forms. Many algorithms were developed by Cremona and Merel (see e.g., [3], [14]), and more recently Stein[16], who made these algorithms and computations more generally available in the modular symbols package for Magma[2]. They are important because of the perfect pairing

$$\Phi : \left(S_k(\Gamma_0(N),\mathbf{C}) \oplus \overline{S_k(\Gamma_0(N),\mathbf{C})}\right) \times \boldsymbol{\mathcal{S}}_k(\Gamma_0(N),\mathbf{C}) \to \mathbf{C},$$

which induces an isomorphisms of Hecke modules. This means that modular symbols can be used to find the coefficients of Hecke eigen forms, and can also be used in computing period of modular forms, and special values of L-series. So computationally, modular symbols are useful for verifying many important conjectures in the theory of modular forms, such as the modularity of elliptic curves, [3], and cases of the Birch–Swinnterton-Dyer conjecture [6] to give a few examples. The new space $\boldsymbol{\mathcal{U}}_k(\Gamma_0(N),\mathbf{Z})$ introduced here should be useful in the future for verifying cases of the Bloch-Kato conjecture, such as extending computations used by Dummigan [5] to higher level cases.

Transportable symbols are of interest because they

- generalize the natural weight 2 phenomena,

$$\{0, g0\} = \{\alpha, g\alpha\}$$

for all α in the upper half complex plane and for all $g \in \Gamma$.
- allow us to compute the intersection pairing of cuspidal modular symbols,
- are naturally contained in $H_{k-1}(\mathcal{W},\mathbf{Z})$.

Transportable modular symbols were introduced in [18] for the first of the above reasons, and in order to generalize weight 2 algorithms of Cremona for computing periods of modular forms [3], to higher weight. In this paper we look at the second point, and we also introduce the space $\boldsymbol{\mathcal{U}}_k(\Gamma_0(N),\mathbf{Z})$.

For any cuspidal Hecke eigenform $f \in S_k(\Gamma_0(N))$, we let I_f be the annihilator of f in the Hecke algebra $\mathbf{T}$. In [16] §2.7 the abelian varieties A_f and $A_f^\vee$ were defined to be given by

$$\begin{aligned} A_f &= \mathrm{Hom}_{\mathbf{C}}(S_k(\Gamma_0(N))[I_f],\mathbf{C})/\Phi_f\boldsymbol{\mathcal{S}}_k(\Gamma_0(N),\mathbf{Z}) \\ A_f^\vee &= \mathrm{Hom}_{\mathbf{C}}(S_k(\Gamma_0(N)),\mathbf{C})[I_f]/\boldsymbol{\mathcal{S}}_k(\Gamma_0(N),\mathbf{Z})[I_f], \end{aligned}$$

where Φ_f is given by $\Phi_f(x) = \Phi(f, x)$. In the weight 2 case, it is known that A_f is a quotient of the Jacobian of $X_0(N)$. In the weight $k > 2$ case the Jacobian of $X_0(N)$ should be replaced by the intermediate Jacobian of the variety $\mathcal{W}$, and the situation is a little more complicated. It is not clear that A_f and $A_f^\vee$ as defined above correctly correspond to the geometry of the situation. Part of the aim of this paper is to find better definitions for A_f and $A_f^\vee$. We propose that it is better to replace $\boldsymbol{\mathcal{S}}$ by $\boldsymbol{\mathcal{U}}$ in the definition of A_f and $A_f^\vee$. This is justified by experimental evidence which leads us to make the following conjecture:

Conjecture. The cup product on $H_{k-1}(\mathcal{W}, \mathbf{Z})$ gives rise to an intersection pairing on $\boldsymbol{\mathcal{S}}_k(\Gamma, \mathbf{Z})$. Let $\mathrm{Int}_{\boldsymbol{\mathcal{U}},f}$ be a matrix describing this intersection pairing on an integral basis of $\boldsymbol{\mathcal{U}}_k(\Gamma, \mathbf{Z})[I_f]$. Then the number

$$\frac{|\boldsymbol{\mathcal{U}}\text{−modular kernel of} f|}{\det(\mathrm{Int}_{\boldsymbol{\mathcal{U}},f})}$$

is equal to 1.

Both $\boldsymbol{\mathcal{U}}$ and the $\boldsymbol{\mathcal{U}}$-modular-kernel of f will be defined in § 4.

In this paper, for simplicity we restrict to the case k even, and trivial character, though the algorithms described easily generalize to odd weight, and arbitrary character.

In writing this paper, the algorithms described were implemented in Magma [2], and the Magma packages written by William Stein [17] form the backbone for the computations.

Acknowledgements

I would like to thank William Stein, Michael Fryers, and Michael Lönne for useful conversations, and Benedict Gross for useful comments when I gave a talk on this topic at Harvard University. I would also like to thank my host Klaus Hulek for his hospitality while I have been at Hannover University. While working on this paper I have been supported by the DFG Schwerpunktprogramm "Globale Methoden in der Komplexen Geometrie", grant Hu 337/5-1. Thanks also to the referees and to Neil Dummigan for pointing out some mistakes in the first version of this paper.

2 Definitions

We use the following notation:

$V_k = \mathbf{Z}_k[X, Y]$ is the space of homogeneous polynomials in X and Y with coefficients in $\mathbf{Z}$.

Γ is a congruence subgroup in $\mathrm{SL}_2(\mathbf{Z})$.

The action of Γ on $\mathbf{P}^1(\mathbf{Q})$ is given by fractional linear transformation, and on V_k by a linear action, so for $g \in \Gamma$, $\alpha \in \mathbf{P}^1(\mathbf{Q})$, and $P(X, Y) \in V_k$ we have

$$gP(X,Y) = P(g^{-1}(X,Y)) = P(dX - bY, -cX + aY),$$
$$g\alpha = \frac{a\alpha + b}{c\alpha + d}, \text{where } g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Boundary symbols $\boldsymbol{\mathcal{B}}_k(\Gamma, \mathbf{Z})$ are defined by:

$$\begin{array}{ll}
\boldsymbol{\mathcal{B}} & = \text{abelian group on } \{\alpha\} \text{ for } \alpha \in \mathbf{P}^1(\mathbf{Q}) \\
\boldsymbol{\mathcal{B}}_k & = V_{k-2} \otimes \boldsymbol{\mathcal{B}} \\
\boldsymbol{\mathcal{B}}_k(\Gamma, \mathbf{Z}) & = \boldsymbol{\mathcal{B}}_k \text{ modulo the relation:} \\
\multicolumn{2}{l}{x \sim \gamma x \quad \forall \gamma \in \Gamma, \text{ and modulo torsion (action of } \Gamma \text{ given below)}}
\end{array}$$

Modular symbols $\boldsymbol{\mathcal{M}}_k(\Gamma, \mathbf{Z})$ are defined by:

$$\begin{array}{ll}
\boldsymbol{\mathcal{M}} & = \text{subgroup of } \boldsymbol{\mathcal{B}} \text{ spanned by} \\
 & \quad \{\alpha, \beta\} := \{\beta\} - \{\alpha\} \\
\boldsymbol{\mathcal{M}}_k & = V_{k-2} \otimes \boldsymbol{\mathcal{M}} \\
\boldsymbol{\mathcal{M}}_k(\Gamma, \mathbf{Z}) & = \boldsymbol{\mathcal{M}}_k \text{ modulo the relation:} \\
\multicolumn{2}{l}{x \sim \gamma x \quad \forall \gamma \in \Gamma, \text{ and modulo torsion (action of } \Gamma \text{ given below)}}
\end{array}$$

Cuspidal modular symbols are defined to be the kernel in following sequence:

$$\begin{tikzcd} 0 \arrow[r] & \boldsymbol{\mathcal{S}}_k(\Gamma, \mathbf{Z}) \arrow[r] & \boldsymbol{\mathcal{M}}_k(\Gamma, \mathbf{Z}) \arrow[r, "\delta"] & \boldsymbol{\mathcal{B}}_k(\Gamma, \mathbf{Z}) \end{tikzcd}$$

where δ is defined by

$$\delta(P\{\alpha, \beta\}) = P\{\beta\} - P\{\alpha\}.$$

We denote elements of $\boldsymbol{\mathcal{M}}_k(\Gamma, \mathbf{Z})$ as sums of elements of the form $P(X, Y)\{\alpha, \beta\}$, with $P(X, Y) \in V_{k-2}$, and $\alpha, \beta \in \mathbf{P}^1(\mathbf{Q})$, omitting the tensor sign. Note that the action of Γ on $\boldsymbol{\mathcal{B}}_k$ and on $\boldsymbol{\mathcal{M}}_k \subset \boldsymbol{\mathcal{B}}_k$ is such that

$$\gamma(P(X, Y)\{\alpha\}) = P(\gamma^{-1}(X, Y))\{\gamma\alpha\}.$$

We call $\boldsymbol{\mathcal{M}}_k(\Gamma, \mathbf{Z})$, $\boldsymbol{\mathcal{B}}_k(\Gamma, \mathbf{Z})$ and $\boldsymbol{\mathcal{S}}_k(\Gamma, \mathbf{Z})$ the modular, boundary, and cuspidal modular symbols of weight k for Γ, respectively. We define $\boldsymbol{\mathcal{M}}_k(\Gamma, \mathbf{Q}) = \boldsymbol{\mathcal{M}}_k(\Gamma, \mathbf{Z}) \otimes \mathbf{Q}$, $\boldsymbol{\mathcal{B}}_k(\Gamma, \mathbf{Q}) = \boldsymbol{\mathcal{B}}_k(\Gamma, \mathbf{Z}) \otimes \mathbf{Q}$ and $\boldsymbol{\mathcal{S}}_k(\Gamma, \mathbf{Q}) = \boldsymbol{\mathcal{S}}_k(\Gamma, \mathbf{Z}) \otimes \mathbf{Q}$. If no coefficient ring is given we mean that the coefficients should be the integers.

If $\mathbf{P}^1(\mathbf{Q})$ is replaced by the upper half plane union the cusps, $\langle^*$ in the definition of modular symbols, we obtain spaces which we denote $\widetilde{\boldsymbol{\mathcal{M}}}_k(\Gamma, \mathbf{Z})$, $\widetilde{\boldsymbol{\mathcal{B}}}_k(\Gamma, \mathbf{Z})$ and $\widetilde{\boldsymbol{\mathcal{S}}}_k(\Gamma, \mathbf{Z})$, and which we refer to as "extended" modular symbols. Though $\widetilde{\boldsymbol{\mathcal{M}}}_k(\Gamma, \mathbf{Z})$ and $\widetilde{\boldsymbol{\mathcal{B}}}_k(\Gamma, \mathbf{Z})$ are uncountable, it turns out that $\widetilde{\boldsymbol{\mathcal{S}}}_k(\Gamma, \mathbf{Z})$ is countable and isomorphic to $\boldsymbol{\mathcal{S}}_k(\Gamma, \mathbf{Z})$, which follow from [18] Lemma 2.3. The advantage of $\widetilde{\boldsymbol{\mathcal{S}}}_k(\Gamma, \mathbf{Z})$ is that modular symbols in this space can be written with end points not in $\mathbf{Q}$, which is useful for the purposes of evaluating period integrals.

2.1 Transportable Modular Symbols

The space of transportable modular symbols is given by

$$\boldsymbol{\mathcal{T}}_k(\Gamma, \mathbf{Z}) := \left\{ \sum P_i\{0, g_i 0\} \middle| g_i \in \Gamma, P_i \in \mathbf{Z}_{k-2}[X, Y], \sum P_i = \sum g_i^{-1} P_i \right\}.$$

It turns out that if we define

$$\begin{aligned}\boldsymbol{\mathcal{B}}'_k(\Gamma) &= \langle P\{a\}\rangle \,/\, \big\langle P\{a\} - \gamma P\{\gamma a\}\big\rangle \\ \boldsymbol{\mathcal{M}}'_k(\Gamma) &= \langle P\{a,b\}\rangle \,/\, \langle P\{a,b\} - \gamma P\{\gamma a, \gamma b\}\rangle,\end{aligned}$$

where in each case P runs over all elements of V_{k-2}, a and b run over all of $\mathbf{P}^1(\mathbf{Q})$, and γ runs over all of Γ, then we have,

$$\boldsymbol{\mathcal{T}}_k(\Gamma, \mathbf{Z}) = \ker\Big(\boldsymbol{\mathcal{M}}'_k(\Gamma) \to \boldsymbol{\mathcal{B}}'_k(\Gamma)\Big)/\text{torsion}$$

(see [19] for a proof), from which is clear that the space of transportable symbols is a Hecke invariant submodule of $\boldsymbol{\mathcal{S}}_k(\Gamma, \mathbf{Z})$. Comparing this with the definition of cuspidal symbols given by

$$\boldsymbol{\mathcal{S}}_k(\Gamma, \mathbf{Z}) = \ker\Big(\boldsymbol{\mathcal{M}}'_k(\Gamma)/\text{torsion} \to \boldsymbol{\mathcal{B}}'_k(\Gamma)/\text{torsion}\Big),$$

we see that the only difference is where we quotient out the torsion.

3 The Index $[\boldsymbol{\mathcal{S}}_k(\Gamma) : \boldsymbol{\mathcal{T}}_k(\Gamma)]$

Results in [18] imply that the index $[\boldsymbol{\mathcal{S}}_k(\Gamma) : \boldsymbol{\mathcal{T}}_k(\Gamma)]$ is finite. The algorithm for finding $\boldsymbol{\mathcal{S}}_k(\Gamma)$ is described in several places e.g., [16], and has been implemented by Stein in MAGMA[2].

To determine a basis for $\boldsymbol{\mathcal{T}}_k(\Gamma)$ we use the following result.

Lemma 1. *Given a fixed finite set of generators G of Γ, any element of $\boldsymbol{\mathcal{T}}_k(\Gamma)$ can be written as*

$$\sum_{g \in G} P_g\{0, g0\}$$

$P_g \in V_{k-2}$ satisfy $\sum_{g\in G}(1 - g^{-1})P_g = 0$.

Proof. Let $\{g_i\}_{i=1\ldots m}$ be a fixed choice of generators for Γ. Given a transportable symbol $\sum_{i=1}^n Q_i\{0, h_i 0\}$, with $\sum(1 - h_i^{-1})Q_i = 0$, we can rewrite each term $Q_i\{0, h_i 0\}$ in terms of symbols of the form $P_j\{0, g_j 0\}$ as follows.

Since $\{g_i\}$ is a set of generators, we can find a sequence $u_i = g_{1_i}$ with $h_1 = \prod_{i=1}^M u_i^{\varepsilon_i}$, where $\varepsilon_i = \pm 1$, and where the product is taken in the order such that $h_1 = u_1^{\varepsilon_1} u_2^{\varepsilon_2} \ldots u_M^{\varepsilon_M}$. Then we have

$$\begin{aligned} Q_1\{0, h0\} &= Q_1\{0, \prod_{i=1}^M u_i^{\varepsilon_i} 0\} \\ &= Q_1\left(\{0, u_{1_1}^{\varepsilon_1} 0\} + \{u_1^{\varepsilon_1} 0, u_1^{\varepsilon_1} u_2^{\varepsilon_2} 0\} + \{u_1^{\varepsilon_1} u_2^{\varepsilon_2} 0, u_1^{\varepsilon_1} u_2^{\varepsilon_2} u_3^{\varepsilon_3} 0\} + \cdots\right) \\ &= \sum_{j=0}^{M-1} Q_1 \left\{ \prod_{i=1}^{j} u_i^{\varepsilon_i} 0, \prod_{i=1}^{j+1} u_i^{\varepsilon_i} 0 \right\} \\ &= \sum_{j=0}^{M-1} \left(\left(\prod_{i=1}^{j} u_i^{\varepsilon_i} \right)^{-1} Q_1 \right) \{0, u_{j+1}^{\varepsilon_{j+1}} 0\} \end{aligned}$$

Now we claim that

$$\sum_{j=0}^{M-1}\left(\left(\prod_{i=1}^{j}u_i^{\varepsilon_i}\right)^{-1}Q_1\right)\{0,u_{j+1}^{\varepsilon_{j+1}}0\}+\sum_{i=2}^{n}Q_i\{0,h_i0\}$$

is still written as a transportable symbol. This is because

$$\begin{aligned}
&\sum_{j=0}^{M-1}\left(\left(\prod_{i=1}^{j}u_i^{\varepsilon_i}\right)^{-1}Q_1\right)-u_{j+1}^{-\varepsilon_{j+1}}\sum_{j=0}^{M-1}\left(\left(\prod_{i=1}^{j}u_i^{\varepsilon_i}\right)^{-1}Q_1\right)\\
&=\sum_{j=0}^{M-1}\left(\left(\prod_{i=1}^{j}u_i^{\varepsilon_i}\right)^{-1}Q_1\right)-\sum_{j=0}^{M-1}u_{j+1}^{-\varepsilon_{j+1}}\left(\left(\prod_{i=1}^{j}u_i^{\varepsilon_i}\right)^{-1}Q_1\right)\\
&=\sum_{j=0}^{M-1}\left(\left(\prod_{i=1}^{j}u_i^{\varepsilon_i}\right)^{-1}Q_1\right)-\sum_{j=0}^{M-1}\left(\left(\prod_{i=1}^{j+1}u_i^{\varepsilon_i}\right)^{-1}Q_1\right)\\
&=Q_1-\left(\left(\prod_{i=1}^{M}u_i^{\varepsilon_i}\right)^{-1}Q_1\right)=Q_1-h_1^{-1}Q_1.
\end{aligned}$$

In this way we can write a transportable symbol to only involve terms of the from $P\{0,g_i^{\pm 1}0\}$. Next, note that if we have a transportable symbol with a term $P\{0,g^{-1}0\}$, for $g\in\Gamma$, we can replace this term with $-gP\{0,g0\}$ since these symbols are equal, and the transportability property is preserved since

$$P-(g^{-1})^{-1}P=P-gP=-(gP-g^{-1}(gP)).$$

It is clear that we can replace any terms like $P\{0,g0\}+Q\{0,g0\}$ by $(P+Q)\{0,g0\}$. Thus we can write a transportable symbol in the required form.

Given the above result, all we need to do to determine an integral basis for $\boldsymbol{\mathcal{T}}_k(\Gamma)$ is to find a set of generators $\{g_i\}_{i=1\ldots m}$, and then find the kernel of

$$\bigoplus_{i=1}^{m}(1-g_i^{-1}):\bigoplus_{i=1}^{m}V_{k-2}\to V_{k-2}.$$

Finding a set of generators can be achieved using the algorithms of Kulkarni[9], and has been implemented in Magma, as described in [20].

Table 1 tabulates values of the index $[D:\boldsymbol{\mathcal{T}}_k(\Gamma_0(N),\mathbf{Z})\cap D]$ of Hecke stable submodules D of $\boldsymbol{\mathcal{S}}_k(\Gamma_0(N),\mathbf{Z})$ corresponding to irreducible Hecke submodules of $S_k(\Gamma_0(N),\mathbf{Z})$. To save space we restrict to the case N prime. Given the data computed, we make the following conjecture.

Conjecture 1. The index $[\boldsymbol{\mathcal{S}}_k(\Gamma_0(N),\mathbf{Z}):\boldsymbol{\mathcal{T}}_k(\Gamma_0(N),\mathbf{Z})]$ is divisible only by primes dividing $(k-2)!N$.

Table 1. Rank of D and index $[D : \boldsymbol{\mathcal{T}}_k(\Gamma_0(N), \mathbf{Z}) \cap D]$ of Hecke stable submodules D of $\boldsymbol{\mathcal{S}}_k(\Gamma_0(N), \mathbf{Z})$ corresponding to irreducible pieces of $S_k(\Gamma_0(N), \mathbf{Z})$

N	rank	index	N	rank	index	N	rank	index	N	rank	index
$k = 4$			$k = 6$			37	14	37	$k = 10$		
17	2	1	13	4	13	”	16	$2^3.3.37^3$	13	8	$2.7.13^3$
”	6	2.17^2	”	6	$2^3.3.13^3$	41	12	41	”	10	$2^6.3^2.5.13^5$
19	2	1	17	1	17	”	20	$2^3.41^3$	17	10	$2.7.17^3$
”	6	19^2	”	1	1	43	16	43	”	14	$2^6.3^2.5.17^5$
23	2	1	”	8	$2^3.17^3$	”	20	$2^3.3.43^3$	19	12	$2^3.7.19^3$
”	8	23^2	19	2	1	$k = 8$			”	16	$2^4.3^2.5.19^5$
29	4	1	”	2	2^3	11	4	5.11^2	5	2	$2^2.5^2$
”	10	2.29^2	”	4	19	13	2	13	”	4	$2^8.3^6.5^5.7^2$
31	4	1	”	7	3.19^3	”	4	$2^2.3^2.13$	”	4	$2^2.5^5$
”	10	31^2	23	6	2.23	”	8	$2^2.5.13^4$	7	4	$2^6.3^6.5^3.7^2$
37	8	1	”	12	$2^2.23^3$	17	2	2.17	”	4	$2^2.3.5.7^4$
”	10	2.37^2	29	8	29	”	6	2.17	”	6	$2^3.7^6$
41	6	1	”	14	$2^3.29^3$	”	12	$2^2.3^2.5.17^4$	11	4	$2^8.3^6.5^3.7^2$
43	8	1	31	10	2.31	19	8	$3^2.5.19^2$	”	6	5.11^3
”	12	43^2	”	16	$2^2.3.31^3$	”	12	$2^3.19^4$	”	10	$2^3.11^5$

4 Intersection Pairing

On $H_{k-1}(\mathcal{W}, \mathbf{Z})$ we have a natural intersection pairing,

$$H_{k-1}(\mathcal{W}, \mathbf{Z}) \times H_{k-1}(\mathcal{W}, \mathbf{Z}) \mapsto H_{2(k-1)}(\mathcal{W}, \mathbf{Z}) \cong \mathbf{Z}$$

$$(\eta, \zeta) \mapsto \eta \cap \zeta.$$

This gives rise to an intersection pairing on the space of modular symbols, which is compatible with the action of the Hecke algebra, in the sense that for any symbols $a, b \in \boldsymbol{\mathcal{S}}_k(\Gamma)$ and for any $T \in \mathbf{T}$ we have $\langle Ta, b\rangle = \langle a, Tb\rangle$.

The algorithm for computing the intersection pairing in weights $k > 2$ is a generalization of the method described by Merel [13] in the case $k = 2$. The introduction of transportable symbols is essential for us to be able to give give this generalization. Exactly as for Merel's method, the computation of the intersection pairing is based on the following two lemmas.

Lemma 2. *All modular symbols in $\boldsymbol{\mathcal{M}}_k(\Gamma, \mathbf{Z})$ can be written as*

$$\sum_{g \in R} g\left(P_g\{0, \infty\}\right)$$

where $P_g \in V_{k-2}$, and R is a set of coset representatives for Γ in $\mathrm{SL}_2(\mathbf{Z})$.

Proof. See [14] §1.2, Proposition 1 and [10], Proposition 1.6. Symbols written in this way are known as *Manin symbols.*

Lemma 3. *All modular symbols in* $\widetilde{\boldsymbol{\mathcal{S}}}_k(\Gamma, \mathbf{Z})$ *can be written as*

$$\sum P_{g_i}\{\rho, g_i\rho\} \tag{1}$$

for some $g_i \in \Gamma$, *and* $P_i \in \mathbf{Q}_{k-2}[X, Y]$, *where* $\rho = (1 + i\sqrt{3})/2$.

Proof. This follow from [18] Lemma 2.3.

Note in particular that this shows that $\widetilde{\boldsymbol{\mathcal{S}}}_k(\Gamma, \mathbf{Z})$ is countable, even though it is the kernel of the boundary map between two uncountable spaces, $\widetilde{\boldsymbol{\mathcal{M}}}_k(\Gamma, \mathbf{Z})$ and $\widetilde{\boldsymbol{\mathcal{B}}}_k(\Gamma, \mathbf{Z})$.

Corollary 1. *All modular symbols in* $\widetilde{\boldsymbol{\mathcal{S}}}_k(\Gamma, \mathbf{Z})$ *can be written as*

$$\sum_{i=1}^{n} g_i\Big(P_i\{\rho, \rho^2\}\Big) \tag{2}$$

for some $g_i \in \mathrm{SL}_2(\mathbf{Z})$ *and* $P_i \in \mathbf{Q}_{k-2}[X, Y]$.

Proof. For any term $P\{\rho, g\rho\}$ in the sum (1), we can write $g = r_1^{e_1} r_2^{e_1} \ldots r_m^{e_m}$ with $r_i \in \{S := \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right), T^{-1} := \left(\begin{smallmatrix} 1 & -1 \\ 0 & 1 \end{smallmatrix}\right)\}$ and $e_i \in \{1, -1\}$. Let $g_i = \prod_{i=1}^{i-1} g_i^{e_i}$. Then

$$P\{\rho, g\rho\} = \sum P\{g_i\rho, g_{i+1}\rho\} = \sum_{i=1}^{m} g_i\Big((g_i^{-1} e_i P)\{\rho, r_i\rho\}\Big).$$

Since $r_i \in \{S, T^{-1}\}$ we have $r_i\rho = \rho^2$, so this sum is in the form of sum (2).

Given the above results, we only need to compute the intersection of symbols of the form $g\big(P\{0, \infty\}\big)$ with those of the form $g\big(Q\{\rho, \rho^2\}\big)$, and extend linearly to obtain the pairing on $\boldsymbol{\mathcal{S}}_k(\Gamma, \mathbf{Z})$.

Remark 1. The intersection of individual symbols $g\big(P\{0, \infty\}\big)$ and $h\big(Q\{\rho, \rho^2\}\big)$ is not well defined. However, when the sum is taken to obtain cuspidal modular symbols, the result is well defined.

Geometrically, weight 2 symbols of the form $g\{\rho, \rho^2\}$ and of the form $g\{0, \infty\}$ correspond to paths as shown in Figures 1, and 2 respectively. It is clear that the intersection of $g\{0, \infty\}$ with $h\{\rho, \rho^2\}$ is non zero if and only if the regions gF and hF are equal under Γ equivalence, where F is the area in the upper half plane with vertices at $0, \rho, \infty, \rho^2$, as shown in the figures. We have $gF = hF \iff gh^{-1} \in \{I, S\}$, so $gF = hF \iff g \sim h$ or $g \sim Sh$ under Γ equivalence. Also, S reverses the direction of the lines, so we have

$$\langle g\{0, \infty\}, h\{\rho, \rho^2\}\rangle = \begin{cases} 1 & \iff g \sim h \\ -1 & \iff g \sim Sh \\ 0 & \iff gF \neq hF \end{cases}$$

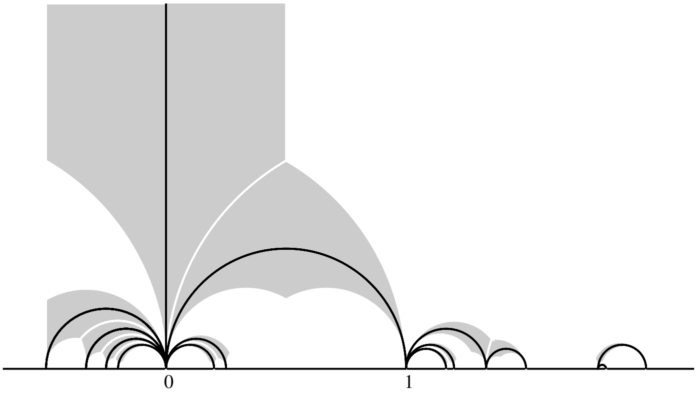


Fig. 1. Images of a path corresponding to $\{0, \infty\}$

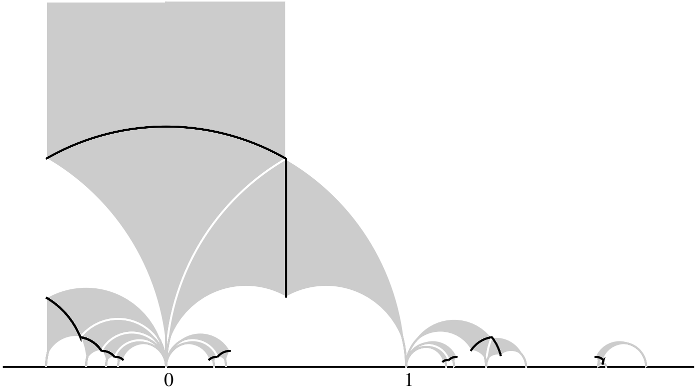


Fig. 2. Images of a path corresponding to $\{\rho, \rho^2\}$

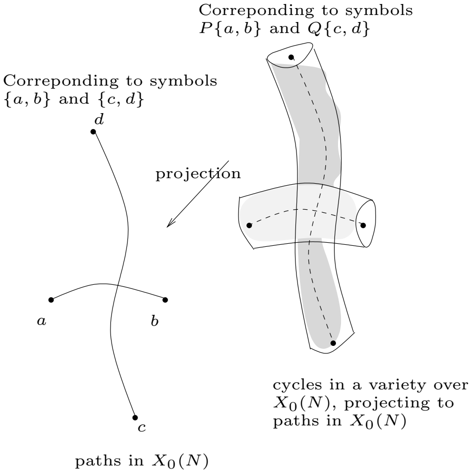


Fig. 3. Geometry corresponding to modular symbols

Geometrically, symbols of weight $k > 2$ correspond to $k-1$ cycles in a variety $\mathcal{W}$. The cycle corresponding to a modular symbol lies over a path on the modular curve, as pictured in Figure 3. This figure shows $\{a,b\},\{c,d\}$ as elements of $\pi_1(X_0(N),\mathbf{P}^1(\mathbf{Q}))$, with $\{a,b\}\cap\{c,d\}=\{x\}$, for some $x\in X_0(N)$. So $\langle\{a,b\},\{c,d\}\rangle=\pm 1$. To determine $\langle P\{a,b\},Q\{c,d\}\rangle$ we must find the intersection of cycles corresponding to P and Q in in the fibre over x.

The fibre of $\mathcal{W}$ over any point τ in the upper half plane is given by the product of $k-2$ copies of an elliptic curve E_τ. If we choose a basis α,β of $H_1(E_\tau,\mathbf{Z})$, with $\langle\alpha,\beta\rangle=1$, then we have that the 2^{k-2} cycles $\delta_1\times\delta_2\times\cdots\times\delta_{k-2}$ for $\delta_i\in\{\alpha,\beta\}$ give a basis for a certain monodromy invariant subspace of

$$H_{k-2}(\overbrace{E_\tau\times E_\tau\times\ldots E_\tau}^{k-2\text{ copies}},\mathbf{Z}),$$

and in terms of this basis, the intersection pairing on $E_\tau\times\ldots E_\tau$ is given by

$$\langle\delta_{1,1}\times\cdots\times\delta_{1,k-2}\ ,\ \delta_{2,1}\times\cdots\times\delta_{2,k-2}\rangle=\prod_{i=1}^{k-2}\langle\delta_{1,i},\delta_{2,i}\rangle\,.$$

So this is represented by a matrix with ±1s on the antidiagonal, and 0 elsewhere.

The space V_{k-2} can be identified with the subspace of $H_{k-2}(E_\tau\times\cdots\times E_\tau,\mathbf{Z})$ given by the symmetric product of the $H_1(E_\tau,\mathbf{Z})$. A monomial X^mY^{k-2-m} corresponds to cycles $\delta_1\times\cdots\times\delta_{k-2}$ where m of the δ_i are equal to α, and the rest are equal to β. There are $\binom{k-2}{m}$ such cycles, which are identified in the symmetric product, so X^mY^{k-2-m} corresponds to $\binom{k-2}{m}^{-1}$ times their sum. Then the pairing on monomials becomes

$$\langle X^mY^{k-2-m},X^{k-2-m}Y^m\rangle=\frac{(-1)^m}{\binom{k-2}{m}}. \tag{3}$$

If $PQ\neq X^{k-2}Y^{k-2}$, then $\langle P,Q\rangle=0$.

Now we have that for $P,Q\in V_{k-2}$ and $g,h\in\mathrm{SL}_2(\mathbf{Z})$

$$\langle g\left(P\{0,\infty\}\right),h\left(Q\{\rho,\rho^2\}\right)\rangle=\begin{cases}\langle P.Q\rangle & \Longleftrightarrow\ g\sim h\\ -\langle P.Q\rangle & \Longleftrightarrow\ g\sim Sh\\ 0 & \Longleftrightarrow\ gF\neq hF,\end{cases}$$

where $\langle P,Q\rangle$ is computed by extending the pairing in 3 linearly to give a symmetric pairing on V_{k-2}.

Proposition 1. *For $P,Q\in V_{k-2}$ and $g\in\Gamma$ and $u\in\mathrm{SL}_2(\mathbf{Z})$, define*

$$\langle P\{\rho,g\rho\},uQ\{u0,u\infty\}\rangle=-\sum_{\substack{k\text{such that}\\ h_kW_k^{\alpha_k}\sim u}}\varepsilon_k\langle P,h_kW_k^{\alpha_k}Q\rangle$$
$$+\sum_{\substack{k\text{such that}\\ h_kW_k^{\alpha_k}\sim uS}}\varepsilon_k\langle P,h_kW_k^{\alpha_k}SQ\rangle,$$

where $g = \prod_{i=1}^{n} W_i^{\varepsilon_i}$, for $W_i \in \{S,T\}$, $\varepsilon_i \in \{1,-1\}$ and $h_k := \prod_{i=1}^{k-1} W_k$, for $k \in \{1..n\}$, and the pairing on V_{k-2} is given by (3). Then extending linearly, this formula gives a Hecke invariant, anti-symmetric intersection pairing on $\boldsymbol{\mathcal{S}}_k(\Gamma, \mathbf{Z})$.

Proof. Let $h = \prod_{i=1}^{n-1} W_i^{\varepsilon_i}$, so that $g = hW_n^{\varepsilon_n}$ Then

$$\begin{aligned}
P\{\rho, g\rho\} &= P\{\rho, hW_n^{\varepsilon}\rho\} \\
&= P\{\rho, h\rho\} - \begin{cases} P\{g\rho, gW_n^{-1}\rho\} \text{ if} \varepsilon = 1 \\ \\ P\{hW_n^{-1}\rho, h\rho\} \text{ if} \varepsilon = -1 \end{cases} \\
&= P\{\rho, h\rho\} + \begin{cases} -g[(g^{-1}P)\{\rho, \rho^2\}] \text{ if} \varepsilon = 1 \\ \\ \quad h[(h^{-1}P)\{\rho, \rho^2\}] \text{ if} \varepsilon = -1 \end{cases} \\
&= P\{\rho, h\rho\} - \varepsilon hW^{\alpha}[((hW^{\alpha})^{-1}P)\{\rho, \rho^2\}]
\end{aligned}$$

where $\alpha = 0$ if $\varepsilon = -1$ and $\alpha = 1$ otherwise.

Repeating this process, we find that

$$P\{\rho, g\rho\} = -\sum_{k=1}^{n} \varepsilon_k h_k W_k^{\alpha} \left[\left((h_k W_k^{\alpha})^{-1} P\right)\{\rho, \rho^2\}\right].$$

If $g \sim u$, then $gu^{-1} \in \Gamma$, so

$$\begin{aligned}
\langle Pg\{\rho, \rho^2\}, uQ\{u0, u\infty\}\rangle &= \langle Pg\{\rho, \rho^2\}, gu^{-1}(uQ\{u0, u\infty\})\rangle \\
&= \langle Pg\{\rho, \rho^2\}, gQ\{g0, g\infty\})\rangle = \langle P, gQ\rangle.
\end{aligned}$$

So,

$$\langle P\{\rho, g\rho\}, uQ\{u0, u\infty\}\rangle = -\sum_{\substack{k \text{suchthat} \\ h_k W_k^{\alpha_k} \sim u}} \varepsilon_k \langle P, h_k W_k^{\alpha_k} Q\rangle.$$

On the other hand, if $g \sim uS$, then $gSu^{-1} \in \Gamma$, so

$$\begin{aligned}
\langle Pg\{\rho, \rho^2\}, uQ\{u0, u\infty\}\rangle &= \langle Pg\{\rho, \rho^2\}, gSu^{-1}(uQ\{u0, u\infty\})\rangle \\
&= \langle Pg\{\rho, \rho^2\}, gSQ\{g\infty, g0\})\rangle = -\langle P, gSQ\rangle.
\end{aligned}$$

So,

$$\langle P\{\rho, g\rho\}, uQ\{u0, u\infty\}\rangle = \sum_{\substack{k \text{suchthat} \\ h_k W_k^{\alpha_k} \sim uS}} \varepsilon_k \langle P, h_k W_k^{\alpha_k} SQ\rangle.$$

This gives the result.

Computations show that the above description does give a pairing that is Hecke invariant (with respect to Hecke operators T_p for $(p, N) = 1$) and anti-symmetric, though sometimes this fails when $5, 13$ or 17 divide the level, which may be due to some as yet undiscovered programming bug.

5 The Space $\boldsymbol{\mathcal{U}}_k(\varGamma_0(N), \mathbf{Z})$

Now the intersection pairing can be computed, we investigate its relationship with the order of various modular kernels. If $k = 2$ then $\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{T}}$ and $\boldsymbol{\mathcal{U}}$ are equal, and the relationship between the modular kernel and the intersection pairing is known, and described for example, by Frey and Müller in [7] §4.2.

5.1 Modular Kernels

Denote by S the space of cusp forms of weight k and level N, and suppose $f \in S$ is a newform. We can assume f has been normalized the coefficients of its q-expansion are algebraic integers. The space S_f is defined to be the subspace of S spanned over $\mathbf{C}$ by the Galois conjugates of f.

We will define several modular kernels corresponding to f. Let $\boldsymbol{\mathcal{R}}$ be either $\boldsymbol{\mathcal{S}}_k(\varGamma_0(N), \mathbf{Z})$, $\boldsymbol{\mathcal{T}}_k(\varGamma_0(N), \mathbf{Z})$ or $\boldsymbol{\mathcal{U}}_k(\varGamma_0(N), \mathbf{Z})$, where $\boldsymbol{\mathcal{U}}_k(\varGamma_0(N), \mathbf{Z})$ will be defined below. These are all lattices of full rank in $\boldsymbol{\mathcal{S}}_k(\varGamma_0(N), \mathbf{Q})$. The pairing $\varPhi$, mentioned in the introduction, defines a map $\boldsymbol{\mathcal{R}} \to \mathrm{Hom}_{\mathbf{C}}(S, \mathbf{C})$, and we denote the image of the period map in $\mathrm{Hom}_{\mathbf{C}}(S_f, \mathbf{C})$ by $\varPhi_f(\boldsymbol{\mathcal{R}})$. We have a commutative diagram with exact columns:

$$\begin{tikzcd}
0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] \\
\boldsymbol{\mathcal{R}}[I_f] \arrow[r] \arrow[d] & \boldsymbol{\mathcal{R}} \arrow[r] \arrow[d] & \varPhi_f(\boldsymbol{\mathcal{R}}) \arrow[d] \\
\mathrm{Hom}_{\mathbf{C}}(S, \mathbf{C})[I_f] \arrow[r] \arrow[d] & \mathrm{Hom}_{\mathbf{C}}(S, \mathbf{C})[I_f] \arrow[r] \arrow[d] & \mathrm{Hom}_{\mathbf{C}}(S, \mathbf{C})[I_f] \arrow[d] \\
A_f^{\vee} \arrow[r] \arrow[d] & J_k(N) \arrow[r] \arrow[d] & A_f \arrow[d] \\
0 & 0 & 0
\end{tikzcd}$$

The diagram is used to define the quotients $A_f^{\vee}$, A_f and $J_k(N)$. For $k > 2$ is is not clear that these complex torii should have any algebraic structure. They can be interpreted as intermediate Jacobians of $\mathcal{W}$. The map $A_f^{\vee} \to A_f$ depends on $\boldsymbol{\mathcal{R}}$, and we refer to its kernel as the $\boldsymbol{\mathcal{S}}$, $\boldsymbol{\mathcal{T}}$ or $\boldsymbol{\mathcal{U}}$-modular kernel depending on the choice of $\boldsymbol{\mathcal{R}}$.

In the case of $\boldsymbol{\mathcal{R}} = \boldsymbol{\mathcal{S}}_k(\varGamma_0(N), \mathbf{Z})$, this is exactly the same as the definition of the modular kernel of f suggested by Stein, [16] § 3.9 Definition 3.34, and § 2.7.

A method for computing the order of the $\boldsymbol{\mathcal{S}}$-modular kernel of f is described in [8], and we have used Stein's MAGMA[17] implementations of this method for finding the degree of the $\boldsymbol{\mathcal{S}}$-modular kernel. A simple modifications of the algorithms described could be used to compute the $\boldsymbol{\mathcal{T}}$-modular-kernel and the $\boldsymbol{\mathcal{U}}$-modular-kernel of f, though actually we have used a simpler method, of simply computing the determinant the matrix formed by the dot products of an integral basis of the kernel of $T_p - a$ with that of the transpose of this matrix, where T_p is the Hecke operator acting on either the $\boldsymbol{\mathcal{T}}$ or $\boldsymbol{\mathcal{U}}$ spaces, and a is an integer such that the kernel of $T_p - a$ acting on $S_k(N)$ is spanned by f. However, this only works in certain cases, (when such a, p exist). So in fact, most of our computations were just done for the $\boldsymbol{\mathcal{S}}$-modular kernel. These numbers can be compared with the determinant of the matrix $\mathrm{Int}_{\boldsymbol{\mathcal{T}},f}$, which describes the intersection pairing given above with respect to an integral basis for $\boldsymbol{\mathcal{T}}$. Computations, with data shown in Table 2 lead us to conjecture:

Conjecture 2. The fraction given by

$$\frac{|\boldsymbol{\mathcal{S}}\text{−modular kernel of } f|}{\det(\mathrm{Int}_{\boldsymbol{\mathcal{T}},f})}$$

is an integer and is divisible only by primes dividing $(k-2)!$.

Table 2. $R_f := |\boldsymbol{\mathcal{S}}$-modular kernel of $f|/\det(\mathrm{Int}_{\boldsymbol{\mathcal{T}},f})$ for new cuspidal Hecke eigen forms $f \in S_k(\Gamma_0(N))$

k	N	R_f	N	R_f
4	18	$2, 4, 1$	19	$1, 4$
	20	$1, 8, 4$		
6	16	$256, 48, 27648, 16$	18	$4, 16, 4608, 48, 384, 4$
	19	$16, 8, 16, 1728$	20	$48, 1, 432, 256, 576, 16$
8	6	$13500, 18, 27000$	7	$36, 2250$
	8	$100, 1458000, 36$	9	$1350, 45, 10800$
	11	$1620, 1620000$	12	$12150000, 75, 162, 182250000, 12$
10	6	$2^5, 2^9 5^1 7^3, 2^{10} 7^3, 2^7 5^1 7^4$	7	$2^{11} 7^2, 2^8 5^1 7^4$
	8	$2^{10} 7^4, 2^6 7^2, 2^6, 2^{14} 5^3 7^3$	9	$2^6 7^1, 2^8 5^2 7^4, 2^6, 2^4 5^1 7^2, 2^{11} 7^2$
	11	$2^{17} 7^2, 2^{16} 5^3 7^7$		
12	6	$2^3 3.7^2, 2^{13} 3^9 5^8 7^4, 2^2 3^2 5^2, 2^2 3^5 5^4, 2^2 3.5^2$		
	7	$2^8 3^6 5^4 7., 2^8 3.5^4, 2^3 3^6 5^4 7^2$		
	8	$2^{11} 3^{12} 5^8 7^4, 2^2 3^2 5^2, 2^5 3^2 5^4, 2^6 3^4 5^2 7^2$		
	9	$2.3^6 5^4, 2^4 3^3 5^2 7,\ 2^{12} 3^7 5^6 7^3,\ 3^3 5^2,\ 2^4 3^5 5^2 7^2$		
	11	$2^8 3^6 5^4 7^2, 2^8 3^4 5^6, 2^7 3^{10} 5^6 7^4$		
14	6	$2^5 3^4, 2^4 3^5 5.11^3, 2^6 3^2 5.11^4, 2^4 3^3 5^3 7.11^4, 2^{12} 3^9 5^3 7.11^7$		

The fact that the primes dividing this ratio divide $(k-2)!$ lead one to expect that in the definition of modular symbols we should replace V_k by some subspace of $V_k \otimes \mathbf{Q}$. We try using the space U_k which is a sublattice of V_k defined by

$$U_k := \left\langle \binom{k}{m} X^{k-m} Y^m,\ m = 0, \dots k \right\rangle \subset V_k.$$

This is exactly the right choice of monomials such that the pairing (3) becomes integrally valued on U_{k-2}. We define

$$\boldsymbol{\mathcal{U}}_k(\Gamma_0(N), \mathbf{Z}) := \left\{ \sum P_i\{0, g_i 0\} \middle| g_i \in \Gamma, P_i \in U_{k-2}, \sum P_i = \sum g_i^{-1} P_i \right\}.$$

Note that $\boldsymbol{\mathcal{U}}_k(\Gamma_0(N), \mathbf{Z})$ can be defined in other ways, as was $\boldsymbol{\mathcal{T}}_k(\Gamma_0(N), \mathbf{Z})$ above, and note that the space of 2 by 2 integral matrices acts on U_k, so $\boldsymbol{\mathcal{U}}_k(\Gamma_0(N), \mathbf{Z})$ is also a Hecke submodule of $\boldsymbol{\mathcal{T}}_k(\Gamma_0(N), \mathbf{Z})$. Now we have the following conjecture:

Conjecture 3. The fraction given by

$$\frac{|\boldsymbol{\mathcal{U}}-\text{modular kernel of} f|}{\det(\mathrm{Int}_{\boldsymbol{\mathcal{U}},f})}$$

is equal to 1.

Supporting numerical evidence is given in Table 3. Data is not given where the pairing computed by the above algorithm does not appear to be symmetric or Hecke invariant.

Table 3. Pairs $\big[|\boldsymbol{\mathcal{S}}\text{-modular kernel of } f|, \det(\mathrm{Int}_{\boldsymbol{\mathcal{U}},f})\big]$ for f new cuspidal Hecke eigen forms of level N, and weight k

N	ratios
	$k = 4$
$6, 7, 8, 9, 11$	$[1, 1]$
10	$[10, 20], [10, 10]$
12	$[12, 12], [12, 12]$
19	$[1444, 1444], [1444, 1444]$
20	$[72, 72], [800, 800], [3600, 3600]$
	$k = 6$
19	$[25542916, 25542916], [4133089, 4133089], [485315404, 485315404], [163743000636976, 491229001910928]$
20	$[14406000000, 14406000000], [332928, 332928], [7372800, 7372800], [61465600, 61465600], [998784, 998784], [4608000000, 4608000000]$
	$k = 8$
8	$[16384, 16384], [2048, 2048], [16384, 16384]$
9	$[17496, 17496], [5832, 5832], [236196, 236196]$
11	$[857435524, 857435524], [857435524, 857435524]$
12	$[207360000, 207360000], [20736, 20736], [331776, 331776], [15360000, 15360000], [41472, 41472]$
	$k = 14$
6	$[77845329, 77845329], [11151360, 11151360], [255977415, 255977415], [10726553600, 10726553600], [135039158100, 135039158100]$

5.2 Computing $\boldsymbol{\mathcal{U}_k(\Gamma_0(N), \mathbf{Z})}$

To compute $\boldsymbol{\mathcal{U}}$ we can apply the same method as described in section § 3, but now computing the integral kernel of

$$\bigoplus_{i=1}^{m}(1 - D^{-1}g_i^{-1}D) : \bigoplus_{i=1}^{m} V_{k-2} \to V_{k-2},$$

where D is the diagonal matrix with diagonal given by the sequence of binomial coefficients $\binom{k-2}{i}$ for $i = 0, \ldots k-2$.

References

1. B. J. Birch, *Elliptic curves over Q: A progress report,* 1969 Number Theory Institute (Proc. Sympos. Pure Math., Vol. XX, State Univ. New York, Stony Brook, N.Y., 1969), pp. 396–400. Amer. Math. Soc., Providence, R.I., 1971.
2. W. Bosma, J. Cannon, and C. Playoust, *The Magma algebra system I: The user language*, 1997, http://www.maths.usyd.edu.au:8000/u/magma/, pp. 235–265.
3. J. E. Cremona, *Algorithms for modular elliptic curves*, second ed., Cambridge University Press, Cambridge, 1997.
4. Exp. 355, P. P. Deligne, *Formes modulaires et representations l-adiques*, Séminaire Bourbaki. Vol. 1968/69: Exposés 347–363. LNM 179, Springer-Verlag, 139–172.
5. N. Dummigan, *Period ratios of modular forms*, Math. Ann. 318 (2000), no. 3, 621–636.
6. E. V. Flynn, F. Leprévost, E. Schaefer, W. Stein, M. Stoll and J. Wetherell. *Empirical evidence for the Birch and Swinnerton-Dyer conjectures for modular Jacobians of genus 2 curves.* Math. Comp. 70 (2001), no. 236, 1675–1697
7. G. Frey, M. Müller, *Arithmetic of modular curves and applications.* Algorithmic algebra and number theory (Heidelberg, 1997), 11–48, Springer, Berlin, 1999.
8. D. R.Kohel and Stein, W. A. *Component groups of quotients of* $J_0(N)$, Algorithmic number theory (Leiden, 2000), 405–412, Lecture Notes in Comput. Sci., 1838.
9. R. S. Kulkarni, *An arithmetic-geometric method in the study of the subgroups of the modular group.* Amer. J. Math. **113** (1991), no. 6, 1053–1133
10. J. I. Manin, *Parabolic points and zeta functions of modular curves*, Izv. Akad. Nauk SSSR Ser. Mat. **36** (1972), 19–66.
11. B. Mazur, *Courbes elliptiques et symboles modulaires*, Séminaire Bourbaki, 24éme année (1971/1972), Exp. No. 414, pp. 277–294. Lecture Notes in Math., Vol. 317, Springer, Berlin, 1973.
12. B. Mazur, P. Swinnerton-Dyer, *Arithmetic of Weil curves*, Invent. Math. 25 (1974), 1–61.
13. L. Merel, *Intersections sur des courbes modulaires.*, Manuscripta Math. **80** (1993), no. 3, 283–289.
14. L. Merel, *Universal Fourier expansions of modular forms*, On Artin's conjecture for odd 2-dimensional representations (Berlin), Springer, 1994, pp. 59–94.
15. V. V. Šokurov, *Modular symbols of arbitrary weight*, Funkcional. Anal. i Priložen. **10** (1976), no. 1, 95–96.
16. W. A. Stein, *Explicit approaches to modular abelian varieties*, U. C. Berkeley Ph.D. thesis (2000).
17. W. A. Stein Modular Symbols, Chapter 88-90 in *The Magma Handbook*, Volume 7, J. Cannon, W. Bosma Eds., (2001).
18. W. A. Stein and H. A. Verrill, *Cuspidal modular symbols are transportable*, LMS Journal of Computational Mathematics **4** (2001), 170–181.
19. H. A. Verrill, *Lattices of higher weight modular symbols*, preprint.
20. H. A. Verrill, *Subgroups of* $\mathrm{PSL}_2(\mathbf{R})$, Chapter in *The Magma Handbook*, Volume 2, J. Cannon, W. Bosma Eds., (2001), 233–254.

Action of Modular Correspondences around CM Points

Jean-Marc Couveignes and Thierry Henocq

Groupe de Recherche en Informatique et Mathématiques du Mirail[*],
Université de Toulouse II, 5 allées Antonio Machado, 31058, Toulouse, France
{couveig, henocq}@univ-tlse2.fr
http://www.univ-tlse2.fr/grimm

Abstract. We study the action of modular correspondences in the p-adic neighborhood of CM points. We deduce and prove two stable and efficient p-adic analytic methods for computing singular values of modular functions. On the way we prove a non trivial lower bound for the density of smooth numbers in imaginary quadratic rings and show that the canonical lift of an elliptic curve over $\mathbb{F}_q$ can be computed in probabilistic time $\ll \exp((\log q)^{\frac{1}{2}+\epsilon})$ under GRH. We also extend the notion of canonical lift to supersingular elliptic curves and show how to compute it in that case.

1 Introduction

Let $X \to X(1)$ be any modular curve seen as a covering of $X(1)$. Let P be a Heegner point on X and let $f \in \bar{\mathbb{Q}}(X)$ be a $\bar{\mathbb{Q}}$-rational function.

For reasonable choices of f, class field theory ensures that $f(P)$ is an algebraic integer. It is a classical algorithmic problem to compute the minimum polynomial of $f(P)$.

The known methods for this rely on complex analytic uniformization of X and provide complex approximations for $f(P)$ and its conjugates f_i. See [5] for a recent general study of this approach.

One then forms and expands the degree h minimal polynomial $\mu(X) = \prod_i (X - f_i)$ the coefficient of which are rational integers.

The difficulty with this method (that appears in quite a range of different contexts) is that it is very hard to control the loss of accuracy while expanding μ.

The only rigorous available evaluations of how many digits are needed are a bit alarming (see [1, Section 7] and [2, Section 9]).

It is thus temptating to look for a p-adic analytic method for computing singular values of modular functions. The reason for that is that the p-adic absolute accuracy is conserved when adding or multiplying two p-adic integers

[*] The GRIMM is supported by the French Ministry of Research through *Action Concertée Incitative CRYPTOLOGIE*, by the Direction Centrale de la Sécurité des Systèmes d'Information and by the Centre Électronique de L'ARmement.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 234–243, 2002.

i.e if one knows a and b up to $O(p^k)$ then one knows $a+b$ and ab up to $O(p^k)$ also.

One may logically look for some p-adic uniformization of X but such an uniformization does not exist in general. In particular it does not exist in the *most important* case of $X = X(1)$.

Instead of that we define and study a representation of the ideal group of an imaginary quadratic order as automorphism group of a p-adic neighborhood of the associated CM points. This representation is quite computational and the CM points are characterized and computed as fixed points of this representation. In this way we also manage to define canonical lifts for supersingular curves.

All this leads to two different proven stable and efficient methods for computing singular values of modular functions.

The reader who is not completely unwilling to read mathematics may also find some intrinsic interest to the p-adic representation itself and to our lemmata.

2 Modular Correspondences in the Neighborhood of CM Points

We refer to [8] for the elementary theory of complex multiplication.

We start with

Definition 1. *Let k be an algebraically closed field and $\mathcal{O}$ the imaginary quadratic order with discriminant $-\Delta$. We denote by $\mathcal{NELL}_\Delta(k)$ the set of isomorphism classes of couples (E,ι) where E is an elliptic curve over k and $\iota : \mathcal{O} \to \mathcal{E}nd(E)$ is a maximal embedding (when E is ordinary ι is an isomorphism). Such a couple is called a normalized elliptic curve. We say that two normalized elliptic curves (E,ι) and (E',ι') are isomorphic if there is an isomorphism $I : E \to E'$ such that $I^{-1}\iota'(X)I = \iota(X)$ for any X in $\mathcal{O}$.*

We denote by $\mathcal{ELL}_\Delta(k)$ the quotient of $\mathcal{NELL}_\Delta(k)$ by the action of complex conjugation. When the characteristic p of k has two primes in the fraction field of $\mathcal{O}$ above it then $\mathcal{ELL}_\Delta(k)$ is the set of isomorphism classes of curves with CM by $\mathcal{O}$.

We now fix an embedding of $\bar{\mathbb{Q}}$ in $\mathbb{C}$. Let $\mathcal{O}$ be a quadratic order with group of units $\{1,-1\}$, class group $\mathcal{C}\ell(\mathcal{O})$, conductor m and discriminant $-\Delta$. Then $\mathcal{ELL}_\Delta(\bar{\mathbb{Q}})$ is the finite set of isomorphism classes of elliptic curves over $\bar{\mathbb{Q}}$ with complex multiplication by $\mathcal{O}$. We may see it as a reduced zero dimensional subvariety in $X(1) = \mathbb{P}^1 - \{\infty\}$, the moduli space of elliptic curves. There is a free faithful action of $\mathcal{C}\ell(\mathcal{O})$ on it.

We fix a prime p and an embedding of $\bar{\mathbb{Q}}$ in $\mathbb{C}_p$ and denote by $\bar{\mathbb{F}}_p$ the residue field of $\mathbb{C}_p$. We assume that p has two primes of $\mathbb{Q}(\sqrt{-\Delta})$ above it. Then $\mathcal{ELL}_\Delta(\bar{\mathbb{Q}})$ splits over $\mathbb{F}_q$ with $q = p^d$ and $d = c\ell(\mathcal{O}')$ where $\mathcal{O}'$ is the order with conductor m' the larger prime to p factor of m. We call $-\Delta'$ the discriminant of $\mathcal{O}'$. We know that reduction modulo p induces a surjection from $\mathcal{ELL}_\Delta(\bar{\mathbb{Q}})$ onto $\mathcal{ELL}_{\Delta'}(\bar{\mathbb{F}}_q)$. This is the set of isomorphism classes of elliptic curves over $\bar{\mathbb{F}}_q$ with CM by $\mathcal{O}'$. It has cardinality $c\ell(\mathcal{O}')$ and is acted on by $\mathcal{C}\ell(\mathcal{O})$. We also assume that $\mathcal{O}'$ has unit group $\{1,-1\}$.

Let $\mathcal{ELL}_\Delta^\circ$ be the set of isomorphism classes of elliptic curves over $\mathbb{C}_p$ that reduce modulo p to an elliptic curve in $\mathcal{ELL}_{\Delta'}(\bar{\mathbb{F}}_q)$. Using the modular invariant j this set can be given an analytic structure and is the disjoint union of $c\ell(\mathcal{O}')$ open p-adic disks of radius 1. Every such disk contains $c\ell(\mathcal{O})/c\ell(\mathcal{O}')$ elements in $\mathcal{ELL}_\Delta(\bar{\mathbb{Q}})$.

To every point in $\mathcal{ELL}_\Delta(\bar{\mathbb{Q}})$ we associate an ideal $\mathfrak{a} \subset \mathcal{O} \subset \bar{\mathbb{Q}} \subset \mathbb{C}$ and a model $E_\mathfrak{a} = \mathbb{C}/\mathfrak{a}$ for the corresponding isomorphism class. This way, all the curves $E_\mathfrak{a}$ share the same endomorphism ring $\mathcal{O}$. The reductions $E_\mathfrak{a}$ mod p provide models for the elements in $\mathcal{ELL}_{\Delta'}(\bar{\mathbb{F}}_q)$. Whenever there is no risk of confusion, we shall denote by $\mathfrak{a}$ a point in $\mathcal{ELL}_\Delta(\bar{\mathbb{Q}})$ or $\mathcal{ELL}_{\Delta'}(\bar{\mathbb{F}}_q)$.

If $\mathfrak{i}$ is a prime to m ideal in $\mathcal{O}$ we denote by $E_\mathfrak{a}[\mathfrak{i}]$ the intersection of kernels of all endomorphisms in $\mathfrak{i}$. Quotienting by this subgroup defines an isogeny $E_\mathfrak{a} \to E_{\mathfrak{a}\mathfrak{i}^{-1}}$. If $\mathfrak{b}$ represents the class of $\mathfrak{a}\mathfrak{i}^{-1}$ we set $\mathfrak{i} \bullet \mathfrak{a} = \mathfrak{b}$. If further $\mathfrak{i}$ is prime to p we similarly define an isogeny from the reduction $\bar{E}_\mathfrak{a}$ modulo p of $E_\mathfrak{a}$.

Thus the group $I(pm)$ of prime to pm ideals of $\mathcal{O}$ acts on both $\mathcal{ELL}_\Delta(\bar{\mathbb{Q}})$ and $\mathcal{ELL}_{\Delta'}(\bar{\mathbb{F}}_q)$ and the reduction map is equivariant for these actions.

We now show how this action extends to a continuous action on $\mathcal{ELL}_\Delta^\circ$. Let x be a point in $\mathcal{ELL}_\Delta^\circ$. Let $\mathfrak{a}$ be a point in $\mathcal{ELL}_\Delta(\bar{\mathbb{Q}})$ which is close to x and let $E_\mathfrak{a} = \mathbb{C}/\mathfrak{a}$ be the corresponding elliptic curve. We denote by $D_\mathfrak{a}$ the disk in $\mathcal{ELL}_\Delta^\circ$ that contains $\mathfrak{a}$ and x. Let E_x be a model for x which is close to $E_\mathfrak{a}$ i.e. an elliptic curve over $\mathbb{C}_p$ such that $j(E_x) = j(x)$ and E_x and $E_\mathfrak{a}$ have equal reductions modulo p (so E_x is the fiber at x in the universal curve over $D_\mathfrak{a}$ and this universal curve exists because $D_\mathfrak{a}$ does not contain $j = 0$ nor $j = 1728$.) Let $\mathfrak{i}$ be an ideal in $I(pm)$ and set $\mathfrak{b} = \mathfrak{i} \bullet \mathfrak{a}$. Let $E_\mathfrak{a}[\mathfrak{i}]$ be the finite subgroup of $E_\mathfrak{a}$ defined by $\mathfrak{i}$. Because $\mathfrak{i}$ is prime to p this group 'lifts' to a group scheme over $D_\mathfrak{a}$ whose fiber at x defines a subgroup $E_x[\mathfrak{i}]$ of E_x. The quotient of E_x by this group defines a point $y = \mathfrak{i} \bullet x$ in $\mathcal{ELL}_\Delta^\circ$ which is close to $\mathfrak{b}$.

For every $\mathfrak{i} \in I(pm)$ the map $[\mathfrak{i}] : x \mapsto \mathfrak{i} \bullet x$ is a continuous map on $\mathcal{ELL}_\Delta^\circ$. Indeed, let $\mathfrak{j}$ be an ideal in $\mathcal{O}$ and α a rational integer such that $\mathfrak{i} = (\alpha)\mathfrak{j}$ and $\mathcal{O}/\mathfrak{j}$ is cyclic of order N. Then $[\mathfrak{i}]$ being the restriction of the level N correspondence is an algebraic map. We recall that the level N correspondence is the divisor on $X(1) \times X(1)$ image of $X_0(N)$ by the map $(E \to E') \mapsto (j(E), j(E'))$. The curve $X_0(N)$ has good reduction modulo p and $\mathfrak{a} \in X_0(N)$ is not p-adically close to any ramification point of j or j'. So $j' - j'(\mathfrak{a})$ is an integral invertible series in $j - j(\mathfrak{a})$ and the radius of convergence of $[\mathfrak{i}]$ is 1. The integer α being inessential we shall assume $\alpha = 1$ and $\mathfrak{i} = \mathfrak{j}$. In that case we say that $\mathfrak{i}$ is *reduced.* The inverse of $[\mathfrak{i}]$ is $[\bar{\mathfrak{i}}]$ given by complex conjugation.

We thus have constructed a morphism ρ from the group $I(pm)$ of prime to pm ideals of $\mathcal{O}$ to the group $\mathrm{Aut}(\mathcal{ELL}_\Delta^\circ)$ of automorphisms of the analytic variety $\mathcal{ELL}_\Delta^\circ$. The restriction of ρ to the group $P(pm)$ of prime to pm principal ideals of $\mathcal{O}$ defines a morphism (still denoted by ρ)

$$\rho : P(pm) \to \mathrm{Aut}^*(\mathcal{ELL}_\Delta^\circ)$$

to the group of automorphisms that fix $\mathcal{ELL}_\Delta(\bar{\mathbb{Q}})$ (the CM points) and therefore stabilize every disk $D_\mathfrak{a}$.

In order to study this morphism we denote by $\delta_{\mathfrak{a}} : \mathrm{Aut}^*(\mathcal{ELL}^\circ_\Delta) \to \mathbb{C}_p^*$ the differentiation at the CM point $\mathfrak{a}$.

From lemma 1 below we deduce that $\delta_{\mathfrak{a}} \circ \rho : P(pm) \to \mathbb{C}_p^*$ is independent of $\mathfrak{a}$, takes values in $\bar{\mathbb{Q}}^*$ and $\delta_{\mathfrak{a}}(\rho((\mathcal{L}))) = \mathcal{L}\mathcal{L}^*$ where $\mathcal{L}^* = \bar{\mathcal{L}}^{-1}$. In particular, the kernel of ρ consists of ideals $(\mathcal{L})$ with $\mathcal{L} \in \mathbb{Q}^*$ prime to pm.

Lemma 1. *Let $\mathcal{O}$ be a quadratic order with group of units $\{1, -1\}$ and conductor m. Let $\mathcal{L} \in \mathcal{O}$ such that $\mathcal{O}/\mathcal{L}$ is cyclic of order N. Let j and j' be the two functions on $X_0(N)$ defined by $j(E \to E') = j(E)$ and $j'(E \to E') = j(E')$. The value of the slope of the tangent $\sigma = \frac{dj'}{dj}$ at all Heegner points with CM by $\mathcal{O}$ and representing multiplication by $\mathcal{L}$ isogenies is $\mathcal{L}\mathcal{L}^*$.*

The order $\mathcal{O}$ has discriminant $-\Delta = -m^2 D$ and basis $(1, m\frac{\sqrt{-D}-D}{2})$ and $\mathcal{L} = a + bm\frac{\sqrt{-D}-D}{2}$ has norm $N = a^2 - abDm + b^2Km^2$ with $K = D(D+1)/4$. Set $\alpha = m\frac{\sqrt{-D}-D}{2}$ and let c be an integer congruent to a/b modulo N. We have $\alpha^2 + Dm\alpha + Km^2 = 0$. Define the two integers $u = \frac{a-bc}{N}$ and $v = b\frac{c^2-cDm+Km^2}{N}$. Note that b is invertible modulo N because $\mathcal{L}$ is reduced. We look for the Smith normal form of $(\mathcal{L}) \subset \mathcal{O}$. Let $\phi : \mathcal{O} \to \mathbb{Z}$ be the linear form defined by $\phi(x+y\alpha) = x - cy$ that induces an isomorphism $\mathcal{O}/\mathcal{L} \overset{\phi}{\to} \mathbb{Z}/N\mathbb{Z}$. Together with the linear form ψ defined by $\psi(x + y\alpha) = y$ this makes a basis (ϕ, ψ) for the dual of $\mathcal{O}$. A dual basis for $\mathcal{O}$ is $(1, \beta)$ with $\beta = c + \alpha$. A basis for $(\mathcal{L})$ is then (N, β) and this is the Smith normal form. The lattice $\mathcal{L}^*\mathcal{O} = \frac{1}{N}(\mathcal{L})$ admits the two basis $(1, \frac{\beta}{N})$ and $(\mathcal{L}^*, \mathcal{L}^*\beta)$ with transition matrix $\mathcal{M} \in \mathrm{PSL}_2(\mathbb{Z})$

$$\begin{pmatrix} \mathcal{L}^*\beta \\ \mathcal{L}^* \end{pmatrix} = \mathcal{M}\begin{pmatrix} \frac{\beta}{N} \\ 1 \end{pmatrix} = \begin{pmatrix} bc + a - bDm & -v \\ b & u \end{pmatrix}\begin{pmatrix} \frac{\beta}{N} \\ 1 \end{pmatrix}.$$

The class of $\tau = \frac{\beta}{N}$ modulo the action of $\Gamma_0(N)$ on the upper half plane represents the N-isogeny $\mathbb{C}/(1,\tau) \overset{\times N}{\to} \mathbb{C}/(1, N\tau) \overset{\times \mathcal{L}^*}{\simeq} \mathbb{C}/(1,\tau)$ which is an endomorphism. So τ is a Heegner point associated to multiplication by $\mathcal{L}$ endomorphism. Since $\frac{dj}{d\tau}$ is a constant times $j\frac{E_6}{E_4}$, the slope $\frac{dj'}{dj}$ is $N\frac{j'}{j}\frac{E_4}{E_4'}\frac{E_6'}{E_6}$ and since $N\tau = \mathcal{M}\tau$ the slope at τ is $N(b\tau + u)^2$ which is easily seen to be independent of c and equal to $\mathcal{L}\mathcal{L}^*$. There are $c\ell(\mathcal{O})$ Heegner points of level N with complex multiplication by $\mathcal{O}$ and representing the multiplication by $\mathcal{L}$ isogeny, all defined over the Hilbert class field of $\mathcal{O}$ and conjugated over $\mathbb{Q}(\sqrt{-\Delta})$.

Since $\mathcal{L}\mathcal{L}^*$ belongs to the later field, the slope is the same at all such Heegner points. □

We observe that the action of a reduced ideal $\mathfrak{i}$ of norm N on a point $x \in \mathcal{ELL}^\circ_\Delta$ can be computed in time polynomial in N, $\log q$, and almost linear in the p-adic accuracy of x i.e. the number of significant terms in its p-adic expansion. One first reduces to the case N is prime (not essential but simpler). One then computes the kernel $\bar{E}_{\mathfrak{a}}[\mathfrak{i}]$ of the isogeny modulo p thanks to Atkin-Elkies techniques (see [15]). This kernel is then lifted on E_x thanks to Hensel's lemma. The isogeny $E_x \to E_y$ follows using Vélu's formulae [18].

We summarize in

Theorem 1. *Let $\mathcal{O}$ be a quadratic order, p a prime and $\mathcal{O}'$ the smallest p-maximal overorder of $\mathcal{O}$. Assume $\mathcal{O}'$ has group of units $\{1, -1\}$. Let m be the conductor of $\mathcal{O}$. The group $P(pm)$ of prime to pm principal ideals of $\mathcal{O}$ has a modular representation ρ as automorphism group of the p-adic disk with radius 1 in $X(1)$ around any point $\mathfrak{a}$ with CM by $\mathcal{O}$. The differentiation of this representation is just $\mathcal{L} \in P(pm) \mapsto \mathcal{L}\mathcal{L}^*$. The action of $\rho(\mathcal{L})$ on a given point can be computed in time polynomial in N, n, $\log q$ and almost linear in k i.e. $k(\log k)^{O(1)}$ where N is the norm of the bigger prime ideal factor of $\mathcal{L}$, and n is the number of such factors with multiplicities, $\mathbb{F}_q$ is the residue field of $\mathfrak{a}$ and k is the desired accuracy of the result.*

Remark 1. If $\mathcal{O}'$ is $\mathbb{Z}[i]$ (resp. $\mathbb{Z}[\rho]$) then the theorem holds with $\mathcal{L}\mathcal{L}^*$ replaced by $(\mathcal{L}\mathcal{L}^*)^2$ (resp. $(\mathcal{L}\mathcal{L}^*)^3$.)

Remark 2. The $\bullet$ action of principal ideals in $\mathcal{O}'$ (not necessarily principal in $\mathcal{O}$) on the set $\mathcal{ELL}_\Delta(\bar{\mathbb{Q}})$ is a Galois action and can be expressed in terms of the Artin map.

3 Computing the Canonical Lift in All Characteristics

In this section we are interested in computing p-adic approximations of the canonical lift of an ordinary elliptic curve over a finite field.

We shall restrict to the case p is prime to the conductor m. So p splits in $\mathcal{O}$. If this is the case the reduction map

$$R : \mathcal{ELL}_\Delta(\bar{\mathbb{Q}}) \to \mathcal{ELL}_\Delta(\bar{\mathbb{F}}_q)$$

is an equivariant bijection.

We shall prove the

Theorem 2. *Assuming GRH, for any positive ϵ there is an algorithm that computes the inverse of the reduction map R at a given point x in $\mathcal{ELL}_\Delta(\bar{\mathbb{F}}_q)$ in probabilistic time*

$$\left[\exp((\log q)^{\frac{1}{2}+\epsilon}) \times \log k\right]^{O(1)} \times k$$

with accuracy k i.e. the error is $O(p^k)$.

In order to prove 2 we give and discuss an algorithm. For fixed ϵ the algorithm goes as follows. We first call E the curve over $\mathbb{F}_q$ associated to the point x. We look for the canonical lift of E.

If the characteristic p of $\mathbb{F}_q$ is less than $2\exp((\log 4q)^{\frac{1}{2}+\epsilon})$ we lift E together with all its conjugates over $\mathbb{F}_p$ using the equations in Lubin and Tate and Serre's work [16,11] and/or the cousin algorithm used in Satoh's algorithm [13]. The running time is polynomial in p and the degree d of $\mathbb{F}_q$ over $\mathbb{F}_p$. The result follows.

If $p > 2\exp((\log 4q)^{\frac{1}{2}+\epsilon})$ we make use of smooth isogenies in the spirit of Oesterlé and Mestre's method [12] and Kohel's thesis [6]. We compute the trace t

of the Frobenius Φ of E using Schoof's algorithm [14]. Let $-\Delta$ be the discriminant of $\mathbb{Z}[\Phi]$ and let $\mathcal{A}$ be the set of prime to $p\Delta$ integers of the form $a+b\Phi$ with $1 \le b \le 2\exp((\log\Delta)^{\frac{1}{2}+\epsilon})$ and $|a+\frac{1}{2}bt| \le \Delta^{\frac{1}{2}}\exp((\log\Delta)^{\frac{1}{2}+\epsilon})$. Let $B = \lfloor \exp(\sqrt{\log\Delta})\rfloor$. We say that an integer in $\mathbb{Z}[\Phi]$ is B-smooth iff all its prime factors have norm bounded by B. We assume Δ is big enough to apply lemma 2. Otherwise we may just read the result in a table. We pick random elements in $\mathcal{A}$ with uniform probability until we find one $\mathcal{L}$ which is B-smooth. By lemma 2 we succeed after $\ll \exp(2(\log\Delta)^{\frac{1}{2}}\log\log\Delta)$ attempts with bounded probability. This is the only probabilistic step in the algorithm. We now choose any lift E_1 of E and call j_1 its j invariant and compute $\mathcal{L}\bullet E_1$. This is done step by step, applying successively all prime factors of $\mathcal{L}$. So the running time is polynomial in B. We denote by $\mathcal{L}\bullet j_1$ the j-invariant of $\mathcal{L}\bullet E_1$ and set

$$j_{k+1} = j_k - \frac{\mathcal{L}\bullet j_k - j_k}{\sigma - 1}$$

for $k \ge 1$ where $\sigma = \mathcal{L}\mathcal{L}^*$.

If j_∞ is the j-invariant of the canonical lift we check that $|j_{k+1} - j_\infty| \le |j_k - j_\infty|^2$. This is just the Newton's tangent method. It is decisive however for this convergence property to hold that $\sigma - 1$ be a p-adic unit. It is a unit indeed otherwise we would have $\mathcal{L} \equiv \bar{\mathcal{L}} \pmod{p}$ so $p|b$ since E is ordinary. But this would contradict our assumption that $p > 2\exp((\log\Delta)^{\frac{1}{2}+\epsilon})$. □

Lemma 2. *Fix an* ϵ *in* $]0,\frac{1}{2}[$. *Let* Φ *be an imaginary quadratic integer and* t *and* q *two integers such that* $\Phi^2 - t\Phi + q = 0$. *Let* $-\Delta = t^2 - 4q$ *be the discriminant of the order generated by* Φ. *Let* $B = \lfloor\exp(\sqrt{\log\Delta})\rfloor$. *Let* $\mathcal{A}$ *be the set of prime to* $q\Delta$ *integers of the form* $a+b\Phi$ *with* $1 \le b \le 2\exp((\log\Delta)^{\frac{1}{2}+\epsilon})$ *and* $|a+\frac{1}{2}bt| \le \Delta^{\frac{1}{2}}\exp((\log\Delta)^{\frac{1}{2}+\epsilon})$. *If GRH holds the proportion of* B*-smooth elements in* $\mathcal{A}$ *is* $\ge \exp(-2(\log\Delta)^{\frac{1}{2}}\log\log\Delta)$ *if* Δ *is big enough (depending on* ϵ*).*

We now prove lemma 2. Call $\mathcal{D}$ the set of prime to $p\Delta$ primes in $\mathbb{Z}[\Phi]$ with degree one and norm less than B. Let $\mathcal{B} \subset \mathcal{D}$ be a system of coset representatives for the action of complex conjugation on $\mathcal{D}$ i.e. $\mathcal{D} = \mathcal{B}\cup\bar{\mathcal{B}}$ and $\mathcal{B}\cap\bar{\mathcal{B}} = \emptyset$. Let $\mathcal{O} = \mathbb{Z}[\Phi]$ and $h = c\ell(\mathcal{O}) < \Delta^{\frac{1}{2}}\log\Delta$ by a result of Lenstra and Pomerance [10]. From Lagarias and Odlyzko [7] the size π of $\mathcal{B}$ is at least $\frac{B}{3\log B}$ if Δ is big enough. Set $u = \lfloor\frac{\sqrt{\log\Delta}}{2} + (\log\Delta)^\epsilon\rfloor$ and let $\mathcal{S}^u\mathcal{B}$ be the u-th symmetric product of $\mathcal{B}$. Let $\kappa : \mathcal{S}^u\mathcal{B} \to \mathcal{C}\ell(\mathcal{O})$ be defined by $\kappa(\{\mathfrak{p}_1, ..., \mathfrak{p}_u\})$ is the class of the product $\prod_{1\le k\le u}\mathfrak{p}_k$. Let $\mathcal{F} \subset \mathcal{S}^u\mathcal{B}\times\mathcal{S}^u\mathcal{B}$ be the subset of couples (V_1, V_2) such that $V_1 \ne V_2$ and $\kappa(V_1) = \kappa(V_2)$. The average size of fibers of κ is $\ge \lfloor\frac{\pi^u}{u!}\rfloor h^{-1} \ge \lfloor\frac{\pi^u}{u!h}\rfloor - 2$ which is bigger than $\exp(\frac{2\log\Delta^{\frac{1}{2}+\epsilon}}{3})$ when Δ is big enough. The size of $\mathcal{F}$ is minimum when all fibers have equal cardinality so the size of $\mathcal{F}$ is at least $(\lfloor\frac{\pi^u}{u!h}\rfloor - 2)(\lfloor\frac{\pi^u}{u!h}\rfloor - 3)h \ge \frac{\pi^{2u}}{2h(u!)^2}$ for Δ big enough. To every couple (V_1, V_2) in $\mathcal{F}$ one associates the product of primes in V_1 together with conjugates of primes in V_2. Let $\mu(V_1, V_2)$ be the unique generator of this ideal of the form $a+b\Phi$ with b positive. We observe that this integer exists because the concerned ideal is principal in $\mathcal{O}$. It has norm $(a+\frac{bt}{2})^2+\frac{\Delta}{4}b^2$ bounded by $\Delta\exp(2(\log\Delta)^{\frac{1}{2}+\epsilon})$ and it is not in $\mathbb{Z}$ because $V_1\neq V_2$. So μ is a map from $\mathcal{F}$ to $\mathcal{A}$. The size of a fiber of μ is bounded by $\binom{2u}{u}$.

So the image of μ which is made of B-smooth elements in $\mathcal{A}$ has size at least $\frac{\pi^{2u}}{2(2u)!h}$. The proportion of B-smooth elements in $\mathcal{A}$ is thus

$$\geq \exp(-\frac{3}{2}(\log\Delta)^{\frac{1}{2}}\log\log\Delta+O((\log\Delta)^{\frac{1}{2}}))$$

which is bigger than $\exp(-2(\log\Delta)^{\frac{1}{2}}\log\log\Delta)$ when Δ is big enough. □

Remark 3. The method of Lubin-Serre-Tate used by Satoh and its variants (especially Mestre's ones using Algebraic Geometrical Means that stresses the underlying dynamical system [3]) use degree p isogenies to compute the canonical lift. We avoid them on the contrary. Firstly because p might be too big and secondly because the slope of a level p correspondence at a CM point is not a p-adic unit. This is not necessarily an inconvenient but it requires a different treatment. Indeed the level p correspondence induces a contracting map on the p-adic neigborhood of CM points that Serre uses to prove the existence and unicity of the canonical lift using the fixed point theorem.

4 Singular Values of Modular Functions

Being able to lift an ordinary elliptic curve we may also lift torsion points on it and this gives a p-adic method for computing p-adic approximations of singular values of any modular function $f\in\bar{\mathbb{Q}}(X)$ at a point P with CM by an order $\mathcal{O}$, provide we are given an ordinary elliptic curve with complex multiplication by $\mathcal{O}$.

This gives a stable and efficient method for computing (ray) class fields.

Indeed, given a negative discriminant $-\Delta$ we first look for the smallest prime to Δ square t^2 such that $t^2+\Delta$ is four times a prime $p=q$. We expect the smallest such t to be quite small (e.g. $(\log\Delta)^{O(1)}$) so that $4q$ is very close to Δ. Even GRH cannot ensure this however.

We then look for an elliptic curve over $\mathbb{F}_q$ with trace t. This is done by choosing random elliptic curves modulo q and requires $q/c\ell(-\Delta)$ trials which is less than $q\Delta^{-\frac{1}{2}+o(1)}$ by Siegel's theorem. Any trial takes time $(\log q)^{O(1)}$ using Schoof's algorithm. This is hopefully $O(\Delta^{\frac{1}{2}+o(1)})$. We then lift this curve using the methods presented above. We thus compute p-adic approximations for all conjugates of an element f in the Hilbert class field of the order with discriminant $-\Delta$ and all this in time $hk^{1+o(1)}\Delta^{o(1)}$ where $h=c\ell(-\Delta)$ is the class number of the order with discriminant $-\Delta$.

If we now want to reconstruct the minimal polynomial of f, we need a bound for the logarithm of coefficients of this polynomial. For reasonable functions (e.g. the modular invariant j see [9, 5.10]) this bound is $O(h^{1+\epsilon})$ so we need accuracy $k=O(h^{1+\epsilon})$ so that the algorithm runs in probabilistic expected time $O(h^{2+\epsilon})$ which is essentially linear in the size of the result and certainly better than the tremendous (but somewhat pessimistic) estimate in [1]. Indeed our method avoids the accuracy problems of the classical one (evaluating modular functions at CM points in the upper half plane). It is compatible with the improvement given by Gee and Stevenhagen in [5] where functions $\eta(Nz)/\eta(z)$ are used (that generalize Weber's functions) together with a rationality criterion deduced from Shimura's reciprocity law.

We now can state the

Theorem 3. *If G.R.H. holds, for any positive ϵ there is an algorithm that computes the Hilbert class polynomial of discriminant $-\Delta$ in probabilistic time $O(\Delta^{1+\epsilon})$.*

The algorithm presented above does not quite prove the theorem since there is no proof that a small enough t exists such that $\Delta + t^2$ is four times a prime.

However, G.R.H. ensures that there exists a principal prime ideal in the Hilbert class field with norm less than a constant times

$$h^2(\log h)^4(\log \Delta)^2(\log\log \Delta)^4$$

which is $O(\Delta(\log \Delta)^8(\log\log \Delta)^4)$ by Lenstra an Pomerance [10].

Therefore there exist $t = \sqrt{\Delta}(\log \Delta)^{4+o(1)}$ and $u = (\log \Delta)^{4+o(1)}$ such that $t^2 + u^2\Delta$ is four times a prime p. Such a pair (t, u) may be found by exhaustive search. The rest of the algorithm goes as above except that in the end we obtain an elliptic curve with CM by an order of discriminant $-u\Delta$. Applying isogenies of degree dividing u we obtain en elliptic curve with CM by the order with discriminant $-\Delta$. □

Remark 4. There is a tentative algorithm for computing CM fields in [2]. This method (Algorithm 3 on page 100) collects information modulo many small primes ℓ by exhaustive search among elliptic curves modulo ℓ for every ℓ. It is overexponential in the class number h however, contrary to the author's claim. The definition field of ordinary elliptic curves used in this method has degree $O(h)$ over $\mathbb{F}_\ell$ and the exhaustive search takes time $O(\ell^h)$ rather than the claimed $O(h^2)$. So this algorithm is worse than any possible one.

It may be possible to turn it into something slightly more sensible by removing step 1 an dealing only with primes with supersingular reductions. Even with this restriction, working with several moduli is not a good idea. See section 5.

5 Canonical Lift of Supersingular Curves

In this section we adapt our ideas to the case of curves with supersingular reduction. We keep the notation of section 2. We assume p has a single prime of $\mathbb{Q}(\sqrt{-\Delta})$ above it. We assume the order $\mathcal{O}$ with discriminant Δ is maximal. In this case the inertia degree d of p in the Hilbert class field is 1 or 2 and $q = p$ or p^2.

Reduction modulo p of curves with CM by $\mathcal{O}$ needs not be injective. However, let $\mathfrak{A}$ be the quaternion algebra ramified at p and ∞ and for every supersingular curve E modulo p let $i_E : \mathfrak{A} \to \mathcal{E}nd(E) \otimes \mathbb{Q}$ be a fixed isomorphism as in

Waterhouse [19]. This way, all endomorphism rings of all supersingular curves are seen as maximal orders inside the same algebra $\mathfrak{A}$. We denote by $\mathcal{E}nd(E)$ the endomorphism ring of E over $\bar{\mathbb{F}}_q$.

Reduction of a normalized curve (E,ι) in $\mathcal{NELL}_\Delta(\bar{\mathbb{Q}})$ thus gives a supersingular curve $\bar{E} = E \mod p$ together with an injection of $\mathcal{O}$ in the maximal order $i_{\bar{E}}(\mathcal{E}nd(\bar{E}))$ of $\mathfrak{A}$.

This is an element of $\mathcal{NELL}_\Delta(\bar{\mathbb{F}}_p)$ the set of isomorphism classes of supersingular curves modulo p normalized with the order $\mathcal{O}$ with discriminant $-\Delta$.

We prove the

Theorem 4. *Let $-\Delta$ be a primitive discriminant and $\mathcal{O}$ the quadratic imaginary maximal order with discriminant $-\Delta$ and p an odd inert prime number in $\mathcal{O}$. The reduction map*

$$R : \mathcal{NELL}_\Delta(\bar{\mathbb{Q}}) \to \mathcal{NELL}_\Delta(\bar{\mathbb{F}}_p)$$

is a bijection.

Its inverse will be called the canonical lift on normalized supersingular curves.

We first observe that the two sets have equal cardinality by one of the many Eichler formulae [4, Proposition 5] and [17, Theorem 2.4.].

We also note that $\mathcal{O}$ has a prime to p element $\mathcal{L}$ such that $\mathcal{L}\mathcal{L}^* \not\equiv 1 \mod p$. This together with theorem 1 and remark 1 implies that R is injective. □

Remark 5. If p ramifies in $\mathcal{O}$ the reduction map is no longer a bijection. It is a two to one surjection. One may define a pair of canonical lifts at p-adic distance $\frac{1}{2}$ of each other.

Remark 6. The theorem above suggests possible generators for the ring of integers of the Hilbert class field.

As for explicit computation of the canonical lift we observe that results and algorithms in section 2 generalize to the case with supersingular reduction.

Let E be a supersingular elliptic curve. Using the graph method of Oesterlé and Mestre we find in probabilistic time $O(p^{1+\epsilon})$ a basis for a sub-order R' of R with index M bounded by $p^{O(1)}$ and the associated quadratic form.

We now assume $\mathcal{O}$ is a maximal imaginary quadratic order where p stays inert and we look for an embedding of $\mathcal{O}$ into R. Since we do not know R we rather look for an embedding in R' of a sub-order $\mathcal{O}'$ of $\mathcal{O}$ with conductor m dividing M.

This boils down to representing $m^2\Delta$ by a positive definite quadratic form of rank three and discriminant $p^{O(1)}$ and is done in time $(p\log\Delta)^{O(1)}\Delta$ by mere exhaustive search and $(p\log\Delta)^{O(1)}$ heuristically by a random search.

This is a competitive approach for computing singular values of modular functions since we can find a very small (e.g. $(\log\Delta)^{O(1)}$ under GRH) inert prime p in $\mathcal{O}$.

The prime p is indeed very small since 3 is fine for half quadratic orders and 5 is fine for half the remaining ones etc. So the endomorphism rings of all supersingular curves modulo small primes can be precomputed together with their norm forms.

References

1. A.O. Atkin and F. Morain. Elliptic curves and primality proving. *Math. Comp.*, 61:29–68, 1993.
2. J. Chao, O. Nakamura, and K. Sobataka. Construction of secure elliptic cryptosystems using CM tests a nd liftings. *ASIACRYPT'98*, 1514:95–109, 1998.
3. Jean-François Mestre. Lettre à P. Gaudry et R. Harley, décembre 2000. *Private communication.*
4. M. Eichler. The basis problem for modular forms and the traces of the hecke operators. *Lecture Notes in Math.*, 320, 1973.
5. Alice Gee and Peter Stevenhagen. Generating class fields using Shimura reciprocity. *Lecture Notes in Computer Science*, 1423:441–453, 1998.
6. David Kohel. *Endomorphism rings of elliptic curves over finite fields.* Thesis. University of California at Berkeley, 1996.
7. J. Lagarias and A. Odlyzko. Effective versions of the Chebotarev density theorem. In A. Fröhlich, editor, *Algebraic Number Fields.* Academic Press, 1977.
8. Serge Lang. *Elliptic functions, second edition.* GTM. Springer, 1987.
9. H. W. Lenstra and A. Lenstra. Algorithms in number theory. *Handbook of Theoretical Computer Science, Algorithms and Complex ity*, A:673–718, 1990.
10. H. W. Lenstra and C. Pomerance. A rigorous time bound for factoring integers. *Journal of the American Mathematical Society*, 5(3):483–516, 1992.
11. J. Lubin, J.-P. Serre, and J. Tate. Elliptic curves and formal groups. *Lecture notes prepared in connection with the seminars held at t he Summer Institute on Algebraic Geometry, Whitney Estate, Woods Hole, Massachu setts, July 6-July 31, 1964*, http://www.ma.utexas.edu/users/voloch/lst.html:1–8, 1964.
12. J.-F. Mestre. La méthode des graphes. exemples et applications. *Proceedings of the international conference on class numbers and fundamental units of algebraic number fields (Katata, 1986)*, pages 217–242, 1986.
13. T. Satoh. The canonical lift of an ordinary elliptic curve over a finite field and its point counting. *J. Ramanujan Math. Soc.*, 15:247–270, 2000.
14. R. Schoof. Elliptic curves over finite fields and the computation of square roots modulo *p*. *Math. Comp.*, 44:183–211, 1985.
15. R. Schoof. Counting points on elliptic curves over finite fields. *Journal de Théorie des Nombres de Bordeaux*, 7:219–254, 1995.
16. J.-P. Serre. Groupes divisibles (d'après John Tate). *Séminaire Bourbaki*, 10(318):73–86, 1966.
17. Thomas R. Shemanske. Ternary quadratic forms and quaternion algebras. *Journal of Number Theory*, 23:203–209, 1986.
18. J. Vélu. Isogénies entre courbes elliptiques. *Comptes rendus à l'Académie des sciences de Paris*, 273, Série A:238–241, 1971.
19. William C. Waterhouse. Abelian varieties over finite fields. *Ann. scient. Ec. Norm. Sup.*, 2(4):521–560, 1969.

Curves $Dy^2 = x^3 - x$ of Odd Analytic Rank

Noam D. Elkies

Department of Mathematics, Harvard University, Cambridge, MA 02138 USA
elkies@math.harvard.edu

Abstract. For nonzero rational D, which may be taken to be a squarefree integer, let E_D be the elliptic curve $Dy^2 = x^3 - x$ over $\mathbf{Q}$ arising in the "congruent number" problem.[1] It is known that the L-function of E_D has sign -1, and thus odd analytic rank $r_{\mathrm{an}}(E_D)$, if and only if $|D|$ is congruent to 5, 6, or 7 mod 8. For such D, we expect by the conjecture of Birch and Swinnerton-Dyer that the arithmetic rank of each of these curves E_D is odd, and therefore positive. We prove that E_D has positive rank for each D such that $|D|$ is in one of the above congruence classes mod 8 and also satisfies $|D| < 10^6$. Our proof is computational: we use the modular parametrization of E_1 or E_2 to construct a rational point P_D on each E_D from CM points on modular curves, and compute P_D to enough accuracy to usually distinguish it from any of the rational torsion points on E_D. In the 1375 cases in which we cannot numerically distinguish P_D from $(E_D)_{\mathrm{tors}}$, we surmise that P_D is in fact a torsion point but that E_D has rank 3, and prove that the rank is positive by searching for and finding a non-torsion rational point. We also report on the conjectural extension to $|D| < 10^7$ of the list of curves E_D with odd $r_{\mathrm{an}}(E_D) > 1$, which raises several new questions.

1 Introduction

1.1 Review: The Curves E_D and Their Arithmetic

For nonzero rational D let E_D be the elliptic curve

$$E_D : Dy^2 = x^3 - x \tag{1}$$

over $\mathbf{Q}$. Since E_D and E_{c^2D} are isomorphic for any nonzero rational c, D, we may assume without loss of generality that D is a squarefree integer. The change of variable $x \leftrightarrow -x$ shows that E_D is also isomorphic with E_{-D}; this may also be seen from the Weierstrass equation $y^2 = x^3 - D^2x$ for E_D.

[1] The problem is: for which D does E_D have nontrivial rational points, or equivalently positive rank? Such D are called "congruent", because they are precisely the numbers that arise as the common difference ("congruum") of a three-term arithmetic progression of rational squares, namely the squares of $(x^2 - 2x - 1)/2y$, $(x^2 + 1)/2y$, and $(x^2 + 2x - 1)/2y$. See the Preface and Chapter XVI of [Di] for the early history of this problem, and [Kob] for a more modern treatment of the curves E_D.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 244–251, 2002.

The arithmetic of the curves E_D has long attracted interest, both for its connection with the classical "congruent number" problem (see [Di, Ch.XVI]; $|D|$ is a "congruent number" if and only if E_D has positive rank) and, more recently, as a paradigmatic example and test case for results and constructions concerning elliptic curves in general (see for instance [Kob]). The curves E_D have some special properties that make them more accessible than general elliptic curves over $\mathbf{Q}$. They have complex multiplication and are quadratic twists of the curve E_1. This led to the computation of the sign of the functional equation of the L-function $L(E_D/\mathbf{Q}, s)$: it depends on $|D|$ mod 8, and equals $+1$ or -1 according as $|D|$ is in $\{1, 2, 3\}$ or $\{5, 6, 7\}$ mod 8. We shall be concerned with the case of sign -1.

The conjecture of Birch and Swinnerton-Dyer (BSD) predicts that the (arithmetic) rank of any elliptic curve E over a number field K, defined as the $\mathbf{Z}$-rank of its Mordell-Weil group $E(K)$, should equal the order of vanishing at $s = 1$ of $L(E/K, s)$, known as the "analytic rank" $r_{\text{an}}(E/K)$. The BSD conjecture implies the "BSD parity conjecture": the arithmetic rank is even or odd according as the functional equation of $L(E/K, s)$ has sign $+1$ or -1. It would follow that if the sign is -1 then E always has positive rank. In our context, where $K = \mathbf{Q}$ and $E = E_D$, this leads to the conjecture that E_D has positive rank (and thus that $|D|$ is a "congruent number") if $|D|$ is any[2] integer of the form $8k+5$, $8k+6$, or $8k + 7$.

1.2 New Results and Computations

We prove:

Theorem 1. *Let D be an integer such that $|D|$ is congruent to* 5, 6, *or* 7 *mod* 8 *and also satisfies $|D| < 10^6$. Then E_D has positive rank over $\mathbf{Q}$.*

In our ANTS-1 paper [E1] we announced such a result for $|D| < 2 \cdot 10^5$. Our main tool for proving Theorem 1 is the same: we use the modular parametrization of E_1 or E_2 to construct a rational point P_D on each E_D from CM points on modular curves, and usually compute P_D to enough accuracy to distinguish it from any of the rational torsion points on E_D. Faster computer hardware and new software were both needed to extend the computation to 10^6. The faster machine made it feasible to compute P_D for more and larger D. Cremona's program MWRANK, not available when [E1] was written, found rational points on the curves E_D on which we could neither distinguish P_D from a torsion point nor find a rational nontorsion point by direct search. This happened for 1375 values of $|D|$ — less than 0.5% of the total, but too many to list here a rational point on E_D for each such D. These tables, and further computational data on the curves E_D, can be found on the Web starting from `<www.math.harvard.edu/~elkies/compnt.html>`.

Our computations also yield conjectural information on the rank of E_D: the rank should equal 1 if and only if P_D is nontorsion. In half the cases, those

[2] We have dropped the hypothesis that D be squarefree because $c^2D \equiv D$ mod 8 for any odd integer c. Our integers D are not divisible by 4, and therefore cannot be of the form c^2D for any even c.

for which $|D|$ or $|D|/2$ is of the form $8k+7$, we obtain this connection from Kolyvagin's theorem [Kol], which gives the "if" direction unconditionally, and the Gross-Zagier formula [GZ], which gives the "only if" direction under the BSD conjecture. Neither Kolyvagin nor Gross-Zagier has been proved to extend to the remaining cases, when $|D|$ or $|D|/2$ is of the form $8k+5$. But we expect that similar results do hold in these cases, and hence that E_D has rank 1 if and only if P_D is nontorsion also when $|D|$ or $|D|/2$ is congruent to 5 mod 8. One piece of evidence in this direction is that whenever we found P_D to be numerically indistinguishable from a torsion point, the Selmer groups for the 2-isogenies between E_D and the curve $Dy^2 = x^3 + 4x$ were large enough for E_D to have arithmetic rank at least 3. We extended the list of curves E_D of conjectural rank ≥ 3 to $|D| < 10^7$ by imposing the 2-descent condition from the start and computing P_D only for those D that pass this test. We find a total of 8740 values of $|D|$. The list not only provides new numerical data on the distribution of quadratic twists of rank > 1 with large $|D|$, but also suggests unexpected biases in the distribution that favor some congruence classes of $|D|$'s.

2 Proof of Theorem 1

Let D be a squarefree integer such that $|D|$ is congruent to 5, 6, or 7 mod 8. Set $K_D = \mathbf{Q}(\sqrt{-|D|}\,)$ if D is odd, and $K_D = \mathbf{Q}(\sqrt{-|D|/2}\,)$ if D is even. Then K_D is an imaginary quadratic field in which the rational prime 2 splits if $D = 8k+7$ or $D = 16k+14$, ramifies if $D = 8k+5$, and is inert if $D = 16k+6$. A point $P \in E_D(\mathbf{Q})$ is equivalent to a K_D-rational point Q of E_1 or E_2 (according as D is odd or even) whose complex conjugate $\overline{Q}$ equals $-Q$. If Q' is any point of E_1 or E_2 over K_D then $Q = Q' - \overline{Q'}$ satisfies $\overline{Q} = -Q$, and thus amounts to a point of E_D over $\mathbf{Q}$. To prove Theorem 1 for E_D, it will be enough to find $Q_D \in E_1(K_D)$ or $E_2(K_D)$ and show that the point $P_D \in E_D(\mathbf{Q})$ corresponding to $Q_D - \overline{Q_D}$ is not in $(E_D(\mathbf{Q}))_{\text{tors}} = E_D[2]$.

We use the modular parametrizations of E_1 and E_2 by the modular curves $\mathrm{X}_0(32)$ and $\mathrm{X}_0(64)$. These curves have "CM points" parametrizing cyclic isogenies of degree 32 or 64 between elliptic curves of complex multiplication by the same order in K_D. If the prime 2 splits in K_D, these points are defined over the class field of K_D; otherwise they are defined over a ray class field. (In the former case, the CM points are often called "Heegner points"; in the latter, [Mo] applies the term "mock Heegner points", though Birch points out that Heegner's seminal paper [He] already used both kinds of points to construct rational points on E_D, and the distinction between the two cases was a later development.) In either case, we obtain a point Q_D defined over K_D by taking a suitable subset of these CM points, mapping them to E_1 or E_2 by the modular parametrization, and adding their images using the group law of the curve. See [Bi1,Bi2,Mo] for more details on these subsets.

Now the key computational point is that the size of each subset is proportional to the class number of K_D, and thus to $|D|^{1/2}$ when averaged over D. This is much smaller than the number of terms of the series needed to numerically estimate $L'(E_D/\mathbf{Q}, 1)$, which is on the order of D: as explained for instance in [BGZ], for a general elliptic curve $E/\mathbf{Q}$ of conductor $N(E)$ it takes $N^{1/2+\epsilon}$ terms to adequately estimate $L'(E/\mathbf{Q}, 1)$, and $N(E_D) = 32D^2$ or $64D^2$ (according as D is odd or even) so $N^{1/2}$ is of order D. As explained in [E1], the numerical computation of each CM point as a point on the complex torus $E_1(\mathbf{C})$ or $E_2(\mathbf{C})$ to within say 10^{-25} takes essentially constant time: find a representative τ in a fundamental domain for the upper half-plane mod $\Gamma_0(32)$ or $\Gamma_0(64)$, and sum enough terms of a power series for $\int_\infty^\tau \varphi \, dq/q$ where φ is the modular form for E_1 or E_2. Thus it takes time $\Delta^{3/2+\epsilon}$ (and negligible space) to approximate Q_D for each $|D| < \Delta$.[3]

We implemented this computation in GP and ran it for $\Delta = 10^6$. For all but 1375 of the 303979 squarefree values of $|D| < 10^6$ congruent to 5, 6, or 7 mod 8, we found that P_D is at distance at least 10^{-8} from the nearest 2-torsion point of E_D, and is thus a rational point of infinite order.

For each of the remaining D, the point P_D is numerically indistinguishable (at distance[4] at most 10^{-20}, usually much less) from a 2-torsion point. We believe that P_D then actually is a torsion point, and thus that we must find a nontorsion rational point on E_D in some other way. We did this as follows. We first searched for rational numbers $x = r/s$ with $|r|, |s| < 5 \cdot 10^7$ such that $s^4 x = rs(r^2 - s^2)$ is D times a square for $|D| < 10^6$. This is a reasonable search since we may assume that $\gcd(r, s) = 1$, require that one of the factors $r, s, r+s, r-s$ of $rs(r^2-s^2)$ have squarefree part $f < (4 \cdot 10^6)^{1/4}$ and that another have squarefree part at most $(4 \cdot 10^6/f)^{1/3}$, and loop over those factors.[5] This took several hours and found points on all but 70 of our 1375 E_D's. The remaining curves were handled by Cremona's MWRANK program, which used a 2-descent on each curve (exploiting its full rational 2-torsion) to locate a rational point. This completed the proof of Theorem 1.

3 Curves E_D of Conjectural Rank ≥ 3

It might seem surprising that we were able to find a rational point on each of the 1375 E_D's for which we could not use P_D. Many curves E_D, even with D well below our upper limit of 10^6, have rank 1 but generator much too large to locate with repeated 2-descents (see for instance [E1]). The reason we could find nontorsion points on the curves E_D with $P_D \in E_D[2]$ is that these are precisely

[3] This computation is particularly efficient in our setting, in which φ is a CM form (so most of its coefficients vanish) and the normalizers of $\Gamma_0(32)$, $\Gamma_0(64)$ in $\mathrm{SL}_2(\mathbf{R})$ can be used to obtain an equivalent τ with imaginary part at least $1/8$ and $\sqrt{3}/16$ respectively. These efficiencies represent a considerable practical improvement, though they contribute negligible factors $O(\Delta^\epsilon)$ to the asymptotic running time of the computation.

[4] Here, as in the preceding paragraph, the distance is measured on the complex torus representing $E_1(\mathbf{C})$ or $E_2(\mathbf{C})$.

[5] In fact we removed the factors of 4 by using the squarefree parts of $(r \pm s)/2$ instead of $r \pm s$ when $r \equiv s \bmod 2$.

the curves E_D of odd sign that should have rank at least 3, which makes the minimal height of a non-torsion point much smaller than it can get in the rank-1 case. We explain these connections below, and then report on our computations that extend to 10^7 the list of $|D|$ such that $r_{\mathrm{an}}(E_D)$ is odd and conjecturally at least 3.

3.1 P_D and the Rank of E_D

Consider first the cases $D = 8k+7$ and $D = 16k+14$. In these cases the prime 2, which is the only prime factor of the conductors of E_1 and E_2, is split in K_D. Therefore the results of Gross-Zagier [GZ] and Kolyvagin [Kol] apply to P_D. The former result gives the canonical height of P_D as a positive multiple of $L'(E_D, 1)$. Therefore $r_{\mathrm{an}}(E_D) > 1$ if and only if P_D is torsion. The latter result shows that if P_D is nontorsion then in fact the arithmetic rank of E_D also equals 1. Hence any E_D of rank 3 or more must be among those for which we could not distinguish P_D from a torsion point.

The hypotheses of the theorems of Gross-Zagier and Kolyvagin are not satisfied in the remaining cases $D = 8k + 5$ and $D = 16k + 6$. However, numerical evidence suggests that both theorems generalize to these cases as well. For instance, when P_D is numerically indistinguishable from a torsion point, E_D seems to have rank 3. For small $|D|$ we readily find three independent points; for all $|D|$ in the range of our search, E_D and each of the curves $Dy^2 = x^3 + 4x$ and $Dy^2 = x^3 - 11x \pm 14$ isogenous with E_D has a 2-Selmer group large enough to accommodate three independent points. When P_D is nontorsion but has small enough height to be recovered from its real approximation by continued fractions, we find that it is divisible by 2 if and only if the 2-Selmer group has rank at least 5, indicating that E_D has either rank ≥ 3 or nontrivial Ш[2]. (The former possibility should not occur, and can often be excluded by 2-descent on one of the curves isogenous to E_D.) Both of these observations are consistent with a generalized Gross-Zagier formula and the conjecture of Birch and Swinnerton-Dyer, and would be most unlikely to hold if the vanishing of P_D had no relation with the arithmetic of E_D. We thus expect that also in these cases E_D should have rank > 1 if and only if P_D is a torsion point.

3.2 Rank and Minimal Nonzero Height

The conjecture of Birch and Swinnerton-Dyer also explains why curves E_D of rank ≥ 3 have nontorsion points of height much smaller than is typical of curves E_D of rank 1. This conjecture relates the regulator of the Mordell-Weil group of E_D with various invariants of the curve, including its real period and the leading coefficient $L^{(r)}(E_D, 1)/r!$ (where $r = r_{\mathrm{an}}(E_D)$). Now the real period is proportional to $|D|^{-1/2}$. The leading coefficient is $\ll |D|^{o(1)}$ under the generalized Riemann hypothesis for $L(E_d, s)$, or even the weaker assumption of the Lindelöf conjecture for this family of L-series (see for instance [IS, p.713]). One expects, and in practice finds, that it is also $\gg |D|^{-o(1)}$ (otherwise $L(E_d, s)$ has zeros $1 + it$ for very small positive t). Thus we expect the regulator to grow as

$|D|^{1/2+o(1)}$, at least if Ш is small, which should be true for most $|D|$. Hence the minimal nonzero height would be at most $|D|^{1/2r}$. When $r = 1$ this grows so fast that already for $|D| < 10^4$ there are many curves E_D with generators much too large to be found by 2-descents.[6] But for $r \geq 3$ the minimal nonzero height is at most $|D|^{1/6+o(1)}$, so $|D|$ must grow much larger before a 2-descent search becomes infeasible.

Remark on curves curves E_D of even sign: For such curves we readily determine whether $r_{\text{an}}(E_D) > 0$ by using the Waldspurger-Tunnell formula [Tu] to compute $L(E_D, 1)$. If $L(E_D, 1) \neq 0$ then $r_{\text{an}}(E_D) = 0$ and E_D also has arithmetic rank 0 by Kolyvagin (or even Coates-Wiles [CW] because E_D has CM). If $L(E_D, 1) = 0$ then $r_{\text{an}}(E_D) \geq 2$, and we can prove that E_D has positive arithmetic rank if we find a nontorsion point. We expect that the minimal height of such a point is $|D|^{1/4+o(1)}$. This grows slower than the $|D|^{1/2+o(1)}$ estimate for rank 1, but fast enough that 2-descent searches fail for $|D|$ much smaller than our bound of 10^6. Even in the odd-rank case that concerns us in this paper, it is the curves of rank 3 that make it hard to extend Theorem 1 much beyond $\Delta = 10^6$: searching for points on those curves take time roughly $\exp \Delta^{1/6}$, which eventually swamps the polynomial time $\Delta^{3/2+\epsilon}$ required to find those curves.

3.3 Computing E_D of Conjectural Rank ≥ 3 with $|D| < 10^7$

We extended to $\Delta = 10^7$ our search for P_D numerically indistinguishable from torsion points. These are the curves that we expect to have rank at least 3. Since we do not expect to extend Theorem 1 to 10^7, we saved time by requiring that the Selmer groups for the isogenies between E_D and $Dy^2 = x^3 + 4x$ be large enough to together accommodate an arithmetic rank of 3. For very large Δ this is a negligible saving because most D pass this test. But it saved a substantial factor in practice for $\Delta = 10^7$: the test eliminated all but 35% of choices of $|D| = 16k + 14$, all but 32.1% of $|D| = 16k + 6$, all but 21.6% of $|D| = 8k + 5$, and all but 16.2% of $|D| = 8k+7$. We found a total of 8740 values of D for which P_D appears to be a torsion point. We expect that each P_D is in fact torsion and that the corresponding E_D all have rank at least 3. Some P_D might conceivably be a nontorsion point very close to $E_D[2]$, but this seems quite unlikely; at any rate no P_D came closer than 10^{-8} but far enough to distinguish from $E_D[2]$. All the curves probably have rank exactly 3: the smallest $|D|$ known for a curve E_D of rank 5 exceeds $4 \cdot 10^9$ [Ro]. At any rate none of our curves with $|D| < 2 \cdot 10^6$ can have rank 5: we applied MWRANK's descents-only mode to each of these E_D and the isogenous curves, and in each case obtained an upper bound of 3 or 4 on the rank. Our curves E_D and the isogenous curves include many examples of conjectural rank 3 and nontrivial Ш[2].

[6] The generators can be obtained using the CM-point construction in time $|D|^{O(1)}$, but not $|D|^{1/2+o(1)}$ because P_D must be computed to high accuracy to recognize its coordinates as rational numbers from their real approximations. Note that in our computations we showed only that P_D is nontorsion and did not attempt to determine it explicitly in $E_D(\mathbf{Q})$.

There are striking disparities in the distribution of our 8740 values of $|D|$ among the allowed congruence classes. The odd classes $8k+5$ and $8k+7$ account for 2338 and 2392 curves E_D of presumed rank 3. But even $|D|$'s are much more plentiful: there are 4010 of them, almost as many as in the two odd classes combined. This might be explained by the behavior of the 2-descent, which depends on the factorization of $|D|$, or the fact that we are twisting a different curve: E_1 for odd D and E_2 for even D. But the 4010 even D's are themselves unequally distributed between the $16k+6$ and $16k+14$ cases, the former being significantly more numerous: 2225 as against 1785. (See Figure 1.) This disparity is much larger than would be predicted by the 2-descent test, which in the range $|D| < 10^7$ favors $16k+16$ but only by a factor of 1.09 whereas 2225 exceeds 1785 by almost 25%. Note too that the 2-descent survival rates would predict a preponderance of $|D| = 8k+7$ over $8k+5$, whereas the two counts are almost identical. Do these disparities persist as $\varDelta$ increases, and if so why? Naturally we would also like to understand the overall distribution of quadratic twists of rank ≥ 3, not only for the "congruent number" family but for an arbitrary initial curve in place of E_D. We hope that the computational data reported here, and

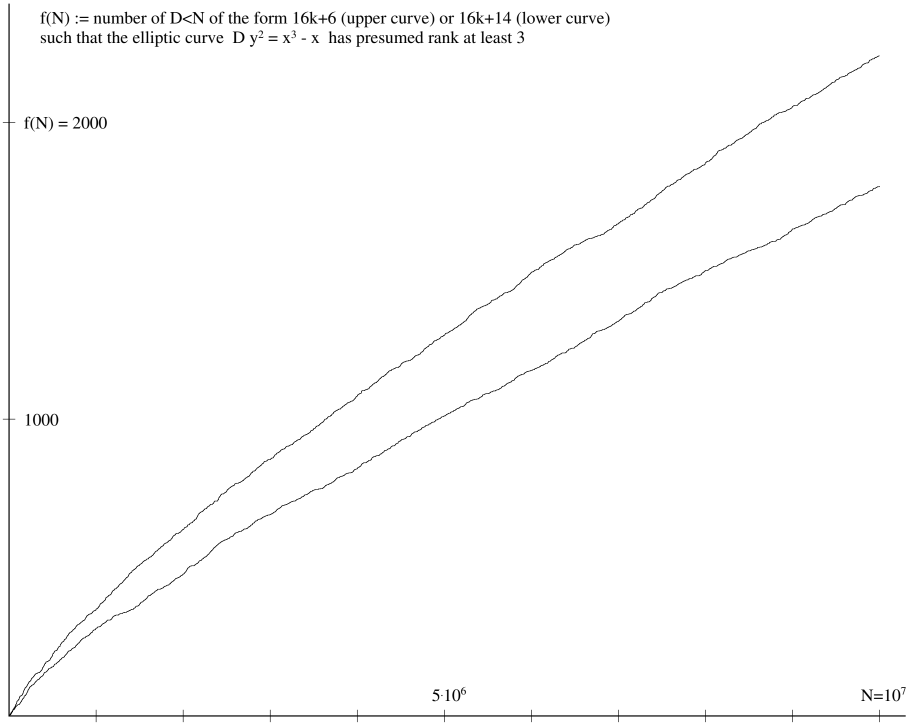


Fig. 1. Twists with $|D| \equiv 6 \bmod 16$ seem to have rank 3 much more often than those with $|D| \equiv 14 \bmod 16$

more fully at `<www.math.harvard.edu/~elkies/compnt.html>`, might suggest reasonable ideas and conjectures in this direction.

Acknowledgments

Thanks to Peter Sarnak for the reference [IS], and to the referee for several suggestions that improved the paper. I am grateful to the Packard Foundation for partial financial support during the preparation of this work.

References

Bi1. Birch, B.J.: Elliptic curves and modular functions. *Symp. Math.* **4** (1970), 27–37.

Bi2. Birch, B.J.: Heegner points of elliptic curves. *Symp. Math.* **15** (1975), 441–445.

BGZ. Buhler, J.P., Gross, B.H., Zagier, D.: On the conjecture of Birch and Swinnerton-Dyer for an elliptic curve of rank 3. *Math. of Computation* **44** (1985) #175, 473–481.

CW. Coates, J., Wiles, A.: On the conjecture of Birch and Swinnerton-Dyer. *Invent. Math.* **39** (1977) #3, 223–251.

Di. Dickson, L.E.: *History of the Theory of Numbers, Vol. II: Diophantine Analysis.* New York: Stechert 1934.

E1. Elkies, N.D.: Heegner point computations, *Lecture Notes in Computer Science* **877** (proceedings of ANTS-1, 5/94), 122–133.

GZ. Gross, B.H., Zagier, D.: Heegner points and derivatives of L-series. *Invent. Math.* **84** (1986), 225–320.

He. Heegner, K.: Diophantische Analysis und Modulfunktionen. *Math. Z.* **56** (1952), 227–253.

IS. Iwaniec, H., Sarnak, P.: Perspectives on the analytic theory of L-functions. *Geom. Funct. Anal.* **2000**, Special Volume (GAFA 2000, Tel Aviv 1999), Part II, 705–741.

Kob. Koblitz, N.: *Introduction to elliptic curves and modular forms.* New York: Springer 1984.

Kol. Kolyvagin, V.A.: Euler systems. Pages 435–483 of *The Grothendieck Festscrhift* Vol. II, Birkhäuser: Boston 1990.

Mo. Monsky, P.: Mock Heegner points and congruent numbers. *Math. Z.* **204** (1990) #1, 45–67.

Ro. Rogers, N.F.: Rank Computations for the Congruent Number Elliptic Curves. *Experimental Mathematics* **9** (200) #4, 591–594.

Si. Silverman, J.H.: *The Arithmetic of Elliptic Curves.* New York: Springer 1986.

Tu. Tunnell, J.B.: A classical Diophantine problem and modular forms of weight 3/2, *Invent. Math.* **72** (1983) #2, 323–334.

Comparing Invariants for Class Fields of Imaginary Quadratic Fields

Andreas Enge and François Morain

Laboratoire d'Informatique (CNRS/UMR 7650), École polytechnique,
91128 Palaiseau Cedex, France
{enge,morain}@lix.polytechnique.fr

Abstract. Class fields of imaginary quadratic number fields can be constructed from singular values of modular functions, called class invariants. From a computational point of view, it is desirable that the associated minimal polynomials be small. We examine different approaches to measure the size of the polynomials. Based on experimental evidence, we compare two families of class invariants suggested in the literature with respect to these criteria. Our results lead to more efficient constructions of elliptic curves for cryptography or in the context of elliptic curve primality proving (ECPP).

1 Introduction

Let K be an imaginary quadratic field of discriminant $-D < 0$ and class number h. The Hilbert class field of K, denoted by K_{H}, is the maximal unramified abelian extension of K. It is known that a minimal polynomial for K_{H}/K has degree h and can be computed using the values of the j-invariant at integers of K. This polynomial, denoted by $H_D[j](X)$, has huge coefficients, and there is a wide variety of functions that can be used in place of j and that lead to polynomials with smaller coefficients. Small polynomials are preferable for several reasons: first of all, the usual way to build them is to use floating point numbers, and the required precision clearly depends on the size of the result. Second, one may want to store the polynomials and although disks are not that expensive today, the smaller the better. The last reason is more recent and is related to the Galois approach described in [15]. For it to succeed, it is necessary to start with coefficients as small as possible.

In this article, we describe the use of two families of η-products: the $\mathfrak{w}_\ell$ family of [23,8] and the $\mathfrak{w}_{p_1,p_2}$ family of [26,9]. We show how to choose a function adapted to one's needs. If one is interested in computing only the minimal polynomial $H_D[u]$ of the invariant u, a minimal logarithmic height seems to be the good notion. If one is interested in precomputing a system of polynomial equations solving the equation $H_D[u](X) = 0$ as described in [15,14], then the size of the largest root of $H_D[u](X)$ should be taken into account.

Our ultimate goal is to build curves with prescribed complex multiplication for ECPP [2]. Although most of our results hold in full generality, we assume for the sake of simplicity that $-D$ is fundamental.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 252–266, 2002.

For our purpose, we do not have to give many details on the theory of elliptic curves. We refer the reader to Silverman's two books [28,29].

2 Complex Multiplication

Let K be an imaginary quadratic field of discriminant $-D$ and class number $h = h(-D)$. The ring of integers of K is $\mathcal{O}_K = \mathbb{Z}[\omega]$ with

$$\omega = \begin{cases} \sqrt{-D/4} & \text{if } D \equiv 0 \bmod 4, \\ (1+\sqrt{-D})/2 & \text{if } D \equiv 3 \bmod 4. \end{cases}$$

We denote by K_{H} the Hilbert class field of K. It can be built using the singular values of the modular invariant j, i.e. the values at certain integers of K. The minimal polynomial of such a singular value is computed from the ideal class group $Cl(\mathcal{O}_K)$, or equivalently from the set $Cl(-D)$ of primitive binary reduced quadratic forms of discriminant $-D$. If $Q = [A, B, C] = AX^2 + BX + C$ is such a form, we put $\tau_Q = \frac{-B+\sqrt{-D}}{2A}$, and the minimal polynomial of $j(\tau_Q)$ is simply

$$H_D[j](X) = \prod_{Q \in Cl(-D)} (X - j(\tau_Q)).$$

The coefficients of $H_D[j]$ are quite large in general, so that it is of interest to consider alternative generators of the class field. These are provided by singular values of other functions, so-called *class invariants*. Most of the known class invariants can be obtained from Dedekind's η function, defined in [7] for a complex variable z by

$$\eta(z) = q^{1/24} \prod_{n \geq 1} (1 - q^n) = q^{1/24} \left(1 + \sum_{n=1}^{\infty} (-1)^n \left(q^{n(3n-1)/2} + q^{n(3n+1)/2} \right) \right)$$

with $q = \exp(2\pi i z)$ and $q^{1/24} = \exp(2\pi i z/24)$.

The functions $\mathfrak{f}(z) = \exp(-\pi i/24)\frac{\eta((z+1)/2)}{\eta(z)}$, $\mathfrak{f}_1(z) = \frac{\eta(z/2)}{\eta(z)}$, $\mathfrak{f}_2(z) = \sqrt{2}\,\frac{\eta(2z)}{\eta(z)}$, $\gamma_2 = \frac{\mathfrak{f}^{24}-16}{\mathfrak{f}^8}$ and $\gamma_3 = \frac{(\mathfrak{f}^{24}+8)(\mathfrak{f}_1^8-\mathfrak{f}_2^8)}{\mathfrak{f}^8}$ are known as Weber's functions, although some of them are already discussed in [17,7]. Weber observed that for many discriminants, powers of these functions are class invariants. For proofs, see [31,3,20,24,25]. More results concerning the values of Weber's functions can be found in [2,32,12,13].

To use functions other than j and Weber's functions for building class fields, we have to solve two problems. First, we need criteria when a given function is a class invariant for a given discriminant. Second, we need a way of computing associated class polynomials, that is of determining a complete set of conjugates under the Galois action. Both questions can be addressed using Shimura's reciprocity law [27]. Singular values of modular functions of level N are contained in the ray class field modulo N. Membership in the Hilbert class field then follows from invariance under the Galois group, formed of Frobenius maps whose actions on the singular values can be derived from Shimura reciprocity. One explicit approach to the reciprocity law is described by Gee and Stevenhagen in [13,12]. In [25], Schertz obtains a very general criterion for singular values to belong to the Hilbert class field. His theorem applies to functions on the modular curve $X_0(N)$ whose q-expansions satisfy certain rationality conditions.

Moreover, Schertz provides an elegant way of obtaining the class polynomial by evaluating the fixed class invariant at a suitably chosen system of quadratic forms, namely an *N-system*. Such a system is defined as a system of representatives $Q_i = [A_i, B_i, C_i]$ for the class group such that the A_i are coprime to N and all the B_i are congruent modulo $2N$. Proposition 2 of [25] shows that N-systems exist for any natural number N and for any discriminant.

3 Simple η Quotients

Let ℓ denote a positive integer and define a 24-th root of Klein's function $\varphi_{\left(\begin{smallmatrix}1 & 0\\ 0 & \ell\end{smallmatrix}\right)}$ (cf. [18, Abschnitt II, §16]) by

$$\mathfrak{w}_\ell(z) = \frac{\eta(z/\ell)}{\eta(z)}.$$

Note that this function generalises Weber's function $\mathfrak{f}_1 = \mathfrak{w}_2$. Although we may use composite numbers ℓ as described in [8], we henceforth focus on prime values. Put $s = 12/\gcd(12, \ell-1)$. Using the q-expansion and the transformation formula of η it is not difficult to obtain the following result, see [8].

Proposition 1. *The function $\mathfrak{w}_\ell^{2s}(z)$ is invariant under $\Gamma^0(\ell)$ and has a q-expansion starting with $q^{-s(\ell-1)/(12\ell)}$.*

We denote the modular equation associated with a modular function u for $\Gamma^0(\ell)$ by $\Phi_\ell[u](U, J)$. The following result is obtained in [8].

Theorem 1. *The degree of $\Phi_\ell[\mathfrak{w}_\ell^{2s}](U, J)$ in J is $\frac{s(\ell-1)}{12}$ and the leading coefficient with respect to J is $-U$. The constant term with respect to U is given by ℓ^s.*

In Table 1, we provide the modular equations of prime level at most 13. They are computed using the methods described for instance in [21,10].

Fricke observed in [11] that when ℓ is split or ramified in K, the singular values of $\mathfrak{w}_\ell^{2s}$ at suitably normalised integers lie in the Hilbert class field K_{H}. When ℓ is a square and coprime to 6, then this already holds for $\mathfrak{w}_\ell$. In [13], Gee and Stevenhagen consider $\mathfrak{w}_3^2$ and work out an example for a particular discriminant; in [12], Gee obtains a general result for resolvents formed with the conjugates of $\mathfrak{w}_5$. By applying the theory of [25] to $\mathfrak{w}_\ell^{2s}$ and lower powers of $\mathfrak{w}_\ell$, the authors prove the following theorem, see [8].

Theorem 2. *Let ℓ be an odd prime and $-D$ a quadratic discriminant such that $\left(\frac{-D}{\ell}\right) \neq -1$. Choose the power $\mathfrak{w}_\ell^e$ and the natural number N, a multiple of ℓ,*

Table 1. Some modular equations of prime degree.

ℓ	$\Phi_\ell[\mathfrak{w}_\ell^{2s}](U,J)$
2	$(U+16)^3 - JU$
3	$(U+27)\,(U+3)^3 - JU$
5	$\left(U^2+10\,U+5\right)^3 - JU$
7	$\left(U^2+13\,U+49\right)\left(U^2+5\,U+1\right)^3 - JU$
13	$\left(U^2+5\,U+13\right)\left(U^4+7\,U^3+20\,U^2+19\,U+1\right)^3 - JU$
11	$U J^5 + J^4(132\,U^3 + 468754\,U^2 + 3732\,U)$ $+J^3\left(-5346\,U^5 + 161201040\,U^4 - 49836805205\,U^3 + 51801406800\,U^2 - 4586706\,U\right)$ $+J^2(67496\,U^7 + 2291468355\,U^6 + 4231762569540\,U^5 + 755793774757450\,U^4$ $+6941543075967060\,U^3 + 214437541826475\,U^2 + 2059075976\,U)$ $+J(-139755\,U^9 + 723797800\,U^8 - 1327909897380\,U^7 + 1036871615940600\,U^6$ $-310557763459301490\,U^5 + 1730954664564250620 0\,U^4$ $-6481517942976139866 0\,U^3 + 77380735840203400\,U^2 - 253478654715\,U)$ $+(U^{12} - 5940\,U^{11} + 14701434\,U^{10} - 19264518900\,U^9 + 13849401061815\,U^8$ $-4875351166521000\,U^7 + 400050977713074380\,U^6 + 12247115445643361580 0\,U^5$ $+651339173406982403161 5\,U^4 + 104264884483130180036700\,U^3$ $+804140494949359194\,U^2 + 2067305393340\,U + 1771561)$

depending on D mod 6 *as specified in the table below. Assume that $Q = [A,B,C]$ is a primitive quadratic form of discriminant $-D$ with* $\gcd(A,N) = 1$, *$B^2 \equiv -D$* (mod 4ℓ) *and B satisfying the additional congruences modulo 3 or 4 as given in the table. If $\tau_Q = \frac{-B+\sqrt{-D}}{2A}$ is the root of Q in the upper complex half plane, then $\mathfrak{w}_\ell^e(\tau_Q) \in K_{\mathrm{H}}$ and its minimal polynomial has coefficients in $\mathcal{O}_K$ and can be computed from an N-system.*

ℓ mod 12	D	invariant	N	B
$\ell = 3$	—	$\mathfrak{w}_3^{12}$	3	—
	$2 \nmid D$	$\mathfrak{w}_3^6$	6	$B \equiv 1 \pmod 4$
1	—	$\mathfrak{w}_\ell^2$	ℓ	—
5	—	$\mathfrak{w}_\ell^6$	ℓ	—
	$3 \nmid D$	$\mathfrak{w}_\ell^2$	3ℓ	$3 \mid B$
7	—	$\mathfrak{w}_\ell^4$	ℓ	—
	$2 \nmid D$	$\mathfrak{w}_\ell^2$	2ℓ	$B \equiv 1 \pmod 4$
11	—	$\mathfrak{w}_\ell^{12}$	ℓ	—
	$2 \nmid D$	$\mathfrak{w}_\ell^6$	2ℓ	$B \equiv 1 \pmod 4$
	$3 \nmid D$	$\mathfrak{w}_\ell^4$	3ℓ	$3 \mid B$
	$\gcd(D,6) = 1$	$\mathfrak{w}_\ell^2$	6ℓ	$B \equiv 9 \pmod{12}$

If furthermore $\ell \mid D$ and the table does not specify any restriction for B mod 4, *then $H_D[\mathfrak{w}_\ell^e] \in \mathbb{Z}[X]$.*

Conjecture 1. The assertions of Theorem 2 also hold in the following cases:

$\ell = 3$	$3 \mid D, D/3 \equiv 4 \pmod{12}$	$\mathfrak{w}_3^4$	$N = 9$	$9 \mid B$
	$3 \mid D, D/3 \equiv 1 \pmod{12}$	$\mathfrak{w}_3^2$	$N = 18$	$3 \mid B$

The minimal polynomial actually depends on the particular choice of the N-system, that is of the initial value of B. If $\left(\frac{-D}{\ell}\right) = 1$, we have two choices for B as a root of $-D$ modulo ℓ, and the associated polynomials are complex conjugates. As a normalisation, we assume in the following that B has been selected as the smallest possible positive value, and define the class polynomial $H_D[\mathfrak{w}_\ell^e]$ as the corresponding minimal polynomial.

4 Double η Quotients

Following [26,9], we let p_1 and p_2 denote two (not necessarily distinct) primes such that $24 \mid (p_1 - 1)(p_2 - 1)$ and define $N = p_1 p_2$ and

$$\mathfrak{w}_{p_1,p_2}(z) = \frac{\eta(z/p_1)\eta(z/p_2)}{\eta(z/(p_1p_2))\eta(z)}.$$

These functions and their associated modular equations are examined in [10], from which we cite the following results.

Proposition 2. *The function $\mathfrak{w}_{p_1,p_2}$ is invariant under $\Gamma^0(N)$ and has a q-expansion starting with $q^{-(p_1-1)(p_2-1)/(24p_1p_2)}$.*

For the proposition to hold, the divisibility of $(p_1 - 1)(p_2 - 1)$ by 24 is crucial. If it is not given, one may use a higher power of the function, namely $\mathfrak{w}_{p_1,p_2}^s$ with $s = 24/\gcd(24, (p_1 - 1)(p_2 - 1))$. However, these invariants lead to larger polynomials (cf. Section 6), whence they are less attractive.

Theorem 3. *The modular polynomial $\Phi_N[\mathfrak{w}_{p_1,p_2}](U, J)$ is a polynomial of degree $\psi(N)$ in U with coefficients in $\mathbb{Z}[J]$. Seen as a polynomial in J, its degree is $\frac{(p_1-1)(p_2-1)}{12}$. If $p_1 \neq p_2$, then $\psi(N) = (p_1 + 1)(p_2 + 1)$, the constant coefficient with respect to U is 1 and the leading coefficient with respect to J is $U^{p_1+p_2}$. If $p_1 = p_2 = p$, then $\psi(N) = p(p + 1)$, the constant coefficient with respect to U is $p^{(p-1)/2}$ and the leading coefficient with respect to J is U^{p-1}.*

Precise conditions under which the values of $\mathfrak{w}_{p_1,p_2}$ provide elements in the Hilbert class field and an algorithm for computing their class polynomials are given in [9], from which the following result is taken.

Theorem 4. *Suppose that p_1 and p_2 are odd primes which split or ramify in $\mathbb{Q}(\sqrt{-D})$; if both ramify, assume furthermore $p_1 \neq p_2$. Let $N = p_1p_2$, $Q = [A, B, C]$ a primitive quadratic form of discriminant $-D$ such that $\gcd(A, N) = 1$ and $N|C$, and $\tau_Q = \frac{-B+\sqrt{-D}}{2A}$ the root of Q in the upper complex half plane. Then $\mathfrak{w}_{p_1,p_2}(\tau_Q) \in K_{\mathrm{H}}$, its minimal polynomial has rational integral coefficients and can be computed from an N-system as defined in Section 2.*

There are up to four possible values for $B \pmod N$; since B and $-B$ yield complex conjugate polynomials and these are in fact real, only up to two distinct polynomials may be obtained. We again assume that the class polynomial $H_D[\mathfrak{w}_{p_1,p_2}]$ is defined from the smallest positive initial value of B.

5 Computing Class Polynomials

Computing the class polynomial for $\mathfrak{w}_\ell^e$ or $\mathfrak{w}_{p_1,p_2}$ amounts to h evaluations of the function at quadratic integers associated to an N-system (see Theorems 2 and 4). In principle, this requires to evaluate the η function $2h$ times for $\mathfrak{w}_\ell^e$ resp. $3h$ times for $\mathfrak{w}_{p,p}$ or $4h$ times for $\mathfrak{w}_{p_1,p_2}$ with $p_1 \neq p_2$. When the class polynomial is real, [9,8] characterise the pairs of quadratic forms leading to complex conjugate values, which almost halves the number of evaluations of the function.

The following observations show that in any case it is sufficient to precompute the values of η only at the h reduced quadratic forms. Let h_1 and h_2 denote the number of ambiguous resp. non-ambiguous reduced forms. Since inverse forms lead to complex conjugate values of η, the precise number of evaluations of η is then reduced to $h_1 + h_2 \approx h/2$.

Consider first the case of $\mathfrak{w}_{p_1,p_2}$ in which $N = p_1 p_2$, and let $Q = [A, B, C]$ with root τ_Q be an element of an N-system chosen according to Theorem 4. Clearly, $\eta(\tau_Q)$ can be obtained by transforming τ_Q into the standard fundamental domain for $\Gamma = \mathrm{Sl}_2(\mathbb{Z})$, which amounts to reducing Q, and looking up the corresponding precomputed value. Notice now that by the definition of an N-system and the special choice of the initial form in Theorem 4, C is divisible by N. Hence, τ_Q/p_1 is a root of the quadratic form $[p_1 A, B, C/p_1]$, which is equally of discriminant $-D$. Furthermore, it is primitive. If it were not, then $\gcd(B, C/p_1) = p_1$ since Q was primitive. But this implies that p_1^2 divides $-D = B^2 - 4AC$, contradicting the fact that $-D$ is fundamental. Thus, $\eta(\tau/p_1)$ can be obtained by reducing $[p_1 A, B, C/p_1]$ and looking up the precomputed value, and the same argumentation holds for $\eta(\tau_Q/p_1)$ and $\eta(\tau_Q/p_1 p_2)$. This reasoning remains valid for $\mathfrak{w}_\ell^e$ with ℓ odd.

We are left with the problem of evaluating $\eta\left(\frac{-B+\sqrt{-D}}{2A}\right)$ for a large number of primitive reduced forms $Q = [A, B, C]$. This amounts to evaluating

$$q_Q = \exp\left(2i\pi \frac{-B + \sqrt{-D}}{2A}\right) = \rho_Q \zeta_Q$$

with $\rho_Q = \exp(-\pi\sqrt{D}/A)$, $\vartheta_Q = -\pi B/A$ and $\zeta_Q = \exp(i\vartheta_Q)$, and to computing the series expansion of η. The root of unity ζ_Q can be obtained via first computing $\cos(-\pi B/A)$ by the arithmetic-geometric mean method [4] and deducing $\sin(-\pi B/A)$ as the negative square root of $1 - \cos^2(-\pi B/A)$ (remember that in general $0 \leq B \leq A$, the case $B < 0$ corresponding to the complex conjugate of $B > 0$).

Evaluating the series expansion of η by Algorithm 6.3.2 of [6] requires five multiplications for two additional terms. Since the series converges quadratically, this part of the algorithm is quite fast, and most of the time is spent with the computation of q_Q. To speed it up, one could imagine to compute first $\exp(-\pi\sqrt{D}/\mathrm{lcm}(A))$ and $\exp(-i\pi/\mathrm{lcm}(A))$ and to recover the other values as some integral powers. However, $\mathrm{lcm}(A)$ is quite large in general, and this approach does not sound very promising. Similarly, setting up a table of the

$\exp(-i\pi/A)$, which are indeed independent of D, and computing ζ_Q by raising this value to the power B does probably not pay off since B can become large.

Notice, however, that the same A may belong to several reduced forms, which then share the same ρ. Also, different forms may have the same ratio B/A and share the same ζ. As an example, the discriminant $-D = -123456799$ has $h = 4790$, $h_1 = 2$ and $h_2 = 2394$, so that 2396 values of η have to be precomputed. But there are only 1281 distinct values of ρ and 1225 distinct values of ζ, so that reusing them saves about half of the time during the computation of the q. One could push this approach even further by looking for families of A resp. B/A that divide each other, computing only one value of ρ resp. ζ and obtaining the others by raising to some power.

6 The Best Choice

For a given $-D$, there are potentially infinitely many invariants that can be used. Which one to choose, then? The first idea is to insist on having polynomials $H_D[u]$ with small coefficients, that is of small height. This is sensible if we want to build K_{H}. For ECPP, we need to solve $H_D[u](X) = 0$ in some large finite field, and we can speed up the algorithm using the Galois decomposition of $Cl(-D)$ if h is composite (see [15,14]). For this approach to be efficient, small roots are preferred.

6.1 Heights

We recall several definitions and facts on heights; for details, see [16]. Let $L/\mathbb{Q}$ be a number field of degree n, and $a = [a_0, \ldots, a_m] \in \mathbb{P}^m(L)$. Then the *logarithmic height* of a is defined as

$$\mathcal{H}(a) = \frac{1}{n} \sum_v \log \max_i |a_i|_v,$$

where $|\cdot|_v$ varies over the absolute values of L, suitably normalised to take inertia and ramification into account. It turns out that $\mathcal{H}(a)$ is in fact invariant under field extensions of L. If $f = \sum a_i X^i \in L[X]$, we define $\mathcal{H}(f) = \mathcal{H}(a)$. For an algebraic number $\alpha \in L$, let $\mathcal{H}(\alpha) = \mathcal{H}(X - \alpha) = \frac{1}{n} \sum_v \log \max(1, |\alpha|_v)$.

When α and the a_i are algebraic integers and f is monic, then all non-archimedian valuations are at most 1 and need not be taken into account. In particular, for α and a_i elements of $\mathbb{Z}$ or $\mathcal{O}_K$, we have $\mathcal{H}(f) = \max |a_i|$ and $\mathcal{H}(\alpha) = |\alpha|$ for the usual absolute value on $\mathbb{C}$.

A slightly different notion of height appears naturally in our context. To correctly round a quadratic integer $a + b\omega$ with a, $b \in \mathbb{Z}$, which is known as a floating point number, we can separate the real and the imaginary part and thus only need sufficient precision to recognise a and b as rational integers. Hence we define the modified height of a polynomial $f = \sum (a_i + b_i\omega) X^i$ as $\mathcal{H}'(f) = \log(\max\{|a_i|, |b_i|\})$. For real polynomials, this notion coincides with the usual height. In our context, $\log|\omega| \in O(\log D)$ is small compared to the maximal value of the $\log|a_i|$ and $\log|b_i|$, which turns out to be rather of the order of $\sqrt{D}$, and the heights differ only marginally. During our experiments, we observed differences only in the third significant digit.

Finally, the Mahler measure of a polynomial $f = \prod_{i=1}^{n}(X - \alpha_i)$ is given by $\mathcal{M}(f) = \sum_{i=1}^{n} \log\max(1, |\alpha_i|)$. Notice that if the α_i are algebraic integers with minimal polynomial f, then $\mathcal{M}(f) = n\mathcal{H}(\alpha_i)$ for any i.

To simplify the notation, we write $\mathcal{H}_D[u]$, $\mathcal{H}'_D[u]$ resp. $\mathcal{M}[u]$ for the corresponding heights of the class polynomial $H_D[u]$.

6.2 Inspecting the Values of j

Let us begin with a review of the properties of j at the root τ_Q of some reduced quadratic form $Q = [A, B, C]$ of discriminant $-D$. Since Q is reduced, we have $A \leq \sqrt{D/3}$ and therefore $|q_Q| \leq \exp(-\pi\sqrt{3}) < 4.34 \cdot 10^{-3}$.

The q-expansion of j, $j(\tau_Q) = q^{-1} + 744 + \sum_{n \geq 1} c_n\, q^n$, is known to satisfy $c_n \sim \frac{e^{4\pi\sqrt{n}}}{\sqrt{2}\, n^{3/4}}$. From [5], the following precise upper bound holds for $n \geq 1$:

$$c_n \leq \frac{1}{\sqrt{2}n^{3/4}} \exp(4\pi\sqrt{n}).$$

It follows that asymptotically $j/q^{-1} \to 1$ ($q \to 0$).

Now consider the different values of A, given in increasing order by $A_1 = 1 < A_2 \leq \cdots$. Hereby, A_1 corresponds to the principal form $Q_1 = [1, 0, D/4]$ or $Q_1 = [1, 1, (D+1)/4]$. Approximating $j(\tau_Q)$ as $1/q_Q$ yields that the largest conjugate is $j(\tau_{Q_1})$. The absolute value of the second largest conjugate is $|j(\tau_{Q_2})| \lessapprox \sqrt{|j(\tau_{Q_1})|}$ since $A_2 \geq 2$, and thus it is much smaller than the largest one. This argumentation can be continued with the next largest values of A.

For most discriminants, there are no small conjugates, and the largest coefficient of $H_D[j] \in \mathbb{Z}[X]$ is the product of all roots. Thus, the height $\mathcal{H}_D[j] = \mathcal{H}'_D[j]$ can be approximated by

$$\widehat{\mathcal{H}}_D := \pi\sqrt{D} \sum_{[A,B,C] \in Cl(-D)} \frac{1}{A},$$

where the sum is taken over a reduced set of representatives for the class group. For instance, one finds $\widehat{\mathcal{H}}_{39} = 45.77822626\ldots$, whereas $\mathcal{H}_{39}[j] \approx 44.48719450$, see Table 3.

Among the class polynomials for all 17702 known discriminants with $2 \leq h \leq 64$, we find 380 ones for which the largest coefficient is that of X, 202 ones where it is in front of X^2 and one discriminant where it occurs for X^3. In most cases, this is due to the fact that there are one, two resp. three conjugates of absolute value less than 1. Omitting these from the product yields a larger term in the elementary symmetric function forming the coefficient in front of X, X^2 resp. X^3. (Among the discriminants with largest coefficient in front of X, 34 had two conjugates and 56 had no conjugate of absolute value less than 1. Notice that when conjugates of absolute value close to 1 occur, the binomial coefficients start to play a role.)

Small conjugates correspond to large values of A close to $\sqrt{D/3}$, since then B and C are approximately $\sqrt{D/3}$ as well, and τ_Q approaches the zero $\frac{-1+\sqrt{-3}}{2}$ of j. For instance, the form $[77, 76, 77]$ of discriminant -17940 yields a j-value of 0.019... Omitting these small conjugates from the product amounts to omitting the corresponding terms $1/A$ in the approximation of the height, which has hardly any influence in practice since only a few very small values are left out. Therefore, the approximation of $\mathcal{H}_D[j]$ given above remains accurate.

6.3 The Largest Roots of Alternative Invariants

The estimation of the largest root is rather straightforward. As in the case of j, in general the value $u(\tau_Q)$ for the invariants u we examine becomes maximal when $Q = Q_1$. Let v be such that $u = q^{-v} + \ldots$, that is, $v = e(\ell - 1)/(24\ell)$ for $u = \mathfrak{w}_\ell^e$ and $v = (p_1 - 1)(p_2 - 1)/(24 p_1 p_2)$ for $u = \mathfrak{w}_{p_1,p_2}$. Then the largest root is closely approximated by $|u_{\max}(D)| \approx |q_{Q_1}|^{-v} = \exp(v\pi\sqrt{D})$. Thus, the invariant with minimal v yields the minimal largest root.

To check the validity of the approximation and to make sure that it holds independently of the class number, we chose two distinct values for the class number and computed the class polynomials of several invariants, namely $\mathfrak{w}_5^2$, $\mathfrak{w}_5^6$, $\mathfrak{w}_7^2$, $\mathfrak{w}_7^4$, $\mathfrak{w}_{11}^2$, $\mathfrak{w}_{11}^4$, $\mathfrak{w}_{11}^6$, $\mathfrak{w}_{11}^{12}$, $\mathfrak{w}_{13}^2$, $\mathfrak{w}_{19}^2$, $\mathfrak{w}_{19}^4$, $\mathfrak{w}_{23}^2$, $\mathfrak{w}_{23}^4$, $\mathfrak{w}_{23}^6$, $\mathfrak{w}_{23}^{12}$, $\mathfrak{w}_{5,7}$, $\mathfrak{w}_{11,13}$ and $\mathfrak{w}_{13,13}$ for (presumably all) discriminants with these class numbers. We examined the 289 discriminants of class number 99 and the 3722 discriminants of class number 128. For each invariant u and each class number we computed the average over all suitable discriminants of the value $\log(|u_{\max}(D)|)/(v\sqrt{D})$. If our approximation were an equality, we would obtain π, and indeed the average values varied between 3.141574 and 3.142383.

6.4 Heights of Alternative Invariants

By analogy with j, one might expect to find a proportional relationship between $\mathcal{H}'_D[u]$ resp. $\mathcal{H}_D[u]$ and $\widehat{\mathcal{H}}_D = \pi\sqrt{D}\sum_{[A,B,C]\in Cl(-D)} \frac{1}{A}$, where the sum is again taken over a reduced set of representatives of the class group. We thus plotted the heights obtained for a given invariant and a given class number. Figure 1 shows the result for $h = 99$ and three invariants. There is indeed a strong linear correlation. For the $\mathfrak{w}_\ell^e$, it looks like a proportional relationship, while for the $\mathfrak{w}_{p_1,p_2}$ there seems to be an additive constant.

So we assume that $\mathcal{H}'_D[u]$ can be approximated by a linear model of the form

$$\mathcal{H}'_D[u] \approx c\widehat{\mathcal{H}}_D + d = c\pi\sqrt{D} \sum_{[A,B,C]\in Cl(-D)} \frac{1}{A} + d$$

with suitably chosen constants c and d that a priori may depend on the invariant and possibly the class number, but not on the discriminant. A linear regression

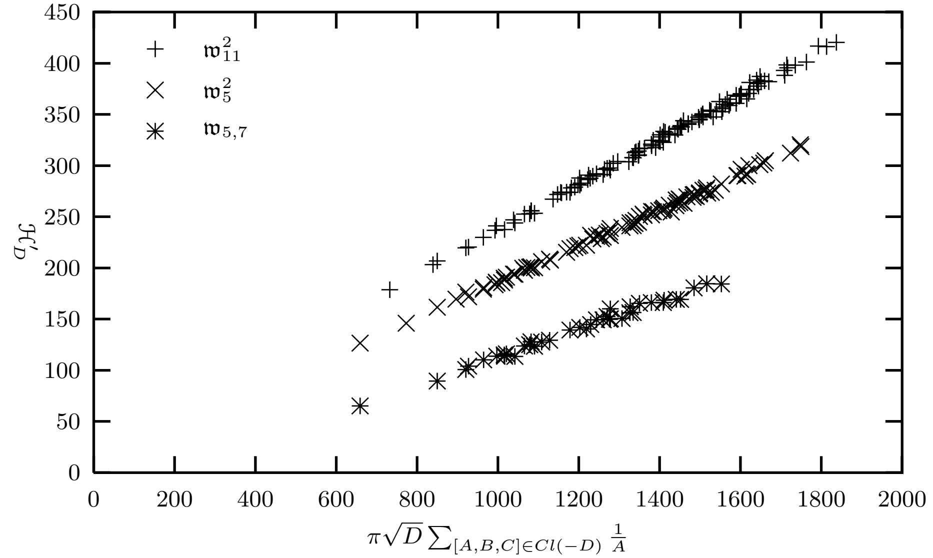


Fig. 1. Heights

for some possible choices of u and h yields the values given in Table 2; the quality of the approximation, measured by the regression coefficient, is at least 0.9943 for all examples.

It turns out that c depends only on u and that it is very close to the quantities $\hat{c}(\mathfrak{w}_\ell^e) = \frac{e(\ell-1)}{24(\ell+1)}$ resp. $\hat{c}(\mathfrak{w}_{p_1,p_2}) = \frac{(p_1-1)(p_2-1)}{12\psi(p_1p_2)}$ with $\psi(p_1p_2) = (p_1+1)(p_2+1)$ for $p_1 \neq p_2$ and $\psi(p^2) = p(p+1)$. Notice that ψ_N is the degree in U of the modular

Table 2. Linear regression for the heights

u	c		$\hat{c}(u)$	d	
	$h = 99$	$h = 128$		$h = 99$	$h = 128$
$\mathfrak{w}_5^2$	0.055549	0.055777	0.055556	11.910318	14.394808
$\mathfrak{w}_5^6$	0.166024	0.167145	0.166667	32.650092	37.297890
$\mathfrak{w}_7^2$	0.063095	0.062701	0.062500	13.020135	17.834842
$\mathfrak{w}_7^4$	0.125302	0.125434	0.125000	27.714413	33.970718
$\mathfrak{w}_{11}^2$	0.069677	0.069628	0.069444	18.690994	23.267310
$\mathfrak{w}_{11}^4$	0.139191	0.139334	0.138889	34.588769	42.494392
$\mathfrak{w}_{11}^6$	0.208889	0.208836	0.208333	50.661632	63.755416
$\mathfrak{w}_{11}^{12}$	0.417707	0.418234	0.416667	101.306985	124.133346
$\mathfrak{w}_{13}^2$	0.071750	0.071681	0.071428	27.799672	36.107055
$\mathfrak{w}_{5,7}$	0.043290	0.041195	0.041667	-22.925045	-23.211023
$\mathfrak{w}_{11,13}$	0.060320	0.058992	0.059523	-48.251076	-56.912405
$\mathfrak{w}_{13,13}$	0.066510	0.066076	0.065934	-56.705532	-71.088596

equation $\Phi_N[u]$ and that in fact $\hat{c}(u) = \frac{\deg_J(\Phi_N[u])}{\deg_U(\Phi_N[u])}$ by Theorems 1 and 3. This can be explained as follows.

Proposition 3. *Let u be a class invariant of level N, and Q a quadratic form of discriminant $-D$ such that $u(\tau_Q) \in K_{\mathrm{H}}$. Then*

$$\frac{\mathcal{M}_D[u]}{\mathcal{M}_D[j]} = \frac{\mathcal{H}(u(\tau_Q))}{\mathcal{H}(j(\tau_Q))} = \frac{\deg_J(\Phi_N[u])}{\deg_U(\Phi_N[u])}(1 + o(1)) = \hat{c}(u)(1 + o(1))$$

for the heights tending to infinity.

Proof. The first equality has already been mentioned in Section 6.1. The tuple $P = (u(\tau_Q), j(\tau_Q))$ is a point on the modular curve of level N defined by $\Phi_N[u]$. Considering u and j as rational functions on this curve, namely as the projections on the coordinates, we have by Proposition B.3.5(b) of [16] that

$$\frac{\mathcal{H}(u(\tau_Q))}{\mathcal{H}(j(\tau_Q))} = \frac{\mathcal{H}(u(P))}{\mathcal{H}(j(P))} = \frac{\deg u}{\deg j}(1 + o(1)) = \frac{\deg_J(\Phi_N[u])}{\deg_U(\Phi_N[u])}(1 + o(1))$$

for the heights tending to infinity.

Replacing the Mahler measures in the formula of the proposition by the heights $\mathcal{H}$ resp. $\mathcal{H}'$, which are basically the same in our context (see Section 6.1), and then replacing $\mathcal{H}_D[j]$ by its approximation $\widehat{\mathcal{H}}_D[j]$, we obtain the observed approximation of $\mathcal{H}'_D[u]$ by $\hat{c}\widehat{\mathcal{H}}_D$. It remains to estimate the error introduced by swapping $\mathcal{M}$ for $\mathcal{H}$. From standard arguments, we obtain $|\mathcal{M}_D[u] - \mathcal{H}_D[u]| \in O(h)$. We feel that it should be possible to show that $\sum \frac{1}{A}$ is in $O(\log h)$. As $\sqrt{D}$ is of the order of h, this implies that the error grows indeed more slowly than $\widehat{\mathcal{H}}_D$, so that asymptotically $\hat{c}\widehat{\mathcal{H}}_D$ is a valid approximation of $\mathcal{H}_D[u]$.

6.5 Numerical Example

As an example, we provide in Table 3 the class polynomials obtained for $D = 39$ with each possible invariant.

7 Applications to ECPP

7.1 Building Elliptic Curves Having Complex Multiplication

In ECPP (cf. [2]), the roots of a class polynomial $H_D[u](X)$ over $\mathbb{Z}/N\mathbb{Z}$ for a probable prime N are used to build an elliptic curve having complex multiplication by $\mathcal{O}_K$. When an invariant u associated to a modular curve of positive genus is employed, then the equation $\Phi[u](U, J)$ serves to recover j as suggested in [9].

```
procedure BuildCMCurve($p$, $D$)
 0. Solve $4p = A^2 + DB^2$ in rational integers $A$ and $B$.
 1. Compute $H_D[u](X)$.
```

Table 3. Class polynomials for $D = 39$

u	$c(u)$	$\mathcal{H}'_D[u]$	$H_D[u]$
$\mathfrak{w}_3^{2*}$	0.0417	2.197	$X^4 + (-1 - \omega)\,X^3 - 6\,X^2 + (-6 + 3\,\omega)\,X + 9$
$(\mathfrak{f}/\sqrt{2})^3$	0.0417	1.386	$X^4 - 3\,X^3 - 4\,X^2 - 2\,X - 1$
$\mathfrak{w}_{5,13}$	0.0476	0.000	$X^4 + X^3 - X^2 - X + 1$
$\mathfrak{w}_{13}^2$	0.0714	5.130	$X^4 + 13\,X^3 + 65\,X^2 + 169\,X + 169$
$\mathfrak{w}_{61}^{2*}$	0.0806	7.021	$X^4 + (-11 - 3\,\omega)\,X^3 + (-86 + 32\,\omega)\,X^2 + (714 + 167\,\omega)\,X - 711 - 1120\,\omega$
$\mathfrak{w}_5^{6*}$	0.167	8.511	$X^4 + (-10 - 9\,\omega)\,X^3 + (-490 - 216\,\omega)\,X^2 + (-2915 - 711\,\omega)\,X - 4355 + 4968\,\omega$
$\mathfrak{w}_{11}^{6*}$	0.208	13.816	$X^4 + (-73 + 27\,\omega)\,X^3 + (-8914 + 1656\,\omega)\,X^2 + (-139058 + 7947\,\omega)\,X + 1000693 - 515016\,\omega$
$\sqrt{-D}\gamma_3$	0.500	30.727	$X^4 + 114660\,X^3 + 108456894\,X^2 + 42553748601\,X - 22104665145927$
j	1	44.487	$X^4 + 331531596\,X^3 - 429878960946\,X^2 + 109873509788637459\,X + 20919104368024767633$

2. Compute a root u_0 of $H_D[u](X) \equiv 0 \bmod p$.
3. Compute the set $\mathcal{J}$ of all roots of $\Phi[u](u_0, J) \equiv 0 \bmod p$ and find one elliptic curve having j-invariant in $\mathcal{J}$ which has cardinality $p + 1 - A$.

Some comments are in order. When the genus is zero (as in the case of the original Weber functions or $\mathfrak{w}_\ell$ for $\ell \in \{3, 5, 7, 13\}$), the polynomial $\Phi[u](X, J)$ has degree 1 in J and there is no cost for finding $j(E)$. This is no longer true for positive genus. A degree of 2 in J is still not very costly. A larger degree, however, means that in general several j-invariants have to be tested before a suitable curve is found. Thus, we fix a maximal degree in J with which we are ready to work in the algorithm.

Whenever $(D, 6) = 1$, we can combine our new invariants with Stark's ideas [30]. We only need to find a relation between γ_2 and our invariant u. From that, we can proceed as explained in [22] to reduce the number of curves to test. It turns out that additionally, the modular equations become smaller than the original ones. For instance, there exists a modular equation between γ_2 and $\mathfrak{w}_{11}^4$, which is smaller than Φ_{11}:

$$X^{12} - 1980\,X^9 + 880\,\gamma_2 X^8 + 44\,{\gamma_2}^2 X^7 + 980078\,X^6 - 871200\,\gamma_2 X^5 + 150040\,{\gamma_2}^2 X^4 \\ + \left(47066580 - 7865\,{\gamma_2}^3\right) X^3 + \left(154\,{\gamma_2}^4 + 560560\,\gamma_2\right) X^2 + \left(1244\,{\gamma_2}^2 - {\gamma_2}^5\right) X + 121.$$

7.2 Using the New Invariants

The implementation described in [2] used only Weber functions, and powers of $\mathfrak{f}$ and $\mathfrak{f}_1$ only for discriminants not divisible by 3. It turns out that the new invariants provide a considerable improvement. We restrict to functions for which the degree in J of the modular polynomial is bounded by 6. For prime ℓ, this means that we only use $\mathfrak{w}_\ell$ for

$$\ell \in \{2, 3, 5, 7, 13\}_1 \cup \{19, 37\}_3 \cup \{17\}_4 \cup \{11, 31, 61\}_5 \cup \{73\}_6$$

(the subscript designates the degree in J). For $\mathfrak{w}_{p_1,p_2}$, this means all (p_1, p_2) for which $(p_1-1)(p_2-1)/12 \le 6$ or

$$\{(3,13),(5,7)\}_2 \cup \{(5,13)\}_4 \cup \{(3,37),(5,19),(7,13)\}_6.$$

Considering the optimal invariant as the one with minimal height of the class polynomial, that is a priori with minimal $\hat{c}(u)$, the functions are chosen in the following order:

$$\begin{array}{l} \mathfrak{w}_2 < \mathfrak{w}_2^2 < \mathfrak{w}_{3,13} < \mathfrak{w}_{3,37} < \mathfrak{w}_{5,7} = \mathfrak{w}_3^2 = \mathfrak{w}_2^3 < \mathfrak{w}_{5,13} < \mathfrak{w}_{5,19} < \mathfrak{w}_{7,13} \\ < \mathfrak{w}_2^4 = \mathfrak{w}_{5,2} < \mathfrak{w}_7^2 < \mathfrak{w}_{11}^2 < \mathfrak{w}_{13}^2 < \mathfrak{w}_{17}^2 < \mathfrak{w}_{19}^2 < \mathfrak{w}_{31}^2 < \mathfrak{w}_{37}^2 < \mathfrak{w}_{61}^2 \\ < \mathfrak{w}_{73}^2 < \mathfrak{w}_3^2 = \mathfrak{w}_2^6 < \mathfrak{w}_3^6 = \mathfrak{w}_7^4 < \mathfrak{w}_{11}^4 < \mathfrak{w}_{19}^4 < \mathfrak{w}_{31}^4 < \mathfrak{w}_5^6 = \mathfrak{w}_2^{12} \\ < \mathfrak{w}_{11}^6 < \mathfrak{w}_{17}^6 < \mathfrak{w}_3^{12} < \mathfrak{w}_{11}^{12} < \gamma_2 < \gamma_3 < j. \end{array}$$

If the criterion of choice is the minimal largest root, i.e. the minimal order v of the pole at infinity, then this order is essentially preserved except for powers of Weber's functions and of $\mathfrak{w}_3$, which become less attractive.

Taking our estimation of the height as optimality criterion, we report in Table 4 how often each invariant is used to build the class field. We hereby consider again the 17702 known fundamental discriminants of class numbers between 2 and 64. We distinguish between cases in which the class polynomial is real and cases in which it has coefficients in $\mathcal{O}_K$; the latter ones are marked by "$*$".

Table 4. Statistics for all D s.t. $2 \le h(-D) \le 64$.

u	#	u	#	u	#	u	#	u	#
$\mathfrak{w}_{3,13}$	2533	$\mathfrak{w}_{7,13}$	893	$\mathfrak{w}_{17}^{2*}$	265	$\mathfrak{w}_{17}^2$	38	$(\mathfrak{f}^4/2)^3$	3
$\mathfrak{f}_1(-D)^2/\sqrt{2}$	1978	$\mathfrak{w}_{5,13}$	884	$\mathfrak{w}_{19}^{2*}$	232	$\mathfrak{w}_3^4$	22	$\mathfrak{w}_{19}^{4*}$	2
$\mathfrak{w}_{5,7}$	1856	$\mathfrak{w}_3^{2*}$	830	$\mathfrak{w}_{61}^{2*}$	166	$\mathfrak{w}_{37}^2$	20	$\mathfrak{w}_{31}^{4*}$	2
$\mathfrak{w}_{3,37}$	1385	$\mathfrak{w}_{5,19}$	599	$\mathfrak{w}_{13}^2$	131	$\mathfrak{w}_3^{6*}$	16	$\mathfrak{w}_3^4$	2
$\mathfrak{w}_7^{2*}$	1105	$\mathfrak{w}_{13}^{2*}$	467	$\mathfrak{w}_{31}^{2*}$	125	γ_2	14	$\mathfrak{w}_7^4$	2
$\mathfrak{w}_{11}^{2*}$	1011	$\mathfrak{f}(-D)^4$	383	$\mathfrak{w}_{73}^{2*}$	75	$\mathfrak{w}_{61}^2$	7	$\mathfrak{w}_7^{4*}$	1
$\mathfrak{f}(-D)^2/\sqrt{2}$	999	$\mathfrak{w}_5^2$	326	$(\mathfrak{f}/\sqrt{2}))^3$	43	$\mathfrak{w}_{73}^2$	4		
$\mathfrak{f}(-4D)/\sqrt{2}$	929	$\mathfrak{w}_5^{2*}$	310	$\mathfrak{w}_{37}^{2*}$	41	$(\mathfrak{f}^2/\sqrt{2})^3$	3		

For the previous implementation of ECPP, the figures are as follows:

u	#	u	#	u	#	u	#
γ_2	8621	$\mathfrak{f}_1(-D)^2/\sqrt{2}$	1978	$\mathfrak{f}(-D)^2/\sqrt{2}$	999	$\mathfrak{f}(-4D)/\sqrt{2}$	929
$\sqrt{-D}\gamma_3$	2967	j	1245	$\mathfrak{f}(-D)^4$	963		

Notice that j and γ_3 disappeared completely from the new table, and that only a few discriminants are left that require γ_2.

All these data are included in the version of ECPP under development by the second author (check his web page).

8 Conclusions

We have shed some light on the use of different invariants for building class fields, which have, for instance, applications to primality proving. We have shown how to choose invariants leading to smaller polynomials and making the computations required in [15] feasible. An open question remains: what would be the best modular equation for our purpose? D. Kohel has suggested [19] Atkin's "optimal" modular equations, already used in the SEA algorithm (see [1,21]), and the impact of his work needs to be seen.

Acknowledgements

We thank D. Kohel and É. Schost for many discussions on our work, J.-L. Nicolas for indicating [5], N. Brisebarre for kindly answering further questions, and an anonymous referee for indications on how to turn our initial heuristics into Proposition 3. The first author gratefully acknowledges being supported by a fellowship within the postdoctoral programme of the German Academic Exchange Service (DAAD). The second author is on leave from the French Department of Defense, Délégation Générale pour l'Armement. This research was partially supported by the French Ministry of Research — ACI Cryptologie.

References

1. A. O. L. Atkin. The number of points on an elliptic curve modulo a prime. Draft, 1988.
2. A. O. L. Atkin and F. Morain. Elliptic curves and primality proving. *Math. Comp.*, 61(203):29–68, July 1993.
3. B. J. Birch. Weber's class invariants. *Mathematika*, 16:283–294, 1969.
4. J. M. Borwein and P. B. Borwein. *Pi and the AGM.* John Wiley, 1987.
5. N. Brisebarre and G. Philibert. Effective lower and upper bounds for the Fourier coefficients of powers of the modular invariant j. In preparation, January 2002.
6. H. Cohen. *Advanced topics in computational number theory*, volume 193 of *Graduate Texts in Mathematics*. Springer-Verlag, 2000.
7. R. Dedekind. Erläuterungen zu den vorstehenden Fragmenten. In R. Dedekind and H. Weber, editors, *Bernhard Riemann's gesammelte mathematische Werke und wissenschaftlicher Nachlaß*, pages 438–447. Teubner, Leipzig, 1876.
8. A. Enge and F. Morain. Further investigations of the generalised Weber functions. In preparation, 2001.
9. A. Enge and R. Schertz. Constructing elliptic curves from modular curves of positive genus. In preparation, 2001.
10. A. Enge and R. Schertz. Modular curves of composite level. In preparation, 2001.
11. Robert Fricke. *Lehrbuch der Algebra*, volume III — Algebraische Zahlen. Vieweg, Braunschweig, 1928.
12. A. Gee. Class invariants by Shimura's reciprocity law. *J. Théor. Nombres Bordeaux*, 11:45–72, 1999.
13. A. Gee and P. Stevenhagen. Generating class fields using Shimura reciprocity. In J. P. Buhler, editor, *Algorithmic Number Theory*, volume 1423 of *Lecture Notes in Comput. Sci.*, pages 441–453. Springer-Verlag, 1998. Third International Symposium, ANTS-III, Portland, Oregon, june 1998, Proceedings.

14. G. Hanrot and F. Morain. Solvability by radicals from a practical algorithmic point of view. Submitted. Available from `http://www.lix.polytechnique.fr/-Labo/Francois.Morain/`, November 2001.
15. G. Hanrot and F. Morain. Solvability by radicals from an algorithmic point of view. In B. Mourrain, editor, *Symbolic and algebraic computation*, pages 175–182. ACM, 2001. Proceedings ISSAC'2001, London, Ontario.
16. Marc Hindry and Joseph H. Silverman. *Diophantine Geometry — An Introduction.* Springer-Verlag, New York, 2000.
17. Carl Gustav Jacob Jacobi. Fundamenta nova theoriae functionum ellipticarum. In *Gesammelte Werke*, pages 49–239. Chelsea, New York, 2 (1969) edition, 1829.
18. Felix Klein. Über die Transformationsgleichung der elliptischen Funktionen und die Auflösung der Gleichungen fünften Grades. *Math. Annalen*, 14:111–172, 1878. *Gesammelte Mathematische Abhandlungen* III:13–75.
19. D. Kohel. CM divisors on modular curves. In preparation, January 2002.
20. C. Meyer. Bemerkungen zum Satz von Heegner-Stark über die imaginärquadratischen Zahlkörper mit der Klassenzahl Eins. *J. Reine Angew. Math.*, 242:179–214, 1970.
21. F. Morain. Calcul du nombre de points sur une courbe elliptique dans un corps fini : aspects algorithmiques. *J. Théor. Nombres Bordeaux*, 7:255–282, 1995.
22. F. Morain. Primality proving using elliptic curves: an update. In J. P. Buhler, editor, *Algorithmic Number Theory*, volume 1423 of *Lecture Notes in Comput. Sci.*, pages 111–127. Springer-Verlag, 1998. Third International Symposium, ANTS-III, Portland, Oregon, june 1998, Proceedings.
23. F. Morain. Modular curves and class invariants. Preprint, June 2000.
24. R. Schertz. Die singulären Werte der Weberschen Funktionen $\mathfrak{f}$, $\mathfrak{f}_1$, $\mathfrak{f}_2$, γ_2, γ_3. *J. Reine Angew. Math.*, 286/287:46–74, 1976.
25. R. Schertz. Weber's class invariants revisited. To appear in J. Théor. Nombres Bordeaux, 2001.
26. Reinhard Schertz. Zur expliziten Berechnung von Ganzheitsbasen in Strahlklassenkörpern über einem imaginär-qudratischen Zahlkörper. *Journal of Number Theory*, 34(1):41–53, January 1990.
27. Goro Shimura. *Introduction to the Arithmetic Theory of Automorphic Functions.* Iwanami Shoten and Princeton University Press, 1971.
28. J. H. Silverman. *The Arithmetic of Elliptic Curves*, volume 106 of *Grad. Texts in Math.* Springer, 1986.
29. J. H. Silverman. *Advanced Topics in the Arithmetic of Elliptic Curves*, volume 151 of *Grad. Texts in Math.* Springer-Verlag, 1994.
30. H. M. Stark. Counting points on CM elliptic curves. *Rocky Mountain J. Math.*, 26(3):1115–1138, 1996.
31. H. Weber. *Lehrbuch der Algebra*, volume III. Chelsea Publishing Company, New York, 1908.
32. N. Yui and D. Zagier. On the singular values of Weber modular functions. *Math. Comp.*, 66(220):1645–1662, October 1997.

A Database of Elliptic Curves – First Report

William A. Stein[1] and Mark Watkins[2]

[1] Harvard University
was@math.harvard.edu
http://modular.fas.harvard.edu
[2] The Pennsylvania State Univerisity
watkins@math.psu.edu
http://www.math.psu.edu/watkins

1 Introduction

In the late 1980s, Brumer and McGuinness [2] undertook the construction of a database of elliptic curves whose absolute discriminant $|\Delta|$ was both prime and satisfied $|\Delta| \leq 10^8$. While the restriction to primality was nice for many reasons, there are still many curves of interest lacking this property. As ten years have passed since the original experiment, we decided to undertake an extension of it, simultaneously extending the range for the type of curves they considered, and also including curves with composite discriminant. Our database can be crudely described as being the curves with $|\Delta| \leq 10^{12}$ which either have conductor smaller than 10^8 or have prime conductor less than 10^{10}—but there are a few caveats concerning issues like quadratic twists and isogenous curves. For each curve in our database, we have undertaken to compute various invariants (as did Brumer and McGuinness), such as the Birch–Swinnerton-Dyer L-ratio, generators, and the modular degree. We did not compute the latter two of these for every curve. The database currently contains about 44 million curves; the end goal is find as many curves with conductor less than 10^8 as possible, and we comment below on this direction of growth of the database. Of these 44 million curves, we have started a first stage of processing (computation of analytic rank data), with point searching to be carried out in a later second stage of computation.

Our general frame of mind is that computation of many of the invariants is rather trivial, for instance, the discriminant, conductor, and even the isogeny structure. We do not even save these data, expecting them to be recomputable quite easily in real time. For instance, for each isogeny class, we store only one representative (the one of minimal Faltings height), as we view the construction of isogenous curves as a "fast" process. It is only information like analytic ranks, modular degrees (both of which use computation of the Frobenius traces l_p), and coordinates of generators that we save; saving the l_p themselves would take too much storage space. It might be seen that our database could be used a "seed" for other more specialised databases, as we can quickly calculate the less time-consuming information and append it to the saved data.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 267–275, 2002.

2 Generating the Curves

While Brumer and McGuinness fixed the a_1, a_2, a_3 invariants of the elliptic curve (12 total possibilities) and then searched for a_4 and a_6 which made $|\Delta|$ small, we instead decided to break the c_4 and c_6 invariants into congruence classes, and then find small solutions to $c_4^3 - c_6^2 = 1728\Delta$. We write $c_4^\star$ for the least nonnegative residue of c_4 modulo 576, and $c_6^\star$ for the least nonnegative residue of c_6 modulo 1728. The work of Connell [3] gives necessary and sufficient conditions on c_4 and c_6 for an elliptic curve with such invariants to exist. We first need that $c_6 \equiv 3 \pmod 4$ (when it follows that c_4 is odd) or $2^4 \mid c_4$ and $c_6 \equiv 0, 8 \pmod{32}$, and secondly we require a local condition at the prime 3, namely that $c_6 \not\equiv \pm 9 \pmod{27}$. Using this information and the fact that $1728 \mid \left(c_4^3 - c_6^2\right)$, this leads to 288 possible $(c_4^\star, c_6^\star)$ pairs.

For each fixed such $(c_4^\star, c_6^\star)$ pair, we can simply loop over c_4 and c_6, finding the curves with $|\Delta| \leq 10^{12}$. Of course, it is only under the ABC-conjecture that we would have an upper bound on c_4 to ensure that we would have found all such curves, and even then the bound would be too large. Our method was to take $c_4 \leq 1.44 \cdot 10^{12}$ in this first step; in any case, curves with larger c_4 are most likely found more easily using the method of Elkies [5].

2.1 Minimal Twists

In the sequel, we shall write E_d for the quadratic twist of E by d. For each (c_4, c_6) pair (again with $c_4 \leq 1.44{\cdot}10^{12}$) which satisfies the $|\Delta| \leq 10^{12}$ condition, we then determine whether this curve is minimal—not only in the traditional sense for its minimal discriminant, but also whether it is has the minimal discriminant in its family of quadratic twists. For $p \geq 5$, this is rather easy to determine; unless $p^6 \mid \Delta$ and $p \mid c_4$, the curve is minimal for quadratic twists (the only difference between this and the standard notion of minimality is that the exponent here is 6 instead of 12). If both the above conditions hold, then we throw the curve out, as $E_{\tilde{p}}$, where $\tilde{p} = \left(\frac{-1}{p}\right) p$, is a curve with lesser discriminant (which will be found by our search procedure). Given that the curve is minimal at a prime divisor $p \geq 5$ of Δ, the local conductor at p is p^2 if $p \mid c_4$ and p^1 otherwise.

The case with $p = 3$ is a bit harder. By Connell's conditions, we see that if $3 \mid c_6$ and $3^9 \mid \left(c_4^3 - c_6^2\right)$ but 3^5 does not exactly divide c_6, then E_{-3} is a curve with invariants $(c_4/9, -c_6/27)$ which has the discriminant reduced by 3^6. This is the only prohibition against the curve being the minimal twist at 3. If $3 \mid\mid c_4$, the curve has good reduction (at 3), while if c_4 is not divisible by 3, the curve has either good or multiplicative reduction. In both cases, the local conductor can be computed readily, it being 3^0 for good reduction and 3^1 for multiplicative. To compute the conductor in the remaining cases of additive reduction (where we know that $3^2 \mid c_4$ and $3^3 \mid c_6$), let $\tilde{c}_4$ be the the least nonnegative residue of $(c_4/9)$ modulo 3, and $\tilde{c}_6$ be the the least nonnegative residue of $(c_6/27)$ modulo 9. Table 1 then gives us the exponent of the local conductor. Here $e = 5$ if $3^4 \mid c_4$ and $e = 4$ if $3^3 \mid\mid c_4$ (note that we must have $3^5 \mid\mid c_6$ in this case for the curve to be twist-minimal, and that the table assumes that the curve is twist-minimal).

Table 1. Local Conductors at 3

$\tilde{c}_4\backslash\tilde{c}_6$	0	1	2	3	4	5	6	7	8
0	e	3	3	5	2	2	5	3	3
1	2	3	4	3	4	4	3	4	3
2	2	3	2	3	3	3	3	2	3

For $p = 2$, the minimality test and conductor computation is much more complicated. We include the prime at infinity (twisting by -1) in the test for $p = 2$. By Connell's conditions, if $2^6 \mid c_4$ and $2^8 \mid c_6$, we see that E_2 is a curve with invariants $(c_4/4, c_6/8)$, and has a lesser discriminant. Also if $2^6 \mid c_4$ and $2^6 \mid\mid c_6$, then one of the twists $E_{\pm 2}$ (the sign depending on whether $c_6/8$ is 8 mod 32) has lesser discriminant. And finally if we have $2^4 \mid\mid c_4$ and $2^6 \mid\mid c_6$ and $2^{18} \mid \left(c_4^3 - c_6^2\right)$, then one of $E_{\pm 1}$ (depending on whether $c_6/64$ is 3 mod 4) is nonminimal (in the standard sense) at 2, and hence can be ignored. If none of these events happens, then the curve is twist-minimal at $p = 2$ and the infinite prime. We next describe how to compute the local conductor at $p = 2$ in terms of congruence conditions. If c_4 is odd, then the local conductor is 2^0 or 2^1, depending on whether 2 divides Δ. In the case where $2^4 | c_4$, when c_6 is 8 mod 32 there is good reduction at 2, and again the local conductor is 2^0. So we are left to consider the cases of additive reduction where $2^4 \mid c_4$ and $2^5 \mid c_6$. Let $\tilde{c}_4$ be the the least nonnegative residue of $(c_4/16)$ modulo 8, and $\tilde{c}_6$ be the the least nonnegative residue of $(c_6/32)$ modulo 8. Table 2 then gives the exponent of the local conductor at 2. In this, the dashed entries simply do not occur. For the entries marked by e, let $\tilde{c}_4$ be the the least nonnegative residue of $(c_4/16)$ modulo 16, and $\tilde{c}_6$ be the the least nonnegative residue of $(c_6/32)$ modulo 16. We then use the further Table 3. All the conductor computations are exercises with Tate's algorithm [12]; again the claims on the conductor need only be valid upon assuming that the curve is twist-minimal.

Table 2. Local Conductors at 2

$\tilde{c}_4\backslash\tilde{c}_6$	0	1	2	3	4	5	6	7
1,5	6	4	e	3	6	4	e	3
2,6	8	3	6	4	7	3	6	4
3,7	5	2	7	4	5	2	7	4
4	6	2	-	4	3	2	-	4
0	6	2	-	4	2	2	-	4

A curve which has minimal discriminant at $p = 2$ will be of minimal conductor at $p = 2$ unless $2^4 \mid\mid N$ or $2^6 \mid\mid N$; we can throw out the curve in the first case, since E_{-1} will be found in the search process (and it has lesser conductor). But in the latter case, we cannot immediately discard the curve, as E_2 will have

Table 3. More of the Same

$\tilde{c}_4 \backslash \tilde{c}_6$	2	6	10	14
1	4	5	5	3
5	3	2	4	4
9	5	3	4	5
13	4	4	3	2

conductor smaller by a factor of 2, but the discriminant is larger by a factor of 64 (this behavior follows from the assumption that E has a twist-minimal discriminant and $2^6 \,||\, N$). So only if $|\Delta| \leq 10^{12}/64$ do we discard the curve; in the alternative case we replace the curve by E_2, so that we have the twist of minimal conductor. Finally, if we have $2^5 \,||\, N$ (possibly after the above twisting by 2), or $2^7 \mid N$, we make the arbitrary decision to discard the curve if $c_6 < 0$, as we will also find E_{-1} in the search, which will have the same conductor and discriminant. This positivity condition on c_6 will be part of our definition of minimal twist.

Using the above method, we can rid ourselves of all curves which are not minimal twists, and simultaneously compute the conductor. If $N > 10^{10}$, we simply ignore the curve; if $N > 10^8$ (and $N \leq 10^{10}$), we check whether N is a strong pseudoprime for 2, 13, 23, and 1662803, this being sufficient to prove primality [6]. At this point, we have a list of curves which meet our size conditions on the discriminant, and which have the minimal conductor in a family of quadratic twists (and minimal discriminant at primes other than $p = 2$).

2.2 Isogenous Curves

The next step will be to get rid of isogenous curves. The process of finding all curves isogenous to a given one is described in [4]. This is a fairly fast process, as most curves will have no nontrivial isogenies. Amongst the isogenous curves, we then take the curve of largest fundamental volume, that is, minimal Faltings height (which is unique by [11]), as our representative. Note that this curve might not have the minimal discriminant in the isogeny class. Our final set of curves is then: the set of elliptic curves E such that E has minimal height in its isogeny class, and has some isogenous curve F (possibly the same as E) for which we have $c_4 \leq 1.44 \cdot 10^{12}$ and either $N \leq 10^{10}$ with $|\Delta|$ prime, or $N \leq 10^8$ with $|\Delta| \leq 10^{12}$ for either the curve F or F_2.

2.3 Future Extension of the Database

As stated above, we would desire to have all minimal twists which have conductor less than 10^8. Cremona's tables have 20726 minimal twists with conductor less than 10^4, and so we might guess there are about 200–250 million minimal twists with conductor less than 10^8, while we only have about 44 million currently. There are many ways of enlarging the database. A first is extending the range on c_4 by using the algorithm of [5], but this will likely add only a small amount of curves. A better way is to find families in which we expect the conductor to be substantially less than the discriminant; for instance, curves with a rational point of order 5 will have a large 5th power dividing the discriminant, which will be reduced to a first power in the conductor. It appears that this technique will add many curves to the database — our results are as yet preliminary, and will be included in a future report on the database. For instance, Cremona's curve 174A given by $[1, 0, 1, -7705, 1226492]$ is not currently in our database, but will be found quickly with parametrisations of 3-torsion. A more simple method for enlarging the database is to extend the discriminant limit to (say) 10^{13} for certain $(c_4^\star, c_6^\star)$ pairs, especially those for which we know ahead of time that we will save significant powers of 2 and 3 in the conductor compared to the discriminant. Consideration of higher powers might allow us to find curves like 11949C (which is $[0, 1, 1, -1218949649, 16380150812351]$) where the discriminant is $-3^{41}7^2569$. However, we will certainly not find all of Cremona's curves, as some like 11770I (which is $[1, -1, 1, -2246050998, 40972734736581]$, and has discriminant $-2^{13}5^3 11^{11}107^4$) will not be found by any of our methods, as the absolute discriminant here is more than 10^{25}. As our database is not meant to be exhaustive, this is not a huge worry; we desire to put as much into the database as possible over as large of ranges as possible, but are not overly worried about exhaustiveness, preferring to include as much useful information as we can, without considering whether our database is "complete" in some sense.

3 Data Computed for Each Curve

One object of interest for an elliptic curve is its algebraic rank. This is hard to compute; indeed, there is no known algorithm to do this, only ones which work conditionally. By the process given in [4], we can try to determine the **analytic rank** of the curve, which is the degree of vanishing of its L-series at the central point. Of course, as there is no way to determine if a computed number is exactly zero, we can only give a good guess as to the analytic rank. The conjecture of Birch and Swinnerton-Dyer asserts that the algebraic rank and the analytic rank are equal, and that the first nonzero derivative of the L-function at the central point has arithmetic significance. For each curve in the database, we computed the suspected analytic rank and first nonzero derivative for both the curve itself and some of its quadratic twists.

Each curve in our database is the curve of minimal Faltings height in its isogeny class. A conjecture of Stevens [11] asserts that this curve should be the **optimal** curve for parametrisations from $X_1(N)$, in the sense that the parametrisations to the isogenous curves factor through the parametrisation to the strong curve (the existence of a modular parametrisation from $X_1(N)$ was proved in [1] following the methods initiated by Wiles [14]). It is sometimes the case that the optimal curve for parametrisations from $X_0(N)$ differs from the curve we find; in [13], a process is given to find the $X_0(N)$-optimal curve, assuming a technical condition, namely that the Manin constant of the optimal curve is 1 (this is similar to the Stevens conjecture). As many of the Frobenius traces were already computed for the analytic rank computation, these can be re-used at this stage. In a section below, we discuss the data obtained.

In the aforementioned paper [13], a process is given to compute the modular degree of an elliptic curve, again assuming that the Manin constant is 1. Compared to the computation of the analytic rank, which requires about the first $\sqrt{N}$ of the Frobenius traces, this method requires on the order of N of these (actually $\tilde{N}$, the symmetric-square conductor; see below). Thus for $N \geq 300000$ or so, it becomes rather time-consuming to compute the modular degree. We therefore compromised, computing the modular degree only if the symmetric-square conductor of the elliptic curve was sufficiently small (if we write $N = \prod_p p^{f_p}$ as a product of local conductors, then the symmetric-square conductor is simply $\tilde{N} = \prod_p p^{\lceil f_p/2 \rceil}$, except possibly when $f_2 = 8$, when the local symmetric-square conductor at 2 might be either 2^3 or 2^4; see [13] for details). We also computed the modular degree in some other interesting cases, for instance, when the rank is large, or in the case where there are differing optimal curves, a topic which we now discuss.

4 Differing Optimal Curves

Here we discuss the question of differing optimal curves for parametrisations from $X_0(N)$ and $X_1(N)$. Note that we do not compute the actual optimal curve for the latter, relying instead on the Stevens conjecture, and compute the optimal curve for $X_0(N)$ only under the assumption that the Manin constant is 1. But the results are still interesting.

There appear to be three families in which the optimal curves differ by a 2-isogeny. One of these, the so-called Setzer-Neumann curves (see [10], [8,9]), was considered by Mestre and Oesterlé in [7]. These curves are parametrised by $c_4 = P - 16$ and $c_6 = u\,(P+8)$, with the discriminant $P = u^2 + 64$ being a prime and u being taken to be congruent to 3 mod 4 to make c_6 be congruent to 3 mod 4 (other authors have taken u to be 1 mod 4). The second family corresponds to taking $c_4 = 16P - 16$ and $c_6 = 4v\,(16P+8)$ with here v being 3 mod 4 and $P = v^2 + 4$ being prime. Here the conductor is $4P$ and the discriminant is $16P$; the differing optimal curves property appears to be preserved upon twisting by -1, which corresponds to negating c_6 (or v). If we take $u = 0$ or $v = 0$, we get the minimal Faltings height curve $[0, 0, 0, -1, 0]$ in the isogeny class 32A, which differs from the $X_0(32)$-optimal curve $[0, 0, 0, 4, 0]$ by a 2-isogeny. Noting that P in this case is a prime power, we can further expand the families to include the isogeny classes 128B/128D which come about from taking $v = \pm 2$ in the second family, and also $u = 15$ in the first family and $v = 11$ in the second family, giving the isogeny classes 17A and 20A respectively. Note that taking $v = -1$ in the second family also gives the isogeny class 20A. Indeed the curve obtained from $v = -1$ is the minimal Faltings height curve $[0, 1, 0, -1, 0]$, while the curve obtained from $v = 11$ differs by a 3-isogeny (since 125 is a third power). Taking $v = 1$ and $v = -11$ leads to similar behavior with the isogeny class 80B.

The class 17A will reappear in our third family; here the curve obtained from taking $u = 15$ differs from the minimal Faltings height curve $[1, -1, 1, -1, 0]$ by a 2-isogeny, and the $X_0(17)$-optimal curve is $[1, -1, 1, -1, -14]$, differing from the $X_1(17)$-optimal curve by a 4-isogeny.

The third family we have found is parametrised by $c_4 = PQ + 16$ and $c_6 = (P + 8)(PQ - 24)$ of discriminant PQ with $Q = P + 16$ and P congruent to 3 mod 4, and with both $|P|$ and $|Q|$ being prime powers, at least one of them being a power of a prime which is congruent to 3 mod 4 (so that $P = 11$ or $P = -2417$ works, but $P = -641$ does not). Upon taking $P = -17$, we obtain the $X_1(17)$-optimal curve for 17A. The isogeny class 15A (where the optimal curves differ by a 4-isogeny) comes about from both $P = -25$ and $P = -1$, the latter giving the minimal Faltings height curve even though $Q = P + 16 = 15$ is not a prime power. Similar to this are some cases where P is even, namely $P = -4$ and $P = -20$, which give 24A and 40A, and the corresponding quadratic twists $P = -12$ and $P = 4$, giving 48A and 80A. Finally there is $P = -8$, which gives 64A, the quadratic twist of 32A. These are all the known examples where the optimal curves differ by a 2-isogeny (and the two examples where they differ by a 4-isogeny); the above-cited work [7] contains the only partial results toward a proof of this classification.

Ignoring the 5-isogeny example of 11A as being spurious, this leaves just the occasions of the optimal curves differing by a 3-isogeny. Here, all known examples are parametrised by

$$c_4 = (n + 3)\left(n^3 + 9n^2 + 27n + 3\right) = (n + 3)^4 - 24(n + 3)$$

and

$$\begin{aligned} c_6 &= -\left(n^6 + 18n^5 + 135n^4 + 504n^3 + 891n^2 + 486n - 27\right) \\ &= -(n + 3)^6 + 36(n + 3)^3 - 216 \end{aligned}$$

where the discriminant is $n\left(n^2 + 9n + 27\right)$. The n's for which the optimal curves differ are (experimentally) precisely those for which $n^2+9n+27$ is a prime power and n has no prime factors congruent to 1 mod 6; else the optimal curves are the same. We have no proof of this.

Within these families with differing optimal curves, we also have conjectures regarding the parity of the modular degree (of the $X_0(N)$-optimal curve). In the first family, if u is 3 mod 8 then the modular degree is odd, while if u is 7 mod 8, the modular degree is even. In work joint with Matt Baker, we have been able to use the recent Refined Eisenstein Theorem of Emerton to prove this observation. In the second family, the modular degree is always odd when v is 3 mod 4 (while the quadratic twist corresponding to $-v$ will have a modular degree greater by a factor of four, and hence be even) — since the conductor here is not prime, our techniques are not applicable, and so we have no proof. In the third family, if P is 7 mod 24, then the modular degree is even, while it is odd if P is 19 mod 24; again we have no proof.

The 3-isogeny family has similar properties regarding the 3-divisibility of the modular degree. The cases where $3|n$ we shall ignore. Also, we ignore $|n| = 8$, where 3 exactly divides the modular degree. Having done this, if n is not a prime power, then 27 divides the modular degree. Else let $|n| = p^r$ and $3^k \mid\mid (p+1)$. We then have that 3^k exactly divides the modular degree, except if $k = 1$, when 3 does not divide the modular degree. We again have no proofs of these experimental data (and few examples where $r \neq 1$ or k is large).

5 Data Obtained

This may seem strange for a comprehensive database project, but we do not dwell on large-scale phemonemon; indeed, the Brumer–McGuinness work is probably already sufficient in this manner, at least for prime conductor. As noted there, telling the difference between a small power of 10^8 (or whatever the upper limit of consideration may be) and a large power of its logarithm is rather hopeless—extending their data by a factor of 5/4 on the logarithmic scale does not help matters much. The Brumer–McGuinness database had 310711 curves (five less than their stated number due to differences in their accounting), though their paper also states that they had actually found 311243 curves but threw some of them out; we have 839 curves which have prime conductor less than 10^8 which are not in their database. We have 11386955 isogeny classes of curves with prime conductor less than 10^{10} in our database (this should grow slightly when curves with $c_4 \geq 1.44 \cdot 10^{12}$ are added). Of these curves with prime conductor, of the ones we have processed, we have that 62.5% of the curves with even functional equation possess rank 0, compared to about 60% for Brumer–McGuinness. It is conjectured that asymptotically this percentage should be 100%. Similarly, 92.5% of the curves with odd functional equation have rank 1, slightly more than the previous results. The least conductor for a rank 5 curve we have found is 34672310 for $[1, -1, 0, -415, 3481]$, and for rank 6 we have $[1, 0, 0, -9227, 340354]$ of conductor 6822208199. These respectively fall short to the best-known (to the authors) examples of $[0, 0, 1, -79, 342]$ of conductor 19047851 and $[0, 0, 1, -7077, 235516]$ of conductor 5258110041 (the former appears in the Brumer–McGuinness database; the latter is due to Tom Womack).

Instead of concentrating on large-scale behavior, we see our database as more of a tool to be used by other mathematicians. For instance, Neil Dummigan queried us concerning examples of strong Weil curves with rank 2 and a rational point of order 5 for which the conductor is not divisible by 5, and we were able to provide him with the example $[0, 1, 1, -840, 39800]$ of conductor 13881 (and modular degree 52000), among other examples which were beyond the range of Cremona's tables (which include $[1, 1, 1, -2365, 43251]$ of conductor 5302). Though we would likely be better able to answer the question after extending our database with parametrisations from $X_0(5)$, the efficacy of our database was evinced. As another example, the second author has conjectured in [13] that 2^r divides the modular degree for any curve (where r is the rank), and perhaps higher powers of 2 should divide the modular degree when the conductor is composite, due to factorisation through Atkin–Lehner involutions. For many large-rank curves in the Brumer–McGuinness database, we verified this. With our extension to curves of composite conductor, we are able to give more evidence for this conjecture. Also, the third 2-isogeny family in the previous section was discovered after looking at our data, as was the parametrisation of the 3-isogeny family, and finally our analytic rank data concerning quadratic twists could be of use.

Acknowledgements

The authors would like to thank Neil Dummigan for the question mentioned in Section V, Noam Elkies for many useful observations, especially concerning Section IV, and Blair Kelly III and Wayne Whitney for providing computing power.

References

1. C. Breuil, B. Conrad, F. Diamond, and R. Taylor, *On the modularity of elliptic curves over* **Q***: Wild 3-adic exercises.* J. Amer. Math. Soc. **14** (2001), 843–939.
2. A. Brumer, O. McGuinness, *The behavior of the Mordell-Weil group of elliptic curves.* Bull. Amer. Math. Soc. (N.S.) **23** (1990), no. 2, 375–382.
3. I. Connell, Lecture Notes from class at McGill University, 1991.
4. J. Cremona, *Algorithms for modular elliptic curves.* Cambridge University Press, Cambridge, 1992. Second edition 1997.
5. N. Elkies, *Rational points near curves and small nonzero* $|x^3 - y^2|$ *via lattice reduction.* In *Algorithmic number theory* (Leiden 2000), 33–63, Lecture Notes in Comput. Sci., 1838, Springer, Berlin, 2000.
6. G. Jaeschke, *On strong pseudoprimes to several bases.* Math. Comp. **61** (1993), no. 204, 915–926.
7. J.-F. Mestre, J. Oesterlé, *Courbes de Weil semi-stables de discriminant une puissance m-ième.* (French) J. Reine Angew. Math. **400** (1989), 173–184.
8. O. Neumann, *Elliptische Kurven mit vorgeschriebenem Reduktionsverhalten. I* (German) Math. Nachr. **49** (1971), 107–123.
9. O. Neumann, *Elliptische Kurven mit vorgeschriebenem Reduktionsverhalten. II* (German) Math. Nachr. **56** (1973), 269–280.
10. B. Setzer, *Elliptic curves of prime conductor.* J. London Math. Soc. (2) **10** (1975), 367–378.
11. G. Stevens, *Stickelberger elements and modular parametrizations of elliptic curves.* Invent. Math. **98** (1989), no. 1, 75–106.
12. J. Tate, *Algorithm for determining the type of a singular fiber in an elliptic pencil.* In *Modular functions of one variable IV*, edited by B. Birch and W. Kuyk, 33–52, Lecture Notes in Math., Vol. 476, Springer, Berlin, 1975.
13. M. Watkins, *Computing the modular degree of an elliptic curve,* preprint, 2001.
14. A. Wiles, *Modular elliptic curves and Fermat's last theorem.* Ann. of Math. (2) **141** (1995), no. 3, 443–551.

Isogeny Volcanoes and the SEA Algorithm

Mireille Fouquet and François Morain*

Laboratoire d'Informatique, École Polytechnique, F-91128 Palaiseau Cedex, France

Abstract. Recently, Kohel gave algorithms to compute the conductor of the endomorphism ring of an ordinary elliptic curve, given the cardinality of the curve. Using his work, we give a complete description of the structure of curves related via rational ℓ-degree isogenies, a structure we call a volcano. We explain how we can travel through this structure using modular polynomials. The computation of the structure is possible without knowing the cardinality of the curve, and that as a result, we deduce information on the cardinality.

1 Introduction

Let E be an elliptic curve over a finite field $\mathbb{F}_q$, where $q = p^r$ with p prime. By Hasse's theorem, the Frobenius π of the curve is an endomorphism of degree 2 with characteristic polynomial $\chi(T) = T^2 - tT + q$ where $|t| \leq 2\sqrt{q}$. It is also known since Deuring [6] that the endomorphism ring of E is either an order in an imaginary quadratic field (the *ordinary* case) or an order in a quaternion algebra (the *supersingular* case). Suppose that E is ordinary and let $d_\pi = t^2 - 4q$ be the discriminant of π. We can write $d_\pi = g^2 d_K$ where d_K is the discriminant of the associated imaginary quadratic field K. To each $f \mid g$ corresponds an order of K and to each such order corresponds an isogeny class of elliptic curves having this particular order as endomorphism ring.

Kohel has shown in his thesis [10] how all these curves are related via isogenies of degree dividing g. Studying this correspondance more closely, we introduce the complete structure of isogenies that we call a *volcano*. Kohel's approach starts from g and finds the conductor f of $\mathrm{End}(E)$, using modular polynomials. We revert this algorithm, using modular polynomials to find g and f. As a consequence, we can come up with an algorithm for computing an elliptic curve of any prescribed conductor $k \mid g$ and in particular the maximal endomorphism ring ($k = 1$), algorithm that is needed in [9].

After introducing some basic notations, we will recall the relevant facts about Kohel's work that describe the structure that grows "under" the isogeny cycles introduced by Couveignes and Morain in [4], forming a volcano. Then we recall the relevant theory of modular polynomials and we are ready to "invert" Kohel's theorem to see the situation from the modular side, which will lead to

* The second author is on the leave from the French Department of Defense, Délégation Générale pour l'Armement. This research was partially supported by the French Ministry of Research – ACI Cryptologie.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 276–291, 2002.

our algorithm. We then give some applications. The first one is related to the computation of t. For a prime $\ell \mid g$, our algorithm gives the ℓ-adic valuation of t and this information can be used in Schoof's algorithm. We can also relate this new structure to the trees that were invented in [3] and use it in the algorithm given in [2] to compute the equation class of an order $\mathcal{O}$. This method is based on the computation of all the j-invariants of curves satisfying certain conditions. The problem is that they never distinguish the curves having an endomorphism ring equal to $\mathcal{O}$ from the others, problem that can be solved using the structure of the volcanoes. Numerical examples are given to illustrate our work.

Although the general theory works for any characteristic, we concentrate on examples where the characteristic is not 2 or 3. The modifications to be made concern formulas for computing isogenous curves, but we do not insist on these in this article.

2 Extending Kohel's Work

2.1 Prerequisites and Notations

If an elliptic curve is not supersingular, then it is known that its ring of endomorphisms is an order in an imaginary quadratic field. Isogenous curves share the same underlying field. In this article, we will consider a set of isogenous curves and the relations between them, so that we can assume that we are dealing with a fixed imaginary quadratic field K of discriminant d_K and maximal order $\mathcal{O}_K$, which can be written as $\mathbb{Z}[\omega_K]$ with $\omega_K = \frac{d_K+\sqrt{d_K}}{2}$. As is well known [1], an order $\mathcal{O}$ in K is completely characterized by its conductor f or equivalently its discriminant. As a matter of fact, $\mathcal{O}$ has finite index in $\mathcal{O}_K$ equal to f and $\mathcal{O} = \mathbb{Z} + f\mathcal{O}_K$. The discriminant of $\mathcal{O}$ is simply $D = f^2 d_K$. Remember also that if $\mathcal{O}_1$ and $\mathcal{O}_2$ are two orders in K of respective discriminants D_1 and D_2, then $\mathcal{O}_1 \subseteq \mathcal{O}_2$ iff there exists a positive integer k such that $D_1 = k^2 D_2$.

The main focus of the article is the relationship between three orders in K related to a given elliptic curve E: $\mathcal{O}_K$, the order $\mathbb{Z}[\pi]$ generated by the Frobenius map π and the endomorphism ring $\mathrm{End}(E)$ of E. These orders are such that $\mathbb{Z}[\pi] \subseteq \mathrm{End}(E) \subseteq \mathcal{O}_K$ or equivalently, $[\mathcal{O}_K : \mathcal{O}] = f$, $[\mathcal{O} : \mathbb{Z}[\pi]] = g$ et $[\mathcal{O}_K : \mathbb{Z}[\pi]] = g/f$.

In his thesis [10], Kohel computes $\mathrm{End}(E)$ starting from the known value of $d_\pi = t^2 - 4q = g^2 d_K$, where t was computed using a polynomial algorithm for point counting [11,13,12,8]. In our case, we deduce from Kohel's work a structure that describes the relations between isogenous curves and their endomorphism rings.

Let us fix the notations that will be used in the rest of the paper. Let $E/\mathbb{F}_q$ be an ordinary elliptic curve and j its j-invariant. Let $\mathcal{O}$ be the endomorphism ring of E, D its discriminant and f its conductor. Let ℓ be a prime different from p.

2.2 Kohel's Theorem

The following proposition justifies the use of ℓ-isogenies of an elliptic curve to determine its endomorphism ring $\mathcal{O}$ (and overall its conductor f).

Proposition 2.1. *[10, Proposition 21] Let $\alpha : E \rightarrow E'$ be an isogeny of prime degree ℓ. Then $\mathcal{O}$ contains $\mathcal{O}'$ or $\mathcal{O}'$ contains $\mathcal{O}$ in K and the index of one in the other divides ℓ.*

This is equivalent to saying $[\mathcal{O} : \mathcal{O}'] = 1$, ℓ or $\frac{1}{\ell}$. We will use the following language when speaking about ℓ-isogenies. A "descending" ℓ-isogeny, denoted by ↓, is an ℓ-isogeny $\alpha : E_1 \rightarrow E_2$ such that $[\mathcal{O}_1 : \mathcal{O}_2] = \ell$ whilst an "ascending" ℓ-isogeny, denoted by ↑, is an ℓ-isogeny $\alpha : E_1 \rightarrow E_2$ such that $[\mathcal{O}_2 : \mathcal{O}_1] = \ell$. In the case where the endomorphim ring is preserved we say that we have an "horizontal" ℓ-isogeny, denoted by →.

Theorem 2.1. *[10, Proposition 23] Table 1 classifies the possibilities for the rational ℓ-isogenies of E defined over $\mathbb{F}_q$.*

Table 1. Number and type of the ℓ-isogenies depending on $[\mathcal{O}_K : \mathcal{O}]$ and $[\mathcal{O} : \mathbb{Z}[\pi]]$.

Case		Number and type	Total number
$\ell \nmid [\mathcal{O}_K : \mathcal{O}]$	$\ell \nmid [\mathcal{O} : \mathbb{Z}[\pi]]$	$1 + \left(\frac{D}{\ell}\right) \ \rightarrow$	$1 + \left(\frac{D}{\ell}\right)$
	$\ell \mid [\mathcal{O} : \mathbb{Z}[\pi]]$	$\left\{\begin{array}{l} 1 + \left(\frac{D}{\ell}\right) \ \rightarrow \\ \ell - \left(\frac{D}{\ell}\right) \ \downarrow \end{array}\right.$	$\ell + 1$
$\ell \mid [\mathcal{O}_K : \mathcal{O}]$	$\ell \nmid [\mathcal{O} : \mathbb{Z}[\pi]]$	$1 \ \uparrow$	1
	$\ell \mid [\mathcal{O} : \mathbb{Z}[\pi]]$	$\left\{\begin{array}{l} 1 \ \uparrow \\ \ell \ \downarrow \end{array}\right.$	$\ell + 1$

2.3 Some Lemmas about the Classification of ℓ-Isogenies

Table 1 gives the keys to understand how the endomorphism rings of isogenous curves are related. We first deduce from these results the relation between an ℓ-isogeny α and its dual denoted by $\hat{\alpha}$.

Lemma 2.1. *Let $\alpha : E \rightarrow E'$ be an ℓ-isogeny and $\hat{\alpha}$ its dual. Then α is an ascending ℓ-isogeny iff $\hat{\alpha}$ is a descending ℓ-isogeny and α is an horizontal ℓ-isogeny iff $\hat{\alpha}$ is an horizontal ℓ-isogeny.*

From these results, we can deduce some properties of the endomorphism rings $\mathcal{O}$ and $\mathcal{O}'$ such that $\alpha : E \rightarrow E'$ is an ℓ-isogeny. With respect to ℓ, we distinguish two cases for the endomorphism rings: the case $\mathbb{Z}[\pi]$ maximal at ℓ, i.e. $\ell \nmid [\mathcal{O}_K : \mathbb{Z}[\pi]]$ or not.

The following lemma ensures that if $\mathbb{Z}[\pi]$ maximal at ℓ, we can only find horizontal ℓ-isogenies.

Lemma 2.2. *Let E be an elliptic curve such that $\mathbb{Z}[\pi]$ is maximal at ℓ. If there exists an ℓ-isogeny of E, then this ℓ-isogeny is an horizontal ℓ-isogeny.*

We suppose now that $\mathbb{Z}[\pi]$ is non-maximal at ℓ.

Lemma 2.3. *[7] If $\ell \mid [\mathcal{O}_K : \mathbb{Z}[\pi]]$ and $\ell \nmid [\mathcal{O} : \mathbb{Z}[\pi]]$, i.e. if $\ell^n \parallel g$ with $n \geq 1$ then $\ell^n \parallel f$, then the only ℓ-isogeny $\alpha : E \to E'$ is such that $\ell \mid [\mathcal{O}' : \mathbb{Z}[\pi]]$, i.e. $\ell^{n-1} \parallel f'$.*

Lemma 2.4. *[7] If $\alpha : E_1 \to E_2$ is a descending ℓ-isogeny and $\ell \mid [\mathcal{O}_2 : \mathbb{Z}[\pi]]$, then for every $\beta : E_2 \to E_3$ such that $\mathcal{O}_3 \not\simeq \mathcal{O}_1$, β is a descending ℓ-isogeny. Moreover, there are ℓ such ℓ-isogenies.*

In other words, if $\beta \neq \hat{\alpha}$, then β is a descending ℓ-isogeny. Since E_2 has $\ell + 1$ ℓ-isogenies, $\hat{\alpha}$ is an ascending ℓ-isogeny and the ℓ others are descending ℓ-isogenies.

Let us now describe a very particular case.

Lemma 2.5. *[7] If there exist two ℓ-isogenies different up to isomorphism from a curve E to a curve E', then they are both horizontal ℓ-isogenies. We can also conclude that ℓ splits in $\mathcal{O}$.*

This peculiar case gives us some informations about the imaginary quadratic field the endomorphism ring is in.

Theorem 2.2. *[13] Suppose there are two ℓ-isogenies α and β distinct up to isomorphism from E to the same curve E'. Then the discriminant D of the endomorphism ring of E is such that $|D| \leq 4\ell^2$.*

This set of lemmas gives us an idea of the graph of ℓ-isogenies of the elliptic curves having the same Frobenius map. It has a structure of a volcano truncated at the level of $\mathbb{Z}[\pi]$. The crater comes from the horizontal ℓ-isogenies (if they exist) that we can find when $\mathcal{O}$ is maximal at ℓ using Table 1 and the rest of the volcanic structure comes from the fact that by Lemmas 2.3 and 2.4, we see that if $\ell \mid [\mathcal{O}_K : \mathcal{O}]$ then E does not have any horizontal ℓ-isogeny. Figure 1 summarizes these ideas.

The level of an elliptic curve in the volcano is the ℓ-adic valuation of its conductor. The height of the volcano is equal to the level of a curve with endomorphism ring isomorphic to $\mathbb{Z}[\pi]$ locally at ℓ.

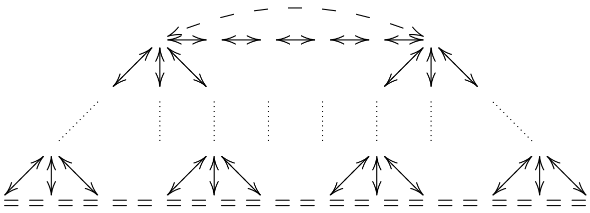


Fig. 1. Isogeny volcano

3 Modular Equations and Isogenies

We remind the reader that there exists a bivariate polynomial $\Phi_\ell(X,Y)$ with integer coefficients with the following property. Two elliptic curves E and E' defined over $\mathbb{F}_q$, are related via a cyclic isogeny α of degree ℓ if and only if $\#E = \#E'$ and $\Phi_\ell(j(E), j(E')) = 0$.

To find the curves related to E via an ℓ-isogeny, we must solve the equation $\Phi_\ell(X, j(E)) = 0$, which gives us their potential invariants. Suppose j^* is one of these roots. The curve E^* we are looking for is known up to twist and we must find an equation for it. Formulas for computing an equation of E^* are given in [13]. These formulas do not work in the case where j or j^* are in $\{0, 1728\}$ or $\partial\Phi_\ell/\partial X(j, j^*) = \partial\Phi_\ell/\partial Y(j, j^*) = 0$. We will call such a curve a *special* curve (or having a special endomorphism ring) and have a procedure detecting this, which is costless, since testing whether $\partial\Phi_\ell/\partial X(j, j^*) = \partial\Phi_\ell/\partial Y(j, j^*) = 0$ costs one polynomial gcd.

We will suppose that we have a procedure IsogenousCurves(E, ℓ) that gives us the list of curves that are ℓ-isogenous to a given curve E when E is not special.

4 Our Algorithm

Let ℓ be a prime number different from p and $\mathcal{N}_\ell(E)$ denote the number of roots of $\Phi_\ell(X, j(E))$ in $\mathbb{F}_q$. Depending on $\mathcal{N}_\ell(E)$, we can determine some properties of $\mathrm{End}(E)$ using Table 1. We summarize them in Table 2.

Table 2. Properties of $\mathcal{O}$ depending on the number and type of the ℓ-isogenies of E.

$\mathcal{N}_\ell(E)$	Type of the ℓ-isogenies		$\left(\frac{D}{\ell}\right)$	$\left(\frac{d_\pi}{\ell}\right)$
0	none	$\ell \nmid [\mathcal{O}_K : \mathcal{O}]$ and $\ell \nmid [\mathcal{O} : \mathbb{Z}[\pi]]$	-1	-1
2	$\rightarrow$	$\ell \nmid [\mathcal{O}_K : \mathcal{O}]$ and $\ell \nmid [\mathcal{O} : \mathbb{Z}[\pi]]$	$+1$	$+1$
1	case 1: $\rightarrow$	$\ell \nmid [\mathcal{O} : \mathbb{Z}[\pi]]$ and $\ell \nmid [\mathcal{O}_K : \mathcal{O}]$	0	0
	case 2: $\uparrow$	$\ell \nmid [\mathcal{O} : \mathbb{Z}[\pi]]$ and $\ell \mid [\mathcal{O}_K : \mathcal{O}]$	0	0
$\ell + 1$	case 1': $\begin{cases} 1 + \left(\frac{D}{\ell}\right) & \rightarrow \\ \ell - \left(\frac{D}{\ell}\right) & \downarrow \end{cases}$	$\ell \mid [\mathcal{O} : \mathbb{Z}[\pi]]$ and $\ell \nmid [\mathcal{O}_K : \mathcal{O}]$	nothing known	0
	case 2': $\begin{cases} 1 & \uparrow \\ \ell & \downarrow \end{cases}$	$\ell \mid [\mathcal{O} : \mathbb{Z}[\pi]]$ and $\ell \mid [\mathcal{O}_K : \mathcal{O}]$	0	0

Kohel [10] uses this approach as one of his methods to compute the endomorphism ring of the elliptic curve E. We use it to compute isogeny volcanoes.

4.1 Goal of the Algorithm

Let $\mathcal{E}$ be a given ordinary elliptic curve defined over a finite field $\mathbb{F}_q$ and $j(\mathcal{E})$ its j-invariant. Let ℓ be a prime different from p. Starting from $\mathcal{E}$, we want to construct a partial isogeny volcano, that is we want to determine the type of the crater of the isogeny volcano and determine a part of the volcano containing $\mathcal{E}$, plus a set of isogenous curves to $\mathcal{E}$ containing a curve with endomorphism ring isomorphic to $\mathbb{Z}[\pi]$ locally at ℓ and one with endomorphism ring isomorphic to $\mathcal{O}_K$ locally at ℓ.

We first give the skeleton of the algorithm and then detail every step.

4.2 Skeleton of the Algorithm

The algorithm is divided into two parts. First, we determine whether $\mathbb{Z}[\pi]$ is maximal at ℓ or not. If not, then we look for a curve E_s in the crater of the isogeny volcano (Figure 1), determine the type of the crater by determining $\epsilon = \left(\frac{d_K}{\ell}\right)$ and then find the height of the volcano using what we call *a full descending path.* Since special curves need a careful treatment, we signal these with an EXIT statement, so as to ligthen the exposition.

Procedure ComputePartialVolcano
Input: An elliptic curve $\mathcal{E}$ and a prime ℓ, $\ell \neq p$.
Output: $\epsilon = \left(\frac{d_K}{\ell}\right)$ and a list $\mathcal{F}$ of full descending paths of the volcano.

```
1. IF $\mathcal{E}$ is special THEN EXIT;
2. $F \leftarrow$IsogenousCurves$(\mathcal{E}, \ell)$;
3. IF $\#F = 0$ THEN $\{\epsilon \leftarrow -1;\ \mathcal{F} \leftarrow \{\mathcal{E}\};$ GOTO 5$\}$
   ELIF $\#F = 2$ THEN $\{\epsilon \leftarrow +1;\ \mathcal{F} \leftarrow \{\mathcal{E}\};$ GOTO 5$\}$
   ELIF $\#F = 1$ THEN
     – $E' \leftarrow F[1]$;
     – IF $E'$ is special THEN EXIT;
       ELIF $\mathcal{N}_\ell(E') = 1$ THEN $\{\epsilon \leftarrow 0;\ \mathcal{F} \leftarrow \{\mathcal{E}\};$ GOTO 5$\}$
       ELSE GOTO 4;
   ELIF $\#F = \ell + 1$ THEN GOTO 4;
4. $(E_s, P, \epsilon, n, \mathcal{F}) \leftarrow$FindFullDescendingPaths$(\mathcal{E}, \ell)$.
5. RETURN $(\epsilon, \mathcal{F})$.
```

4.3 Special Curves

If our original curve $\mathcal{E}$ has its j-invariant equal to 0 or 1728, then we cannot build any part of the volcano. We do not know how to distinguish the curves that are isogenous to $\mathcal{E}$ over $\mathbb{F}_q$ from the ones which are only isogenous to $\mathcal{E}$ over the algebraic closure of $\mathbb{F}_q$. If we encounter such a curve during the construction of the volcano, we know that this curve is in the crater of the volcano and we can deduce from this a full descending path and ϵ. But we will not be able to construct the whole volcano.

If at any moment in the construction, we encounter a curve E having two distinct ℓ-isogenies to a curve E', then we deduce that E is in the crater and the type of the crater. We will not be able to construct the entire volcano since we do not have the equation of E' but we can still get the complete subtree below E and therefore a full descending path.

4.4 The Case $\mathcal{N}_\ell(\mathcal{E}) \neq \ell + 1$

• $\mathcal{N}_\ell(\mathcal{E}) = 0$**:** In this case, if we refer to Table 2, we see that there is no ℓ-isogeny from $\mathcal{E}$ to another elliptic curve and that ℓ is inert in $\mathbb{Z}[\pi]$. We can also deduce that $\mathcal{O}_{K_\ell} \simeq \mathrm{End}(\mathcal{E})_\ell \simeq \mathbb{Z}[\pi]_\ell$.
• $\mathcal{N}_\ell(\mathcal{E}) = 2$**:** Referring to Table 2, we see that ℓ splits in $\mathbb{Z}[\pi]$. This case has already been treated by Couveignes, Dewaghe and Morain ([4], [3]). Using Lemma 2.2, we know that for every elliptic curve E' such that $\alpha : \mathcal{E} \to E'$ with α ℓ-isogeny then $\mathcal{O}' \simeq \mathrm{End}(\mathcal{E})$. We can also deduce that $\mathcal{O}_{K_\ell} \simeq \mathrm{End}(\mathcal{E})_\ell \simeq \mathbb{Z}[\pi]_\ell$.
• $\mathcal{N}_\ell(\mathcal{E}) = 1$**:** In this case, ℓ ramifies in $\mathbb{Z}[\pi]$. In Table 2, we see that this is a dual case. By dual, we mean that we may be in a case where $\mathbb{Z}[\pi]$ is maximal at ℓ or not. We need to distinguish those two cases. In order to do so, we will need its isogenous curve E' and $\mathcal{N}_\ell(E')$.

Case 1: $\mathcal{N}_\ell(E') = 1$**.** Suppose that $\mathbb{Z}[\pi]$ is not maximal at ℓ. Referring to Table 2, we know that $\ell \nmid [\mathrm{End}(\mathcal{E}) : \mathbb{Z}[\pi]]$, $\ell \mid [\mathcal{O}_K : \mathcal{O}]$ and the ℓ-isogeny $\alpha : \mathcal{E} \to E'$ is an ascending ℓ-isogeny. Therefore applying Lemma 2.3, we have $\ell \mid [\mathcal{O}' : \mathbb{Z}[\pi]]$. Thus, referring to Table 1, $\mathcal{N}_\ell(E') = \ell + 1$, which contradicts what we first found for $\mathcal{N}_\ell(E')$. Therefore, $\mathbb{Z}[\pi]$ is maximal at ℓ.

Case 2: $\mathcal{N}_\ell(E') = \ell + 1$**.** Suppose that $\mathbb{Z}[\pi]$ is maximal at ℓ, i.e. $\ell \nmid [\mathrm{End}(\mathcal{E}) : \mathbb{Z}[\pi]]$ and $\ell \nmid [\mathcal{O}_K : \mathrm{End}(\mathcal{E})]$. Referring to Table 2, we know that the ℓ-isogeny $\alpha : \mathcal{E} \to E'$ is an horizontal ℓ-isogeny and $(D_\mathcal{E}/\ell) = 0$. Therefore $\mathcal{O}'$ has the same conductor as $\mathrm{End}(\mathcal{E})$, i.e. $\ell \nmid [\mathcal{O}' : \mathbb{Z}[\pi]]$, $\ell \nmid [\mathcal{O}_K : \mathcal{O}']$ and $(D'/\ell) = 0$. Referring to Table 1, we see that $\mathcal{N}_\ell(E') = 1 + \left(\frac{D'}{\ell}\right) = 1$ which contradicts the result we first found for $\mathcal{N}_\ell(E')$. Therefore, $\mathbb{Z}[\pi]$ is not maximal at ℓ.

In this case, we can already make some conclusion about $\mathcal{O}$: $\mathcal{O}_{K_\ell} \not\simeq \mathrm{End}(\mathcal{E})_\ell$ and $\mathrm{End}(\mathcal{E})_\ell \simeq \mathbb{Z}[\pi]_\ell$, i.e. there exists an $n > 1$ such that $\ell^n \parallel g$ and $\ell^n \parallel f$.

4.5 The General Case $\mathcal{N}_\ell(\mathcal{E}) = \ell + 1$

By looking at the skeleton of the algorithm in Section 4.2, we see that this case is the most interesting one.

From now on, we assume that E is of level r, $r \in \mathbb{N}$, and $\mathcal{N}_\ell(E)$ equals $\ell + 1$. In fact, we have the equality $\mathcal{N}_\ell(E_i) = \ell + 1$ until we find the ending point of our recurrence that we recognize by $\mathcal{N}_\ell(E_i) = 1$.

This part of the algorithm is based on finding an elliptic curve E_s such that E_s is in the crater, using *descending paths*. First we precise this notion.

Descending paths.

Definition 4.1. *A* descending path *of an elliptic curve E is a path $E = E_0 \to E_1 \to E_2 \to \cdots \to E_{m-1} \to E_m$ of elliptic curves such that the map $E_i \to E_{i+1}$, for $i \in [0, \ldots, m[$, is a descending ℓ-isogeny and $\ell \nmid [\mathcal{O}_m : \mathbb{Z}[\pi]]$. We will say that we have a* full descending path *if E is in the crater of the volcano.*

Lemma 4.1. *With the notations of Definition 4.1, if E is of level r then E_i is of level $r + i$.*

Proof: We prove this lemma by induction. $E_0 = E$ is of level r. Let us suppose that the result is true for E_j, with $0 \leq j < m$. We know that the map $E_j \to E_{j+1}$ is a descending ℓ-isogeny. Therefore, since the level of E_j is $r+j$, i.e. $\ell^{r+j} \parallel [\mathcal{O}_K : \mathcal{O}_j]$ and by definition of a descending ℓ-isogeny, then $\ell^{r+j+1} \parallel [\mathcal{O}_K : \mathcal{O}_{j+1}]$. Thus E_{j+1} is of level $r + (j + 1)$. □

The main goal of finding a descending path starting from an elliptic curve E is to locate the endomorphism ring of E in the volcanic structure (see Figure 1) with respect to $\mathbb{Z}[\pi]$.

Corollary 4.1. *Let $\mathcal{P}$ be a descending path starting from E and let $m = \#\mathcal{P}-1$. Then E is of level $(n - m)$ where n is the height of the volcano.*

Now that we have defined this notion and its interest, we will show how to compute a descending path. We first give the algorithm and then prove its correctness.

Procedure FindDescendingPath
Input: A non special elliptic curve E such that $\ell \mid [\mathcal{O}_K : \mathbb{Z}[\pi]]$.
Output: A descending path starting from E.

```
1. $F \leftarrow$IsogenousCurves$(E, \ell)$;
2. IF $\#F = 1$ THEN $\{P[1] \leftarrow \{E\}$; $i_0 \leftarrow 1$; GOTO 6$\}$;
3. FOR $i := 1$ TO 3 DO
   (a) $P[i] \leftarrow \{E\} \cup \{F[i]\}$; $G[i] \leftarrow E$; $G'[i] \leftarrow F[i]$;
   (b) IF $G'[i]$ is special THEN $S[i] \leftarrow \emptyset$
         ELSE $S[i] \leftarrow$IsogenousCurves$(G'[i], \ell)$;
4. $i_0 \leftarrow -1$
5. WHILE $(i_0 = -1)$ DO
      FOR $i := 1$ TO 3 DO (at this point, $G'[i]$ is one of the curves isogenous
      to $G[i]$ and $S[i]$ contains a list of curves isogenous to $G'[i]$)
         IF $S[i] = \emptyset$ THEN use next $i$;
         IF $\#S[i] = 1$ THEN $\{i_0 \leftarrow i$; (we have found the base of the volcano)$\}$
         ELSE
         (a) IF $(j(S[i][1]) = j(G[i]))$ THEN $\{$(we must not use the dual of
               the preceding isogeny) $G[i] \leftarrow G'[i]$; $G'[i] \leftarrow S[i][2]$;$\}$;
               ELSE $\{G[i] \leftarrow G'[i]$; $G'[i] \leftarrow S[i][1]$;$\}$;
         (b) $P[i] \leftarrow P[i] \cup \{G'[i]\}$;
         (c) IF $G'[i]$ is special THEN $S[i] \leftarrow \emptyset$
               ELSE $S[i] \leftarrow$IsogenousCurves$(G'[i], \ell)$;
6. RETURN $P[i_0]$.
```

By Lemma 2.4, we know that whenever we have an ℓ-isogeny $\alpha : E \to E'$ that is a descending ℓ-isogeny, every ℓ-isogeny $\beta : E' \to E''$ such that $\mathrm{End}(E'') \not\simeq \mathrm{End}(E)$ is a descending ℓ-isogeny. Therefore, inductively, if we start a path of ℓ-isogenies with a descending ℓ-isogeny, we will get a descending path.

To find such an ℓ-isogeny to start the path, we will compute in parallel three different paths starting from any three different curves isogenous to E.

Having three different starting curves ensures us of having a path starting with a descending ℓ-isogeny and therefore a non-empty path.

Since a non-descending path is composed of a path of non-descending ℓ-isogenies and a descending path, a non-descending path is longer than a descending path. Therefore, the first path that stops is a descending path.

Lemma 4.2. *The time complexity of the algorithm* FindDescendingPath *is $O(m\mathcal{F}(\ell))$, where m is the height of E and $\mathcal{F}(\ell)$ the time to find three roots of a modular polynomial.*

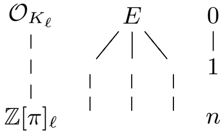
Proof: To calculate each one of the three paths, it takes $m+1$ partial factorizations of the modular equation. □

Why do we need a curve in the crater? If we have a curve E_s in the crater and a full descending path $E_s \to E_1 \to E_2 \to \cdots \to E_{m-1} \to E_m$, we get the height of the volcano and then using the algorithms that are given to find a partial volcano, we can move easily in the volcano and construct the rest of it if we want. To find such a curve E_s we need to know how to recognize that a curve is in the crater.

Detecting the crater and thus determining ϵ. From Table 2, we see that a curve in the crater has $1 + \left(\frac{D}{\ell}\right)$ horizontal ℓ-isogenies and $\ell - \left(\frac{D}{\ell}\right)$ descending ℓ-isogenies. We detect these three different cases in three different ways.

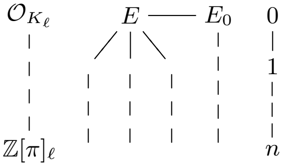
Suppose E is in the crater and let n be the height of the volcano. Then one of the following conditions will be met.

- Case a: There is no horizontal ℓ-isogeny. Considering the fact that we are in the crater, we have $\ell + 1$ descending ℓ-isogenies. Then all the descending paths starting from the $\ell+1$ isogenous curves to E have the same length. The following graph characterizes this situation.

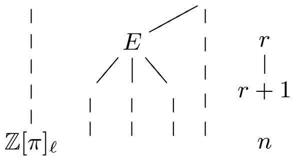


The length of the descending paths is $n-1$ because all the curves corresponding to the $\ell+1$ roots of $\Phi_\ell(X, j)$ are at level 1. We can also deduce that ℓ is inert in $\mathcal{O}_K$ and thus $\epsilon = -1$.

- Case b: There is exactly one horizontal ℓ-isogeny and there are also ℓ descending ℓ-isogenies. Then one of the descending paths starting from the $\ell+1$ isogenous curves to E is of length n (let us say that this path starts on E_0) and the other ℓ ones are of length $(n-1)$. The following graph characterizes this situation and makes the parallel with the normal situation.



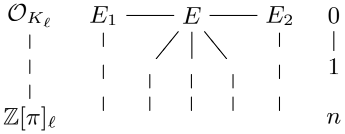
Horizontal case



“Normal” case

We cannot confuse this case with the "normal" case of one ascending ℓ-isogeny and ℓ descending ℓ-isogenies, because in the horizontal case, the difference between the length of the path starting on E_0 and the other paths is 1 whereas in the "normal" case this difference is 2. We know also that ℓ ramifies in $\mathcal{O}_K$ and therefore $\epsilon = 0$.

- Case c: There are two horizontal ℓ-isogenies and there are also $\ell - 1$ descending ℓ-isogenies. Then two of the descending paths starting from the $\ell + 1$ isogenous curves to E are of length n (let us say that these two paths start on E_1 and E_2) and the other $\ell - 1$ ones are of length $n - 1$. The following graph characterizes this situation.



The difference with the preceding case is that we find two paths longer than the others instead of just one. So no confusion with the "normal" case is possible. We know also that ℓ splits in $\mathcal{O}_K$ and therefore $\epsilon = +1$.

How to find a curve in the crater. The algorithm finding a curve in the crater is exactly the inverse of the one finding a descending path. We want to construct an *ascending path* starting from $\mathcal{E}$.

Definition 4.2. *An* ascending path *of an elliptic curve E is a path $E = E_0 \rightarrow E_{-1} \rightarrow E_{-2} \rightarrow \cdots \rightarrow E_{-(s-1)} \rightarrow E_{-s}$ of elliptic curves such that the map $E_{-i} \rightarrow E_{-(i+1)}$, for $i \in [0, \ldots, s-1[$, is an ascending ℓ-isogeny and $\ell \nmid [\mathcal{O}_K : \mathcal{O}_{-s}]$.*

We will say that we have a full ascending path *if $\mathcal{O}_\ell \simeq \mathbb{Z}[\pi]_\ell$.*

Lemma 4.3. *Using the same notations as in Definition 4.2, if E is of level r then E_{-i} is of level $r - i$.*

Corollary 4.2. *If the length of the ascending path starting on E is $r + 1$, then E is at level r.*

At every step of this algorithm, we want to find a curve at an inferior level than E i.e. the unique ascending ℓ-isogeny of E. To do so, we will compute a descending path for every curve isogenous to E and compare their sizes. We reiterate this until we detect a curve in the crater.

Procedure DetectSurface
Input: A list of descending paths $\mathfrak{P}$ and the curve E_{cur}.
Output: $(\epsilon, i_{max}, \lambda, \mathcal{F})$ such that

- $\epsilon = 0$, i_{max} such that $\#\mathfrak{P}[i_{max}]$ is maximal and $\lambda = \#\mathfrak{P}[i_{max}]$
- OR $\epsilon = (d_K/\ell)$, $i_{max} = -1$ and λ is the height of the volcano if we detect that E_{cur} is in the crater;
- $\mathcal{F}$ is a list of (some) full descending paths.

```
1. $\epsilon \leftarrow 0$; $\mathcal{F} \leftarrow \emptyset$;
2. Find $i_{max}$ such that $\#\mathfrak{P}[i]$ is maximal;
```

```
3. $I \leftarrow \{i$ s.t. $i \neq i_{max}$ and $\#\mathfrak{P}[i] = \#\mathfrak{P}[i_{max}]\}$;
4. /* Case where the crater is detected and $\left(\frac{d_K}{\ell}\right) = -1$ (case a) */
   IF $\#I = \ell$ THEN $\{\epsilon \leftarrow -1;$
   $\lambda \leftarrow \#\mathfrak{P}[i_{max}]$; $i_{max} \leftarrow -1$; $\mathcal{F} \leftarrow \{\{E_{cur}, \mathfrak{P}[1]\}\}$; $\}$
5. /* Case where the crater is detected and $\left(\frac{d_K}{\ell}\right) = +1$ (case c)*/
   IF $\#I = 1$ THEN $\{$ $i_{max2} \leftarrow I[1]$; $\epsilon \leftarrow 1$; $\lambda \leftarrow \#\mathfrak{P}[i_{max}] - 1$; $i_0 \leftarrow$ any
   index distinct from $i_{max}$ and $i_{max2}$; $\mathcal{F} \leftarrow \{\{E_{cur}, \mathfrak{P}[i_0]\}, \mathfrak{P}[i_{max}], \mathfrak{P}[i_{max2}]\}$;
   $i_{max} \leftarrow -1$; $\}$
6. IF $\#I = 0$ THEN
   (a) IF $i_{max} = 1$ THEN $i_0 \leftarrow 2$; ELSE $i_0 \leftarrow 1$;
   (b) IF $\#\mathfrak{P}[i_{max}] - \#\mathfrak{P}[i_0] = 1$ /* Case where the crater is detected and
       $\left(\frac{d_K}{\ell}\right) = 0$ (case b) */
       THEN $\{\epsilon \leftarrow 0$; $\lambda \leftarrow \#\mathfrak{P}[i_{max}]-1$; $\mathcal{F} \leftarrow \{\{E_{cur}, \mathfrak{P}[i_0]\}, \mathfrak{P}[i_{max}]\}$; $i_{max} \leftarrow$
       $-1$; $\}$
       ELSE $\{\lambda \leftarrow \#\mathfrak{P}[i_{max}] - 1;\}$
7. RETURN $(\epsilon, i_{max}, \lambda, \mathcal{F})$.
```

Procedure FindFullDescendingPaths
Input: A non-special elliptic curve E such that $\ell \mid [\mathcal{O}_K : \mathbb{Z}[\pi]]$.
Output: $(E_s, P, \epsilon, n, \mathcal{F})$ such that E_s is in the crater, isogenous to E, P is an ascendin path from E to E_s, $\epsilon = (d_K/\ell)$, n the height of the volcano and $\mathcal{F}$ is a list of (some) full descending paths.

```
1. $E_{cur} \leftarrow E$;
2. $F \leftarrow$IsogenousCurves$(E_{cur}, \ell)$;
3. $P \leftarrow \{E_{cur}\}$;
4. IF $\#F = 1$ THEN $\{E_{cur} \leftarrow F[1]$; IF $E_{cur}$ is special THEN EXIT; ELSE
   $\{P \leftarrow P \cup \{F[1]\};\}\}$
5. $i_0 \leftarrow 0$;
6. WHILE $i_0 \neq -1$ DO
   (a) $F \leftarrow$IsogenousCurves$(E_{cur}, \ell)$;
   (b) FOR $i := 1$ TO $\ell + 1$ DO
       IF $F[i]$ is special THEN EXIT;
       $\mathfrak{P}[i] \leftarrow$ FindDescendingPath$(F[i])$;
   (c) $(\epsilon, i_0, \lambda, \mathcal{F}) \leftarrow$DetectSurface$(\mathfrak{P})$;
   (d) IF $i_0 \neq -1$ THEN $\{E_{cur} \leftarrow F[i_0]$; $P \leftarrow P \cup \{E_{cur}\};\}$
7. $E_s \leftarrow E_{cur}$;
8. RETURN $(E_s, P, \epsilon, \lambda, \mathcal{F})$;
```

Lemma 4.4. *The complexity of the algorithm* FindFullDescendingPaths *is* $O(n^2 \ell \mathcal{F}(\ell))$*, with* $\mathcal{F}(\ell)$ *the time to calculate all the roots of a modular polynomial.*

Proof: To go from level μ to level $\mu - 1$, we need to calculate $\ell + 1$ descending paths. This takes $O(\mu \ell \mathcal{F}(\ell))$ operations, for a total of $\Sigma_{\mu=1}^{n} \mu \mathcal{F}(\ell) = \frac{n(n+1)}{2} \mathcal{F}(\ell)$. Therefore it takes $O(n^2 \ell \mathcal{F}(\ell))$ operations to compute an ascending path. □

The following theorem gives the complexity of the algorithm to compute a partial volcano.

Theorem 4.1. *It takes $O(n^2\ell\mathcal{F}(\ell))$ operations to compute a partial volcano of ℓ-isogenies, with $n \leq \frac{\log_2(|d_K|)}{\log_2(\ell)}$ and $\mathcal{F}(\ell)$ the time to calculate all the roots of a modular polynomial.*

Proof: The whole algorithm is based on the computation of an ascending path starting from $\mathcal{E}$. □

5 Number of Isogeny Volcanoes

We define the endomorphism class of E denoted by $\mathcal{C}(E)$ to be a set of curves isogenous but non isomorphic having the same endomorphism ring $\mathcal{O}$. There exists a bijection between $C(\mathcal{O})$ and $\mathcal{C}(E)$. If there exists a unique ℓ-isogeny volcano then we can compute the set of $h(\mathcal{O})$ elliptic curves in $\mathcal{C}(E)$ using this volcano. Therefore we use properties of $h(\mathcal{O})$ to compute the number of ℓ-isogeny volcanoes.

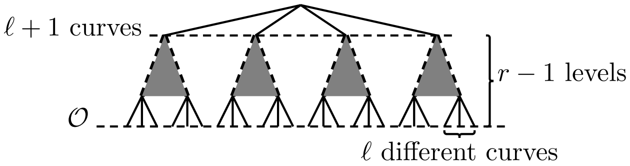
Theorem 5.1. *The number of different volcanoes of ℓ-isogenies is*

$$h({f'}^2 d_K)/\mathrm{ord}(\mathfrak{l})$$

where $\mathrm{ord}(\mathfrak{l})$ *is the order of the ideal* $\mathfrak{l}$ *which is a prime ideal of norm ℓ.*

Proof: We treat separately the different types of volcanoes.

Case where $\left(\frac{d_K}{\ell}\right) = -1$. In this situation, every ℓ-isogeny volcano is of the form:

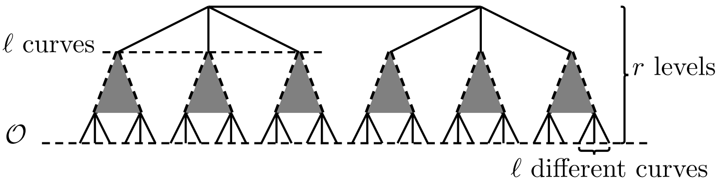


In this type of volcano we have found $\ell^r + \ell^{r-1}$ of the $h(\mathcal{O})$ curves isogenous to E having the same endomorphism ring $\mathcal{O}$. We have

$$h(m^2D) = \frac{h(D)m}{[\mathcal{O}_1^* : \mathcal{O}_2^*]} \prod_{p|m} \left(1 - \left(\frac{D}{p}\right)\frac{1}{p}\right)$$

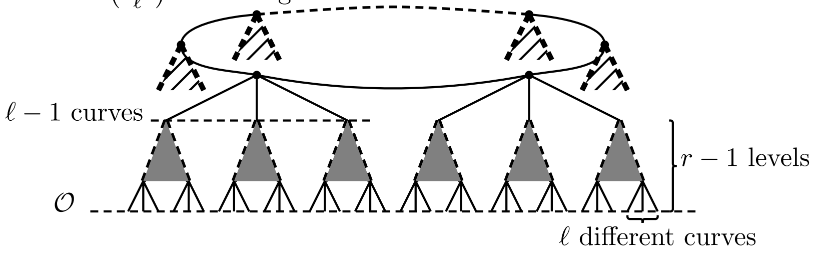
where $\mathcal{O}_1$ and $\mathcal{O}_2$ are the orders of discriminant D and m^2D ([5, Coro 7.28]) and when D is different from -4 and -3, $[\mathcal{O}_1^* : \mathcal{O}_2^*]$ is equal to 1. In our case we consider $m = \ell^r$ where r is the ℓ-adic valuation of the conductor f of $\mathcal{O}$. We set $f = f'\ell^r$. Then $h(f^2 d_K) = h({f'}^2 D)\ell^r \left(1 - \left(\frac{D}{\ell}\right)\frac{1}{\ell}\right) = h({f'}^2 D)\ell^r(1 + 1/\ell) = h({f'}^2 D)(\ell^r + \ell^{r-1})$. Then there are $h({f'}^2 D)$ distinct volcanoes of this type.

Case where $\left(\frac{d_K}{\ell}\right) = 0$. In this situation, every ℓ-isogeny volcano is of the form:



In such a volcano, we get $2\ell^r$ curves in $\mathcal{C}(E)$. In this case, it is also clear that there are $h(f'^2 D_K)/2$ distinct volcanoes (reusing the preceding notations).

Case where $\left(\frac{d_K}{\ell}\right) = 1$. We get a volcano of the form:



For each one of the graph under the crater we get $(\ell-1)\ell^{r-1}$ curves in $\mathcal{C}(E)$. We now have to determine the size of the crater. If we consider the set of the curves in the crater lifted in $\mathbb{C}$, we get the following cycle $\mathcal{E}_0 \to \mathcal{E}_1 \to \cdots \mathcal{E}_{s-1} \to \mathcal{E}_s \simeq \mathcal{E}_0$ where $\mathcal{E}_i \simeq \mathbb{C}/\mathfrak{a}_i$. Since we consider ℓ-isogenies we have $\mathfrak{a}_i = \mathfrak{a}_{i+1}\mathfrak{l}$ where $\mathfrak{l}$ is a prime ideal of norm ℓ. Therefore $\mathfrak{a}_0 = \mathfrak{a}_s = \mathfrak{l}^s\mathfrak{a}_0$ i.e. $\mathfrak{l}^s$ is a principal ideal of $\mathcal{O}_K$ and thus s is the order of $\mathfrak{l}$ in $\mathcal{O}_K$ and s is the size of the crater. Therefore the number of different volcanoes we can build is $h(f'^2 d_K)/\text{ord}(\mathfrak{l})$ where $\text{ord}(\mathfrak{l})$ is the order of the ideal $\mathfrak{l}$ which is a prime ideal of norm ℓ.

Using the type of decomposition of the ideal $\ell\mathcal{O}_K$, we can generalise this last formula to all the types of volcanoes. □

6 Application to Point Counting

First, we suppose that $\ell \neq 2$ and that we have not encountered a special curve (for these cases see [7]).

If $\mathcal{N}_\ell(\mathcal{E})$ is equal to 1 or $\ell+1$, then we can deduce that ℓ ramifies in $\mathbb{Z}[\pi]$ i.e. $\left(\frac{d_\pi}{\ell}\right) = 0$ and therefore we immediately know that $t^2 \equiv 4q \pmod{\ell}$. Our idea is to explain how a more precise result can be found, namely the ℓ-adic valuation of $t^2 - 4q$ that we note ν_ℓ. We will determine n such that $\ell^n \parallel g$, i.e. the height of the isogeny volcano, and since $t^2 - 4q = g^2 d_K$, we get $t^2 \equiv 4q \pmod{\ell^{2n+\delta}}$ and therefore $\nu_\ell \geq 2n+\delta$. The value of δ is determined by the Legendre symbol $\left(\frac{d_K}{\ell}\right)$. If it is equal to 0, then we deduce that $\ell \mid d_K$, therefore $\delta = 1$. Otherwise, $\delta = 0$. By definition of the fundamental discriminant d_K, we have in fact $\nu_\ell = 2n + \delta$ (except maybe in the case $\ell = 2$, see [7]).

6.1 Finding t mod ℓ^ν

In general (that is except in the cases where we happened to find a special case), our algorithm has given us $t^2 \equiv 4q \bmod \ell^\nu$, we may want $t \bmod \ell^\nu$. Suppose

$\ell \neq 2$. Then there are only two squareroots of $4q$ modulo ℓ^ν. To find the sign of t, it is enough to find the sign of $t_1 \equiv t \bmod \ell$. Finding t_1 is done via the determination of an eigenspace of π and the associated eigenfactor of the ℓ-th division polynomial Ψ_ℓ *à la* Elkies. This will determine the eigenvalue, which turns out to be $t_1/2 \bmod \ell$ in that case.

6.2 Finding t mod $\ell^{\nu+1}$

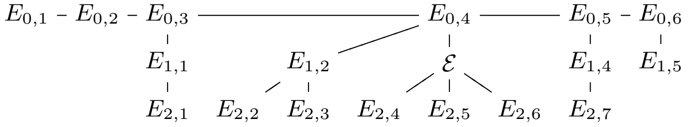
Now that we have t mod ℓ^ν, is it possible to find t mod $\ell^{\nu+1}$? When $(d_K/\ell) \neq +1$, we cannot do anything, since we already explored all possible isogenies. In the case where $(d_K/\ell) = +1$, the head of the volcano is an isogeny cycle and the ideas of [4] apply there too (see [7]).

Further applications are given in [7]. In particular, we solve a problem of Lercier encountered in [11].

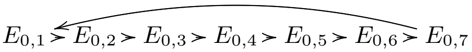
7 Numerical Examples

The reader can find a more complete set of examples in [7].

Example 1 (Normal case, ℓ splits in $\mathcal{O}_K$ i.e. $\left(\frac{d_K}{\ell}\right) = +1$): Let $p = 10009$ and $\mathcal{E} = [7478, 1649]$. The j-invariant of $\mathcal{E}$ is $j_\mathcal{E} = 83$. Using $\ell = 3$, we find

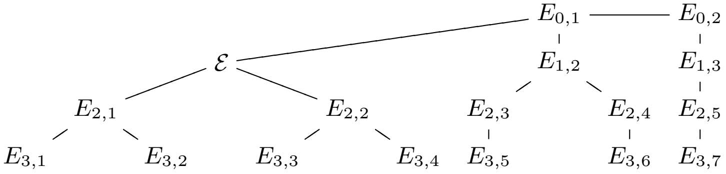


Therefore, $n = 2$, $\left(\frac{d_K}{\ell}\right) = 1$ thus $\delta = 0$ and $t^2 \equiv 4p \pmod{3^4}$ and in fact $t \equiv 34 \bmod 3^4$. Moreover, in this case, we are able to construct at the surface a cycle of isogenies. We get the following graph:



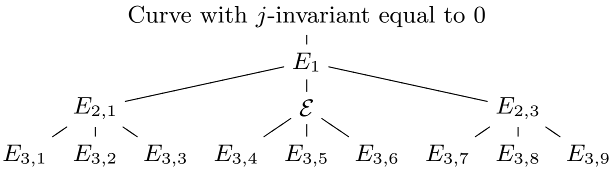
Using this cycle, we find that $t \equiv -47 \bmod 3^5$. As a matter of fact, $t = -47$.

Example 2 (Incomplete case for $\ell = 2$ from [3]): Let $p = 1009$ and $\mathcal{E} = [1, 3]$. The j-invariant of $\mathcal{E}$ is $j_\mathcal{E} = 269$. For $\ell = 2$, one gets



Therefore, $n = 3$, $\left(\frac{d_K}{\ell}\right) = 0$ thus $\delta = 2$ and $t^2 \equiv 4p \pmod{2^8}$. As a matter of fact, $t = -50$, therefore $d_K = -24$, $g = 2^3$ and $(-50)^2 \equiv 4 \times 1009 \pmod{2^9}$. In this case, we only get a lower bound of the valuation.

Example 3 (Case where the curve E_s has j-invariant equal to 0): Let $p = 1009$ and $\mathcal{E} = [363, 690]$. The j-invariant of $\mathcal{E}$ is $j_\mathcal{E} = 433$. Consider $\ell = 3$:



Therefore, $n = 3$, $\left(\frac{d_K}{\ell}\right) = 0$ thus $\delta = 1$ and $t^2 \equiv 4p \pmod{3^7}$. As a matter of fact, $t = 43$.

8 Conclusion

We have found an answer to several problems encountered while implementing various algorithms for elliptic curves over finite fields. The volcano structure is an important point of view on the isogeny class of a curve and may therefore become an important tool for that type of studies. It would be interesting to study more closely the relationships between distinct volcanoes of same prime ℓ. Another direction would be to look at volcanoes of composite indices.

Acknowledgments. We would like to thank D. Kohel for useful discussions on isogenies and for anticipating some of the results on the volcano structure. Special thanks also to P. Gaudry for useful remarks concerning this work.

References

1. Z. I. Borevitch and I. R. Chafarevitch. *Théorie des nombres.* Gauthiers-Villars, Paris, 1967.
2. J. Chao, O. Nakamura, K. Sobataka, and S. Tsujii. Construction of secure elliptic cryptosystems using CM tests and liftings. In K. Ohta and D. Pei, editors, *Advances in Cryptology – ASIACRYPT'98*, volume 1514 of *Lecture Notes in Comput. Sci.*, pages 95–109. Springer-Verlag, 1998. Beijing, China.
3. J.-M. Couveignes, L. Dewaghe, and F. Morain. Isogeny cycles and the Schoof-Elkies-Atkin algorithm. Research Report LIX/RR/96/03, LIX, April 1996. Available at `http://www.lix.polytechnique.fr/Labo/Francois.Morain/`.
4. J.-M. Couveignes and F. Morain. Schoof's algorithm and isogeny cycles. In *ANTS-I*, 1994.
5. D. H. Cox. *Primes of the Form $x^2 + ny^2$.* Wiley-Interscience, 1989.
6. M. Deuring. Die Typen der Multiplikatorenringe elliptischer Funktionenkörper. *Abh. Math. Sem. Hamburg*, 14:197–272, 1941.
7. M. Fouquet. *Anneau d'endomorphismes et cardinalité des courbes elliptiques : aspects algorithmiques.* Thèse, École polytechnique, December 2001. Available at `http://www.lix.polytechnique.fr/Labo/Mireille.Fouquet/`.
8. M. Fouquet, P. Gaudry, and R. Harley. An extension of Satoh's algorithm and its implementation. *J. Ramanujan Math. Soc.*, December 2000.
9. S.D. Galbraith, F. Hess, and N.P. Smart. Extending the GHS weil descent attack. `http://eprint.iacr.org/`, 2001.
10. D. Kohel. *Endomorphism rings of elliptic curves over finite fields.* Phd thesis, University of California, Berkeley, 1996.

11. R. Lercier. *Algorithmique des courbes elliptiques dans les corps finis.* Thèse, École polytechnique, June 1997.
12. T. Satoh. The canonical lift of an ordinary elliptic curve over a finite field and its point counting. *J. Ramanujan Math. Soc.*, 15:247–270, December 2000.
13. R. Schoof. Counting points on elliptic curves over finite fields. *J. Théor. Nombres Bordeaux*, 1995.

Fast Elliptic Curve Point Counting Using Gaussian Normal Basis

Hae Young Kim[1], Jung Youl Park[1], Jung Hee Cheon[2], Je Hong Park[1],
Jae Heon Kim[3], and Sang Geun Hahn[1]

[1] Department of Mathematics
Korea Advanced Institute of Science and Technology(KAIST)
Daejon, Republic of Korea
{hykim,jungyoul,arttex,sghahn}@mathx.kaist.ac.kr
http://crypt.kaist.ac.kr

[2] International Research Center for Information Security (IRIS)
Information and Communications University (ICU)
Daejon, Republic of Korea
jhcheon@icu.ac.kr
http://vega.icu.ac.kr/~jhcheon

[3] National Security Research Institute (NSRI)
Daejon, Republic of Korea
jaeheon@etri.re.kr

Abstract. In this paper we present an improved algorithm for counting points on elliptic curves over finite fields. It is mainly based on Satoh-Skjernaa-Taguchi algorithm [SST01], and uses a Gaussian Normal Basis (GNB) of small type $t \leq 4$. In practice, about 42% (36% for prime N) of fields in cryptographic context (i.e., for $p = 2$ and $160 < N < 600$) have such bases. They can be lifted from $\mathbb{F}_{p^N}$ to $\mathbb{Z}_{p^N}$ in a natural way. From the specific properties of GNBs, efficient multiplication and the Frobenius substitution are available. Thus a fast norm computation algorithm is derived, which runs in $O(N^{2\mu} \log N)$ with $O(N^2)$ space, where the time complexity of multiplying two n-bit objects is $O(n^{\mu})$. As a result, for all small characteristic p, we reduced the time complexity of the SST-algorithm from $O(N^{2\mu+0.5})$ to $O(N^{2\mu+\frac{1}{\mu+1}})$ and the space complexity still fits in $O(N^2)$. Our approach is expected to be applicable to the AGM since the exhibited improvement is not restricted to only [SST01].

Keywords: elliptic curve, Gaussian normal basis, order counting

1 Introduction

Elliptic curve cryptography was independently proposed by Koblitz [Kob87] and Miller [Mil87] in 1985. Because it runs with a smaller key size than an RSA-type cryptosystem, it is possible to implement a fast and compact cryptosystem. As a result a vast amount of research has been done on its secure and efficient implementations. One of the important issues on studying elliptic curve cryptosystems is to count the number of points on an elliptic curve E over a finite

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 292–307, 2002.

field $\mathbb{F}_q$ with $q = p^n$. In 1985, Schoof [Sch85] gave the first polynomial-time algorithm whose complexity is $O(\log^{3\mu+2} q)$. Later, Elkies [Elk98] and Atkin [Art92] improved this to so-called Schoof-Elkies-Atkin (SEA) algorithm with running time $O(\log^{2\mu+2} q)$ for a large characteristic. SEA-algorithm was extended to small characteristics by Couveignes [Cou96]. In 2000, Satoh [Sat00] proposed an algorithm running in $O(N^{2\mu+1})$ time and $O(N^3)$ space for the small characteristic $p \geq 5$. Fouquet, Gaudry and Harley [FGH00] extended Satoh's algorithm for the cases $p = 2, 3$. Skjernaa [Skj00] independently extended it for the case $p = 2$. In 2001, Vercauteren, Preneel and Vandewalle [VPV01] presented a modified memory-efficient version of the algorithm whose space complexity fell to $O(N^2)$. The most recent counting algorithm, suggested by Satoh, Skjernaa and Taguchi [SST01], uses the Frobenius substitution to reduce the number of arithmetic operations over a p-adic number field with full precision. This algorithm runs in $O(N^{2\mu+0.5})$ time with $O(N^2)$ memory for $p = 2$, and $O(\max\{N^{\mu+2}, N^{2\mu+0.5}\} \log N)$ time with $O(N^{2.5})$ memory for $p \geq 3$. Harley, Mestre and Gaudry [HMG01] announced a totally different algorithm, based on the AGM (arithmetic geometric mean) iteration with a fast norm algorithm, which, as far as the authors know, has not been published yet.

Our contribution is the improvement of Satoh-Skjernaa-Taguchi (SST) algorithm. The time complexity of our algorithm is $O(N^{2\mu+\frac{1}{\mu+1}})$. We focus on a finite field with the GNB of type t. For the practical reason, we restrict $t \leq 4$. In spite of such a restriction, our cases cover about 42% (36% for prime N) of the fields in a cryptographical contexts, i.e., for $p = 2$ and $160 < N < 600$. It is known that multiplication is performed efficiently in the finite field with the GNB of small types [Sil99], [VPV01]. So we lift the GNB from the finite field to the p-adic number field in a natural way to utilize the benefits of GNBs, for the SST-algorithm mainly works over a p-adic number field. Thus a fast norm computation algorithm for the p-adic number field is derived. It runs in $O((NM)^\mu \log N)$ time with $O(NM)$ space to get precision M, while that of the SST-algorithm runs in $O((NM)^\mu M^{0.5})$ time with $O(NM)$ space. Additionally, M is about $N/2$ in point counting algorithm. As a result, for all small characteristic p, we reduced the time complexity of the SST-algorithm from $O(N^{2\mu+0.5})$ to $O(N^{2\mu+\frac{1}{\mu+1}})$ and the space complexity still fits in $O(N^2)$. As to the large-scale computation with the smallest type, our algorithm takes only about 1 day and 10 hours to count the number of points on the elliptic curve defined on $\mathbb{F}_{2^{12010}}$. Since the AGM method uses multiplication and the norm computation over p-adic field, we also expect that our methods speed up the AGM algorithm.

This paper is organized as follows: First, we set up the notation and terminology at the end of section 1, then in section 2 we briefly review Satoh-Skjernaa-Taguchi [SST01] algorithm. We introduce the notion of a Gauss period and a normal basis representation in section 3, which leads us to compute multiplication and the Frobenius substitution efficiently as described in section 4. In section 5 we present an algorithm to compute the norm with fewer operations. Followed by section 6 we describe how our algorithm can be applied to point counting. We exhibit our practical results and notes for implementation in section 7. Finally this paper ends up with some comments in section 8.

Notation. Throughout this paper, p is assumed to be a small prime. We put that q is a power of p and N is a positive integer. We denote the unramified extension of degree N of $\mathbb{Q}_p$ by $\mathbb{Q}_{p^N}$, and its valuation ring by $\mathbb{Z}_{p^N}$. In general, σ stands for the Frobenius substitution in $\mathrm{Gal}(\mathbb{Q}_{p^N}/\mathbb{Q}_p)$, and π is the reduction map by p from $\mathbb{Q}_{p^N}$ to $\mathbb{F}_{p^N}$. Given a positive integer M, an operation is said to be with precision M if it is done modulo p^M. For the rest of this paper, E is a non-supersingular elliptic curve over $\mathbb{F}_{p^N}$ and $j(E)$ denotes its j-invariant. We assume that $j(E) \in \mathbb{F}_{p^N} - \mathbb{F}_{p^2}$.

2 Satoh-Skjernaa-Taguchi Algorithm

In this section, we briefly review the SST-algorithm. We assume that $j(E) \notin \mathbb{F}_{p^2}$. Furthermore, the case $j(E) \in \mathbb{F}_{p^2}$ can be easily handled by counting points over a tiny subfield. It is well known that for T, the trace of the Frobenius endomorphism, $\#E(\mathbb{F}_{p^N}) = p^N + 1 - T$.

The canonical lift $E^{\uparrow}$ of a non-supersingular elliptic curve E from $\mathbb{F}_{p^N}$ to $\mathbb{Z}_{p^N}$ is an elliptic curve over $\mathbb{Q}_{p^N}$ which satisfies $\pi(E^{\uparrow}) = E$ and $\mathrm{End}(E) \cong \mathrm{End}(E^{\uparrow})$. Moreover, the canonical lift is unique up to isomorphism [Deu41]. Satoh [Sat00] showed that once we obtain the lifted j-invariant $j^{\uparrow}$ and the dual of the Frobenius endomorphism (Verschiebung) of $E^{\uparrow}$, we can calculate T, the trace of the Frobenius endomorphism, from the lifted data. By Hasse's theorem, we have $|T| \leq 2\sqrt{p^N}$. Therefore, it suffices to lift all the data with precision $M = N/2 + O(1)$. The SST-algorithm [SST01] is outlined as follows.

SST-Algorithm

(1) Compute the j-invariant of the canonical lift of E modulo p^M.
(2) Calculate the square of the leading coefficient, c_1, of the homomorphism induced by the lifted p-th Verschiebung on the formal group of $E^{\uparrow}$.
(3) Find an integer T satisfying $T^2 \equiv \mathrm{Norm}_{\mathbb{Q}_{p^N}/\mathbb{Q}_p}(c_1^2) \mod p^M$ and $|T| \leq 2\sqrt{p^N}$, and determine the sign of T.

2.1 Computing the Canonical Lift

To compute the $j^{\uparrow}$, the p-th modular polynomial $\Phi_p(X,Y)$ plays an important role. By a result of Lubin-Serre-Tate [LST64], the canonical lift is characterized as follows: let $j \in \mathbb{F}_{p^N} - \mathbb{F}_{p^2}$, then the solution J of $\Phi_p(\sigma^{-1}(J), J) = 0$ with $J \equiv j \mod p$ is unique in $\mathbb{Z}_{p^N}$, and $J = j^{\uparrow}$. To calculate the j-invariant of the canonical lift of E, Satoh's original algorithm [Sat00] lifts all conjugates of j simultaneously, which requires $O(N^3)$ memory. Later Vercauteren *et al.* [VPV01] improved this algorithm to reduce the space complexity to $O(N^2)$ by the direct computation of $j^{\uparrow}$. However, it still takes many evaluations of the modular polynomial $\Phi_p(X, Y)$ and inversions of elements in $\mathbb{Z}_{p^N}$. For efficiency, the SST-algorithm [SST01] used the following lemma which is a slight modification of the above result of Lubin *et al.*

Lemma 1. *For $j \in \mathbb{F}_{p^N} - \mathbb{F}_{p^2}$, let $y \in \mathbb{Z}_{p^N}$ with $y \equiv j^\uparrow \mod p^i$ for some $i \geq 1$, and let $\eta \in \mathbb{Z}_{p^N}$ be the element with $\Phi_p(\sigma^{-1}(y), \eta) \equiv 0 \mod p^{i+1}$ and $\eta \equiv y \mod p$, then $\eta \equiv j^\uparrow \mod p^{i+1}$.*

From the above Lemma, we see that for given $j^\uparrow$ with precision $i \geq 1$, we can raise the precision one by one, by updating Φ_p for every bit. Suppose that we have y satisfying $\Phi_p(\sigma^{-1}(y), y) \equiv 0 \mod p^i$ for some $i \geq 1$. Then it suffices to find $\delta_y \in \mathbb{Z}_{p^N}$ such that $\Phi_p(\sigma^{-1}(y), y+\delta_y) \equiv 0 \mod p^{i+1}$. Since $\Phi_p(\sigma^{-1}(y), y+\delta_y) = \Phi_p(\sigma^{-1}(y), y) + \delta_y \partial_Y \Phi_p(\sigma^{-1}(y), y) + O(\delta_y^2)$, we take

$$\delta_y \equiv -\Phi_p(\sigma^{-1}(y), y)(1/\partial_Y \Phi_p(\sigma^{-1}(y), y)) \mod p^{i+1}.$$

Moreover, it is enough to obtain $(1/\partial_Y \Phi_p(\sigma^{-1}(y), y))$ with precision 1 by the condition of y. The SST-algorithm uses a more refined technique; let $W := O(M^{\frac{\mu}{\mu+1}})$[1]. After obtaining $j^\uparrow$ with precision W by the above method, one can raise the precision by a similar computation based on the following observation:

$$\begin{aligned}&\Phi_p(x + p^{mW+i}\Delta_X, y + p^{mW+i}\Delta_Y)\\ &\qquad \equiv \Phi_p(x, y) + p^{mW+i}(\partial_X \Phi_p(x, y)\Delta_X + \partial_Y \Phi_p(x, y)\Delta_Y) \mod p^{(m+1)W}\end{aligned}$$

for $i \geq 0$ and $m \geq 1$. One can easily find that all of the operations between parentheses can be done within precision W. Furthermore, the use of σ^{-1} reduced many redundant evaluation of Φ_p, while [VPV01] did not. Computing the j-invariant of the canonical lift of E takes $O\big(N^{2\mu+\frac{1}{\mu+1}}\big)$ bit operations using $O(N^2)$ memory, where W subjects to $O\big(N^{\frac{\mu}{\mu+1}}\big)$.

2.2 Computing the Leading Coefficient Associated with p-th Verschiebung

We determine the kernel of the lifted p-th Verschiebung, and then compute the square of leading coefficient of the homomorphism induced by the lifted p-th Verschiebung on the formal group. It can be performed by the algorithms described in [Sat00] for $p \geq 5$, [FGH00] for $p = 2, 3$, and [VPV01] or [Skj00] for $p = 2$.

2.3 Norm Computation over the p-adic Number Fields

For $p = 2$, Satoh *et al.* presented a new algorithm to compute the norm of an element in $1 + 2^2\mathbb{Z}_2$ modulo 2^M, which is suitable for point counting of elliptic curves over $\mathbb{F}_{2^N}$. It is an analytic method using $\mathrm{Norm}_{\mathbb{Q}_{2^N}/\mathbb{Q}_2}(A) =$

[1] For cryptographic application, a word size of the CPU is recommended for W.

$\exp(\mathrm{Tr}_{\mathbb{Q}_{2^N}/\mathbb{Q}_2}(\log A))$, for $A \in 1 + 2^2\mathbb{Z}_2$. They computes the norm with precision $N/2 + O(1)$ in $O(N^{2\mu+0.5})$ time and $O(N^2)$ space by developing a fast method to obtain $\mathrm{Tr}_{\mathbb{Q}_{2^N}/\mathbb{Q}_2}$.

For $p > 2$, they use Kedlaya [Ked01] together with the Paterson-Stockmeyer algorithm [PS73]. It runs in $O(\max\{N^{2+\mu}, N^{2\mu+1/2}\} \log N)$ time with $O(N^{2.5})$ memory.

3 Gauss Periods and Normal Bases in Finite Fields

Let us recall that there are two most-common bases of an extension field : a normal basis (NB) and a polynomial basis (PB). When L/K be a finite Galois extension of degree N, a basis of L over K is called a normal basis if it is of the form $(\lambda\alpha)_{\lambda\in\mathrm{Gal}(L/K)}$ for some $\alpha \in L$. Any such α is called a *normal element.* A basis is called a polynomial basis if it is of the form $(\omega^i)_{0\le i<N}$ for some $\omega \in L$. In this section, we concentrate our interest on a normal basis, especially which is generated by a Gauss period which is defined below.

Definition 1. [Men2] *Let q be a prime or prime power, and let N, t be positive integers such that $Nt+1$ is a prime not dividing q. Let τ be any primitive t-th root of unity in $\mathbb{Z}/(Nt+1)\mathbb{Z}$. Let γ be a primitive $(Nt+1)$-th root of unity in some extension field of $\mathbb{F}_q$. A Gauss period of type (N, t) over $\mathbb{F}_q$ is defined as*

$$\alpha = \sum_{i=0}^{t-1} \gamma^{\tau^i}.$$

Let us call a NB induced by the Gauss period of type (N,t) as the *Gaussian normal basis of type t* and denote GNB of type t. It is easy to see that the Gauss period of type (N, t) belongs to $\mathbb{F}_{q^N}$. GNBs are very practical for the cryptographic application because their representations have the computational advantage that both squaring and multiplication can be done very simply. There is a simple criterion for a Gauss period to be a normal element.

Theorem 1. [Men2] *Let q, N and t be positive integers in Definition 1. Let e be the order of q modulo $Nt+1$. Then $\gcd(Nt/e, N) = 1$ if and only if the Gauss period of type (N, t) over $\mathbb{F}_q$ generates the normal basis for $\mathbb{F}_{q^N}$ over $\mathbb{F}_q$.*

We need to know the following lemma to develop the next section.

Lemma 2. [Men2] *Let $Nt+1$ be a prime, and $\gcd(Nt/e, N) = 1$ where e denotes the order of q modulo $Nt+1$. Let τ be a primitive t-th root of unity in $\mathbb{Z}/(Nt+1)\mathbb{Z}$. Then every non-zero element k in $\mathbb{Z}/(Nt+1)\mathbb{Z}$ can be written uniquely in the form*

$$k = q^i\tau^j, \quad 0 \le i \le N-1, 0 \le j \le t-1.$$

It is known that a representation with respect to a GNB of type 1 can be considered as an ordinary polynomial by a suitable change of indices, and Blake *et al.*[BRS98] showed that this idea could be extended to a GNB of type 2 by using a symmetric polynomial of double length. Refer [Sil99], [BRS98] and [BSS00] for more details about GNB of type 1 and 2 over finite fields. Moreover, their idea is extendable to GNBs of all types after a slight modification. We will deal with this extension over p-adic number fields in details in Section 4.

There is a famous conjecture of Artin that for each square-free integer $g \neq -1$, there exist infinitely many primes which have g as a primitive root. Hooley proved that this conjecture is true under the Extended Riemann Hypothesis [Hoo67], [Mur88]. Therefore, assuming the Extended Riemann Hypothesis or Artin's conjecture, it is expected that there are infinitely many finite fields with GNBs of type $t \leq 2$.

Remark 1. Note that for N prime, the type t has to be even. Furthermore, there are 26 values of prime N between 160 and 600 which there is an GNB of type 2 or 4 of $\mathbb{F}_{2^N}$ over $\mathbb{F}_2$. It covers 36% primes between 160 and 600 (For type 2, N= 173, 179, 191, 233, 239, 251, 281, 293, 359, 419, 431, 443, 491, 509, 593, For type 4, N= 163, 193, 199, 277, 307, 373, 409, 433, 487, 499, 577.)

4 *p*-adic Lift of Gauss Periods over Finite Fields

In this section, we consider p-adic fields taking advantage of both polynomial and normal basis representations. In this family of fields, both of multiplication and the Frobenius substitution can be done efficiently.

Theorem 2. *Let (K, v) be a complete discrete valuation field, and let L/K be a finite unramified extension of degree N. Let R_L (resp. R_K) denote the valuation ring of L (resp. K) and let $p \in K$ be a prime element of K which also prime in L. We also denote the residue class field of K and L by k_K and k_L, respectively. If $\mathcal{B}$ is a k_K-basis of k_L, then for any lift $\tilde{\mathcal{B}}$ of $\mathcal{B}$ in R_L, $\tilde{\mathcal{B}}$ is a K-basis of L. Furthermore, $\tilde{\mathcal{B}}$ is a R_K-basis of R_L.*

Proof. Let $\mathcal{B} = \{b_1, b_2, \ldots, b_N\}$ and let $\tilde{\mathcal{B}} = \{r_1, r_2, \ldots, r_N\}$. Denote π be the reduction modulo p map. If we have a non-trivial K-linear relation $c_1r_1 + c_2r_2 + \cdots + c_Nr_N = 0$, then without loss of generality, we can assume $c_i \in R_K$ and for at least one i, $\pi(c_i) \neq 0$. So we obtain $\pi(c_1)b_1 + \pi(c_2)b_2 + \cdots + \pi(c_N)b_N = 0$ in k_L, which is a contradiction. By comparing dimension, we see that $\tilde{\mathcal{B}}$ is a K-basis of L. For the second statement, it suffices to show that R_L is represented by R_K-linear sum of $\tilde{\mathcal{B}}$. Suppose that we have an element c in R_L which is not represented by R_K-linear sum of $\tilde{\mathcal{B}}$. Let $c := c_1r_1 + c_2r_2 + \cdots + c_Nr_N$, with $c_i \in K - R_K$ for some i. Put $z = \min_j(v(c_j))$; then $z < 0 \leq v(c)$ and $\pi(cp^{-z}) = 0$. By multiplying p^{-z} to c, we can write $\pi(c_1p^{-z})b_1 + \pi(c_2p^{-z})b_2 + \cdots + \pi(c_Np^{-z})b_N = 0$ in k_L which is a contradiction.

Corollary 1. *Let q be a prime or prime power, and let N, t be positive integers such that $Nt+1$ is a prime not dividing q. Let γ be a primitive $(Nt+1)$-th root of unity in some extension field of $\mathbb{Q}_q$. If* $\gcd(Nt/e, N) = 1$ *where e denotes* *the order of* q *modulo* $Nt+1$*, then for any primitive* t*-th root of unity* τ *in* $\mathbb{Z}/(Nt+1)\mathbb{Z}$*,*

$$\alpha := \sum_{i=0}^{t-1} \gamma^{\tau^i}$$

generates a normal basis over $\mathbb{Q}_q$*. Furthermore,* $[\mathbb{Q}_q(\alpha) : \mathbb{Q}_q] = N$*.*

Proof. From $\gcd(Nt+1, q) = 1$, it follows that $\pi(X^{Nt+1} - 1)$ is square free polynomial in $\mathbb{F}_q[X]$. Then $\mathbb{Q}_q(\gamma)$ is the unramified extension of $\mathbb{Q}_q$ [[Lan94], Prop. II.4.7], and so is $\mathbb{Q}_q(\alpha)$. Therefore, $[\mathbb{Q}_q(\alpha) : \mathbb{Q}_q] = [\mathbb{F}_q(\pi(\alpha)) : \mathbb{F}_q]$. Clearly, $\pi(\alpha)$ is a Gauss period of type (N, t) over $\mathbb{F}_q$. Thus $[\mathbb{F}_q(\pi(\alpha)) : \mathbb{F}_q] = N$ from the Theorem 1. By the Theorem 2, the desired conclusion can be shown.

From Corollary 1, the Gauss period can be lifted from the finite field $\mathbb{F}_{q^N}$ to $\mathbb{Q}_{q^N}$ and so we will use the family of fields that have the residue class field with the Gauss period. Let us extend the notion of a GNB from a finite field to a p-adic number field naturally, still denoting it by a GNB over a p-adic field. Note that γ defined Corollary 1 satisfies $\sigma(\gamma) = \gamma^q$ for the Frobenius substitution $\sigma \in \mathrm{Gal}(\mathbb{Q}_q(\gamma)/\mathbb{Q}_q)$ because $\gamma^{Nt+1} = 1$ and $q^{Nt} \equiv 1 \mod Nt+1$.

For convenience, we consider only the case of $p = q$. Furthermore, the following arguments hold for any other q where $q = p^l$ for some l. In the remainder of this section, we will describe how elements in $\mathbb{Z}_{p^N}$ represented with respect to the GNB of type t are expanded to elements in $\mathbb{Z}_{p^{Nt}}$ with respect to the PB, and how we can easily multiply two elements together and get the Frobenius substitution $\sigma \in \mathrm{Gal}(\mathbb{Q}_{p^N}/\mathbb{Q}_p)$.

4.1 *p*-adic Number Fields with GNBs of Type 1

We assume the condition in Corollary 1 with $t = 1$. In this case, $\alpha \in \mathbb{Q}_{p^N}$ is equal to γ and a normal element. Furthermore, $\pi(\alpha)$ is a normal element generating the GNB of type 1 over the field $\mathbb{F}_p$.

To get a type 1 GNB over p-adic number fields, consider the minimal polynomial $F(X) = X^N + X^{N-1} + \cdots + X + 1 \in \mathbb{Z}_p[X]$ of α. Since $\sigma(\alpha) = \alpha^p$ and p is primitive in $\mathbb{Z}/(N+1)\mathbb{Z}$, we have the NB,

$$\tilde{\mathcal{B}} = \{\alpha, \sigma(\alpha), \ldots, \sigma^{N-1}(\alpha)\} = \{\alpha, \alpha^p, \ldots, \alpha^{p^{N-1}}\} = \{\alpha, \alpha^2, \ldots, \alpha^N\}.$$

Similar to extension fields with GNBs of type 1 over finite fields [Sil99], multiplication can be handled through polynomial arithmetic. Moreover, the Frobenius substitution can be done by applying simple permutation.

Multiplication. For multiplication in $\mathbb{Z}_{p^N} := \mathbb{Z}_p[X]/\langle F(X)\rangle$, we will use the ring $R' = \mathbb{Z}_p[X]/\langle X^{N+1} - 1\rangle$. A lift of elements in $\mathbb{Z}_p[X]/\langle F(X)\rangle$ into R' is given as follows;

$$\sum_{i=0}^{N-1} a_i X^i \mapsto \sum_{i=0}^{N-1} a_i X^i + 0X^N.$$

Conversely, a projection from R' to $\mathbb{Z}_p[X]/\langle F(X)\rangle$ is given by the reduction modulo F. It implies that for arbitrary elements A, B in $\mathbb{Z}_p[X]/\langle F(X)\rangle$,

$$A \cdot B \equiv (A \cdot B \mod X^{N+1} - 1) \mod F.$$

Specifically, multiply A by B modulo $X^{N+1} - 1$, and then take the remainder modulo F to obtain $A \cdot B \in \mathbb{Z}_p[X]/\langle F(X)\rangle$. This multiplication is just that of two polynomials with degrees less than or equal to N; hence the complexity is clearly $O((NM)^{\mu})$ to get precision M.

Frobenius Substitution. Let $A(X) = a_0 + a_1X + \cdots + a_NX^N$ be an element of R'. By substituting $X = \alpha$, we obtain the Frobenius substitution of $A(\alpha)$ by

$$\sigma(\sum_{i=0}^{N} a_i\alpha^i) = \sum_{i=0}^{N} a_i\alpha^{pi} = a_0 + \sum_{j=1}^{N} a_{j/p}\alpha^j, \quad \text{where } j/p \in (\mathbb{Z}/(N+1)\mathbb{Z})^*.$$

Similarly, for all $k \in \mathbb{Z}$

$$\sigma^k(\sum_{i=0}^{N} a_i\alpha^i) = \sum_{i=0}^{N} a_i\alpha^{p^k i} = a_0 + \sum_{j=1}^{N} a_{j/p^k}\alpha^j, \quad \text{where } j/p^k \in (\mathbb{Z}/(N+1)\mathbb{Z})^*.$$

So we can compute $\sigma^k(A)$ by simple permutation on the set $(\mathbb{Z}/(N+1)\mathbb{Z})^*$, which needs $O(N)$ bit operations in a naive implementation and $O(1)$ bit operations with some elaborate implementation.

4.2 *p*-adic Number Fields with GNBs of Type 2

We assume the condition in Corollary 1 with $t = 2$. Then $\alpha(= \gamma + \gamma^{-1})$ is the Gauss period of type $(N, 2)$ in $\mathbb{Z}_{p^N}$, and normal in $\mathbb{Q}_{p^N}$. Since $\sigma'(\gamma) = \gamma^p$ for the Frobenius substitution σ' in $\mathrm{Gal}(\mathbb{Q}_p(\gamma)/\mathbb{Q}_p)$ and $\sigma'|_{\mathbb{Q}(\alpha)} = \sigma$, we obtain the normal basis

$$\tilde{\mathcal{B}} = \{\alpha, \sigma(\alpha), \ldots, \sigma^{N-1}(\alpha)\} = \{\gamma + \gamma^{-1}, \gamma^p + \gamma^{-p}, \ldots, \gamma^{p^{N-1}} + \gamma^{-p^{N-1}}\}.$$

For every $0 \leq i \leq N-1$, exactly one element in the pair $(p^i, -p^i)$ can be written as $j \mod 2N + 1$, for some $1 \leq j \leq N$ (see Lemma 2). Thus we have

$$\tilde{\mathcal{B}} = \{\gamma + \gamma^{-1}, \gamma^2 + \gamma^{-2}, \ldots, \gamma^N + \gamma^N\} = \{\gamma + \gamma^{2N}, \gamma^2 + \gamma^{2N-1}, \ldots, \gamma^N + \gamma^N\}.$$

From the last equality, the PB representation is naturally induced.

Multiplication. First, we consider a representation of an element in $\mathbb{Z}_{p^N} := \mathbb{Z}_p[X]/\langle F(X)\rangle$, where $F(X)$ is a minimal polynomial of α. We denote $\|x\| := \min\{|y| \mid y \equiv x \mod 2N + 1\}$, where x, $y \in \mathbb{Z}$. Then we can define a bijection $f : \{0, 1, \ldots, N-1\} \rightarrow \{1, 2, \ldots, N\}$ by $f(i) = \|p^i\|$.

For an element $A = \sum_{i=0}^{N-1} a_i \sigma^i(\alpha)$ in $\mathbb{Z}_{p^N}$, we can rewrite A with respect to γ as follows;

$$\begin{aligned} A &= \sum_{i=0}^{N-1} a_i(\gamma^{p^i} + \gamma^{-p^i}) = \sum_{j=1}^{N} a_{f^{-1}(j)}(\gamma^j + \gamma^{-j}) \\ &= \sum_{j=1}^{N} a_{f^{-1}(j)}\gamma^j + \sum_{j=1}^{N} a_{f^{-1}(j)}\gamma^{2N+1-j}. \end{aligned}$$

By replacing γ by X, we obtain the polynomial

$$A(X) = \sum_{j=1}^{N} a_{f^{-1}(j)}X^j + \sum_{j=1}^{N} a_{f^{-1}(j)}X^{2N+1-j}.$$

Thus, an element in $\mathbb{Z}_{p^N}$ can be uniquely represented as a polynomial in the ring $\mathbb{Z}_p[X]/\langle X^{2N+1} - 1\rangle$ uniquely determined modulo $X^{2N} + X^{2N-1} + \cdots + 1$. Note that this polynomial has a special property, $A(1/X)X^{2N+1} = A(X)$.

Definition 2. *For a given ring R, we call a polynomial $A(X) \in R[X]$ semi-palindromic if $A(X)$ is of the form*

$$A(X) = a_0 + \sum_{i=1}^{N} a_i(X^i + X^{2N+1-i}), \quad \textit{where } a_i \in R \textit{ for } 0 \le i \le N.$$

A semi-palindromic polynomial with $a_0 = 0$ is called palindromic.

Let S be a set of all semi-palindromic polynomials over $\mathbb{Z}_p$, then it is actually the set of all polynomials modulo $X^{2N+1} - 1$ representing elements in $\mathbb{Z}_{p^N}$. The addition is defined as the ordinary polynomial addition of elements in S, and the product of two polynomials $A(X)$, $B(X) \in S$ is the unique polynomial $C(X) \in S$ such that

$$C(X) \equiv A(X) \cdot B(X) \mod X^{2N+1} - 1. \tag{4.1}$$

Equation (4.1) yields that multiplication can be implemented using the standard polynomial multiplication with modular reduction. Indeed, if we substitute $X = \gamma$, then we see that $C(\gamma) \in \mathbb{Z}_{p^N}$ and so $C(X) \in S$.

Let us consider more efficient method to multiply two elements. For an element $A(X) \in S$, we can easily eliminate the constant term using the equality

$$A(\gamma) = \sum_{i=0}^{2N} a_i\gamma^i = \sum_{i=1}^{2N}(a_i - a_0)\gamma^i.$$

Thus, it is enough to consider multiplication of two palindromic polynomials in S. Given two palindromic polynomials $A(X)$ and $B(X)$, we can write them as

$$A(X) = A_1(X)X + A_2(X)X^{N+1} \text{ and } B(X) = B_1(X)X + B_2(X)X^{N+1},$$

where both $\deg(A_i)$ and $\deg(B_i)$ are less than N. Additionally, A_1 is of the symmetric form of A_2, that is $X^{N-1}A_1(1/X) = A_2(X)$ and the same holds for B_1 and B_2. We can easily show that a similar relation holds for pairs $(A_1B_1,\, A_2B_2)$ and $(A_1B_2,\, A_2B_1)$ (i.e. $X^{2N-2}A_1(1/X)B_1(1/X) = A_2(X)B_2(X)$). Since $A(X) \cdot B(X)$ is given by

$$A(X) \cdot B(X) \equiv A_1B_1X^2 + (A_1B_2 + A_2B_1)X^{N+2} + A_2B_2X \mod X^{2N+1} - 1,$$

the multiplication in $\mathbb{Z}_{p^N}$ with respect to palindromic representation can be done by two multiplications, A_1B_1 and A_1B_2 , of polynomials of degree less than N. Hence the complexity is $O(2(NM)^{\mu})$. Note that it is two times slower than the case of $t = 1$.

Frobenius Substitution. Let $A(X) = a_0 + \sum_{i=1}^{N} a_iX^i + \sum_{i=1}^{N} a_iX^{2N+1-i}$ be an element in S. When we substitute $X = \gamma$, we obtain $A(\gamma) = a_0 + \sum_{i=1}^{N} a_i(\gamma^i + \gamma^{-i})$ and so the k-th Frobenius substitution σ^k of $A(\gamma)$ by

$$\sigma^k(A(\gamma)) = a_0 + \sum_{i=1}^{N} a_i(\gamma^{p^k i} + \gamma^{-p^k i}) = a_0 + \sum_{j=1}^{N} a_{\|j/p^k\|}(\gamma^j + \gamma^{-j}) \qquad \text{for all } k \in \mathbb{Z}.$$

Thus we can get the k-th Frobenius substitution by simple permutation on the set $(\mathbb{Z}/(N+1)\mathbb{Z})^*$.

4.3 p-adic Number Fields with GNBs of Type $t > 2$

We generalize the p-adic lift of GNBs of type 1 and 2. We assume that $q\ (= p)$, N and t satisfies the condition of Corollary 1.

When t is even, for an element A in $\mathbb{Z}_{p^N}$, we can express A with respect to γ by

$$\begin{aligned} A &= \sum_{i=0}^{N-1} a_i\sigma^i(\alpha) = \sum_{i=0}^{N-1} a_i(\sum_{j=0}^{t-1} \gamma^{p^i\tau^j}) \\ &= \sum_{i=0}^{N-1} a_i(\sum_{j=0}^{t/2-1} \gamma^{p^i\tau^j} + \gamma^{-p^i\tau^j}) = \sum_{i=1}^{Nt/2} c_i\gamma^i + \sum_{i=1}^{Nt/2} c_i\gamma^{tN+1-i}, \end{aligned}$$

where $\{c_i \,|\, 1 \le i \le N\}$ is a bijective image of $\{a_j \,|\, 0 \le j \le N-1\}$. If we replace γ by X, then elements of $\mathbb{Z}_{p^N}$ can be represented as palindromic polynomials in the polynomial ring modulo $X^{Nt+1} - 1$. To get multiplication with this representation, we need two multiplications of polynomials of degree less than or equal to $(Nt/2) - 1$ as done in the case of type 2. Therefore, the complexity is $O(2(tNM/2)^{\mu})$. The Frobenius substitution can be done by a suitable permutation in the same manner as the case of $t = 2$.

When t is odd, an element A in $\mathbb{Z}_{p^N}$ can be represented by the polynomial modulo $X^{Nt+1} - 1$. The multiplication in this representation can be done by multiplication of polynomials with degree less than or equal to Nt. Therefore the time complexity of the multiplication is $O((tNM)^{\mu})$. The Frobenius substitution can be done in the same manner as the case of $t = 1$.

Remark 2. If t is even, then multiplication with a GNB of type t is slower than that of type 1 by a constant factor of $2(t/2)^{\mu}$. Similarly, if t is odd, then it is slower by a constant factor of t^{μ}. Thus for $t \geq 5$, it is slower by at least 10 times. With this practical reason, hereafter we restrict the choice of t so that $t \leq 4$.

5 Norm Computation Algorithm

In this section, we develop an algorithm to compute $\mathrm{Norm}_{\mathbb{Q}_{p^N}/\mathbb{Q}_p}(A) \mod p^M$ for $A \in \mathbb{Z}_{p^N}$, where $\mathbb{F}_{p^N}$ has the GNB of type $t \leq 4$. We will use the representation for elements in $\mathbb{Z}_{p^N}$ as described in previous section. By using the 2-adic expansion of N, our algorithm requires fewer multiplications and more Frobenius substitutions. Let $N = \sum_{i=0}^{l} n_i 2^i$ with $n_i \in \{0, 1\}$ and $n_l = 1$. Denote it by $[n_0, n_1, \ldots, n_l]_2$. Since $\mathrm{Gal}(\mathbb{Q}_{p^N}/\mathbb{Q}_p)$ is generated by σ, we obtain that

$$\mathrm{Norm}_{\mathbb{Q}_{p^N}/\mathbb{Q}_p}(A) = A(\sigma A) \cdots (\sigma^{N-1} A) = M_{l-1} \cdot \prod_{i=0}^{l-2} (\sigma^{N-[n_0,n_1,\ldots,n_i]_2} M_i)^{n_i},$$

where $M_i = (\sigma^{2^{i-1}} M_{i-1}) M_{i-1},\ M_0 = A$. The following norm computation algorithm is derived from the above expression.

Algorithm ComputeNorm

```
Input : $A \in \mathbb{Z}_{p^N},\ N = [n_0, n_1, \ldots, n_l]_2, n_l = 1$
Output : $\mathrm{Norm}_{\mathbb{Q}_{p^N}/\mathbb{Q}_p}(A)$.
Begin
    1. $M \leftarrow A$;
    2. If $n_0 = 1$ then Temp $\leftarrow \sigma^{N-1} A$;
       Else Temp $\leftarrow 1$;
    3. For $i = 1$ to $l - 1$ do
       (a) $M \leftarrow (\sigma^{2^{i-1}} M)M$;
       (b) If $n_i = 1$ then Temp $\leftarrow$ (Temp)$\cdot(\sigma^{n_{i+1}2^{i+1}+\cdots+n_l 2^l} M)$;
    4. $M \leftarrow (\sigma^{2^{l-1}} M)(M)$;
    5. $M \leftarrow M\cdot$Temp;
    6. Return $M$;
End
```

ComputeNorm requires at most $2\lfloor \log_2 N \rfloor$ times multiplications over $\mathbb{Z}_{p^N}$ and at most $2\lfloor \log_2 N \rfloor$ times σ^i substitutions. Since the Frobenius substitution requires at most $O(N)$ bit operations for the field with a GNB of type $t \leq 4$ as described in section 4, it requires $O((\log N)(NM)^{\mu})$ time and $O(NM)$ space to get precision M.

Example 1. If $N = 10 = 2 + 2^3$, then

$$\mathrm{Norm}_{K/\mathbb{Q}_p}(A) = A(\sigma A) \cdots (\sigma^9 A) = (A(\sigma A) \cdots (\sigma^{2^3-1} A))(\sigma^{2^3}(A(\sigma A)))$$

1. $M_1 \leftarrow (\sigma A)A$
   (a) Temp$\leftarrow \sigma^{2^3} M_1$
2. $M_2 \leftarrow (\sigma^2 M_1)M_1$
3. $M_3 \leftarrow (\sigma^{2^2} M_2)M_2$
4. $\mathrm{Norm}_{K/\mathbb{Q}_p}(A) = M_3 \cdot$Temp

6 Application to Point Counting

Now we describe how our algorithm can be applied to point counting based on SST. Before explaining the application, first we consider two basic operations: multiplication and the Frobenius substitution. Since the SST-algorithm uses a polynomial basis generated by ψ satisfying $\psi^{p^N-1} = 1$, so in general, the reduction polynomial $f(X)$ is dense. For the given polynomial $f(X)$ of degree N, $A(X) \mod f(X)$ is given by $A - (((A/X^N)Z)/X^{N-2})f(X)$ for $\deg A \leq 2N-2$ where Z is precomputed as $Z := X^{2N-2}/f$. Hence the multiplication in $\mathbb{Z}_p[X]/\langle f(X)\rangle$ is about three times slower than in $\mathbb{Z}_p/\langle X^{N+1}-1\rangle$. In the case of a type 1 GNB, our reduction polynomial is exactly $X^{N+1}-1$, so our multiplication is three times faster than that of the SST-algorithm. In the case of type 2 GNB, it is 1.5 times faster than that of the SST-algorithm for the similar reason. With type 3 or 4, our multiplication may be slower than that of the SST-algorithm, since the polynomial representation is lengthy. For the Frobenius substitution, our method requires almost nothing, while the SST-algorithm requires $p-1$ multiplications and $p-1$ additions over $\mathbb{Z}_{p^N}$ (see [SST01]).

We will show that our algorithm improves the complexity of the SST-algorithm to $O(N^{2\mu+\frac{1}{\mu+1}})$ in time and $O(N^2)$ in space, while the SST-algorithm runs in $O(N^{2\mu+0.5})$ time and at a minimum of $O(N^2)$ space. Recall that the SST-algorithm works with precision $M := N/2 + O(1)$. It was previously proved in [SST01] that it takes $O(N^{2\mu+\frac{1}{\mu+1}})$ time and $O(N^2)$ space in step (1) and (2) of the algorithm in Section 2. In step (3), applying algorithm 1 in section 5 to the norm computation, the time complexity dropped from $O(N^{2\mu+0.5})$ to $O(N^{2\mu}\log N)$, while the space complexity remains fixed to $O(N^2)$ for all small p. Hence the total complexities in time and space can be obtained.

For a detailed description, since all Frobenius substitution requires almost nothing, there is at least a 10% speed-up in Step (1) (See [SST01] for details). Moreover, as our multiplication is much faster in the case of $t = 1$ or 2, the total running time is roughly reduced by a constant factor of 3 for type 1, and by 1.5 for type 2 at least.

7 Implementation and Results

In this section, we show experimental running time of our version of the SST-algorithm for $p = 2$. For comparison, we also present the recent results of the

SST-algorithm in [SST01], which is from [SST01]. Both algorithms have been implemented in the *C* programming language for the most part, and some assembly for most basic operations on multi-precision integers. Satoh *et al.* obtained their result on a 32bit Pentium III-866 MHz processor, while ours was on a Pentium III-800MHz processor with a 128MB RAM of the main memory, running Linux Mandrake 2.2.17 and compiled using gcc compiler version 2.95.3 with options optimized to Pentium III processors including '-O3'. Since two platforms are different, an exact comparison between the two running results can be ambiguous. Therefore one has to regard this as a reference. Before providing our actual results, we will briefly comment on the implementation of our algorithm.

First, for efficiency we used a constant value of 32 for $162 \leq N \leq 302$, a word size of a Pentium III processor, for W in the algorithm described in Section 2.1, hence in many steps operations are performed within one-word precision. It allows us to eliminate much of the loop overhead by using an unrolled version of operations. All elements of $\mathbb{Z}_{2^N}$ are represented as polynomials as in Section 4. For GNBs of even types we used a palindromicity to store only half of the polynomial, while elements of $\mathbb{F}_{2^N}$ are always represented as full-size polynomials. Multiplication of two elements in $\mathbb{Z}_{2^N}$ is implemented using Karatsuba's method. We use naive multiplication, so called pencil-and-paper method, for the coefficients.

In the Table 1, we present the running time of both algorithms for finite fields $\mathbb{F}_{2^N}$ where N is between 160 and 600. For our results, we used finite fields with GNBs of type 1, 2, 3 or 4. It shows that our improvement largely enhances the speed as that of Satoh *et al.* in the case of type 1 and 2. We also present the result of AGM method for a rough comparison.

For a researching interest, we also show out results for large N for GNBs of type 1 and 2 in the Table 2, with varing W. These results are obtained on the same machine environment, but the compiler gcc version 3.0.3 is used instead of ver 2.95.3.

8 Conclusion

In this paper, we reduced the time complexity of the original the SST-algorithm from $O(N^{2\mu+0.5})$ to $O(N^{2\mu+\frac{1}{\mu+1}})$ with some restrictions on N, while the space complexity still remains fixed to $O(N^2)$ for any small p. We also developed a fast algorithm for the norm computation with $O((NM)^\mu \log N)$ time and $O(NM)$ space to get precision M. In addition, our algorithm refined the running time by a maximum constant factor of 3. In a cryptographic context (i.e., for $p = 2$ and $160 < N < 600$), about 42% (36% for prime N) of fields have such bases. Because of the reduced complexity, our method works well for a large N. As shown in Section 7, it takes merely 17.38 minutes to count the number of points on an arbitrary elliptic curve defined on the finite field $\mathbb{F}_{2^{3010}}$, and about 1 day and 10 hours on $\mathbb{F}_{2^{12010}}$ by Pentium III-800 MHz computer.

Furthermore, our improvement is not only restricted to the SST-algorithm. It can also work with all algorithms working on p-adic number fields, which

Table 1. Timings(in sec) for computations of j-invariant, Norm and Order counting. The time table of AGM method through Alpha 750 MHz is published on the homepage of Argo Tech(http://argote.ch).

N	Type	j-inv	Norm	Total	Note	N	Type	j-inv	Norm	Total	Note
162	1	0.076	0.018	**0.110**		233	-	1.72	0.29	2.24	[SST01]
163	-	-	-	0.07	AGM	**233**	2	0.433	0.142	**0.743**	
163	-	0.58	0.10	0.76	[SST01]	235	4	1.318	0.481	2.513	
163	4	0.390	0.124	0.766		236	3	1.264	0.520	2.365	
166	3	0.350	0.154	0.691		239	-	1.86	0.45	2.54	[SST01]
173	2	0.192	0.070	**0.336**		**239**	2	0.432	0.168	**0.771**	
193	-	0.98	0.19	1.31	[SST01]	239	-	-	-	0.24	AGM
193	4	0.680	0.206	1.281		244	3	1.354	0.560	2.539	
194	2	0.228	0.068	**0.383**		265	4	1.672	0.505	3.066	
196	1	0.121	0.035	**0.201**		**268**	1	0.284	0.083	**0.474**	
197	-	-	-	0.14	AGM	279	4	1.807	0.662	3.453	
197	-	1.04	0.20	1.38	[SST01]	283	-	2.97	0.73	4.13	[SST01]
199	4	0.749	0.270	1.445		286	3	1.767	0.829	3.442	
204	3	0.754	0.265	1.328		**292**	1	0.306	0.096	**0.523**	
209	2	0.325	0.083	**0.511**		**293**	2	0.598	0.210	**1.059**	
210	1	0.168	0.042	**0.262**		307	4	2.548	0.991	4.893	

Table 2. Timings for computations of Norm and Order counting for large N.

N	Type	Norm	Total	W
3010	1	2.63 min	17.38 min	96
3005	2	7.93 min	41.03 min	96
6010	1	34.33 min	2 hr 59.25 min	128
6005	2	1hr 31.68 min	7 hr 7.25 min	128
12010	1	6 hr 45 min	1 day 10 hr 24 min	192

multiplications and the Frobenius substitutions play dominant roles. It is known that the AGM method uses norm computation over a p-adic number field; hence we expect that our norm computation algorithm can be combined with the AGM method to give faster point counting.

Acknowledgement

The authors of this paper would like to thank Takakazu Satoh for his many insightful comments and interesting discussions on this work.

References

Art92. A. O. L. Atkin, The number of points on an elliptic curve modulo a prime, Series of e-mails to the NUMBERTHRY mailing list, 1992.

BRS98. L. F. Blake, R. M. Roth, and G. Seroussi, Efficient Arithmetic in GF(2^n) through Palindromic Representation, Tech. Rep. **HPL-98-134**, Hewlett Packard, 1998.

BSS00. I. Blake, G. Seroussi, and N. Smart, *Elliptic Curves in Cryptography*, Cambridge Univ. Press, 2000.

Cou96. J. M. Couveignes, Computing l-isogenies using the p-Torsion, *Algorithmic number theory - ANTS-II*, LNCS **1122**, pp. 59–66, Springer-Verlag, 1996.

Deu41. M. Deuring, Die Typen der Multiplikatorenringe elliptischer Funktionenkörper. *Abh. Math. Sem. Univ. Hamburg*, **14**, pp. 197–272, 1941.

Elk98. N. D. Elkies, Elliptic and modular curves over finite fields and related computational issues, In D.A. Buell and eds. J.T. Teitelbaum, editors, *Computational perspective on number theory*, AMS/IP Stud. Adv. Math., **7**, pp. 21–78, Province, RI: AMS, 1998. Proceedings of a Conference in Honor of A.O.L. Atkin.

FGH00. M. Fouquet, P.Gaudry, and R. Harley, On Satoh's algorithm and its implementation, *J. Ramanujan Math. Soc.*, **15**, pp. 281–318, 2000.

HMG01. R. Harley, Counting points with the arithmetic-geometric mean(joint work with J. F. Mestre and P. Gaudry), Eurocrypt 2001, Rump session, 2001.

Hoo67. C. Hooley, On Artin's conjecture, *J. Reine Angew Math.*, **225**, pp. 209–220, 1967.

Ked01. K. Kedlaya, Counting points on hyperelliptic curves using Monsky-Washnitzer cohomology, available at `http://arXiv.org/abs/math/0105031`.

Kob87. N. Koblitz, Elliptic curve cyptosystem, *Math. Comp.*, **48**(177), pp. 203-209, 1998.

Lan94. S. Lang, *Algebraic Number Theory*, Springer-Verlag, 1994.

LST64. J. Lubin, J. P. Serre, and J. Tate. Elliptic curves and formal group. *Lecture notes in prepared in connection with the seminars held at the Summer institute on Algebraic Geometry, Whitney Estate, Woods Hole, Massachusetts*, 1964.

Men1. A. Menezes, *Elliptic Curve Public Key Cryptosystems*, Kluwer Academic Publishers, 1993.

Men2. A. Menezes, *Application of Finite Fields*, Kluwer Academic Publisher, 1993.

Mil87. V. Miller, Use of elliptic curves in cryptography. *Crypto'86*, LNCS **263**, pp. 417–426, 1987.

Mur88. M. R. Murty, Artin's conjecture for primitive roots, *Math. Intelligencer*, **10** (4), pp. 59–67, 1988.

PS73. M. S. Parterson and L. J. Stockmeyer, On the number of nonscalar multiplications necessary to evaluate polynomials. *SIMA J. Comput.*, **2**, pp. 60–67, 1973.

Sat00. T. Satoh, The canonical lift of an ordinary elliptic curve over a finite field and its point counting, *J. Ramanujan Math. Soc.*, **15**, pp. 247–270, 2000.

Sch85. R. Schoof, Elliptic curves over finite fields and the computation of square roots mod p, *Math. Comput.*, **44**, pp. 483–494, 1985.

Sil99. J. H. Silverman, Fast Multiplication in Finite Fields GF(2^N), *Crytographic Hardware and Embedded Systems - CHES'99*, LNCS **1717**, pp. 122-134, Springer-Verlag, 1999.

Skj00. B. Skjernaa, Satoh Point Counting in characteristic 2. To appear in *Math. Comp.*

SST01. T. Satoh, B. Skjernaa, and Y. Taguchi, Fast Computation of Canonical Lifts of Elliptic curves and its Application to Point Counting, *Preprint*, 2001.

VPV01. F. Vercauteren, B. Preneel, and J. Vandewalle, A Memory Efficient Version of Satoh's Algorithm. *Advances in Cryptology - Eurocrypt 2001*, LNCS **2045**, pp. 1–13, Springer-Verlag, 2001.

An Extension of Kedlaya's Algorithm to Artin-Schreier Curves in Characteristic 2

Jan Denef[1] and Frederik Vercauteren[2,3,⋆]

[1] Department of Mathematics
University of Leuven
Celestijnenlaan 200B, B-3001 Leuven-Heverlee, Belgium
jan.denef@wis.kuleuven.ac.be

[2] Department of Electrical Engineering
University of Leuven
Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, Belgium
frederik.vercauteren@esat.kuleuven.ac.be

[3] Computer Science Department
University of Bristol
Woodland Road, Bristol BS8 1UB, United Kingdom
frederik@cs.bris.ac.uk

Abstract. In this paper we present an extension of Kedlaya's algorithm for computing the zeta function of an Artin-Schreier curve over a finite field $\mathbb{F}_q$ of characteristic 2. The algorithm has running time $O(g^{5+\varepsilon} \log^{3+\varepsilon} q)$ and needs $O(g^3 \log^3 q)$ storage space for a genus g curve. Our first implementation in MAGMA shows that one can now generate hyperelliptic curves suitable for cryptography in reasonable time. We also compare our algorithm with an algorithm by Lauder and Wan which has the same time and space complexity. Furthermore, the method introduced in this paper can be used for any hyperelliptic curve over a finite field of characteristic 2.

Keywords: Hyperelliptic curves, Monsky-Washnitzer cohomology, Kedlaya's algorithm, Lauder & Wan algorithm, cryptography

1 Introduction

Computing the zeta function of abelian varieties over finite fields is one of the most important problems in computational algebraic geometry and has many applications [24], e.g. the construction of cryptosystems based on Jacobians of curves. The most important systems use elliptic curves as introduced by Miller [18] and Koblitz [13] or hyperelliptic curves which were proposed by Koblitz [14]. More general, but less practical systems work in the Jacobian of superelliptic curves [9] and of $\mathcal{C}_{ab}$ curves [1].

[⋆] F.W.O. research assistant, sponsored by the Fund for Scientific Research - Flanders (Belgium).

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 308–323, 2002.

The problem of counting the number of points on elliptic curves over finite fields of any characteristic can be solved in polynomial time using Schoof's algorithm [26] and its improvements due to Atkin [2] and Elkies [6]. An excellent account of the resulting SEA-algorithm can be found in [3] and [17]. For finite fields of small characteristic, Satoh [25] described an algorithm based on p-adic methods which is asymptotically faster than the SEA-algorithm. Skjernaa [27] and Fouquet, Gaudry and Harley [8] extended the algorithm to characteristic 2 and Vercauteren [29] presented a memory efficient version. Recently Mestre and Harley proposed a variant of Satoh's algorithm based on the Arithmetic-Geometric Mean, which has the same asymptotic behaviour as [29], but is faster by some constant.

The equivalent problem for higher genus curves seems to be much more difficult. Pila [23] described a theoretical generalisation of Schoof's approach, but the algorithm is not practical, not even for genus 2 as shown by Gaudry and Harley [11]. An extension of Satoh's method to higher genus curves needs the Serre-Tate canonical lift of the Jacobian of the curve, which need not be a Jacobian itself and thus is difficult to compute with. The AGM method does generalise to hyperelliptic curves, but currently only the genus 2 case is practical.

Recently Kedlaya [12] described a p-adic algorithm to compute the zeta function of hyperelliptic curves over finite fields of small *odd* characteristic, using the theory of Monsky-Washnitzer cohomology. The running time of the algorithm is $O(g^{5+\varepsilon} \log^{3+\varepsilon} q)$ for a hyperelliptic curve of genus g. The algorithm readily generalises to superelliptic curves as shown by Gaudry and Gurel [10]. A related approach by Lauder and Wan [15] is based on Dwork's proof of the rationality of the zeta function and leads to a polynomial time algorithm for computing the zeta function of an arbitrary variety over a finite field. Despite its polynomial complexity, the algorithm in its most general form is not practical. Using Dwork cohomology, Lauder and Wan [16] adapted their original algorithm for the special case of Artin-Schreier curves, resulting in an $O(g^{5+\varepsilon} \log^{3+\varepsilon} q)$ time algorithm.

In this paper we extend Kedlaya's algorithm to Artin-Schreier curves defined by an equation of the form $y^2 - x^m y - f(x) = 0$ over some finite field $\mathbb{F}_q$ of characteristic 2. The resulting algorithm has running time $O(g^{5+\varepsilon} \log^{3+\varepsilon} q)$ and needs $O(g^3 \log^3 q)$ storage space for a genus g curve. We have implemented our algorithm as well as Lauder & Wan's algorithm in the MAGMA computer algebra system and present a comparison of the efficiency of both algorithms.

Finally we remark that using the ideas introduced in this paper, we recently extended Kedlaya's algorithm to *all* hyperelliptic curves defined over a finite field of characteristic 2. More details can be found in the forthcoming paper [5].

The remainder of the paper is organised as follows: after recalling the formalism of Monsky-Washnitzer cohomology in Section 2, we show in Section 3 how to extend Kedlaya's algorithm to the aforementioned Artin-Schreier curves. Section 4 contains a ready to implement description of the resulting algorithm. In Section 5, we present running times of an implementation of both algorithms in MAGMA and compare their efficiency.

2 Monsky-Washnitzer Cohomology

In this section we briefly recall the definition and main properties of Monsky-Washnitzer cohomology. More details can be found in the seminal papers by Monsky and Washnitzer [19,20,21], the lectures by Monsky [22] and the survey by van der Put [28].

Let $\overline{X}$ be a smooth affine variety over a finite field $k := \mathbb{F}_q$ with coordinate ring $\overline{A}$. Let R denote a complete discrete valuation ring with uniformizer π, residue field $R/\pi R = k$ and fraction field K of characteristic 0. Elkik [7] showed that one can always find a smooth finitely generated R-algebra A such that $A/\pi A \cong \overline{A}$. To compute the zeta function of $\overline{X}$ one needs to lift the Frobenius endomorphism $\overline{F}$ on $\overline{A}$ to the R-algebra A, but in general this is not possible. Note that for elliptic curves, Satoh solves this problem by using the Serre-Tate canonical lift which does admit a lift of the Frobenius endomorphism. To remedy this difficulty one could work with the π-adic completion A^∞ of A. But again we run into difficulties since the de Rham cohomology of A^∞ is larger than that of A. As an example, consider the affine line over $\mathbb{F}_p$, so $A = R[x]$, then each term in $\sum_{n=0}^\infty p^n x^{p^n-1} dx$ is an exact differential form, but its sum is not, since $\sum_{n=0}^\infty x^{p^n}$ is not in A^∞. The fundamental problem is that the series $\sum_0^\infty p^n x^{p^n-1}$ does not converge fast enough for its integral to converge as well. Monsky and Washnitzer solve this problem by working with a subalgebra $A^\dagger$ of A^∞, whose elements satisfy growth conditions. This *dagger ring* or *weak completion* $A^\dagger$ is defined as follows: write $A := R[x_1, \ldots, x_n]/(f_1, \ldots, f_m)$, then

$$A^\dagger := R\langle x_1, \ldots, x_n\rangle^\dagger/(f_1, \ldots, f_m),$$

where $R\langle x_1, \ldots, x_n\rangle^\dagger$ consists of power series

$$\left\{\sum a_\alpha x^\alpha \in R[[x_1, \ldots, x_n]] \mid \exists C, \rho \in \mathbb{R}, C > 0, 0 < \rho < 1, \forall \alpha : |a_\alpha| \le C\rho^{|\alpha|}\right\},$$

where $\alpha := (\alpha_1, \ldots, \alpha_n)$, $x^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ and $|\alpha| := \sum_{i=0}^n \alpha_i$. Equivalently, $R\langle x_1, \ldots, x_n\rangle^\dagger$ can be defined as the set of overconvergent power series, i.e. elements of $R[[x_1, \ldots, x_n]]$ which converge in a polydisc

$$\{(x_1, \ldots, x_n) \in K^n \mid |x_1| \le \rho_1, \ldots, |x_n| \le \rho_n\}$$

with all $\rho_i > 1$. The ring $A^\dagger$ clearly satisfies $A^\dagger/\pi A^\dagger = \overline{A}$, is weakly complete, i.e. is equal to its weak completion and is flat over R. A finitely generated algebra which satisfies these three properties is called a *lift* of $\overline{A}$. One can show that if $\overline{A}$ is smooth and finitely generated, there always exists a lift $A^\dagger$ of $\overline{A}$ and that every lift of $\overline{A}$ is R-isomorphic to $A^\dagger$. Furthermore, let $\overline{B}/k$ be smooth and finitely generated, with lift $B^\dagger$ and let $g : \overline{A} \to \overline{B}$ be a morphism of k-algebra's, then there exists an R-homomorphism $G : A^\dagger \to B^\dagger$ lifting g. The last property implies that we can lift the q-power Frobenius from $\overline{A}$ to $A^\dagger$.

For $A^\dagger$ we can define the universal module $D^1(A^\dagger)$ of differentials

$$D^1(A^\dagger) := (A^\dagger\ dx_1 + \cdots + A^\dagger\ dx_n)/(\sum_{i=1}^m A^\dagger(\frac{\partial f_i}{\partial x_1}\ dx_1 + \cdots + \frac{\partial f_i}{\partial x_n}\ dx_n)).$$

Let $D^i(A^\dagger) := \bigwedge^i D^1(A^\dagger)$ be the i-th exterior product of $D^1(A^\dagger)$ and denote with $d_i : D^i(A^\dagger) \to D^{i+1}(A^\dagger)$ the exterior differentiation. Since $d_{i+1} \circ d_i = 0$ we get the de Rham complex $D(A^\dagger)$

$$0 \longrightarrow D^0(A^\dagger) \xrightarrow{d_0} D^1(A^\dagger) \xrightarrow{d_1} D^2(A^\dagger) \xrightarrow{d_2} D^3(A^\dagger) \cdots$$

The i-th cohomology group of $D(A^\dagger)$ is defined as $H^i(\overline{A}/R) := \text{Ker } d_i/\text{Im } d_{i-1}$ and $H^i(\overline{A}/K) := H^i(\overline{A}/R) \otimes_R K$ finally defines the i-th Monsky-Washnitzer cohomology group. One can prove that for smooth, finitely generated k-algebra's $\overline{A}$ the map $\overline{A} \mapsto H^\bullet(\overline{A}/K)$ is well defined and functorial, which justifies the notation. Let F be a lift of the q-power Frobenius endomorphism of $\overline{A}$ to $A^\dagger$, then F induces an endomorphism F_* on the cohomology groups. The main theorem of Monsky-Washnitzer cohomology is that the $H^i(\overline{A}/K)$ satisfy a Lefschetz fixed point formula.

Theorem 1 (Lefschetz fixed point formula). *Let $\overline{X}/\mathbb{F}_q$ be a smooth affine variety of dimension n, then the number of $\mathbb{F}_q$-rational points on $\overline{X}$ equals*

$$\sum_{i=0}^{n} (-1)^i \text{Tr}\left(q^n F_*^{-1} | H^i(\overline{A}/K)\right).$$

3 Cohomology of Artin-Schreier Curves over $\mathbb{F}_{2^n}$

Let $\mathbb{F}_q$ be a finite field with $q = 2^n$ elements and fix an algebraic closure $\overline{\mathbb{F}}_q$. Let K be a degree n unramified extension of $\mathbb{Q}_2$ and let R be its valuation ring with residue field $R/2R = \mathbb{F}_q$. The Artin-Schreier curves we will consider are defined by an affine equation of the form

$$\overline{C}_{m,\overline{f}} : y^2 - x^m y - \overline{f}(x) = 0, \tag{1}$$

with $0 \leq m \leq g$, $\overline{f} \in \mathbb{F}_q[x]$ monic of degree $2g+1$ and such that $\overline{C}_{m,\overline{f}}$ is non-singular. Let $\mathfrak{p} : \overline{C}_{m,\overline{f}}(\overline{\mathbb{F}}_q) \to \mathbb{A}^1(\overline{\mathbb{F}}_q)$ be the projection map on the x-axis, then the branch locus of $\mathfrak{p}$ is empty if and only if $m = 0$ and consists of the singleton $\{0\}$ if and only if $m > 0$. Without loss of generality we may assume that $f(0) = 0$ if $m > 0$, i.e. that $(0,0)$ is the unique ramification point of $\mathfrak{p}$. Indeed, the isomorphism defined by $x \mapsto x$ and $y \mapsto y + \overline{f}(0)^{1/2}$ shows that we can replace $\overline{f}(x)$ with $\overline{f}(x) - \overline{f}(0) + x^m \overline{f}(0)^{1/2}$, which clearly is divisible by x if $m > 0$. Note that since $\overline{C}_{m,\overline{f}}$ is non-singular we have $\overline{f}'(0) \neq 0$.

Let $\overline{H}_m(x)$ be defined as $\overline{H}_0(x) := 1$ and $\overline{H}_m(x) := x$ for $m > 0$, i.e. $\overline{H}_m(\theta)$ is zero if and only if the point with x-coordinate θ ramifies. Let $\overline{C}'_{m,\overline{f}}$ be the curve obtained from $\overline{C}_{m,\overline{f}}$ by deleting the support of $\overline{H}_m(x)$. Then the coordinate ring of $\overline{C}'_{m,\overline{f}}$ is given by $\overline{A}_{m,\overline{f}} := R[x, y, (\overline{H}_m(x))^{-1}]/(y^2 - x^m y - \overline{f}(x))$.

Take any lift $f \in R[x]$ of $\overline{f}$, with the restrictions that f should be monic and of degree $2g+1$ and that $f(0) = 0$ for $m > 0$. Let $H_0(x) := 1$ and $H_m(x) := x$ for

$m > 0$ and let $C'_{m,f}$ be the curve obtained from $C_{m,f} : y^2 - x^m y - f(x) = 0$ by deleting the support of $H_m(x)$. Note that the point $(0, 0)$ still is a ramification point on $C_{m,f}$, which explains why we need the extra restriction on f if $m > 0$. The coordinate ring of $C'_{m,f}$ is $A_{m,f} := R[x, y, (H_m(x))^{-1}]/(y^2 - x^m y - f(x))$ and there exists an involution $\imath$ on $A_{m,f}$ which sends x to x and y to $-y + x^m$.

Let $A^{\dagger}_{m,f}$ be the dagger ring of $A_{m,f}$. Using the equation of the curve we can always represent elements of $A^{\dagger}_{m,f}$ as a series $\sum_{l=-\infty}^{+\infty}(a_l + b_l y)x^l$ with $a_l, b_l \in R$. If $m = 0$ then all a_l, b_l with $l < 0$ are zero. Furthermore, the growth condition implies that there exists some real numbers δ and $\epsilon > 0$ such that $v_2(a_l) \geq \epsilon\cdot|l|+\delta$ and $v_2(b_l) \geq \epsilon \cdot |l + 1| + \delta$.

Lift the p-power Frobenius $\overline{\sigma}$ on $\mathbb{F}_q$ to the Frobenius substitution σ on R. We extend σ to an endomorphism of $A^{\dagger}_{m,f}$ by mapping x to x^2 and y to y^σ, with

$$(y^\sigma)^2 - x^{2m} y^\sigma - f(x)^\sigma = 0 \text{ and } y^\sigma \equiv y^2 \bmod 2.$$

Using Newton iteration we can compute the solution to the above equations as an element of the 2-adic completion of $A_{m,f}$, but it is not immediately clear that there exists a solution in $A^{\dagger}_{m,f}$. The existence of such a solution follows immediately from a theorem by Bosch [4], but since we need an explicit estimate of the rate of convergence, we prove the following lemma.

Lemma 1. *For $k \geq 1$, let $W_k(x, y) := \sum_{l=-L_k}^{A_k} a_l x^l + \sum_{l=-L_k}^{B_k} b_l x^l y \in A_{m,f}$ satisfy*

$$W_k(x, y)^2 - x^{2m} W_k(x, y) - f(x)^\sigma \equiv 0 \bmod 2^k \text{ and } W_k(x, y) \equiv y^2 \bmod 2$$

with $a_{A_k} \neq 0$, $b_{B_k} \neq 0$, $a_{-L_k} \neq 0$ or $b_{-L_k} \neq 0$ and such that $a_l = 0$ or $v_2(a_l) < k$ for $-L_k \leq l \leq A_k$ and $b_l = 0$ or $v_2(b_l) < k$ for $-L_k \leq l \leq B_k$. Then the degrees A_k, B_k and L_k can be bounded for $k \geq 2$ as

$$\begin{aligned} A_k &\leq 2(k-1)(\deg f - 2m) + 2m, \\ B_k &\leq 2(k-2)(\deg f - 2m) + \deg f - m, \\ L_k &\leq 2(k-1)(2m) - 2m. \end{aligned} \tag{2}$$

`Proof:` An easy calculation shows that $W_1(x, y) = f(x) + x^m y$ and

$$W_2(x, y) = \frac{(f(x)^2 - f(x)^\sigma) + x^{2m} f(x)}{x^{2m}} + y\frac{2x^m f(x) + x^{3m}}{x^{2m}},$$

so that W_2 indeed satisfies the lemma.

Newton iteration on $Y^2 - x^{2m} Y - f(x)^\sigma = 0$ gives

$$W_{k+1} \equiv W_k - \frac{W_k^2 - x^{2m} W_k - f(x)^\sigma}{2W_k - x^{2m}} \bmod 2^{k+1} \equiv \frac{W_k^2 - f(x)^\sigma}{x^{2m}} \bmod 2^{k+1}.$$

Let $\alpha_k(x) := \sum_{l=-L_k}^{A_k} a_l x^l$, $\beta_k(x) := \sum_{l=-L_k}^{B_k} b_l x^l$ such that $W_k = \alpha_k + \beta_k y$. Note that $W_k \equiv W_{k-1} \bmod 2^{k-1}$, so we can define

$$\Delta_{\alpha,k}(x) := \frac{\alpha_k(x) - \alpha_{k-1}(x)}{2^{k-1}} \qquad \text{and} \qquad \Delta_{\beta,k}(x) := \frac{\beta_k(x) - \beta_{k-1}(x)}{2^{k-1}},$$

for $k \geq 1$ with $\mathit{\Delta}_{\alpha,0}(x) := \mathit{\Delta}_{\beta,0}(x) := 0$. It is clear that W_k can be written as

$$W_k = \mathit{\Delta}_{\alpha,1} + 2\mathit{\Delta}_{\alpha,2} + \cdots + 2^{k-1}\mathit{\Delta}_{\alpha,k} + y\left(\mathit{\Delta}_{\beta,1} + 2\mathit{\Delta}_{\beta,2} + \cdots + 2^{k-1}\mathit{\Delta}_{\beta,k}\right).$$

Plugging this into the Newton iteration gives the following equation

$$x^{2m}W_{k+1} \equiv \sum_{\substack{1 \leq i < j \\ i+j-1 < k+1}} 2^{i+j-1}\left(\mathit{\Delta}_{\alpha,i}\mathit{\Delta}_{\alpha,j} + (f(x) + x^m y)\,\mathit{\Delta}_{\beta,i}\mathit{\Delta}_{\beta,j}\right) - f(x)^\sigma +$$

$$y \sum_{i+j-1<k+1} 2^{i+j-1}\mathit{\Delta}_{\alpha,i}\mathit{\Delta}_{\beta,j} + \sum_{2(i-1)<k+1} 2^{2(i-1)}\left(\mathit{\Delta}_{\alpha,i}^2 + (f(x) + x^m y)\mathit{\Delta}_{\beta,i}^2\right) \bmod 2^{k+1}.$$

Since $\deg \mathit{\Delta}_{\alpha,i} \leq A_i$ and $\deg \mathit{\Delta}_{\beta,i} \leq B_i$, we get that A_{k+1} is less or equal than

$$\max\Bigg(\deg f^\sigma, \max_{i+j<k+2}(A_i + A_j, B_i + B_j + \deg f),$$
$$\max_{2i<k+3}(2A_i, 2B_i + \deg f)\Bigg) - 2m.$$

Using the bounds given in (2) for A_i and B_i we see that A_{k+1} also satisfies the bounds (2). Note that we have to take into account the values for $\deg \mathit{\Delta}_{\alpha_1} = \deg f$ and $\deg \mathit{\Delta}_{\beta_1} = m$ since these do not satisfy the bounds (2), but this does not cause any problems. A similar reasoning for B_{k+1} and L_{k+1} shows that these also satisfy the given bounds. □ **Remark.** If we want to compute an approximation $W_N(x,y)$ of y^σ modulo 2^N for a certain precision N, then the *total degree* $A_N + L_N$ of the Laurent polynomials is bounded by $2(N-1)\deg f$.

The above lemma indeed shows that we can lift the q-power Frobenius $\overline{F}$ to an endomorphism F on the dagger ring $A^\dagger_{m,f}$; it suffices to take $F := \sigma^n$. Before we can actually compute the zeta function using the Lefschetz fixed point theorem, we need to determine a basis of the K-vectorspace $H^1(\overline{A}_{m,\overline{f}}/K)$.

We first prove that $x^i y\,dx$ with $i = 0, \ldots, 2g-1$, and $\frac{dx}{x}$ if $m > 0$, form a basis for the algebraic de Rham cohomology $H^1_{DR}(A_{m,f}/K)$ of $A_{m,f}$. The extra $\frac{dx}{x}$ for $m > 0$ is caused by the fact that we removed the point $(0,0)$ from the curve $C_{m,f}$. Every element of $H^1_{DR}(A_{m,f}/K)$ can be written as a linear combination of differentials of the form $x^k y^l\,dx$, $x^k y^l\,dy$ with $k \in \mathbb{Z}$, $l \in \mathbb{N}$ and $k \geq 0$ if $m = 0$. Using the equation of the curve, we can reduce to the case $l = 0$ or 1. Since $d(x^k y)$ and $d(x^k y^2)$ are exact, we conclude that $H^1_{DR}(A_{m,f}/K)$ is generated by differentials of the form $x^i y\,dx$ with $i \in \mathbb{Z}$ ($i \in \mathbb{N}$ for $m = 0$) and $\frac{dx}{x}$ if $m > 0$.

Rewriting the equation of the curve as $(2y - x^m)^2 = 4f + x^{2m}$ and differentiating gives the equality $(2y - x^m)\,d(2y - x^m) = (2f' + mx^{2m-1})\,dx$. For all $k > 0$ we therefore have

$$\begin{aligned} x^k(2f' + mx^{2m-1})(2y - x^m)\,dx &= x^k(2y - x^m)^2\,d(2y - x^m) \\ &\equiv -\frac{k}{3}x^{k-1}(2y - x^m)^3\,dx \\ &= -\frac{k}{3}x^{k-1}(4f + x^{2m})(2y - x^m)\,dx, \end{aligned} \qquad (3)$$

where $\equiv$ means equality modulo exact differentials. Thus we conclude that $\left[x^k(2f' + mx^{2m-1}) + \frac{k}{3}x^{k-1}(4f + x^{2m})\right] y\ dx$ is exact. The polynomial between brackets has degree $2g + k$ and non-zero leading coefficient $2(2g + 1) + 4\frac{k}{3} \neq 0$. A similar argument for $k = 0$ shows that $(2f' + mx^{2m-1})y\ dx$ is exact and clearly has degree $2g$ in x. With these formulae we can express $x^{2g+k}y\ dx$ for $k \geq 0$ as a linear combination of $x^iy\ dx$ with $0 \leq i < 2g$.

For $m > 0$ and $k < 0$ the formulae 3 are still valid, but now the conclusion is that $\left[x^k(2f' + mx^{2m-1}) + \frac{k}{3}x^{k-1}(4f + x^{2m})\right] y\ dx + \beta\frac{dx}{x}$ is exact for some suitable element $\beta \in K$. The Laurent polynomial between brackets has valuation at zero k, since $f(0) = 0$, but $f'(0) \neq 0$. The term x^k has coefficient $2f'(0)(1+\frac{2k}{3})$ which clearly is different from zero, since $f'(0) \neq 0$. Therefore we can express all differentials of the form $x^ky\ dx$ with $k < 0$ as a linear combination of $x^iy\ dx$ for $i = 0, \ldots, 2g - 1$ and $\frac{dx}{x}$ if $m > 0$.

A consequence of Lemma 3 will be that all these differential forms are linear independent and thus form a basis for the algebraic de Rham cohomology $H^1_{DR}(A_{m,f}/K)$. To show that this is also a basis of the Monsky-Washnitzer cohomology $H^1(\overline{A}_{m,\overline{f}}/K)$, we need to bound the denominators which are introduced during the reduction process. Therefore we prove the following two lemmata.

Lemma 2. *Let $A := R[x, y]/(y^2 - x^my - f(x))$ and suppose that*

$$x^ry\ dx = \sum_{i=0}^{2g-1} a_ix^iy\ dx + ds, \tag{4}$$

with $r \in \mathbb{N}$, $a_i \in K$ and $s \in A \otimes K$. Then $2^ca_i \in R$, $2^{c'}s - \beta \in A$, where $c = 3 + \lfloor\log_2(r + g + 1)\rfloor$, $c' = 1 + c + \lfloor\log_2(2g + m)\rfloor$ and β some suitable element in K.

Proof: The proof has two distinct parts. The first part is similar to Kedlaya's argument in [12, Lemma 3], and is based on a local analysis around the point at infinity of the curve $C_{m,f}$. Put $t = x^g/y$, then one easily verifies that

$$x = t^{-2}\left(1 + \sum_{j=1}^{\infty}\alpha_jt^j\right), y = t^{-2g-1}\left(1 + \sum_{j=1}^{\infty}\beta_jt^j\right), \tag{5}$$

with $\alpha_j, \beta_j \in R$. To see this, put $z = 1/x$, rewrite the equation of the curve $C_{m,f}$ as $z - z^{g-m+1} - t^2z^{2g+1}f(1/z) = 0$ and write z as a power series in t using Newton iteration. The relation (4) can be rewritten as

$$2^{c-1}x^r(2y - x^m)\ dx = \sum_{i=0}^{2g-1} 2^{c-1}a_ix^i(2y - x^m)\ dx + \ dS, \tag{6}$$

with $S \in A \otimes K$. Considering the involution of A which sends x to x and $2y - x^m$ to $-(2y - x^m)$, we see that we can write $S = \sum_{i=0}^{N} A_ix^i(2y - x^m)$, with N big enough and $A_i \in K$. This yields

$$2^{c-1}x^r(2y-x^m)\ dx - \sum_{i=0}^{2g-1} 2^{c-1}a_ix^i(2y-x^m)\ dx = d\left(\sum_{i=0}^{N} A_ix^i(2y - x^m)\right). \tag{7}$$

In the above equation we express x and y in terms of t using equalities (5). Since $x^i y = t^{-2i-2g-1} + \cdots$, we get $x^i(2y - x^m)\ dx = (-4t^{-2i-2g-4} + \cdots)\ dt$, which yields

$$2^{c-1} \sum_{j=-\max(2r+2g+4,6g+2)} \gamma_j t^j\ dt = d\left(\sum_{i=0}^{N} 2A_i(t^{-2i-2g-1} + \cdots) - A_i(t^{-2i-2m} + \cdots)\right),$$

with $\gamma_j \in K$ for all j and $\gamma_j \in R$ when $j < -2(2g-1) - 2g - 4 = -6g - 2$. Integrating with respect to t and dividing by 2 gives

$$\sum_{j \geq -\max(2r+2g+3,6g+1)} \gamma_j' t^j = \sum_{i=0}^{N} A_i(t^{-2i-2g-1} + \cdots) - \sum_{i=0}^{N} \frac{A_i}{2}(t^{-2i-2m} + \cdots), \quad (8)$$

with $\gamma_j' \in K$ for all j and $\gamma_j' \in R$ when $j < -6g - 1$. Indeed the integration process introduces denominators which become integral after multiplication with $2^{\lfloor \log(2r+2g+2) \rfloor} = 2^{c-2}$ if $r \geq 2g - 1$. A first consequence of (8) is that $A_i = 0$ for all $i > \max(r+1, 2g)$. We claim that (8) implies that $A_i \in R$ for all $i > 2g$. Suppose the claim is false. Then let i_0 be the largest integer with $i_0 > 2g$ and $A_{i_0} \notin R$. Note that $-2i_0 - 2g - 1 < -6g - 1$, since $i_0 > 2g$. Hence the monomials in the left hand side of (8) with degree $\leq -2i_0 - 2g - 1$ have coefficients in R. Moreover the monomials of degree $< -2i_0 - 2g - 1$, in the first sum in the right hand side of (8) also have coefficients in R, but this is false for the monomial of degree $-2i_0 - 2g - 1$. Hence the second sum in the right hand side of (8) contains a monomial of degree $-2i_0 - 2g - 1$ whose coefficient is not in R. That means that there is a maximal i_1 with $A_{i_1}/2 \notin R$ and $-2i_1 - 2m \leq -2i_0 - 2g - 1$. Because of parity we have that $-2i_1 - 2m < -2i_0 - 2g - 1$. Hence the right hand side of (8) contains a monomial of degree $-2i_1 - 2m < -2i_0 - 2g - 1$ whose coefficient is not in R. But this contradicts what we said about the left hand side. This finishes the claim that $A_i \in R$ for all $i > 2g$.

We now turn to the second part of the proof. Note that $(2y - x^m)^2 = v(x)$ with $v(x) := 4f + x^{2m}$. Moreover, $d(2y - x^m) = \frac{w(x)}{2y - x^m}\ dx$, where $w(x) := 2f' + mx^{2m-1}$. We will use these formulae to deduce from (7) a relation which does not involve y. For this purpose we multiply (7) with $\frac{2y-x^m}{dx} = \frac{w(x)}{d(2y-x^m)}$ obtaining

$$2^{c-1}x^r v(x) - \sum_{i=0}^{2g-1} 2^{c-1} a_i x^i v(x) = \sum_{i=0}^{N} A_i i x^{i-1} v(x) + \sum_{i=0}^{N} A_i x^i w(x).$$

We rewrite this in the form

$$\left(\sum_{i=0}^{2g-1} 2^{c-1} a_i x^i\right) v(x) + \left(\sum_{i=0}^{2g} A_i i x^{i-1}\right) v(x) + \left(\sum_{i=0}^{2g} A_i x^i\right) w(x) = F(x), \quad (9)$$

where

$$F(x) := 2^{c-1}x^r v(x) - \sum_{i=2g+1}^{N} A_i i x^{i-1} v(x) - \sum_{i=2g+1}^{N} A_i x^i w(x) \qquad (10)$$

is a polynomial over R, since $A_i \in R$ for all $i > 2g$. To get rid of the disturbing factor 2 in the definition of $w(x)$, we consider $u(x) := \frac{1}{2}(xw(x) - mv(x)) = xf' - 2mf$. We rewrite (9) in such a way that $w(x)$ gets replaced by $u(x)$:

$$\left(\sum_{i=0}^{2g}(2^{c-1}a_{i-1} + iA_i + mA_i)x^i\right) v(x) + \left(\sum_{i=0}^{2g} 2A_i x^i\right) u(x) = xF(x), \qquad (11)$$

with the convention that $a_{-1} = 0$. We consider (11) as a linear system of $4g+2$ equations in the unknowns $2^{c-1}a_{i-1} + iA_i + mA_i$ and $2A_i$ for $i = 0, \ldots, 2g$. The determinant of this system is the resultant $\mathrm{Res}(v, u)$ of u and v, because $\deg v(x) = \deg u(x) = 2g + 1$. Since the leading coefficient of u is a unit, we have $\mathrm{Res}(u, v) = \text{ unit } \cdot \prod_{u(\theta)=0} v(\theta)$, where θ ranges over all roots of u in the algebraic closure of K. All these roots θ have non-negative valuation.

Suppose first that $m = 0$. Then $\mathrm{Res}(u, v)$ is a unit in R since $v(\theta) = 4f(\theta) + 1$ is a unit for each root θ of u. The determinant of the system being a unit, we conclude that $2A_i$ and $2^{c-1}a_{i-1} + iA_i + mA_i$ are in R for $i = 0, \ldots, 2g$. Hence $2^c a_i \in R$ and $2S \in A$. So for $m = 0$ the lemma then follows directly from (6).

Suppose now that $m \geq 1$. The restrictions on f imply that $f(0) = 0$ and $f'(0) \neq 0 \bmod 2$. Hence 0 is a common zero of u and v and $\mathrm{Res}(u, v) = 0$. From (10) it follows that $F(0) = 0$, hence $A_0 = 0$ by (9), since $w(0) = 2f'(0) \neq 0$. We now consider (11) divided by x^2 as a linear system of $4g$ equations in the unknowns $2^{c-1}a_{i-1} + iA_i + mA_i$ and $2A_i$ for $i = 1, \ldots, 2g$. The determinant of this system is the resultant $\mathrm{Res}(\frac{v}{x}, \frac{u}{x})$. Let θ be a root of $u/x = f'(x) - 2mf(x)/x$, then θ has valuation zero since $f'(0) \neq 0 \bmod 2$. Hence $v(\theta) = 4f(\theta) + \theta^{2m}$ is a unit. Thus $\mathrm{Res}(\frac{v}{x}, \frac{u}{x})$ is a unit and both $2A_i$ and $2^{c-1}a_{i-1} + iA_i + mA_i$ are in R for $i = 1, \ldots, 2g$. We now continue as in the case $m = 0$. This ends the proof of the lemma. □

Remark. Lemma 2 remains valid when we replace $\sum_{i=0}^{2g-1}$ by $\sum_{i=\kappa}^{2g-1+\kappa}$ whenever $r \geq \kappa \in \mathbb{N}$. The proof is the same, except that we also have to show that $A_i = 0$ for all $i < \kappa$. This follows from (7) by a local analysis at a point on the curve with $x = 0$.

Lemma 3. *With the above notation and $m > 0$, suppose that*

$$x^{-r}y\, dx = \sum_{i=0}^{2g-1} a_i x^i y\, dx + b\frac{dx}{x} + ds, \qquad (12)$$

where $r \in \mathbb{N}$, $a_i, b \in K$ and $s \in A_{m,f} \otimes K$. Then $2^c a_i \in R$, $2^{c'} b \in R$, $2^{c'} s - \beta \in A_{m,f}$, with $c = 3 + \lfloor \log_2(r+1) \rfloor$, $c' = 1 + c + \lfloor \log_2(2g+m) \rfloor$ and $\beta \in K$.

Remark. Actually one can take $c = 3 + \lfloor \log_2(r-2) \rfloor$ when $r \geq 3$ and $c = 0$ when $0 \leq r \leq 2$.

`Proof:` The proof again consists of two distinct parts. The first part is similar to Kedlaya's argument in [12, Lemma 2] and is based on a local analysis around the ramification point $(0, 0)$ on the curve. In the completion of the local ring of the curve at $(0,0)$ we can write

$$x = \gamma_2 y^2 + \sum_{j \geq 3} \gamma_j y^j, \tag{13}$$

with $\gamma_j \in R$ and γ_2 a unit in R. Indeed, to see this use the equation of the curve and the conditions $f(0) = 0$, $f'(0) \neq 0 \bmod 2$, to express x as a power series in y using Newton iteration.

Considering the involution as in the proof of Lemma 2, we can transform relation (12) to the form

$$2^{c-1}x^{-r}(2y - x^m)\,dx - \sum_{i=0}^{2g-1} 2^{c-1}a_i x^i(2y - x^m)\,dx = d\left(\sum_{i=-N}^{M} A_i x^i(2y - x^m)\right),$$

with N and M large enough integers. Using the expansion at infinity given by the formulae (5) in the proof of Lemma 2 and substituting them in the above equation, one verifies that we can take $M = 2g$.

Expressing x in terms of y using (13) and dividing by 2 we obtain

$$2^{c-2} \sum_{j \geq -2r+2} \gamma'_j y^j\, dy =$$
$$d\Big(\sum_{i=-N}^{2g} A_i(\gamma_2^i y^{2i+1} + \cdots)\Big) - d\Big(\sum_{i=-N}^{2g} \frac{A_i}{2}(\gamma_2^{i+m} y^{2i+2m} + \cdots)\Big),$$

with $\gamma'_j \in K$ for all j and $\gamma'_j \in R$ when $j \leq 0$. Integrating the left hand side of this equation with respect to y yields a series whose terms of degree ≤ 1 have coefficients in R. Thus the same argument as in the proof of Lemma 2 shows that $A_i \in R$ for all $i \leq 0$. Moreover if $r = 0$, then $A_i = 0$ when $i \leq 0$. This terminates the first part of the proof.

We still have to proof that $2^c a_i \in R$ for $i = 0, \ldots, 2g-1$ and that $2A_i \in R$ for $i = 1, \ldots, 2g$. This follows by the same argument as in the second part of the proof of Lemma 2. However, in the present situation A_0 might not be zero, but we proved already that $A_0 \in R$. Therefore we bring the terms which contain A_0 to the other side in equation (11) from the proof of Lemma 2. This then ends the proof of Lemma 3. □

Remark. Lemma 3 remains valid when we replace $\sum_{i=0}^{2g-1}$ by $\sum_{i=-\kappa}^{2g-1-\kappa}$ whenever $r \geq \kappa \in \mathbb{N}$. The proof is exactly the same.

Remark. If $r = 0$, then in the above proof the A_i are zero for all $i \leq 0$, and for $0 \leq i \leq 2g-1$ the a_i are completely determined by (11) as we saw by considering resultants. This shows that the $x^i y\, dx$ for $i = 0, \ldots, 2g-1$ and $\frac{dx}{x}$ are linearly independent in $H^1_{DR}(A_{m,f}/K)$.

Lemma 2 and 3 show that the basis for $H^1_{DR}(A_{m,f}/K)$ is a generating set for $H^1(\overline{A}_{m,\overline{f}}/K)$, since the reduction process converges. Indeed, for $a_k x^k y \in A^{\dagger}_{m,f}$ the valuation of a_k grows as a linear function of $|k|$, while the valuation of the denominators introduced during the reduction of $a_k x^k y\, dx$ are only logarithmic in $|k|$.

The Monsky-Washnitzer cohomology $H^1(\overline{A}_{m,\overline{f}}/K)$ is the direct sum of the $\imath$-invariant part $H^1(\overline{A}_{m,\overline{f}}/K)^+$ on which $\imath$ acts trivially and the $\imath$-anti invariant part $H^1(\overline{A}_{m,\overline{f}}/K)^-$ on which $\imath$ acts by multiplication by -1. Note that $\frac{dx}{x}$ is a basis for the invariant part $H^1(\overline{A}_{m,\overline{f}}/K)^+$ for $m > 0$ and the Frobenius acts on it by multiplication with q. Hence for $m > 0$ the Lefschetz fixed point theorem yields

$$\begin{aligned}\#\overline{C}_{m,\overline{f}}(\mathbb{F}_{q^k}) &= 1 + \#\overline{C}'_{m,\overline{f}}(\mathbb{F}_{q^k})\\ &= 1 + \mathrm{Tr}\left(q^k F_*^{-k} | H^0(\overline{A}_{m,\overline{f}}/K)\right) - \mathrm{Tr}\left(q^k F_*^{-k} | H^1(\overline{A}_{m,\overline{f}}/K)\right)\\ &= 1 + q^k - \mathrm{Tr}\left(q^k F_*^{-k} | H^1(\overline{A}_{m,\overline{f}}/K)^+\right)\\ &\qquad - \mathrm{Tr}\left(q^k F_*^{-k} | H^1(\overline{A}_{m,\overline{f}}/K)^-\right)\\ &= q^k - \mathrm{Tr}\left(q^k F_*^{-k} | H^1(\overline{A}_{m,\overline{f}}/K)^-\right).\end{aligned}$$

Let $\widetilde{C}_{m,\overline{f}}$ be the unique smooth projective curve birational to $\overline{C}_{m,\overline{f}}$, then

$$\#\widetilde{C}_{m,\overline{f}}(\mathbb{F}_{q^k}) = q^k + 1 - \mathrm{Tr}\left(q^k F_*^{-k} | H^1(\overline{A}_{m,\overline{f}}/K)^-\right) = q^k + 1 - \sum_{i=1}^{2g} \alpha_i^k,$$

where α_i are the eigenvalues of qF_*^{-1} on $H^1(\overline{A}_{m,\overline{f}}/K)^-$. By the Weil conjectures there exists a polynomial $\chi(t) \in \mathbb{Z}[t]$ of the form $t^{2g} + a_1 t^{2g-1} + \cdots + a_{2g}$, whose roots $\beta_1, \ldots, \beta_{2g}$ satisfy $\beta_i \beta_{g+i} = q$ for $i = 1, \ldots, g$, $|\beta_i| = \sqrt{q}$ for $i = 1, \ldots, 2g$ and $\#\widetilde{C}_{m,\overline{f}}(\mathbb{F}_{q^k}) = q^k + 1 - \sum_{i=1}^{2g} \beta_i^k$ for all $k > 0$. This implies that we can label the β's such that $\alpha_i = \beta_i$ for $i = 1, \ldots, 2g$. Since $\alpha_i \alpha_{g+i} = q$, the α_i are also the eigenvalues of F_* on $H^1(\overline{A}_{m,\overline{f}}/K)^-$. It is well known that the zeta function $Z(\widetilde{C}_{m,\overline{f}}/\mathbb{F}_q; t)$ can be written as $Z(\widetilde{C}_{m,\overline{f}}/\mathbb{F}_q; t) = \frac{t^{2g}\chi(1/t)}{(1-t)(1-qt)}$. Therefore, it is sufficient to compute $\chi(t)$ as the characteristic polynomial of F_* on $H^1(\overline{A}_{m,\overline{f}}/K)^-$.

4 Detailed Algorithm and Complexity

Using the results of the previous section, we describe an algorithm for computing the characteristic polynomial of Frobenius $\chi(t)$ and the zeta function of a smooth projective Artin-Schreier curve $\widetilde{C}_{m,\overline{f}}$ of genus g over $\mathbb{F}_q$ with $q = 2^n$. We have shown that we can compute $\chi(t) = t^{2g} + a_1 t^{2g-1} + \cdots + a_{2g}$ as the characteristic polynomial of F_* on $H^1(\overline{A}_{m,\overline{f}}/K)^-$. The Weil conjectures imply that $q^{g-i} a_i = a_{2g-i}$, so it suffices to compute $a_1, \ldots, a_g$, and that for $i = 1, \ldots, g$ the a_i can be bounded by

$$|a_i| \le \binom{2g}{i} q^{i/2} \le \binom{2g}{g} q^{g/2} \le 2^{2g} q^{g/2}.$$

Thus to determine the zeta function, we have to compute the action of F_* on a basis of $H^1(\overline{A}_{m,\overline{f}}/K)^-$ modulo 2^B with $B \geq \left\lceil \log_2 \left(2\binom{2g}{g} q^{g/2} \right) \right\rceil$. However, we need to take into account the loss of precision caused by the reduction process.

Combining Lemmata 1-3 one can prove that it is sufficient to compute with a precision N which satisfies $N - 3 - \lfloor \log_2(2N \deg f + g) \rfloor \geq B$.

Algorithms 1-3 contain a detailed description of the most important subroutines of our algorithm. The function `Artin_Schreier_Zeta_Function` essentially computes an approximation M of the matrix through which the p-th power Frobenius acts on a basis of $H^1(\overline{A}_{m,\overline{f}}/K)^-$. The function `Lift_p_Frobenius_y` computes a sufficiently precise approximation of y^σ using a Newton iteration on the equation $Y^2 - x^{2m}Y - f(x)^\sigma = 0$ and `Series_Invert` computes the inverse of an invertible element of $A^\dagger_{m,f}$ up to precision N. In step 4 of Algorithm 1 we call `Reduce_MW_Cohomology` to express a differential $Gy\ dx$ with $G \in R[x, x^{-1}]$ on a basis of $H^1(\overline{A}_{m,\overline{f}}/K)^-$. The result of this function is a polynomial $S \in K[x]$, with $\deg S < 2g$ such that for a given precision B we have the following equivalence modulo exact forms and invariant forms $G(x, x^{-1})y\ dx \sim R(x)y\ dx \mod 2^B$, where $\mod 2^B$ means modulo $2^B(Ry\ dx + \cdots + Rx^{2g-1}y\ dx)$. Once we have found the matrix M, we compute `Norm(`M`)` $= MM^\sigma \cdots M^{\sigma^{n-1}}$ which is an approximation of the action of Frobenius on $H^1(\overline{A}_{m,f}/K)^-$. Finally, we determine its characteristic polynomial with precision $\left\lceil \log_2 \left(2\binom{2g}{g} q^{g/2} \right) \right\rceil$ and recover the characteristic polynomial of Frobenius $\chi(t)$ from the first g coefficients. Note that M is not necessarily defined over R, so we have to increase B if necessary to obtain the desired precision.

The complexity analysis of the algorithm is similar to Kedlaya's algorithm in [12, Section 5], except that in our case the reduction takes $O(g^{5+\varepsilon}n^{3+\varepsilon})$ time instead of $O(g^{4+\varepsilon}n^{3+\varepsilon})$ time. A detailed complexity analysis can be found in [5], which proves that the zeta function of a genus g Artin-Schreier curve $\widetilde{C}_{m,\overline{f}}$ over a finite field $\mathbb{F}_q$ with $q = 2^n$ elements, can be computed deterministically in $O(g^{5+\varepsilon}n^{3+\varepsilon})$ bit operations with space complexity $O(g^3n^3)$.

5 Implementation and Numerical Results

In this section we compare the efficiency of our algorithm with an algorithm by Lauder and Wan [16], which also runs in $O(g^{5+\varepsilon}n^{3+\varepsilon})$ bit operations and needs $O(g^3n^3)$ storage space. As far as we know, Lauder & Wan's algorithm has not been implemented before.

Table 1 presents running times of our algorithm and Lauder & Wan's algorithm for genus 2 and genus 3 Artin-Schreier curves over various finite fields of characteristic 2 obtained on a Sun UltraSparc III 600 MHz running Solaris 5.8 and MAGMA V2.8-1. In these examples we have taken $B = \left\lceil \log_2 \left(2\binom{2g}{g} q^{g/2} \right) \right\rceil$ and the results were verified by checking the group order of the Jacobian.

Algorithm 1 (Artin_Schreier_Zeta_Function).

IN: *Artin-Schreier curve* $\overline{C}_{m,\overline{f}}$ *over* $\mathbb{F}_q$ *given by equation* $y^2 - x^m y = \overline{f}(x)$.
OUT: *The zeta function* $Z(\widetilde{C}_{m,\overline{f}}/\mathbb{F}_q; t)$.

```
1. Compute $N \in \mathbb{N}$ with $N - 3 - \lfloor \log_2(2N \deg f + g) \rfloor \geq B$;
2. $\overline{f} = \overline{f} - \overline{f}(0) + \sqrt{\overline{f}(0)} x^m$; $f = R[x] \leftarrow \overline{f} \bmod 2^N$;
3. $\alpha_N(x), \beta_N(x) =$ Lift_p_Frobenius_y($m, f, N$);
4. For $i = 0$ To $2g - 1$ Do
    4.1. $Red_i(x) =$ Reduce_MW_Cohomology($2x^{2i+1}\beta_N(x), m, f, B$);
    4.2. For $j = 0$ To $2g - 1$ Do $M[j][i] =$ Coeff($Red_i, j$);
5. $Norm_M = MM^{\sigma} \cdots M^{\sigma^{n-1}} \bmod 2^B$;
6. $\chi(t) =$ Characteristic_Pol($Norm_M$) $\bmod 2^B$;
7. For $i = 0$ To $i = g$ Do
    7.1. If Coeff($\chi, 2g - i$) $>$ $\binom{2g}{i} q^{i/2}$ Then Coeff($\chi, 2g - i$) $-= 2^B$;
    7.2. Coeff($\chi, i$) $= q^{g-i}$ Coeff($\chi, 2g - i$);
8. Return $Z(\widetilde{C}_{m,\overline{f}}/\mathbb{F}_q; t) = \dfrac{t^{2g}\chi(1/t)}{(1-t)(1-qt)}$.
```

Algorithm 2 (Lift_p_Frobenius_y).

IN: *Artin-Schreier curve* $C_{m,f}$ *over* R *and precision* N.
OUT: $\alpha_N, \beta_N \in R[x, x^{-1}]$ *with* $y^{\sigma} \equiv \alpha_N(x, x^{-1}) + \beta_N(x, x^{-1}) y \bmod 2^N$.

```
1. If $N = 1$ Then $\alpha_N = f(x)$; $\beta_N = x^m$;
2. Else
    2.1. $N' = \lceil \frac{N}{2} \rceil$;
    2.2. $\alpha_{N'}, \beta_{N'} =$ Lift_p_Frobenius_y($m, f(x), N'$);
    2.3. $\gamma_N, \delta_N =$ Series_Invert($1 - \frac{2(\alpha_{N'}(x) + \beta_{N'}(x) y)}{x^{2m}}, m, f(x), N$);
    2.4. $\mu_N \equiv -\alpha_{N'} + x^{-2m}(\alpha_{N'}^2 + \beta_{N'}^2 f(x) - f(x)^{\sigma}) \bmod 2^N$;
    2.5. $\nu_N \equiv -\beta_{N'} + x^{-2m}(2\alpha_{N'}\beta_{N'} + \beta_{N'}^2 x^m) \bmod 2^N$;
    2.6. $\alpha_N \equiv \alpha_{N'} + \mu_N \gamma_N + \nu_N \delta_N f(x) \bmod 2^N$;
    2.7. $\beta_N \equiv \beta_{N'} + \mu_N \delta_N + \nu_N(\gamma_N + \delta_N x^m) \bmod 2^N$;
3. Return $\alpha_N, \beta_N$.
```

Algorithm 3 (`Reduce_MW_Cohomology`).

IN: *Artin-Schreier curve* $C_{m,f}$*, precision* B *and element* $G \in R[x, x^{-1}]$.
OUT: $S \in K[x]$*, with* $\deg S < 2g$ *such that* $Sy\,dx \sim Gy\,dx \bmod 2^B$.

```
1. Compute $N \in \mathbb{N}$ with $N - 3 - \lfloor \log_2(2N \deg f + g) \rfloor \geq B$;
2. $D =$ Degree($G$); $V =$ Valuation($G$); $T = G$;
3. For $i = D$ To $2g$ By $-1$
    3.1. $P \equiv x^{i-2g}(2f' + mx^{2m-1}) + \frac{i-2g}{3} x^{i-2g-1}(4f + x^{2m}) \bmod 2^N$;
    3.2. $T \equiv T -$ (Coeff($T, i$)$\cdot P$)$/(2(2g+1) + \frac{4(i-2g)}{3}) \bmod 2^N$;
4. For $i = V$ To $-1$
    4.1. $P \equiv x^i(2f' + mx^{2m-1}) + \frac{i}{3} x^{i-1}(4f + x^{2m}) \bmod 2^N$;
    4.2. $T \equiv T -$ (Coeff($T, i$)$\cdot P$)$/(2(1 + \frac{2i}{3})f'(0)) \bmod 2^N$;
5. Return $S \equiv T \bmod 2^B$.
```

Table 1. Running times for genus 2 and genus 3 Artin-Schreier curves over $\mathbb{F}_{2^n}$ of Denef-Vercauteren (D-V) vs. Lauder-Wan (L-W) algorithm.

Genus 2 curves			Genus 3 curves		
Field Size	Time D-V (s)	Time L-W (s)	Field Size	Time D-V (s)	Time L-W (s)
13 bits	2.7	6.0	11 bits	7.0	24.3
23 bits	12.9	22.9	17 bits	29.6	85.1
37 bits	93.5	141	23 bits	76.2	219
47 bits	178	259	31 bits	189	501
59 bits	347	465	41 bits	663	1231
71 bits	983	973	47 bits	1067	1773
83 bits	1207	1493	59 bits	1724	3156

6 Conclusion

We have presented an extension of Kedlaya's algorithm to Artin-Schreier curves over finite fields of characteristic 2. The resulting algorithm runs in $O(g^{5+\varepsilon} n^{3+\varepsilon})$ bit operations and needs $O(g^3 n^3)$ storage space for a genus g curve over $\mathbb{F}_{2^n}$. The ideas presented in this paper can also be used to devise an algorithm for computing the zeta function of an arbitrary hyperelliptic curve over a finite field of characteristic 2 as shown in [5].

References

1. S. Arita. Algorithms for computations in jacobians of C_{ab} curve and their application to discrete-log-based public key cryptosystems. In *Proceedings of Conference on The Mathematics of Public Key Cryptography*, Toronto, June 1999.
2. A.O.L. Atkin. The number of points on an elliptic curve modulo a prime. *Series of e-mails to the* NMBRTHRY *mailing list*, 1992.
3. I.F. Blake, G. Seroussi, and N.P. Smart. *Elliptic curves in cryptography.* volume 265 of *London Mathematical Society Lecture Note Series*, 1999.
4. S. Bosch. A rigid analytic version of M. Artin's theorem on analytic equations. *Math. Ann.*, 255:395–404, 1981.
5. J. Denef and F. Vercauteren. An extension of Kedlaya's algorithm to hyperelliptic curves in characteristic 2. *Preprint*, 2002.
6. N. Elkies. Elliptic and modular curves over finite fields and related computational issues. *Computational Perspectives on Number Theory*, pages 21–76, 1998.
7. R. Elkik. Solutions d'équations a coefficients dans un anneau henselien. *Ann. Scient. Ec. Norm. Syp.*, 6(4):553–604, 1973.
8. M. Fouquet, P. Gaudry, and R. Harley. On Satoh's algorithm and its implementation. *J. Ramanujan Math. Soc.*, 15:281–318, 2000.
9. S. Galbraith, S. Paulus, and N. Smart. Arithmetic on superelliptic curves. *Math. Comp.*, 71(237):393–405, 2002.
10. P. Gaudry and N. Gürel. An extension of Kedlaya's algorithm for counting points on superelliptic curves. In *Advances in Cryptology - ASIACRYPT 2001*, Lecture Notes in Computer Science, 2001.
11. P. Gaudry and R. Harley. Counting points on hyperelliptic curves over finite fields. *Bosma, Wieb (ed.), ANTS-IV, Lect. Notes Comput. Sci. 1838, 313-332* , 2000.
12. K.S. Kedlaya. Counting points on hyperelliptic curves using Monsky-Washnitzer cohomology. Preprint 2001.
13. N. Koblitz. Elliptic curve cryptosystems. *Math. Comp.*, 48:203–209, 1987.
14. N. Koblitz. Hyperelliptic cryptosystems. *J. Cryptology*, 1(3):139–150, 1989.
15. A.G.B. Lauder and D. Wan. Counting points on varieties over finite fields of small characteristic. Preprint 2001.
16. A.G.B. Lauder and D. Wan. Computing zeta functions of Artin-Schreier curves over finite fields. Preprint 2001.
17. R. Lercier. *Algorithmique des courbes elliptiques dans les corps finis.* PhD thesis, Laboratoire d'Informatique de l'École polytechnique (LIX), 1997.
18. V. Miller. Uses of elliptic curves in cryptography. *Advances in Cryptology - ASIACRYPT '91, Lecture notes in Computer Science*, 218:460–469, 1993.
19. P. Monsky and G. Washnitzer. Formal cohomology. I. *Ann. of Math.*, 88:181–217, 1968.
20. P. Monsky. Formal cohomology. II: The cohomology sequence of a pair. *Ann. of Math.*, 88:218–238, 1968.
21. P. Monsky. Formal cohomology. III: Fixed point theorems. *Ann. of Math.*, 93:315–343, 1971.
22. P. Monsky. *p-adic analysis and zeta functions.* Lectures in Mathematics, Department of Mathematics Kyoto University. 4. Tokyo, Japan, 1970.
23. J. Pila. Frobenius maps of abelian varieties and finding roots of unity in finite fields. *Math. Comp.*, 55(192):745–763, 1990.
24. B. Poonen. Computational aspects of curves of genus at least 2. *Cohen, Henri (ed.), ANTS-II, Lect. Notes Comput. Sci. 1122, 283-306* , 1996.

25. T. Satoh. The canonical lift of an ordinary elliptic curve over a finite field and its point counting. *J. Ramanujan Math. Soc.*, 15:247–270, 2000.
26. R. Schoof. Elliptic curves over finite fields and the computation of square roots mod p. *Math. Comp.*, 44:483–494, 1985.
27. B. Skjernaa. Satoh's algorithm in characteristic 2. *To appear in Math. Comp.*, 2000.
28. M. van der Put. The cohomology of Monsky and Washnitzer. *Mém. Soc. Math. France*, 23:33–60, 1986.
29. F. Vercauteren, B. Preneel, and J. Vandewalle. A memory efficient version of Satoh's algorithm. In *Advances in Cryptology - EUROCRYPT 2001*, number 2045 in Lecture Notes in Computer Science, pages 1–13, 2001.

Implementing the Tate Pairing

Steven D. Galbraith[1,*], Keith Harrison[2], and David Soldera[2]

[1] Mathematics Department, Royal Holloway University of London,
Egham, Surrey TW20 0EX, UK.
Steven.Galbraith@rhul.ac.uk

[2] Hewlett-Packard Laboratories, Bristol,
Filton Road, Stoke Gifford, Bristol BS34 8QZ, UK.
keith_harrison@hp.com, David_Soldera@hplb.hpl.hp.com

Abstract. The Tate pairing has found several new applications in cryptography. This paper provides methods to quickly compute the Tate pairing, and hence enables efficient implementation of these cryptosystems. We also give division-free formulae for point tripling on a family of elliptic curves in characteristic three. Examples of the running time for these methods are given.

1 Introduction

The Weil and Tate pairings have recently been used to construct cryptosystems, such as the identity-based key exchange and signature schemes of Sakai, Ohgishi and Kasahara [13], the tripartite Diffie-Hellman protocol of Joux [9], the escrow El Gamal system of Verheul [15] (see [3] for a better solution), the identity-based encryption scheme of Boneh and Franklin [3], the credential scheme of Verheul [16], the short signature scheme of Boneh, Lynn and Shacham [4] and many more.

For most of these applications either the Weil pairing or Tate pairing may be used (these pairings both provide good functionality for use in cryptosystems). In practice, as has been observed in [4,7], the Tate pairing is more efficient for computation (we give some timings in Section 10.1 which show how much slower the Weil pairing is). If these cryptosystems are to be adopted for practical applications it is essential to provide methods which improve the performance of Tate pairing computations.

In this paper we give techniques which enable efficient computation of the Tate pairing for cryptographic applications. Some of these techniques are familiar from the literature on fast point exponentiation for elliptic curve cryptography, but most of them are specific to the cryptographic application of the Tate pairing.

We now summarise the paper. Sections 2 and 3 describe the basics of the Tate pairing and Miller's algorithm. Section 4 indicates how the Tate pairing is used in cryptosystems. Section 5 contains the core observations which dictate the development of our later techniques. Section 6 shows how properties of the

[*] This author thanks Hewlett-Packard Laboratories, Bristol for support.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 324–337, 2002.

group order (namely, the size of the large prime, and small Hamming weights) may be used to give improved performance. Section 7 introduces formulae for elliptic curve point tripling in characteristic three, and shows how this leads to an efficient base-three Miller's algorithm. Section 8 and 9 discuss the implementation of the finite field arithmetic. Section 10 contains some of our timing results.

We must note that Barreto, Kim, Lynn and Scott [1] have independently obtained many fine results on this topic.

2 The Tate Pairing

The Weil pairing was introduced into cryptography by Menezes, Okamoto and Vanstone [12] who used it to attack the elliptic curve discrete logarithm problem on certain elliptic curves. The Tate pairing was introduced into cryptography by Frey and Rück [5] in their extension of the work of Menezes, Okamoto and Vanstone.

Let E be an elliptic curve over a finite field $\mathbb{F}_q$. We write $\mathcal{O}_E$ for the point at infinity on E. Let l be a positive integer which is coprime to q. In most applications l is a prime and $l|\#E(\mathbb{F}_q)$. Let k be a positive integer such that the field $\mathbb{F}_{q^k}$ contains the lth roots of unity (in other words, $l|(q^k-1)$). Let $G = E(\mathbb{F}_{q^k})$ and write $G[l]$ for the subgroup of points of order l and G/lG for the quotient group (which is also a group of exponent l). Then the Tate pairing is a mapping

$$\langle\cdot,\cdot\rangle : G[l] \times G/lG \to \mathbb{F}_{q^k}^*/(\mathbb{F}_{q^k}^*)^l. \tag{1}$$

The quotient group on the right hand side of (1) can be thought of as the set of equivalence classes of $\mathbb{F}_{q^k}^*$ under the equivalence relation $a \equiv b$ if and only if there exists $c \in \mathbb{F}_{q^k}^*$ such that $a = bc^l$. We call this relation 'equivalence modulo lth powers'.

The Tate pairing satisfies the following properties [5]:

1. (Well-defined) $\langle\mathcal{O}_E, Q\rangle \in (\mathbb{F}_{q^k}^*)^l$ for all $Q \in G$ and $\langle P, Q\rangle \in (\mathbb{F}_{q^k}^*)^l$ for all $P \in G[l]$ and all $Q \in lG$.
2. (Non-degeneracy) For each point $P \in G[l] - \{0\}$ there is some point $Q \in G$ such that $\langle P, Q\rangle \notin (\mathbb{F}_{q^k}^*)^l$.
3. (Bilinearity) For any integer n, $\langle[n]P, Q\rangle \equiv \langle P, [n]Q\rangle \equiv \langle P, Q\rangle^n$ modulo lth powers.

The Tate pairing is defined as follows. Given the point P there is a function g such that the divisor of g is equal to $l(P) - l(\mathcal{O}_E)$ (see [14] for an introduction to divisors). There is a divisor D which is equivalent to $(Q) - (\mathcal{O}_E)$ such that the support of D is disjoint from the support of g. Then the value of the Tate pairing (up to lth powers) is

$$\langle P, Q\rangle = g(D)$$

where $g(D) = \prod_i g(P_i)^{n_i}$ if $D = \sum_i n_i(P_i)$.

We emphasise that the Tate pairing is only defined up to a multiple by an lth power in $\mathbb{F}_{q^k}^*$. For most applications in cryptography a unique value is required, and so it is necessary to exponentiate the value of the Tate pairing to the power $(q^k - 1)/l$ (since raising to this power eliminates all multiples of order l).

3 Miller's Algorithm

The Tate pairing can be computed using an algorithm first proposed by Miller [11] in the context of the Weil pairing. This algorithm is also described in [5,6]. Miller's algorithm is basically the usual 'double and add' algorithm for elliptic curve point multiplication combined with an evaluation of certain intermediate functions which are the straight lines used in the addition process.

Before giving the details of this algorithm we recall the elliptic curve addition law (for more details see [2,14]).

Let P and Q be points on an elliptic curve E. Let l_1 be the line through P and Q (if $P = Q$ then l_1 is taken to be the tangent to the curve E at P, if one of P or Q is $\mathcal{O}_E$ then l_1 is a 'vertical line' through the affine point). Then l_1 intersects the cubic curve E at one further point, say R_1. Now let l_2 be the line between R_1 and $\mathcal{O}_E$ (which is a 'vertical line' when R_1 is not equal to $\mathcal{O}_E$). Then l_2 intersects E at a third point R_2 which is defined to be the sum of P and Q.

The lines l_1 and l_2 can be thought of as functions on the curve, and the corresponding principal divisors are

$$(l_1) = (P) + (Q) + (R_1) - 3(\mathcal{O}_E) \text{ and } (l_2) = (R_1) + (R_2) - 2(\mathcal{O}_E).$$

It follows that we have the following equality of divisors

$$(P) - (\mathcal{O}_E) + (Q) - (\mathcal{O}_E) = (R_2) - (\mathcal{O}_E) + (l_1/l_2).$$

Let E be an elliptic curve over $\mathbb{F}_q$ and let P and Q be given points of prime order l for which we want to compute $\langle P, Q \rangle$. Miller's algorithm is given in Figure 1.

To understand how this algorithm works, first note that the divisor $(Q') - (S)$ is in the same divisor class as the divisor $(Q) - (\mathcal{O}_E)$ and, since S was chosen randomly, it is likely that the points Q' and S in the support of $(Q') - (S)$ do not appear in any intermediate computations in the algorithm. Secondly, note that at each stage in the algorithm T_1 is the point obtained by computing $[m]P$ where m is the integer whose binary expansion is an initial segment (most significant digits) of the binary expansion of l. The value f_1 is the evaluation at the divisor $(Q') - (S)$ of the function f defined such that

$$m(P) - m(\mathcal{O}_E) = (T_1) - (\mathcal{O}_E) + (f).$$

Hence, at the end of the algorithm we have $T_1 = \mathcal{O}_E$ and f_1 is the evaluation at $(Q') - (S)$ of the function g such that $l(P) - l(\mathcal{O}_E) = (g)$, as required from the definition of the Tate pairing.

```
1. Choose a random point $S \in E(\mathbb{F}_{q^k})$ and compute $Q' = Q + S \in E(\mathbb{F}_{q^k})$.
2. Set $n = \lfloor \log_2(l) \rfloor - 1$, $T_1 = P$, $f_1 = 1$.
3. While $n \geq 0$ do
     – Calculate the equations of the straight lines $l_1$ and $l_2$ arising in a doubling
       of $T_1$. Set $T_1 = [2]T_1$ and $f_1 = f_1^2(l_1(Q')l_2(S))/((l_2(Q')l_1(S))$.
     – If the $n$th bit of $l$ is one then Calculate the equations of the straight lines
       $l_1$ and $l_2$ arising in an addition of $T_1$ and $P$ (in the case $n = 0$ we have
       $l_2 = 1$). Set $T_1 = T_1 + P$ and set $f_1 = f_1(l_1(Q')l_2(S))/((l_2(Q')l_1(S))$.
     – Decrement $n$.
4. Return $f_1$.
```

Fig. 1. Miller's Algorithm.

4 The Cryptographic Applications

We do not discuss the cryptographic applications of the Tate pairing in detail since we are interested in implementation issues which are common to all schemes. We simply note that:

1. Cryptosystems based on the Weil pairing may be modified to use the Tate pairing, and this will improve their computational performance.
2. In many of these schemes the calculation of the Tate pairing is one of the dominant computational tasks.

In most applications of the Weil and Tate pairing to cryptography we consider an elliptic curve E over $\mathbb{F}_q$ with number of points divisible by some prime l. It is necessary that l have at least 160 bits for security, and for efficiency it is desired that l and q not be too large. Also important for these applications is the finite field $\mathbb{F}_{q^k}$ where k is defined to be the smallest integer such that $l|(q^k - 1)$. It is necessary that q^k have at least 1000 bits for security, and for good efficiency it is desired that q^k not be too large. Further discussion about these matters may be found in [7], but the conclusion is that there are three cases particularly relevant for cryptography:

1. Supersingular elliptic curves over certain prime fields $\mathbb{F}_p$ where p has 512 bits (in this case $k = 2$). For example the curve $y^2 = x^3 + 1$ used in [3] when $p \equiv 2 \pmod 3$.
2. Supersingular elliptic curves of the form $y^2 + y = x^3 + x + b$ ($b \in \{0, 1\}$) over $\mathbb{F}_2$ considered as a group over $\mathbb{F}_{2^m}$ where m is prime of size around 250 (in this case $k = 4$).
3. Supersingular elliptic curves of the form $y^2 = x^3 - x \pm 1$ over $\mathbb{F}_3$ considered as a group over $\mathbb{F}_{3^m}$ where m is prime of size around 110 (in this case $k = 6$).

For the cryptographic applications the basic operation is to compute the value of the Tate pairing $\langle P, Q \rangle$ where $P \in E(\mathbb{F}_q)$ and where $Q \in E(\mathbb{F}_{q^k})$ (usually Q is the image of some multiple of P under a non-rational endomorphism or ‘distortion map’). We stress that since a unique value is required for the cryptographic applications we must also raise the value of the Tate pairing to the power $(q^k - 1)/l$.

5 Efficient Computation of the Tate Pairing

Our analysis begins in this section, where we make three general comments about efficient computation of the Tate pairing in the specific application we have in mind.

The most important observation is that we compute $\langle P, Q\rangle$ where $P \in E(\mathbb{F}_q)$ and where $Q \in E(\mathbb{F}_{q^k})$. In practice, this means that the coefficients of the lines l_i in Miller's algorithm (Figure 1) are all elements of the smaller field $\mathbb{F}_q$ while the large field $\mathbb{F}_{q^k}$ is only used for computing the value f_1.

This observation is the most fundamental observation in the paper and most of the implementation details arise from trying to make the most of it. In particular, to benefit from this observation, one should work with an efficient representation of $\mathbb{F}_q$ for all operations involving the elliptic curve E, the points T_1 and T_2, and the straight lines l_i. One should then implement efficient operations for $\mathbb{F}_{q^k}$ which allow fast scalar multiplication by elements in $\mathbb{F}_q$. The natural way to proceed is to represent $\mathbb{F}_{q^k}$ as a degree k extension of $\mathbb{F}_q$. We give many more details in Section 9. We comment that this is different to the approach proposed by Boneh, Lynn and Shacham [4].

A further example of working in subfields whenever possible is to consider the choice of the random point S in Miller's algorithm (Figure 1). As stated, $S \in E(\mathbb{F}_{q^k})$ but in fact we may take $S \in E(\mathbb{F}_q)$ and this reduces the number of operations in $\mathbb{F}_{q^k}$. See [1] for further consequences of this choice.

It is interesting at this point to consider the relationship between the Weil pairing and the Tate pairing. We write $e_l(P, Q)$ for the Weil pairing. In most situations the Weil pairing is related to the Tate pairing by the equation

$$e_l(P, Q) = \langle P, Q\rangle / \langle Q, P\rangle$$

(and no exponentiation is required to get a unique value). This is the way the Weil pairing is usually computed. Other methods to compute the Weil pairing (such as Section III.8 of [14]) seem to be even less efficient. This leads to the often quoted statement "the Weil pairing is just two applications of the Tate pairing". However, in the case that $P \in E(\mathbb{F}_q)$ but $Q \in E(\mathbb{F}_{q^k})$ then these two Tate pairing operations require very different computation times. Hence, the Weil pairing seems to require much more than twice the running time of the Tate pairing in the cryptographic applications.

Our second observation relates to the well-known fact that divisions are more expensive than multiplications. This statement is particularly true for divisions in the large field $\mathbb{F}_{q^k}$ since we are representing it as a degree k extension of the field $\mathbb{F}_q$. Hence it is desirable to reduce the number of divisions in $\mathbb{F}_{q^k}$ in Miller's algorithm. Consider the divisions which are required in the large field $\mathbb{F}_{q^k}$ when computing the value f_1. It is obvious that these divisions can all be gathered into a single division at the conclusion of the algorithm by representing the value f_1 as a quotient f_1/f_2 and using multiplications to update the f_i.

Our third general observation is that, as with elliptic curve point exponentiation algorithms, there is a significant improvement by using window methods (see [2], [8]). These methods employ a precomputation stage which computes the values $[n]P$ for all values n in a 'window' of 3 or 4 bits. Miller's algorithm then proceeds by performing addition operations according to windows in the binary expansion of the exponent l instead of bit by bit. This does not change the number of doubling operations, but it does reduce the number of addition operations. The methods are completely standard (see [2,8]) and it is not necessary to repeat them here.

Note that in Section 6 we describe a class of groups which are particularly efficient for the Tate pairing computation, and the window methods are no longer useful for these groups.

Finally we mention homogenizing Miller's algorithm and using projective coordinates to remove divisions. With the algorithm as developed in this paper it did not seem to be useful to use such techniques. However, when methods of [1] are incorporated, then homogenizing Miller's algorithm becomes worthwhile when done carefully.

6 Choice of Groups

As noted by Boneh and Franklin [3] it is not necessary that the prime order l be of the same size as the field q. For instance, when working with supersingular elliptic curves over $\mathbb{F}_p$ where $p > 3$ it is necessary that p have at least 512 bits, but l may be chosen to have 160 bits.

This technique of working in a smaller subgroup has a huge impact on the complexity of Miller's algorithm, since the number of iterations depends on $\log_2(l)$. This technique may be used in characteristic two and three as well, whenever the group order of $E(\mathbb{F}_q)$ has factors of a suitable size.

A further method which speeds up the Tate pairing very significantly is to choose the prime l such that it has very low hamming weight (or, more generally, so that it has low hamming weight in a signed binary representation, or in a signed base-three representation in characteristic three). This greatly reduces the number of addition operations in Miller's algorithm. Note that this technique means that window methods are no longer required, and so there is no precomputation step in this case.

The system of Boneh and Franklin [3] for large prime characteristic can be trivially modified to employ primes l of low Hamming weight. In characteristic two an example of such a group order is the following: Let E be the elliptic curve $y^2 + y = x^3 + x + 1$ over $\mathbb{F}_{2^{283}}$. Then $\#E(\mathbb{F}_{2^{283}}) = l$ where l is the prime $l = 2^{283} + 2^{142} + 1$, which has Hamming weight 3. There are other cases in characteristic two with prime number of points which have the same property of their (signed) binary expansion. Similarly, supersingular curves with a prime number of points in characteristic three will have low Hamming weight of the signed base-three expansion of l.

For several examples in characteristic two and three the group order N has small Hamming weight, but the large prime factor l is a quotient of N by a small cofactor h and so it does not have small Hamming weight. In practice one can compute the Tate pairing of the points P and Q of order l with respect to the group order N (and then raise to the exponent $(q^k - 1)/N$ which also has low Hamming weight). In this case the small Hamming weight of N provides computational savings in Miller's algorithm. This technique is used for the implementation results in Section 10 and it reduces the running time by at least 30%.

We now show that the value computed by Miller's algorithm is the same in both cases. Let g be a function such that $(g) = l(P) - l(\mathcal{O}_E)$ and let g' be a function such that $(g') = N(P) - N(\mathcal{O}_E)$ where $N = hl$. Then $(g') = h(g) = (g^h)$. If D is a divisor in the same divisor class as $(Q) - (\mathcal{O}_E)$ with support disjoint from (g) then

$$g'(D)^{(q^k-1)/N} = g(D)^{h(q^k-1)/(hl)} = \langle P, Q\rangle^{(q^k-1)/l}.$$

7 Specific Advantages in Characteristic Two and Three

In this section we discuss certain features of elliptic curves in small characteristic. In particular, we discuss certain arithmetic operations which are particularly efficient, such as point tripling in characteristic three.

7.1 Doubling in Characteristic Two

It is well-known in elliptic curve cryptography that there are performance advantages available in characteristic two, particularly when implementing elliptic curve exponentiation directly in hardware. For a survey of point exponentiation methods in characteristic two see Hankerson, Hernandez and Menezes [8]. These methods can all be used to improve Miller's algorithm in characteristic two, and it follows that cryptosystems based on the Tate pairing on supersingular curves in characteristic two have good performance. Note that, for the field sizes we are considering, Karatsuba multiplication does not seem to provide any benefit. All the relevant methods from [8] were used to obtain the timings in Section 10.

7.2 Tripling in Characteristic Three

In characteristic three for our supersingular elliptic curves (and, more generally, for curves over $\mathbb{F}_{3^m}$ with equations of the form $y^2 = x^3 + Ax + B$) it happens that the tripling operation can be performed extremely efficiently.

Indeed, one can give tripling formulae which do not require divisions! For the Tate pairing computation it is necessary to obtain the equations of the straight lines used for the addition rule, and so one division is unavoidable.

We give all the details of the tripling operations and the straight lines below:

Let $P = (x_1, y_1)$ be a point on $E : y^2 = x^3 - x + b$ over $\mathbb{F}_{3^m}$. The tangent to E at P has slope $\lambda_2 = 1/y_1$ and the equation of the tangent line is

$$l_1 : y - \lambda_2 x + (\lambda_2 x_1 - y_1) = 0.$$

The point $(x_2, y_2) = [2]P$ has coordinates $x_2 = \lambda_2^2 + x_1$ and $y_2 = -\lambda_2^3 - y_1$. The equation of the vertical line is $l_2 : x - x_2 = 0$.

The line between (x_1, y_1) and (x_2, y_2) has slope $\lambda_3 = y_1^3 - \lambda_2$ and its equation is

$$l'_1 : y + (\lambda_2 - y_1^3)x + (y_1^3 x_1 - \lambda_2 x_1 - y_1) = 0.$$

The point $(x_3, y_3) = [3](x_1, y_1)$ has coordinates $x_3 = x_1 + y_1^2 + y_1^6$ and $y_3 = -y_1^9$. The equation of the vertical line is $l'_2 : x - x_3 = 0$. Note that these formulae provide a division-free algorithm for tripling on these elliptic curves in characteristic three.

Also note that cubing is very fast in characteristic three (especially in hardware, or if a normal basis representation is used) and so computing y_1^3, y_1^6 and y_1^9 is cheap from y_1 and y_1^2.

These formulae for point tripling are very efficient and so it is prudent to re-write Miller's algorithm to utilise a signed base-3 representation of the exponent l. Recall that a signed base-3 representation is an expression

$$l = \sum_{n=0}^{m} l_n 3^n$$

where $l_n \in \{-1, 0, 1\}$ and we may assume that $l_m = 1$. We call each l_n a 'trit'. There should be no confusion between the notation l_n for trits and the lines l_1 and l_2. We sketch the details in Figure 2. We stress that, in practice, care must be taken to implement the formulae for l_1 and l'_1 above so that the number of multiplications in the large field $\mathbb{F}_{q^k}$ is minimised.

```
1. Choose a random point $S \in E(\mathbb{F}_q)$ and compute $Q' = Q + S \in E(\mathbb{F}_{q^k})$.
2. Compute the value $f_2$ of the function $f = 1/(x - x_P)$ evaluated at the divisor
   $(Q') - (S)$. (The function $f$ satisfies $-(P) + (\mathcal{O}_E) = (P) - (\mathcal{O}_E) + (f)$.)
3. Let $n$ be such that $l$ has a signed base-3 representation $l = \sum_{j=0}^{n+1} l_j 3^j$ with
   $l_{n+1} = 1$. Set $T_1 = P$ and $f_1 = 1$.
4. While $n \geq 0$ do
   – Perform a tripling of $T_1$, i.e., compute the equations for the lines $l_1, l_2, l'_1, l'_2$
     above, set $T_1 = [3]T_1$, and update the value of $f_1$ via
         $f_1 = f_1^3(l_1/l_2 \cdot l'_1/l'_2)((Q') - (S))$.
   – If the $n$th trit in the signed base-3 expansion of $l$ is 1 then set $T_1 = T_1 + P$
     and set $f_1 = f_1(l_1/l_2)((Q') - (S))$ where $l_1$ and $l_2$ are the lines appearing
     in the point addition.
   – If the $n$th trit in the signed base-3 expansion of $l$ is $-1$ then set $T_1 = T_1 - P$
     and set $f_1 = f_1 f_2(l_1/l_2)((Q') - (S))$ where $l_1$ and $l_2$ are the lines appearing
     in the point addition and $f_2$ is from Step 2 above.
   – Decrement $n$.
5. Return $f_1$.
```

Fig. 2. Miller's Algorithm in base three.

Note that the efficient tripling formulae are valuable for efficient implementation of the system proposed by Koblitz in [10].

8 Efficient Implementation of Characteristic Three Fields

It is essential to have an efficient implementation of arithmetic in the finite field $\mathbb{F}_{3^m}$. A lot of research has been done into efficient implementation of characteristic two finite fields, and also for large characteristic p, but characteristic three does not seem to have been studied in detail. Either polynomial bases or normal bases may be used (see [2] for details).

The conventional wisdom for representing values in characteristic two is to represent each coefficient by a single bit and to pack 32 coefficients into a single computer word. In this way, the addition of two values can be performed efficiently by using an exclusive-or machine instruction to add 32 coefficients at a time. Most finite field packages treat characteristic two as a special case and then degenerate to using a bignum implementation for odd characteristic. This can be improved upon.

We note that a coefficient in characteristic 3 has the values 0, 1, or 2. That is, we need two bits to represent such a value. Rather than pack sixteen 2-bit coefficients into a 32 bit word, we pack the high order bits into one word array and the low order bits into a separate word array.

In other words, we write the 16 coefficients modulo 3 as $a = a_{\mathrm{lo}} + 2a_{\mathrm{hi}}$. This gives the following advantages:

1. Doubling a value can be performed by swapping the high and low order bit arrays. Note: negation is identical to doubling in characteristic three.
2. Adding two values $r = a + b$ leads to

$$(r_{\mathrm{hi}}, r_{\mathrm{lo}}) = (a_{\mathrm{hi}}, a_{\mathrm{lo}}) + (b_{\mathrm{hi}}, b_{\mathrm{lo}})$$

   where

$$\begin{aligned} r_{\mathrm{lo}} &= ((a_{\mathrm{lo}} \wedge b_{\mathrm{lo}}) \& (\sim (a_{\mathrm{hi}} | b_{\mathrm{hi}}))) | (a_{\mathrm{hi}} \& b_{\mathrm{hi}}) \\ r_{\mathrm{hi}} &= ((a_{\mathrm{hi}} \wedge b_{\mathrm{hi}}) \& (\sim (a_{\mathrm{lo}} | b_{\mathrm{lo}}))) | (a_{\mathrm{lo}} \& b_{\mathrm{lo}}). \end{aligned}$$

   Here, as usual, $\sim$ means bitwise complement, & means bitwise and, | means bitwise or, and $\wedge$ means bitwise exclusive-or. In other words, we can add 32 coefficients with 12 boolean operations.
3. Cubing is performed analogously to squaring in characteristic two by using a ‘thinning’ algorithm with a reduction operation (this is just a shift if a normal basis is used).
4. Subtracting two values is performed using addition:

$$(r_{\mathrm{hi}}, r_{\mathrm{lo}}) = (a_{\mathrm{hi}}, a_{\mathrm{lo}}) - (b_{\mathrm{hi}}, b_{\mathrm{lo}}) = (a_{\mathrm{hi}}, a_{\mathrm{lo}}) + (b_{\mathrm{lo}}, b_{\mathrm{hi}}).$$

9 Efficient Computation in Extension Fields

We now describe some implementation details for finite field extensions. These issues arise because of our choice of field representation, which in turn is motivated by the benefit of working in subfields wherever possible.

The two most important cases are the elliptic curves $y^2 = x^3 - x \pm 1$ over extensions of $\mathbb{F}_3$ and $y^2 + y = x^3 + x + b$ over extensions of $\mathbb{F}_2$. In practice it is necessary to be able to work efficiently with finite fields $\mathbb{F}_{3^{6m}}$ and $\mathbb{F}_{2^{4m}}$ where m is prime. We give further details about how to achieve this.

9.1 Characteristic Two

We represent $\mathbb{F}_{2^{4m}}$ by a tower of two quadratic extensions of $\mathbb{F}_{2^m}$. To be precise, let $F = \mathbb{F}_{2^m}$ and denote

$$F_1 = F[x]/(x^2 + x + 1) \cong \mathbb{F}_{2^{2m}}$$

and

$$F_2 = F_1[y]/(y^2 + (x + 1)y + 1) \cong \mathbb{F}_{2^{4m}}.$$

A general element of F_2 can be written as $a + bx + cy + dxy$ with $a, b, c, d \in F$.

The naive way to perform multiplication of two elements $(u_1 + yv_1)$ and $(u_2 + yv_2)$ of F_2 (where $u_i, v_i \in F_1$) to obtain the product

$$(u_1u_2 + v_1v_2) + y(u_1v_2 + u_2v_1 + (x + 1)v_1v_2)$$

would require 4 multiplications in F_1 (plus the 'special' multiplication by the term $(x + 1)$). A more efficient multiplication process is to compute the three products $t_1 = u_1u_2, t_2 = v_1v_2$ and $t_3 = (u_1 + v_1)(u_2 + v_2)$. The desired product is then recovered as $(t_1 + t_2) + y(t_3 - t_1 + xt_2)$ which requires 3 multiplications in F_1 plus the 'special' multiplication xt_2 (which is shown below to be just a single addition).

Similarly, multiplication of general elements $(u_1 + xv_1)(u_2 + xv_2)$ in F_1 can be performed with just 3 multiplications in F, plus one 'special' multiplication.

Finally, note that the result of the special multiplication $x(u + xv)$ is equal to $v + x(u + v)$, which is computed by a single addition.

In conclusion, the cost of a general multiplication in F_2 is reduced from 16 (or more) multiplications in F to only 9 multiplications in F.

Division in F_2 can be reduced to a single division in F by using conjugates. The details are straightforward, and since there is only one division in F_2 in our algorithm this is not worth discussing in depth here.

9.2 Characteristic Three

We represent $\mathbb{F}_{3^{6m}}$ by a tower of extensions of $\mathbb{F}_{3^m}$. To be precise, let $F = \mathbb{F}_{3^m}$ and denote

$$F_1 = F[a]/(a^3 - a + 1) \cong \mathbb{F}_{3^{3m}}$$

and

$$F_2 = F_1[b]/(b^2+1) \cong \mathbb{F}_{3^{6m}}.$$

(Note that $i = b$, $\alpha = a$ and $\beta = -a$ in the notation of Section 3.9 of [7])

As in the previous subsection, a multiplication of general elements in F_2 can be performed with fewer multiplications than the naive method. The details are as follows:

To multiply $(u_1 + bv_1)(u_2 + bv_2)$ where $u_1, u_2, v_1, v_2 \in F_1$ we compute $t_1 = u_1u_2$, $t_2 = v_1v_2$ and $t_3 = (u_1 + v_1)(u_2 + v_2)$. The product is then recovered as

$$(t_1 - t_2) + b(t_3 - t_1 - t_2).$$

The product of $(u_1 + av_1 + a^2w_1)$ and $(u_2 + av_2 + a^2w_2)$ for u_1,u_2,v_1,v_2,w_1, $w_2 \in F$ is

$$(u_1u_2 + v_1w_2 + w_1v_2) + a(u_1v_2 + v_1u_2 + v_1w_2 + w_1v_2 + w_1w_2)$$
$$+a^2(u_1w_2 + v_1v_2 + w_1u_2 + w_1w_2).$$

To compute this in fewer than 9 multiplications compute $t_1 = u_1u_2$, $t_2 = u_1w_2$, $t_3 = v_1v_2$, $t_4 = v_1w_2$, $t_5 = w_1u_2$, $t_6 = w_1v_2$, $t_7 = w_1w_2$ and $t_8 = (u_1 + v_1 + w_1)(u_2 + v_2 + w_2)$. The product is recovered as

$$(t_1 - t_4 - t_6) + a(t_8 - t_1 - t_2 - t_3 - t_5 + t_7) + a^2(t_2 + t_3 + t_5 + t_7).$$

Hence we have reduced the cost of multiplication in F_2 from 36 to 24 multiplications in F. Barreto has observed that this can be reduced to 18 multiplications by considering a single extension of degree 6 rather than the quadratic and cubic extensions separately.

Again, inversion can be reduced to a single inversion in F by using conjugates. The details are straightforward (the conjugates of $(u + av + a^2w)$ are simply $(u + (a+1)v + (a+1)^2w)$ and $(u + (a+2)v + (a+2)^2w)$).

Finally, the exponentiation operation in the finite field F_2 is performed using the signed base-3 expansion of the exponent (which has low Hamming weight in most of our examples and so window methods are not necessary).

9.3 Timing Results

In summary, we have the following timing results for field operations. We record the cost in terms of the number of multiplications in the ground field. Let F be $\mathbb{F}_{2^m}$ or $\mathbb{F}_{3^m}$ respectively and F_2 be $\mathbb{F}_{2^{4m}}$ or $\mathbb{F}_{3^{6m}}$. Here, for instance, $F * F_2$ means the cost of multiplying a general element of F_2 by an element of F and $1/F$ means the cost of inverting an element of F.

	Characteristic two		Characteristic three	
m	241	283	97	163
$F * F$	1M	1M	1M	1M
$F * F_2$	4M	4M	6M	6M
$F_2 * F_2$	9M	9M	24M	24M
$1/F$	13.85M	9.25M	5.36M	5.05M
$1/F_2$	44.85M	40.25M	107.36M	107.05M

Notes:

1. The field extension inversion was not heavily optimised because it is only invoked once in the computation of a Tate or Weil pairing.
2. In characteristic three it is cheaper to perform a field inversion in F than to compute a field by field extension multiplication. We attribute this to the inefficiency of multiplication, rather than to any special benefit of inversion in characteristic three (it is an open problem to provide more efficient multiplication algorithms in characteristic three).
3. It is always worth tracking whether a value is in the field or in the field extension - and performing the appropriate operation.
4. The costs of performing the field inversion were established by timing 100,000 field inversions and 100,000 field multiplications. The other costs were established by examination of the code.

10 Timing Results

We have implemented the Tate pairing using the methods given above. All timings were performed on a 1 GHz Pentium III with 256Mb RAM (an HP VISUALISE NT workstation). The language used was C. The compiler was Microsoft Visual C++ V6.0 with speed optimisations on.

10.1 Characteristic Two Timings

We give a few timings for characteristic two. Due to the numerous techniques available for efficient characteristic two arithmetic and elliptic curve operations it follows that characteristic two is the best choice for fast implementations of the Tate pairing.

Example 1:
Consider the elliptic curve $E : y^2 + y = x^3 + x + 1$ over

$$\mathbb{F}_{2^{241}} = \mathbb{F}_2[x]/\langle x^{241} + x^{70} + 1\rangle.$$

The large prime order is $l = 2^{241} - 2^{121} + 1$
Consider points $P \in E(\mathbb{F}_{2^{241}})$ and $Q \in E(\mathbb{F}_{2^{964}})$ of order l.
Weil Pairing $e_l(P, Q)$ time: 140.9 ms.
Tate Pairing $\langle P, Q\rangle^{(2^{964}-1)/l}$ time: 32.50 ms (including the finite field exponentiation).

Example 2:
Consider the elliptic curve $E : y^2 + y = x^3 + x + 1$ over

$$\mathbb{F}_{2^{283}} = \mathbb{F}_2[x]/\langle x^{283} + x^{194} + x^{129} + x^{65} + 1\rangle.$$

The large prime l is $2^{283} + 2^{142} + 1$
Consider points $P \in E(\mathbb{F}_{2^{283}})$ and $Q \in E(\mathbb{F}_{2^{1132}})$ of order l.

Weil Pairing $e_l(P,Q)$ time: 175.8 ms.

Tate Pairing $\langle P,Q\rangle^{(2^{1132}-1)/l}$ time : 57.19 ms (including the finite field exponentiation).

Notes:

1. These times show that cryptosystems based on the Tate pairing are completely practical for PC-based applications.
2. As explained in Section 5, the Weil pairing takes longer than twice the running time of the Tate pairing for the cryptographic applications.

10.2 Characteristic Three Timings

We now give timings for characteristic three.

Example 3:
Consider the elliptic curve $E : y^2 = x^3 - x + 1$ over

$$\mathbb{F}_{3^{97}} = \mathbb{F}_3[x]/\langle x^{97} + x^{16} + 1\rangle.$$

The group order is $N = 7l = 3^{97} + 3^{49} + 1$.

We took points $P \in E(\mathbb{F}_{3^{97}})$ and $Q \in E(\mathbb{F}_{3^{582}})$ of order l and computed the Tate pairing of order $7l$.

Tate Pairing: 168 ms (including finite field exponentiation)

Example 4:
Consider the elliptic curve $E : y^2 = x^3 - x + 1$ over

$$\mathbb{F}_{3^{163}} = \mathbb{F}_3[x]/\langle x^{163} + x^{80} + 2\rangle.$$

The group order is $N = 7l = 3^{163} - 3^{82} + 1$.

We took points $P \in E(\mathbb{F}_{3^{163}})$ and $Q \in E(\mathbb{F}_{3^{978}})$ of order l and computed the Tate pairing with respect to the order $7l$.

Tate Pairing: 581 ms (including the finite field exponentiation)

11 Further Topics

We refer to [1] for further significant implementation techniques. In particular, our timings have been improved by a factor of 3 in characteristic two and 6 in characteristic three by using those methods plus further optimisations.

It is likely that multiplication of finite field elements in characteristic three can be significantly improved. This is an avenue for further research.

Acknowledgements

The authors would like to thank Paulo Barreto, Dan Boneh, Kenny Paterson, Mike Scott and an anonymous referee for helpful comments.

References

1. P. S. L. M. Barreto, H. Y. Kim, B. Lynn and M. Scott, Efficient algorithms for pairing-based cryptosystems, Cryptology ePrint archive: Report 2002/008 (February 6, 2002).
2. I.F. Blake, G. Seroussi and N.P. Smart, *Elliptic Curves in Cryptography*, Cambridge University Press, 1999.
3. D. Boneh and M. Franklin, Identity-based encryption from the Weil pairing, in J. Kilian (ed.), Crypto 2001, Springer LNCS 2139 (2001) 213–229.
4. D. Boneh, B. Lynn and H. Shacham, Short signatures from the Weil pairing, in C. Boyd (ed.), Asiacrypt 2001, Springer LNCS 2248, (2001) 514–532.
5. G. Frey and H.-G. Rück, A remark concerning m-divisibility and the discrete logarithm in the divisor class group of curves, *Math. Comp.*, **62**, No.206 (1994) 865–874.
6. G. Frey, M. Müller and H.-G. Rück, The Tate pairing and the discrete logarithm applied to elliptic curve cryptosystems, *IEEE Trans. Inform. Theory*, **45**, no. 5 (1999) 1717–1719.
7. S. D. Galbraith, Supersingular curves in cryptography, in C. Boyd (ed.), Asiacrypt 2001, Springer LNCS 2248 (2001) 495–513.
8. D. Hankerson, J. Hernandez and A. J. Menezes, Software implementation of elliptic curve cryptography over binary fields, Proceedings of CHES 2000, Springer LNCS 1965 (2000), 1-24.
9. A. Joux, A one round protocol for tripartite Diffie-Hellman, in W. Bosma (ed.), ANTS-IV, Springer LNCS 1838 (2000) 385–393.
10. N. Koblitz, An elliptic curve implementation of the finite field digital signature algorithm, in H. Krawczyk (ed.), Crypto '98, Springer LNCS 1462 (1998) 327–337.
11. V. Miller, Short programs for functions on curves, unpublished manuscript 1986.
12. A. J. Menezes, T. Okamoto and S. A. Vanstone, Reducing elliptic curve logarithms to logarithms in a finite field, *IEEE Trans. Inf. Theory*, **39**, No. 5 (1993) 1639–1646.
13. R. Sakai, K. Ohgishi and M. Kasahara, Cryptosystems based on pairing, in SCIS 2000, Okinawa, Japan, January 2000.
14. J. H. Silverman, The arithmetic of elliptic curves, Springer GTM 106, 1986.
15. E. R. Verheul, Evidence that XTR is more secure than supersingular elliptic curve cryptosystems, in B. Pfitzmann (ed.), Eurocrypt 2001, Springer LNCS 2045 (2001), 195–210.
16. E. R. Verheul, Self-blindable credential certificates from the Weil pairing, in C. Boyd (ed.), Asiacrypt 2001, Springer LNCS 2248 (2001) 533–551.

Smooth Orders and Cryptographic Applications

Carl Pomerance[1] and Igor E. Shparlinski[2]

[1] Department of Fundamental Mathematics, Bell Laboratories
Murray Hill, NJ 07974-0636, USA
carlp@research.bell-labs.com

[2] Department of Computing, Macquarie University
Sydney, NSW 2109, Australia
igor@ics.mq.edu.au

Abstract. We obtain rigorous upper bounds on the number of primes $p \leq x$ for which $p-1$ is smooth or has a large smooth factor. Conjecturally these bounds are nearly tight. As a corollary, we show that for almost all primes p the multiplicative order of 2 modulo p is not smooth, and we prove a similar but weaker result for almost all odd numbers n. We also discuss some cryptographic applications.

1 Introduction

We recall that an integer $k \geq 1$ is called y-smooth if it is divisible only by primes $p \leq y$. Here we obtain reasonably good upper bounds on the number of primes $p \leq x$ for which $p-1$ is y-smooth and also for primes $p \leq x$ for which $p-1$ has a large y-smooth factor.

We apply these bounds to show that for almost all primes p the multiplicative order $l(p)$ of 2 modulo p is not smooth. In particular, we show that for any function $\varepsilon(p) \to 0$, for almost all primes p, $l(p)$ has a prime divisor $q \geq p^{\varepsilon(p)}$. We also prove a similar statement for the multiplicative order $l(n)$ of 2 modulo almost all odd integers n.

Besides being a natural question, it also has some cryptographic motivations which we discuss in Section 4.

As usual, $\varphi(m)$ denotes the Euler function. We use log to denote the natural logarithm. Throughout the paper the implied constants in symbols 'O', '$\gg$' and '$\ll$' are absolute (the notations $U \ll V$ and $V \gg U$ are equivalent to $U = O(V)$ for positive functions U, V). The symbol '$\sim$' indicates the asymptotic relation is uniform over all parameters in their stated ranges.

2 Smooth Divisors of $p-1$

Let $P(n)$ denote the largest prime divisor of the integer $n \geq 2$, and let $P(1) = 1$. Let $\pi(x, y)$ denote the number of primes $p \leq x$ with $P(p-1) \leq y$. Let $\psi(x, y)$ denote the number of positive integers $n \leq x$ with $P(n) \leq y$. It seems reasonable to conjecture that a random integer in the interval $[1, x]$ is about as likely to be

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 338–348, 2002.

y-smooth as is a random integer of the form $p-1$ where p is a prime in $[1,x]$, at least if y is not too small. That is, it may be that

$$\frac{1}{x}\psi(x,y) \ \sim\ \frac{1}{\pi(x)}\pi(x,y), \tag{1}$$

for $y \le x$ and $y \to \infty$. This possibility is explicitly raised in [18], but the thought goes back at least to [6]. Through the years there has been progress towards the weaker assertion

$$\pi(x,y) \ \gg\ \psi(x,y)/\log x,$$

but only in the range $x^{\vartheta} \le y \le x$, where $\vartheta > 0$ is fixed. A recent paper of Baker and Harman [2] has the champion value of ϑ, namely 0.2961, but they have the inequality in the somewhat weaker form

$$\pi(x,y) \ \ge\ \psi(x,y)/(\log x)^{O(1)}.$$

Earlier papers on this subject are the already-cited [6] and [18], as well as papers by Wooldridge, Balog, Fouvry and Grupp, and Friedlander. In [1] there is a proof that $\pi(x,y)$ is proportional to $\pi(x)$ when $\log x/\log y$ is bounded, conditional on a reasonable hypothesis on the distribution of primes in arithmetic progressions. In addition, Granville (see [8]) has an unpublished argument that (1) holds when $\log x/\log y$ is bounded, conditional on the Elliott-Halberstam conjecture. In [15] a connection of (1) to a strong form of the generalized twin prime conjecture is demonstrated.

There are highly nontrivial *upper* bounds for $\pi(x,y)$ by Fouvry and others when $y > x^{1/2}$, and here the quest is to find the largest value of ϑ for which you can prove there is some $c > 0$ with $\pi(x,x^{\vartheta}) \le (1-c+o(1))\pi(x)$, or even just $\pi(x)-\pi(x,x^{\vartheta}) \to \infty$. Such a quest may be considered a back-door attack on the conjecture that there are infinitely many Sophie Germain primes, namely primes q where $2q+1$ is also prime. However, the results in our paper are more aimed at smaller values of y; we make no new contribution towards the problem of a nontrivial upper bound for $\pi(x,y)$ when y is large. Finally, we remark that there is at least one paper [16] (brought to our attention by the referee) that gives an upper bound for the number of primes up to x for which the order of a given element is y-smooth when $y > x^{1/2}$.

Let $\rho(u)$ denote the Dickman–de Bruijn function which is defined by

$$\rho(u) \ =\ 1, \qquad 0 \le u \le 1,$$

and

$$\rho(u) \ =\ 1 - \int_1^u \frac{\rho(v-1)}{v}dv, \qquad u > 1.$$

We recall that $\rho(u) = u^{-u+o(u)}$ as $u \to \infty$. For these and other properties of $\rho(u)$, see [25].

It is known that $\psi(x,y) \sim \rho(u)x$ in a wide range, and so, in light of the above comments, it seems appropriate to compare $\pi(x,y)$ with $\rho(u)\pi(x)$. In fact we give an upper bound for $\pi(x,y)$ that is nearly this sharp.

We begin with the following lemma which is perhaps of independent interest.

Lemma 1. *For* $\exp\big((\log\log x)^2\big) \le y \le x$, *we have*

$$\sum_{m\le x,\, P(m)\le y} \frac{m}{\varphi(m)} \ \sim\ \frac{\zeta(2)\zeta(3)}{\zeta(6)}\psi(x,y)$$

where $u = \log x/\log y$ *and where* $\zeta(s)$ *denotes the Riemann zeta function.*

Proof. Let $z = \log y$ and assume that $\exp\big((\log\log x)^2\big) \le y \le x$. We have

$$\begin{aligned}\sum_{m\le x,\, P(m)\le y} \frac{m}{\varphi(m)} \ &=\ \sum_{m\le x,\, P(m)\le y}\ \sum_{d\mid m} \frac{\mu^2(d)}{\varphi(d)}\\ &=\ \sum_{d\le x,\, P(d)\le y} \frac{\mu^2(d)}{\varphi(d)} \sum_{m\le x/d,\, P(m)\le y} 1\\ &=\ \sum_{d\le z,\, P(d)\le y} \frac{\mu^2(d)}{\varphi(d)}\psi(x/d,y)\ +\ \sum_{z<d\le x,\, P(d)\le y} \frac{\mu^2(d)}{\varphi(d)}\psi(x/d,y).\end{aligned}$$

Since $\psi(x,y) \sim \rho(u)x$ uniformly for $y \ge \exp\big((\log\log x)^{5/3+\varepsilon}\big)$, a result of Hildebrand (see [25], Chapter III.5, Corollary 9.3) and since

$$\rho(\log(x/d)/\log y)\ \sim\ \rho(u)$$

for $y \ge \exp\big((\log\log x)^2\big)$ and $d \le z$, we have

$$\begin{aligned}\sum_{d\le z,\, P(d)\le y} \frac{\mu^2(d)}{\varphi(d)}\psi(x/d,y) \sim\ \rho(u)x \sum_{d\le z,\, P(d)\le y} \frac{\mu^2(d)}{d\varphi(d)}\ &\sim\ \rho(u)x \sum_{P(d)\le y} \frac{\mu^2(d)}{d\varphi(d)}\\ &\sim\ \rho(u)x \sum_{d\ge 1} \frac{\mu^2(d)}{d\varphi(d)} = \frac{\zeta(2)\zeta(3)}{\zeta(6)}\rho(u)x.\end{aligned}$$

Let $j_0 = \lfloor \log z \rfloor$, so that

$$\begin{aligned}\sum_{z<d\le x,\, P(d)\le y} \frac{\mu^2(d)}{\varphi(d)}\psi(x/d,y)\ &\le\ \sum_{j_0\le j<\log x}\ \sum_{e^j<d\le e^{j+1},\, P(d)\le y} \frac{\mu^2(d)}{\varphi(d)}\psi(x/d,y)\\ &\ll x \sum_{j_0\le j<\log x}\ \sum_{e^j<d\le e^{j+1},\, P(d)\le y} \frac{\mu^2(d)}{d\varphi(d)}\rho\left(u - \frac{j+1}{\log y}\right)\\ &\ll\ x \sum_{j_0\le j<\log x} e^{-j}\rho\left(u - \frac{j+1}{\log y}\right)\\ &\ll\ xe^{-j_0}\rho\left(u - \frac{j_0+1}{\log y}\right)\ =\ o(\rho(u)x),\end{aligned}$$

by the choice of z. This completes the proof. □

Theorem 1. *For* $\exp\left(\sqrt{\log x \log\log x}\right) \le y \le x$, *we have*

$$\pi(x,y) \ll u\rho(u)\pi(x)$$

where $u = \log x/\log y$.

Proof. In the following the letter q runs over prime numbers. Let $\pi_q(x)$ denote the number of primes $p \le x$ with $P(p-1) = q$. Let $z = \exp\left((\log\log x)^2\right)$, and assume $z \le Y \le x$. We have

$$\begin{aligned}\pi(x,Y) - \pi(x,Y/e) &= \sum_{Y/e<q\le Y} \pi_q(x) = \sum_{Y/e<q\le Y} \sum_{\substack{m\le(x-1)/q,\, P(m)\le q\\ mq+1 \text{ prime}}} 1\\ &\le \sum_{\substack{m\le ex/Y\\ P(m)\le Y}} \sum_{\substack{Y/e<q\le Y\\ mq+1 \text{ prime}}} 1 \ll \sum_{\substack{m\le ex/Y\\ P(m)\le Y}} \frac{m}{\varphi(m)}\cdot\frac{Y}{\log^2 Y},\end{aligned}$$

where we use Brun's method (see [9], Theorem 2.2, page 68) for the last inequality. We thus have by Lemma 1,

$$\begin{aligned}\pi(x,Y) - \pi(x,Y/e) &\ll \rho\left(\frac{\log x - \log Y + 1}{\log Y}\right)\frac{x}{Y}\cdot\frac{Y}{\log^2 Y}\\ &\le \frac{x}{\log^2 Y}\rho\left(\frac{\log x}{\log Y} - 1\right).\end{aligned}$$

Now assume y is as in the theorem and let $i_0 = \lfloor \log z\rfloor$. Then, by the above estimate,

$$\begin{aligned}\pi(x,y) &\le \pi(x,z) + \sum_{i=0}^{i_0}\left(\pi(x,y/e^i) - \pi(x,y/e^{i+1})\right)\\ &\ll \psi(x,z) + x\sum_{i=0}^{i_0}\frac{1}{(\log y - i)^2}\rho\left(\frac{\log x}{\log y - i} - 1\right).\end{aligned}$$

The function $f(t) = \rho(\log x/(\log y - t) - 1)/(\log y - t)^2$ is decreasing for $0 \le t \le i_0$, so that

$$\begin{aligned}&\sum_{i=0}^{i_0}\frac{1}{(\log y - i)^2}\rho\left(\frac{\log x}{\log y - i} - 1\right)\\ &\le \frac{\rho(u-1)}{\log^2 y} + \int_0^{i_0}\frac{1}{(\log y - t)^2}\rho\left(\frac{\log x}{\log y - t} - 1\right)dt.\end{aligned}$$

The integral is equal to

$$\begin{aligned}\frac{1}{\log x}\int_{u-1}^{\frac{\log x}{\log y - i_0}-1}\rho(s)\,ds &< \frac{1}{\log x}\int_{u-1}^{\infty}\rho(s)\,ds = -\frac{1}{\log x}\int_u^{\infty} t\rho'(t)\,dt\\ &= \frac{1}{\log x}\left(u\rho(u) + \int_u^{\infty}\rho(t)\,dt\right) \ll u\rho(u)/\log x.\end{aligned}$$

Thus,

$$\pi(x,y) \ll \psi(x,z) + \rho(u-1)x/\log^2 y + u\rho(u)x/\log x. \qquad (2)$$

Note that $\rho(u-1) \sim \rho(u)u\log u$, see (61) in Chapter III.5 of [25]. We have then that $\rho(u-1)/\log^2 y \ll u\rho(u)/\log x$ in the stated range for y. In addition, by the choice of z, the term $\psi(x,z)$ is negligible in comparison to $u\rho(u)x/\log x$. This completes the proof of the theorem. □

We remark that but for the factor u in Theorem 1, the estimate is likely to be best possible. It is reasonable to conjecture that $\pi(x,y) = o(\psi(x,y))$ uniformly for $x \to \infty$ and $y \ge 2$. Theorem 1 implies this result for $y \ge \exp\left(\sqrt{\log x \log\log x}\right)$, and (2) does so in the wider range $y \ge \exp\left((\log x)^{1/3+\varepsilon}\right)$. That $\pi(x,2) = o(\psi(x,2))$ is essentially due to Fermat, but already for $y = 3$, the conjecture that $\pi(x,3) = o(\psi(x,3))$ seems difficult. Hooley [11] has shown, under assumption of several unproved hypotheses, including the Generalised Riemann Hypothesis, that the set of integers n with $2^n - 3$ prime has density 0. It is likely the same proof would go through for primes of the form $3 \cdot 2^n + 1$. Thus, there may be a conditional proof that $\pi(x,3) = o(\psi(x,3))$, and if so, it is likely that a similar proof would work for $\pi(x,y) = o(\psi(x,y))$ with y fixed or growing slowly.

There is another approach to Theorem 1 through direct sieving. That is, for any parameter z with $1 \le z \le y$ we have

$$\pi(x,y) - \pi(z) \le \sum_{P(d)\le z} \mu(d) \sum_{\substack{n\le x,\, P(n)\le y \\ n\equiv -1 \pmod d}} 1.$$

The inner sum has been studied somewhat, see [7], and using such results, plus sieve methods, may yield a larger range of validity in Theorem 1.

Now, let $\pi(x,y,w)$ denote the number of primes $p \le x$ such that $p-1$ has a divisor $m > w$ with $P(m) \le y$.

Theorem 2. *For* $\exp\left(\sqrt{\log x \log\log x}\right) \le y \le w \le x$, *we have*

$$\pi(x,y,w) \ll \frac{u\rho(v)}{\log(2v)}\pi(x) + u\rho(u)\pi(x),$$

where $u = \log x/\log y$, *and* $v = \log w/\log y$.

Proof. Let $Q(n)$ denote the least prime factor of n, if the integer $n > 1$, and let $Q(1) = +\infty$. For a positive integer m, let $\pi_m(x,y)$ denote the number of primes $p \le x$ such that $m|p-1$ and such that all prime factors of $(p-1)/m$ exceed y, that is, $Q((p-1)/m) > y$. Note that

$$\pi(x,y,w) = \sum_{m>w,\, P(m)\le y} \pi_m(x,y).$$

Therefore, by Brun's method, see [9],

$$\pi(x,y,w) - \pi(x,y) \le \sum_{w<m<x/y,\, P(m)\le y} \pi_m(x,y)$$

$$\begin{aligned}
&\leq \sum_{w<m<x/y,\, P(m)\leq y} \ \sum_{\substack{n\leq(x+1)/m,\, Q(n)>y\\ nm+1 \text{ prime}}} 1\\
&\ll \sum_{w<m<x/y,\, P(m)\leq y} \frac{x/m}{\log y \log(x/m)} \cdot \frac{m}{\varphi(m)}\\
&\leq \frac{x}{\log^2 y} \sum_{w<m<x/y,\, P(m)\leq y} \frac{1}{\varphi(m)}.
\end{aligned}$$

Now,

$$\begin{aligned}
\sum_{w<m<x,\, P(m)\leq y} \frac{1}{\varphi(m)} &= \sum_{w<m<x,\, P(m)\leq y} \frac{1}{m}\cdot\frac{m}{\varphi(m)}\\
&= \frac{1}{x}\sum_{w<m<x,\, P(m)\leq y} \frac{m}{\varphi(m)} + \int_w^x \frac{1}{t^2} \sum_{w<m\leq t,\, P(m)\leq y} \frac{m}{\varphi(m)}\, dt.
\end{aligned}$$

Using Lemma 1, we have

$$\sum_{w<m\leq t,\, P(m)\leq y} \frac{m}{\varphi(m)} \leq \sum_{m\leq t,\, P(m)\leq y} \frac{m}{\varphi(m)} \ll \rho\left(\frac{\log t}{\log y}\right) t,$$

so that

$$\begin{aligned}
\sum_{w<m<x,\, P(m)\leq y} \frac{1}{\varphi(m)} &\ll \rho(u) + \int_w^\infty \frac{1}{t}\rho\left(\frac{\log t}{\log y}\right) dt\\
&= \rho(u) + \log y \int_v^\infty \rho(s)\, ds\\
&\ll \frac{\log y}{\log(2v)}\rho(v).
\end{aligned}$$

The last estimate follows from a similar integral calculation in the proof of Theorem 1, and from the fact that $\rho(s)/\rho(s+1) \sim s\log s$ as $s\to\infty$.

Putting this estimate into our earlier estimate, and using $\log x = u\log y$, we have that

$$\pi(x,y,w) - \pi(x,y) \ll \frac{\rho(v)}{\log(2v)}\cdot\frac{x}{\log y} = \frac{u\rho(v)}{\log(2v)}\cdot\frac{x}{\log x}.$$

This estimate, combined with Theorem 1, completes the proof. □

3 Smooth Orders of 2

For an odd integer n, let $l(n)$ denote the multiplicative order of 2 modulo n.

Let $\mathcal{L}(x,y)$ denote the set of odd primes $p\leq x$ with $l(p)$ being y-smooth, and let $L(x,y) = |\mathcal{L}(x,y)|$ be the cardinality of $\mathcal{L}(x,y)$.

Theorem 3. *For* $\exp\left(\sqrt{\log x \log\log x}\right) \le y \le x$, *we have*

$$L(x,y) \ll \frac{u\rho(u/2)}{\log(2u)}\pi(x),$$

where $u = \log x/\log y$.

Proof. Let $z = \log y$. We first consider only primes p with $l(p) > x^{1/2}/z$. Note that if $p \le x$ is such that $l(p)$ is y-smooth and $l(p) > x^{1/2}/z$, then $p-1$ has a y-smooth divisor which exceeds $x^{1/2}/z$. But, by Theorem 2, we have

$$\pi(x,y,x^{1/2}/z) \ll \frac{u\rho(u/2-\log z/\log y)}{\log(2u)}\pi(x) + u\rho(u)\pi(x) \sim \frac{u\rho(u/2)}{\log(2u)}\pi(x),$$

by our choice of z. Now let us estimate L_0, the number of primes p with $l(p)$ a y-smooth integer bounded by $x^{1/2}/z$. For each integer j, the number of primes p with $l(p)=j$ is evidently at most j, so that

$$L_0 \le \sum_{j\le x^{1/2}/z,\, P(j)\le y} j \le \frac{x^{1/2}}{z}\psi\left(\frac{x^{1/2}}{z}, y\right) \sim \frac{x}{z^2}\rho(u/2).$$

Since $x/z = xu/\log x \sim u\pi(x)$, and $\log(2u) = o(z)$ in the stated range for y, we have

$$L_0 \ll x\rho(u/2)/z^2 = o\left((u\rho(u/2)/\log(2u))\pi(x)\right),$$

which, with our earlier calculation, completes the proof. □

In particular, we see that for any function $\varepsilon(x) \to 0$, the number of primes $p \le x$ for which $l(p)$ is $x^{\varepsilon(x)}$-smooth is $o(\pi(x))$.

Now we show that Theorem 3 combined with known sieve estimates implies that the order of 2 modulo n is not smooth for almost all integer n.

Let $\mathcal{N}(x,y)$ denote the set of odd integers $n \le x$ with $l(n)$ being y-smooth, and let $N(x,y) = |\mathcal{N}(x,y)|$ be the cardinality of $\mathcal{N}(x,y)$.

Theorem 4. *For* $\exp\left(\sqrt{\log x \log\log x}\right) \le y \le x$, *we have*

$$N(x,y) \ll x/u$$

where $u = \log x/\log y$.

Proof. If $l(n)$ is y-smooth, then clearly each prime factor p of n must have $l(p)$ being y-smooth. By Brun's method (Theorem 2.2, p. 68 of [9])

$$\begin{aligned} N(x,y) &\ll x \prod_{p\le x,\, p\notin\mathcal{L}(x,y)}\left(1-\frac{1}{p}\right) \ll \frac{x}{\log x}\prod_{p\in\mathcal{L}(x,y)}\left(1-\frac{1}{p}\right)^{-1} \\ &\ll \frac{x}{u}\prod_{p\in\mathcal{L}(x,y),\, p>y}\left(1+\frac{1}{p}\right). \end{aligned}$$

It is now enough to show that

$$\sum_{p\in\mathcal{L}(x,y),\, p>y} \frac{1}{p} \ll 1.$$

By Theorem 3 and partial summation, we have

$$\begin{aligned}\sum_{p\in\mathcal{L}(x,y),\, p>y} \frac{1}{p} &= \frac{1}{x}\left(L(x,y)-\pi(y)\right) + \int_y^x \frac{1}{t^2}\left(L(t,y)-\pi(y)\right)\,dt\\ &\ll \int_y^x \frac{1}{t\log y}\rho\left(\frac{\log t}{2\log y}\right)\,dt\\ &= \int_{1/2}^{u/2} 2\rho(s)\,ds \ll 1,\end{aligned}$$

which completes the proof. □

In particular, we see that for any function $\varepsilon(x) \to 0$, the number of odd integers $1 \le n \le x$ for which $l(n)$ is $x^{\varepsilon(x)}$-smooth is $o(x)$.

4 Cryptographic Applications

We remark that it is well known that primes p for which $p-1$ is smooth are not suitable for cryptographic applications which rely on the hardness of the discrete logarithm problem modulo p. Our Theorem 1 implies that there are very few such primes. This fact has never been doubted in practice but our results provide its rigorous confirmation and a quantitative form of this statement. Unfortunately it also means that the polynomial factorization algorithm of [23] almost never runs in polynomial time. A similar remark pertains to integer factorization via the $p-1$ method of Pollard (cf. [20]). Both of these applications to smooth values of $p-1$ are actually to *very* smooth values, and so the more delicate calculations of the current paper are not really necessary to deduce that usually the algorithms are not polynomial.

It is clear to see that using 2 as the generator for exponentiation-based cryptographic constructions, such as the Diffie-Hellman key exchange scheme, the El Gamal cryptosystem, the Digital Signature Algorithm and so on (these and many other examples can be found in [14,24]) reduces the cost of exponentiation. Indeed using repeated squaring type algorithms to compute $g^a \pmod p$ requires a substantial number of multiplications by g, see Section 9.3 of [5] or Chapter 14 of [14]. Thus using $g = 2$ reduces this stage to merely one bit-shift and, possibly, one subtraction of the modulus (only in 50% of the cases), for example, see Section 14.81 of [14].

We remark that it is often recommended to work in groups of prime order, which 2 may not necessarily generate. In this case one can select a large prime divisor q of the order $l(p)$ of 2 modulo p and then compute $g \equiv 2^r \pmod p$, where $r = l(p)/q$. Obviously g generates a group of order q. Now, to compute $g^x \pmod p$ one just computes $y \equiv rx \pmod q$ and then

$$g^x \equiv 2^y \pmod p.$$

There is also one more reason to use 2 as the base. It has been shown in [4] that in this case a slight modification of the corresponding Diffie-Hellman key exchange scheme has a very important property of bit security (provided the whole scheme is secure in the traditional sense). More precisely, it has been shown in [4] that recovering even a certain bit of information about the modified secret Diffie-Hellman key modulo p (deciding whether it belongs to the interval $[0, (p-1)/2]$) is as hard as the recovering the whole key.

On the other hand, if the multiplicative order of 2 modulo p is smooth then the Pohlig–Hellman algorithm can be used to efficiently solve the discrete logarithm problem in base 2, see Section 3.6.4 of [14] or Section 5.1 of [24]. We recall that based on our current knowledge we may conclude that the hardness of the discrete logarithm problem modulo p in base g, for an integer g, is majorised

1. by $q^{1/2}$ where q is the largest prime divisor of the multiplicative order of g modulo p, see [14,24];
2. by $\mathbf{L}_p(1/2, 2^{1/2})$ for a rigorous unconditional algorithm, see [19];
3. by $\mathbf{L}_p\left(1/3, (64/9)^{1/3}\right)$ for the heuristic number field sieve algorithm, see [21,22],

where as usual we denote by $\mathbf{L}_m(\alpha, \gamma)$ any quantity of the form

$$\mathbf{L}_m(\alpha, \gamma) \; = \; \exp\left((\gamma + o(1))(\log m)^{\alpha}(\log\log m)^{1-\alpha}\right),$$

with the "$o(1)$" expression tending to 0 as the variable m tends to ∞.

The problem is: If the prime p is selected at random, what are the chances that the running time $q^{1/2}$ of the Pohlig–Hellman algorithm 1 is smaller than the running time of, say, algorithm 2? It follows from Theorem 3 that the chances of this occurring are vanishingly small. Thus, our result implies that for $g = 2$ and a randomly selected prime p, with probability exponentially close to 1, the security of the discrete logarithm to base $g = 2$ is as high as when a "safe" prime p is deliberately chosen (namely, a prime p where $p-1$ is twice a prime).

For the suggested modifications in [4] of the ElGamal public key cryptosystem, it is also important that the order of 2 modulo p is not smooth and thus the discrete logarithm problem in the corresponding group is hard. On the other hand, as in [4], we have to warn that small generators are not suitable for using with the ElGamal signature scheme, see [3]. However, the results of this paper can be extended to multiplicative orders of any fixed integer $g \ge 2$.

5 Remarks

We remark that it is likely to be true that $L(x, y) \ll \rho(u)\pi(x)$ in the stated range for y. The slightly weaker estimate $L(x, y) \ll u\rho(u)\pi(x)$ is likely to be provable assuming the Generalised Riemann Hypothesis, using the tools that Hooley [10] has used to prove Artin's conjecture on the Generalised Riemann Hypothesis.

Studying other arithmetic properties of $l(p)$, for example, the number of prime and integer divisors, is of interest as well. A recent paper on this subject is [17] (also see [13]).

Finally, having in mind applications to elliptic curve cryptography, one can ask how often a given elliptic curve defined over $\mathbb{Q}$ has a smooth order modulo a prime p. This subject is considered in [12], the paper of Lenstra where elliptic curve factoring is first introduced.

References

1. W. R. Alford, A. Granville and C. Pomerance, 'There are infinitely many Carmichael numbers,' *Annals Math.* **140** (1994), 703-722.
2. R. C. Baker and G. Harman, 'Shifted primes without large prime factors,' *Acta Arith.* **83** (1998), 331–361.
3. D. Bleichenbacher, 'Generating ElGamal signatures without knowing the secret key,' *Lect. Notes in Comp. Sci.*, Springer-Verlag, Berlin, **1070** (1996), 10–18.
4. D. Boneh and R. Venkatesan, 'Hardness of computing the most significant bits of secret keys in Diffie–Hellman and related schemes,' *Lect. Notes in Comp. Sci.*, Springer-Verlag, Berlin, **1109** (1996), 129–142.
5. R. Crandall and C. Pomerance, *Prime numbers: a computational perspective*, Springer-Verlag, New York, 2001.
6. P. Erdős, 'On the normal number of prime factors of $p-1$ and some other related problems concerning Euler's ϕ-function,' *Quart. J. Math. (Oxford Ser.)* **6** (1935), 205–213.
7. A. Granville, 'Integers without large prime factors, in arithmetic progressions. II,' *Philos. Trans. Roy. Soc. London Ser. A* **345** (1993), 349–362.
8. A. Granville, 'Smooth numbers: computational number theory and beyond,' *Proc. MSRI Conf. Algorithmic Number Theory: Lattices, Number Fields, Curves, and Cryptography, Berkeley, 2000*, J. Buhler and P. Stevenhagen, eds., Cambridge University Press, to appear.
9. H. Halberstam and H.–E. Richert, *Sieve methods*, Academic Press, London, 1974.
10. C. Hooley, 'On Artin's conjecture,' *J. Reine Angew. Math.* **225** (1967), 209–220.
11. C. Hooley, *Applications of sieve methods to the theory of numbers*, Cambridge Tracts in Mathematics, No. 70, Cambridge University Press, Cambridge-New York-Melbourne, 1976.
12. H. W. Lenstra, Jr., 'Factoring integers with elliptic curves,' *Ann. of Math.* **2** (1987), 649–673.
13. S. Li and C. Pomerance, 'On generalizing Artin's conjecture on primitive roots to composite moduli,' *Preprint*, 2001.
14. A. J. Menezes, P. C. van Oorschot and S. A. Vanstone, *Handbook of Applied Cryptography*, CRC Press, Boca Raton, FL, 1996.
15. G. Martin, 'An asymptotic formula for the number of smooth values of a polynomial,' *J. Number Theory* **93** (2002), 108–182.
16. P. Moree, 'A note on Artin's conjecture,' *Simon Stevin* **67** (1993), 255–257.
17. M. R. Murty and F. Saidak, 'Non-abelian generalizations of the Erdős–Kac theorem,' *Preprint*, 2001.

18. C. Pomerance, ‘Popular values of Euler’s function,’ *Mathematika* **27** (1980), 84–89.
19. C. Pomerance, ‘Fast, rigorous factorization and discrete logarithm algorithms,’ *Discrete Algorithms and Complexity*, Academic Press, 1987, 119–143
20. C. Pomerance and J. Sorenson, ‘Counting the integers factorable via cyclotomic methods,’ *J. Algorithms* **19** (1995), 250–265.
21. O. Schirokauer, ‘Discrete logarithms and local units,’ *Philos. Trans. Roy. Soc. London, Ser. A* **345** (1993), 409–423.
22. O. Schirokauer, D. Weber and T. Denny, ‘Discrete logarithms: The effectiveness of the index calculus method,’ *Lect. Notes in Comp. Sci.*, Springer-Verlag, Berlin, **1122** (1996), 337–362.
23. V. Shoup, ‘Smoothness and factoring polynomials over finite fields,’ *Inform. Proc. Letters*, **38** (1991), 39–42.
24. D. R. Stinson, *Cryptography: Theory and Practice*, CRC Press, Boca Raton, FL, 1995.
25. G. Tenenbaum, *Introduction to analytic and probabilistic number theory*, University Press, Cambridge, UK, 1995.

Chinese Remaindering for Algebraic Numbers in a Hidden Field

Igor E. Shparlinski[1] and Ron Steinfeld[2]

[1] Department of Computing, Macquarie University Sydney, NSW 2109, Australia
igor@ics.mq.edu.au
[2] School of Network Computing, Monash University Frankston 3199, Australia
ron.steinfeld@infotech.monash.edu.au

Abstract. We use lattice reduction to obtain a polynomial time algorithm for Chinese Remaindering in algebraic number fields in the case when the field itself is unknown.

1 Introduction

It is well known that if we are given the residues of a rational number $\alpha = r/s \in \mathbb{Q}$ modulo sufficiently many rational primes $p_1, \ldots, p_m$, which do not divide the denominator s, then the Chinese Remainder Theorem can be used to recover α. If α is an integer $\alpha \in \mathbb{Z}$ then the result is immediate, otherwise one need to use continued fractions to recover α, see [4,7]. For algebraic number fields the situation looks more complicated. Although the Chinese Remainder Theorem can easily be extended to these fields, the actual algorithm to recover $\alpha \in \mathbb{K}$ where $\mathbb{K}$ is a given algebraic number fields, is not straightforward even if α is an algebraic integer. Having cryptographic applications in mind, one could assume that if the field $\mathbb{K}$ itself is unknown then the problem becomes intractable. One of the possible cryptographic scenarios could be fixing several algebraic numbers $\alpha_1, \ldots, \alpha_k$ and then (using linear algebra) finding a system of k polynomials

$$H_i(X_1, \ldots, X_k) \in \mathbb{Z}[X_1, \ldots, X_k], \qquad i = 1, \ldots, k,$$

of degree d with at most t coefficients (for reasonably small values of d, t and coefficient size) and such that

$$H_i(\alpha_1, \ldots, \alpha_k) = 0, \qquad i = 1, \ldots, k.$$

Making these polynomials public, one can prove the knowledge of the elements $\alpha_1, \ldots, \alpha_k$ by giving their residues $a_1, \ldots a_k$ modulo some prime ideal $\mathfrak{p}$. The verifier can easily check that

$$H_i(a_1, \ldots, a_k) \equiv 0 \pmod{\mathfrak{p}}, \qquad i = 1, \ldots, k, \tag{1}$$

(because these polynomials are sparse). On the other hand the forgery attempt via solving the system of congruences (1) seems to be infeasible. In particular no

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 349–356, 2002.

known algorithm takes advantage of sparsity of the involved polynomials. In fact all known algorithms run in time polynomial in the $(d+1)^k$, thus exponential in k, see [5,6].

Of course in this case the prime ideals which are used at each identification round in the above scheme should admit a description over $\mathbb{Q}$ and thus should not reveal (at least explicitly) any information about $\mathbb{K}$ (otherwise an attack based on the Chinese Remainder Theorem immediately applies). However, we show that even if the field $\mathbb{K}$ is unknown, using lattice reduction one can recover in polynomial time (separately for each $i \in \{1, \ldots, k\}$) the minimal polynomial of α_i from sufficiently many rational integer residues of α_i modulo prime ideals.

To be more precise we need some definitions and basic facts about prime ideals in algebraic number fields. They can all be found in [11] and in many other sources.

Let $\mathbb{Z}_{\mathbb{K}}$ be the ring of integers of $\mathbb{K}$. We recall that the degree of a prime ideal p is the integer d such that the norm $\mathrm{Nm}\,(\mathsf{p}) = p^d$, where p is a rational prime number which is divisible by p, see Section 1 of Chapter 4 of [11]. It is also known that residue ring $\mathbb{Z}_{\mathbb{K}}/\mathsf{p}$ is isomorphic to the finite field of p^d elements, $\mathbb{Z}_{\mathbb{K}}/\mathsf{p} \simeq \mathbb{F}_{p^d}$.

Let α be a root of an irreducible polynomial

$$F_\alpha(X) = f_n X^n + \ldots + f_1 X + f_0 \in \mathbb{Z}[X], \quad f_n \neq 0,$$

which is the minimal polynomial of α over $\mathbb{Z}$. We denote by $\mathcal{P}_\alpha$ the set of all rational prime numbers p with $\gcd(p, f_n) = 1$ and which have a prime divisor p of first degree in $\mathbb{Z}_{\mathbb{K}}$. In particular, $\mathbb{Z}_{\mathbb{K}}/\mathsf{p} \simeq \mathbb{F}_p$ for $p \in \mathcal{P}_\alpha$.

It is known that p has a prime divisor p of first degree if and only if $F_\alpha(X)$ has a root modulo p, see Theorem 4.12 of [11].

We also remark that almost all prime ideal of $\mathbb{Z}_{\mathbb{K}}$ are of first degree, and the density of primes which are divisible by a prime ideal of first degree is $1/n$, see Theorem 7.11 of [11].

Then for each prime number $p \in \mathcal{P}_\alpha$ and a prime ideal p of first degree which divides p, there exists an unique integer a, $0 \le a \le p-1$, such that $\alpha \equiv a \pmod{\mathsf{p}}$, where the congruence is considered in $\mathbb{Z}_{\mathbb{K}}$.

Here we prove that given m pairs (a_i, p_i) where a_i is the residue of α modulo a prime ideal p_i and $p_i = \mathrm{Nm}\,(\mathsf{p}_i) \in \mathcal{P}_\alpha$, $i = 1, \ldots, m$, one can recover α in polynomial time, provided that the product $P = p_1 \ldots p_m$ is sufficiently large. We also present some numerical results showing the effectiveness of our algorithm.

Throughout this paper, $\log z$ means the logarithm of z in base 2

2 Lattices

As we have mentioned, our algorithm is based on lattice reduction. Here recall some definitions and relevant results. For general references on lattice theory and its applications to cryptography see [9,12,13].

Let $\{\mathbf{b}_1, \ldots, \mathbf{b}_s\}$ be a set of linearly independent vectors in $\mathbb{R}^s$. The set of vectors

$$L = \left\{ \sum_{i=1}^{s} n_i \mathbf{b}_i \mid n_i \in \mathbb{Z} \right\},$$

is called an s-dimensional full rank lattice. The set $\{\mathbf{b}_1, \ldots, \mathbf{b}_s\}$ is called a *basis* of L, and L is said to be spanned by $\{\mathbf{b}_1, \ldots, \mathbf{b}_s\}$. We refer to [8] for the general background on lattices.

A basic lattice problem is the *shortest vector problem*: given a basis of a lattice L in $\mathbb{R}^s$, find a nonzero lattice vector $\mathbf{v} \in L$ of the smallest possible Euclidean norm among all lattice vectors. The shortest vector problem generally refers to the Euclidean norm, but of course, other norms are possible as well. Although the shortest vector problem appears to **NP**-hard various approximate polynomial time algorithms can be designed, see [9,12,13] for references.

We use the best known approximation polynomial-time result for the shortest vector problem given in Corollary 15 of [1].

Lemma 1. *For any constant $\tau > 0$, there exists a randomised polynomial-time algorithm which, given an s-dimensional full rank lattice L, finds a lattice vector $\mathbf{v}$ satisfying with probability exponentially close to 1 the inequality*

$$\|\mathbf{v}\| \le 2^{\tau s \log\log s / \log s} \min\{\|\mathbf{z}\| \ : \ \mathbf{z} \in L\}.$$

Proof. By taking $k = \lceil c \log n \rceil$ in Corollary 15 of [1] where $c > 0$ is a sufficiently large constant, we obtain a randomised polynomial-time algorithm which approximates the shortest vector within $2^{\tau s \log\log s / \log s}$ for any constant $\tau > 0$. □

The best deterministic polynomial-time algorithm known for this problem has a slightly larger approximation factor $2^{\tau s \log^2 \log s / \log s}$, see [14].

3 Algorithm

Assume we are given m primes $p_i \in \mathcal{P}_\alpha$ and m rational integers $a_i \in \mathbb{Z}$ with $\alpha \equiv a_i \pmod{\mathsf{p}_i}$, where $\mathrm{Nm}(\mathsf{p}_i) = p_i$, $i = 1, \ldots, m$. To recover the minimal polynomial F_α we consider the lattice $\mathcal{L}$ formed by the solutions $(x_0, \ldots, x_n) \in \mathbb{Z}^{n+1}$ of the system of the congruences

$$x_0 + x_1 a_i + \ldots + x_n a_i^n \equiv 0 \pmod{p_i}, \qquad i = 1, \ldots, m.$$

Equivalently, $\mathcal{L}$ consists of the solutions $(x_0, \ldots, x_n) \in \mathbb{Z}^{n+1}$ of the congruence

$$x_0 + x_1 A_1 + \ldots + x_n A_n \equiv 0 \pmod{P},$$

where $P = p_1 \ldots p_m$ and A_j, $0 \le A_j \le P - 1$, is defined by

$$A_j \equiv a_i^j \pmod{p_i}, \qquad i = 1, \ldots, m.$$

for $j = 1, \dots, n$. One verifies that $\mathcal{L}$ is spanned by the rows of the following $(n+1) \times (n+1)$ matrix

$$B = \begin{pmatrix} P & 0 \dots 0\, 0 \\ -A_1 & 1 \dots 0\, 0 \\ \vdots & \vdots \ddots \vdots\, \vdots \\ -A_{n-1} & 0 \dots 1\, 0 \\ -A_n & 0 \dots 0\, 1 \end{pmatrix}. \qquad (2)$$

It is easy to see that $\mathcal{L}$ contains the vector of coefficients

$$\mathbf{f} = (f_0, \dots, f_n)$$

of F_α. Thus, if the determinant $\det B = P$ is sufficiently large compared to $\|\mathbf{f}\|$ then one can hope that all other lattice vectors, which are not parallel to $\mathbf{f}$ are much longer. Hence a lattice reduction algorithm has to output a multiple of $\mathbf{f}$. Below we show that indeed under certain natural conditions this property can be rigorously established.

We define the *size* $s(G)$ of a polynomial

$$G(X) = g_n X^n + \dots + g_1 X + g_0 \in \mathbb{Z}[X]$$

as

$$s(G) = \left(\sum_{j=0}^{n} |g_j|^2 \right)^{1/2}.$$

Theorem 1. *Given integers $a_1, \dots, a_m \in \mathbb{Z}$ and primes $p_1, \dots, p_m \in \mathcal{P}_\alpha$ with $\alpha \equiv a_i \pmod{\mathsf{p}_i}$, where p_i is a prime ideal of first degree which divides p_i, $i = 1, \dots, m$, and such that for some constant $\tau > 0$*

$$p_1 \dots p_m > 2^{\tau n^2 \log\log(n+1)/\log(n+1)} s(F_\alpha)^{2n},$$

one can find α in randomised polynomial in n and $\log p_1, \dots, \log p_m$ time with probability exponentially close to 1.

Proof. We use the Chinese Remainder Theorem (over the integers) to find $A_1, \dots, A_n$ and thus to construct the matrix B given by (2) and the corresponding lattice $\mathcal{L}$. Obviously this can be done in time polynomial in n and $\log p_1, \dots, \log p_m$. Assume that the algorithm of Lemma 1 outputs a vector $\mathbf{g} = (g_0, \dots, g_n) \in \mathcal{L}$.

Suppose that $\mathbf{g}$ is not parallel to $\mathbf{f}$. Let us consider the polynomial

$$G(X) = g_n X^n + \dots + g_1 X + g_0 \in \mathbb{Z}[X].$$

Let us denote by R the resultant of the polynomials G and F_α. Because F_α is irreducible and G is not a multiple of F we conclude $R \neq 0$.

We have $F_\alpha(a_i) \equiv F_\alpha(\alpha) \equiv 0 \pmod{\mathsf{p}_i}$, and using that $F_\alpha(a_i) \in \mathbb{Z}$ we derive $F_\alpha(a_i) \equiv 0 \pmod{p_i}$, $i = 1, \ldots, m$.

We also see that $G(a_i) \equiv 0 \pmod{p_i}$ thus the polynomials G and F_α have a common root modulo p_i which implies $R \equiv 0 \pmod{p_i}$, $i = 1, \ldots, m$.

Therefore $R \equiv 0 \pmod{p_1 \ldots p_m}$ which implies $|R| \geq p_1 \ldots p_m$.

On the other hand, from Lemma 1 we conclude that

$$s(G) \leq 2^{\tau n \log\log(n+1)/\log(n+1)} s(F_\alpha).$$

Using the Hadamard inequality, we derive

$$|R| \leq s(G)^n s(F_\alpha)^n \leq 2^{\tau n^2 \log\log(n+1)/\log(n+1)} s(F_\alpha)^{2n}$$

which is impossible because of the assumption of the theorem. Therefore $\mathbf{g}$ is parallel to $\mathbf{f}$ and one can now easily find $\mathbf{f}$. □

Certainly if an oracle for exact solving the shortest vector problem is available then the condition of Theorem 1 simplifies to

$$p_1 \ldots p_m > s(F)^{2n}. \tag{3}$$

We remark that if n is small thus such exact algorithms of [1,10] become feasible.

4 Numerical Experiments

We ran several experiments to test the practical performance of our algorithm. For these experiments we used the integer-arithmetic implementation of the LLL lattice reduction algorithm from [15]. We also used the algorithm of Berlekamp from [15] to test whether a prime p has a prime ideal divisor p of first degree and also to find a residue a, $0 \leq a \leq p-1$, of α modulo p.

Our results are tabulated in Table 1 found in the Appendix. In these experiments, we swept n, m, and $\log s(F_\alpha)$ through the tabulated values. For each triple $(n, m, \log s(F_\alpha))$, we searched for the smallest prime product $P = p_1 \ldots p_m$ such that the algorithm still succeeded to recover the F_α in at least one of three successive runs with independently chosen primes. In Table 1 we denote the (base 2) logarithm of this smallest prime product by $\log P_{min}$. We also tabulate for comparison our proven upper bound on $\log P_{min}$ with respect to a shortest vector oracle, as obtained from Theorem 1, namely $2n \log s(F_\alpha)$. Note that for P we always used m equal length random primes in the interval $[2^{\ell-0.25}, 2^{\ell+0.25}]$, and we varied the parameter ℓ to vary P. The tabulated running times are for a 600MHz Pentium PC running Windows-NT.

The results in Table 1 show that our algorithm performed better than our proven bound of Theorem 1 and even better than predicted by its modification (3) for a shortest vector oracle. In particular, the experimental value for $\log P_{min}$ was in most cases only slightly over half of the upper bound of Theorem 1, that is approximately $n \log s(F_\alpha)$. This demonstrates that typically the lattice $\mathcal{L}$ (as this usually happens for a random lattice) has at most one

Table 1. Experimental results.

n	m	$\log s\,(F_\alpha)$	$\log P_{min}$	$2n \log s\,(F_\alpha)$	Run Time (sec)
2	2	60	178	240	0.01
2	2	80	240	320	0.02
2	2	100	298	400	0.05
2	4	60	180	240	0.02
2	4	80	240	320	0.02
2	4	100	300	400	0.04
2	8	60	176	240	0.02
2	8	80	240	320	0.02
2	8	100	304	400	0.03
6	6	60	414	720	0.64
6	6	80	558	960	1.41
6	6	100	696	1200	2.42
6	12	60	420	720	0.74
6	12	80	564	960	1.54
6	12	100	696	1200	2.49
6	24	60	432	720	0.76
6	24	80	552	960	1.38
6	24	100	696	1200	2.40
10	10	60	660	1200	6.6
10	10	80	880	1600	13.39
10	10	100	1100	2000	24.08
10	20	60	660	1200	6.44
10	20	80	880	1600	14.05
10	20	100	1100	2000	24.68
10	40	60	680	1200	7.07
10	40	80	880	1600	13.61
10	40	100	1120	2000	26.02
14	28	20	308	560	1.95
14	28	40	616	1120	11.73
14	28	60	896	1680	29.88
14	28	80	1204	2240	65.13
18	36	20	396	720	6.49
18	36	40	756	1440	33.41
18	36	60	1116	2160	90.3
18	36	80	1512	2880	210.91
22	44	20	484	880	16.64
22	44	40	924	1760	86.24
22	44	60	1364	2640	241.07
22	44	80	1848	3520	556.97
60	60	80	5340	9600	87021.5

vector (up to a multiplication by a constant) of length significantly less than $|\det B|^{1/n} = P^{1/n}$. We recall that the Minkowski theorem, see Theorem 5.3.6 in Section 5.3 of [8] guarantees the existence at least one vector $\mathbf{v} \in \mathcal{L}$ of length $\|\mathbf{v}\| \le \gamma_n^{1/2} |\det B|^{1/n}$, where γ_n is the Hermite constant, for which we have

$$\frac{1}{2\pi e} n + o(n) \le \gamma_n \le \frac{1.744}{2\pi e} n + o(n), \qquad n \to \infty,$$

see the inequality (48) in Chapter 1 [3]. On the other hand, heuristically, if there is a vector which is substantially shorter than that, this happens only because of some special reason (which in our case happens for the vector $\mathbf{f}$ of coefficients of $\mathbb{F}_\alpha$). It is also useful to remember that typically lattice reduction algorithms behave much better than their theoretical prediction. Thus this explains why the algorithm works correctly already for the values $\log P \approx n \log s\,(F_\alpha)$. It would be interesting to improve the bound of Theorem 1 and bring it closer to this heuristic observation.

The observed running time of our algorithm was also increasing less quickly as a function of n than expected from theoretical bounds. In particular, it is known [2] that the running time of LLL on a lattice of dimension n is upper bounded as $O\left(n^6 \log^3 H\right)$, if the input basis vectors have Euclidean norm at most H. For our lattice $H = O\left(P_{min}\right) = O\left(s(F_\alpha)^n\right)$ so the running time bound for LLL in our algorithm is $O\left(n^9 \log^3 s\,(F_\alpha)\right)$. The observed running times in Table 1 behave approximately as only $O\left(n^5 \log^3 s\,(F_\alpha)\right)$. Again this is in complete agreement with the well known fact that lattice reduction algorithms usually exhibit better characteristics (in both running time and the precision of the result) than their theoretical prediction.

Finally we remark that it is natural to ask why in the Introduction we describe a scheme which we immediately break. We believe that the idea of making calculations in a hidden field deserves further exploration and it is possible that a safe cryptographic construction, based on this idea may exist.

Acknowledgement

The authors thank Amin Shokrollahi for sharing his idea of using a hidden number field in cryptography.

References

1. M. Ajtai, R. Kumar and D. Sivakumar, 'A sieve algorithm for the shortest lattice vector problem', *Proc. 33rd ACM Symp. on Theory of Comput.*, Crete, Greece, July 6-8, 2001, 601–610.
2. H. Cohen, *A course in computational algebraic number theory*, Springer-Verlag, Berlin, 1993.
3. J. H. Conway and N. J. A. Sloan, *Sphere packings, lattices and groups*, Springer-Verlag, Berlin, 1998.
4. C. Ding, D. Pei and A. Salomaa, *Chinese Remainder Theorem: Applications in computing, coding, cryptography*, World Scientific, Singapore, 1996.

5. J. von zur Gathen, 'Irreducibility of multivariate polynomials', *J. Comp. and Syst. Sci.*, **31** (1985), 225–264.
6. J. von zur Gathen and E. Kaltofen, 'Factoring sparse multivariate polynomials', *J. Comp. and Syst. Sci.*, **31** (1985), 265–287.
7. R. T. Gregory and E. V. Krishnamurthy, *Methods and applications of error-free computation*, Springer-Verlag, Berlin, 1984.
8. M. Grötschel, L. Lovász and A. Schrijver, *Geometric algorithms and combinatorial optimization*, Springer-Verlag, Berlin, 1993.
9. A. Joux and J. Stern, 'Lattice reduction: A toolbox for the cryptanalyst', *J. Cryptology*, **11** (1998), 161–185.
10. R. Kannan, 'Algorithmic geometry of numbers', *Annual Review of Comp. Sci.*, **2** (1987), 231–267.
11. W. Narkiewicz, *Elementary and analytic theory of algebraic numbers*, Polish Sci. Publ., Warszawa, 1990.
12. P. Q. Nguyen and J. Stern, 'Lattice reduction in cryptology: An update', *Lect. Notes in Comp. Sci.*, Springer-Verlag, Berlin, **1838** (2000), 85–112.
13. P. Q. Nguyen and J. Stern, 'The two faces of lattices in cryptology', *Lect. Notes in Comp. Sci.*, Springer-Verlag, Berlin, **2146** (2001), 146–180.
14. C. P. Schnorr, 'A hierarchy of polynomial time basis reduction algorithms', *Theor. Comp. Sci.*, **53** (1987), 201–224.
15. V. Shoup, 'NTL: A library for doing number theory (version 5.2b)', `http://-shoup.net/ntl/`, 2001.

Appendix

This Appendix contains the tabulation of experimental results (Table 1) we have referred to in Section 4.

In our largest example we recovered a polynomial of degree $n = 60$ with 80–bit coefficients in just over 24 hours (this is one we have only one such example). We remark that in this case we did not find P_{min} exactly but just made one attempt with a fixed value for P. This value of $P \approx 1.1n \log s\,(F_\alpha)$ was predicted to equal P_{min} for the chosen parameters ($n = 60$, $\log s\,(F_\alpha) = 80$), based on our numerous experiments with smaller polynomials.

For our smaller examples in every case the running time was less than 10 minutes and we found exact values of P_{min}.

An Algorithm for Computing Weierstrass Points

Florian Hess

Computer Science Department,
Woodland Road, University of Bristol, BS8 1UB, UK

Abstract. We develop algorithms for computing differentiations and Weierstrass points of algebraic curves in any characteristic. As an application we explain how this can be used to compute special models of curves together with a map to $\mathbb{P}^1$ of low degree.

1 Introduction

A Weierstrass point on a non-singular irreducible algebraic curve is a point P for which there exist functions on the curve with unusual pole orders at P and no poles everywhere else. The finite set of Weierstrass points forms an important invariant of a curve which is of particular use for the study of automorphisms. Using Weierstrass points it can for example be shown that the automorphism group of a curve is finite.

The theory of Weierstrass points over the base field $\mathbb{C}$ dates back to the 19th century. The generalization of the theory to base fields of any characteristic was carried out around 1935 in a series of works [4,5,6,13,14,18], most notably by F. K. Schmidt. Over finite fields as base fields a variation of the theory yields a proof of the Riemann hypothesis for curves and several improvements on it [17].

In this paper we will focus on algorithmic aspects of the theory in arbitrary characteristic. Using the framework in [7] we describe algorithms for the computation of differentiations (alias higher derivatives) and Weierstrass places, speaking in terms of the function field of a given curve. As an application of these algorithms we devise an algorithm for computing special models of curves together with a map to $\mathbb{P}^1$ of low degree. In practice, using simplified instead of unwieldy models speeds up computations considerably. One among many examples for this is the integration of elliptic and hyperelliptic functions, see the discussion in [8]. The algorithms of this work have been implemented in Kash [11] and Magma [1,2].

We also give a brief exposition of the main statements of the theory of Weierstrass points in arbitrary characteristic as in [6,13,14,15,17]. For the convenience of the reader we do provide proofs since the prior expositions are different or somewhat unaccessible.

2 Preliminaries

For the purpose of the paper we will focus on the function fields of curves rather than the curves themselves. By F/k we denote throughout an algebraic function

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 357–371, 2002.

field of transcendence degree one over the exact constant field k. We refer to [16] for the theory of algebraic function fields. We assume further that F/k has a separating element and is conservative, i.e. its genus g is invariant under constant field extensions. There are further restrictions on k for the algorithms in [7] to work, but k perfect is a sufficient condition. By [7] we further assume that we have algorithms to compute in F/k as a field and $k(x)$-vector space for x a separating element, that we can compute with places, divisors and Riemann-Roch spaces $\mathcal{L}(D) = \{a \in F^\times \mid (a) + D \geq 0\} \cup \{0\}$ for divisors D. In addition we will need to compute with differentials in F/k. Implementations of such algorithms are available in Kash [11] and Magma [1,2].

The algebraic closure of k is denoted by $\bar{k}$ and the conorm map from F/k to a constant field extension Fk_1/k_1 by $\mathrm{con}_{Fk_1/F}$. Also, $i(D)$ is the index of speciality of the divisor D.

3 Weierstrass Places

In this section we state the main definitions and theorems about Weierstrass places. More generally we will consider D-Weierstrass places for D a divisor, which occurred first in [12].

Definition 1. *Let D be a divisor and P a place of degree one of F/k. The number $\mu \in \mathbb{Z}^{\geq 1}$ is called D-gap number of P if $\mathcal{L}(D + (\mu - 1)P) = \mathcal{L}(D + \mu P)$ and it is called D-pole number of P if equality does not hold.*

In other words, an integer μ is a D-pole number of P precisely if there is an element $a \in F$ such that $v_{P'}(a) + v_{P'}(D) \geq 0$ for all places $P' \neq P$ and $v_P(a) + v_P(D) + \mu = 0$. Typically one thinks of $D = 0$ in which case D-gap and D-pole numbers are simply called gap and pole numbers respectively.

We remark that if μ is not a D-gap number then $\dim(D + (\mu - 1)P) = \dim(D + \mu P) - 1$ since $\deg(P) = 1$. Also, linearly equivalent divisors have the same gap numbers for every P.

Theorem 2. *All but finitely many places of degree one from constant field extensions Fk_1/k_1 have the same $\mathrm{con}_{Fk_1/F}(D)$-gap numbers.*

Proof. Follows from Corollary 20 and the remark after Algorithm 31.

Definition 3. *The D-gap numbers of F/k are defined to be the numbers common to almost all places in Theorem 2. A place of degree one of F/k is called D-Weierstrass place if its D-gap numbers are different from the D-gap numbers of F/k.*

One of the main tasks of this paper is to explain how to compute D-gap numbers and D-Weierstrass places of F/k.

Proposition 4. *Every D-gap number μ of P satisfies $1 \leq \mu \leq 2g - 1 - \deg(D)$. There are $i(D)$ many D-gap numbers of P.*

Proof. Follows from Proposition 12 and Proposition 14.

The usual Weierstrass places and gap numbers of F/k are the D-Weierstrass places and D-gap numbers for $D = 0$ the zero divisor. The numbers in $\mathbb{Z}^{\geq 1}$ which are not gap numbers at a place P are called pole numbers of P. They occur as pole orders of elements of F/k at P and form an additive semigroup, the so called Weierstrass semigroup at P.

Theorem 5. *There exists at least one Weierstrass place of $F\bar{k}/\bar{k}$ for $g \geq 2$. For $g \in \{0, 1\}$ there are no Weierstrass places.*

Proof. Follows from Corollary 21 (the ramification divisor of F/k is zero if and only if $g \in \{0, 1\}$).

4 Differentiations

Classically Weierstrass places are related with differential calculus and higher derivatives by the Wronskian determinant, to be explained in the next section. In positive characteristic p every j-th derivative $d^j a/dx^j$ for $a, x \in F$ and x separating vanishes identically for $j \geq p$, making them useless for forming the Wronskian determinant. Thus j-th derivatives have to be defined differently in positive characteristic, leading to differentiations. The rest of the section is however valid in any characteristic.

Definition 6. *Let S/R be an unitary extension of entire rings. A differentiation of R is a homomorphism $D : R \to S[[t]]$, $a \mapsto \sum_{i=0}^{\infty} D^{(i)}(a)t^i$ such that $D^{(0)}(a) = a$ for all $a \in R$. The differentiation D is called iterative if its image is contained in $R[[t]]$ and $D^{(i)} \circ D^{(j)} = \binom{i+j}{i} \cdot D^{(i+j)}$.*

We remark that $D^{(i)}$ will take over the role of an i-th derivative. Let $D : R \to R[[t]]$ be a differentiation. Upon identifying R with $D(R)$ we obtain a differentiation $D' : D(R) \to D(R)[[t']]$, $b \mapsto \sum_{j=0}^{\infty} D(D^{(j)}(D^{-1}(b)))t'^j$. On the other hand we get a differentiation $D_t : D(R) \to R[[t]][[t']]$ defined by $t \mapsto t + t'$. We have that D is iterative if and only if $D' = D_t$. Indeed, let $a \in R$ and $b = D(a) = \sum_{i=0}^{\infty} D^{(i)}(a)t^i$. Clearly $D'(b) = \sum_{j=0}^{\infty} t'^j \sum_{i=0}^{\infty} D^{(i)}(D^{(j)}(a))t^i$. On the other hand a straightforward calculation shows $D_t(b) = \sum_{j=0}^{\infty} t'^j \sum_{i=j}^{\infty} \binom{i}{j} D^{(i)}(a)t^{i-j}$. Substituting i by $i + j$ yields the equivalence. A trivial example of an iterative differentiation is $D(a) = a$, where every element in R can be regarded as an absolute differentiation constant.

Let S be a field. A differentiation of $D : R \to S[[t]]$ can be extended in precisely one way to the field of fractions $Q(R)$ of R because of $D^{(0)}(a) = a$. Let R be a field and $\alpha \in S$. Assume α is a root of the monic and separable polynomial $f \in R[x]$. Denote by $D(f) \in D(R)[x]$ the polynomial obtained from f by applying D coefficientwise. By Hensel's lemma, $D(f)$ has a unique root $\alpha' \in S[[t]]$ such that $\alpha' = \alpha \bmod t$. Hence there is precisely one extension of D to $R[\alpha]$, given by $\alpha \mapsto \alpha'$. In both cases, if D is iterative then its extension is also iterative. To see this assume that $D : R \to R[[t]]$ is iterative and let $\hat{D}$ be the extension of D. The image of $\hat{D}$ is contained in $Q(R)[[t]]$ and $R[\alpha][[t]]$ respectively. Since D is iterative we have $D' = D_t$. But $\hat{D}'$ and $\hat{D}_t$ are extensions of D' and D_t respectively. Hence $\hat{D}' = \hat{D}_t$ because of the proven uniqueness properties, and $\hat{D}$ is iterative. If α is transcendental over R we can extend D in more than one way. The main example is to extend by $\alpha \mapsto \alpha+t$. Then $D^{(0)}(\alpha) = \alpha$, $D^{(1)}(\alpha) = 1$ and $D^{(j)}(\alpha) = 0$ for $j > 1$. Hence $D'(\alpha+t) = (\alpha+t)+t'$. But $D_t(\alpha+t) = \alpha+(t+t')$ so this extension is iterative if D on R is.

Definition 7. *A differentiation of a function field F/k is a differentiation of F such that $D(a) = a$ for all $a \in k$. The differentiation is called with respect to x and written D_x if $x \in F$ is a separating element and $D(x) = x + t$.*

Lemma 8. *For every separating element $x \in F$ there is exactly one differentiation D_x with respect to x. Furthermore, D_x is iterative and $D_x^{(1)} = d/dx$.*

Proof. The first statements follow from the above discussion, extending the trivial differentiation $D(a) = a$ from k to F via $x \mapsto x + t$. For the last we have $D_x^{(1)} = d/dx$ on $k[x]$ because of $x \mapsto x+t$. This is then also true on the separable extension $F/k(x)$ by the uniqueness of extending the derivation d/dx, since $D_x^{(1)}$ is also a derivation on F.

Lemma 9. *Let P be a place of degree one of F/k and $\pi \in F$ a local uniformizer at P. Let $\phi : F \longrightarrow k((\pi))$ be the homomorphism which maps elements of F to their P-adic expansions. Then $\phi(D_\pi^{(j)}(a)) = \sum_{i=i_0}^{\infty} \binom{i}{j} a_i \pi^{i-j}$ for $\phi(a) = \sum_{i=i_0}^{\infty} a_i \pi^i$ and $a \in F$.*

Proof. We obtain a differentiation on $k((\pi))$ by $\pi \mapsto \pi + t$ which restricts to a differentiation of F via the embedding ϕ and which extends the differentiation D_π on $k[\pi]$. Since π is separating, both differentiations must coincide. The result now follows from the binomial series of $(\pi + t)^j$.

By Lemma 8 and the iterativity property we have $D_x^{(j)} = j!^{-1} d^j/dx^j$ in characteristic zero. Lemma 9 also shows $D_\pi : \mathfrak{o}_P \to \mathfrak{o}_P[[t]]$ and that $\phi : \mathfrak{o}_P \to k[[\pi]]$ is obtained from following D_π by the coefficientwise reduction mod P and substituting π for t.

Consider the field of fractions $\widetilde{F}$ of the Dedekind domain $F \otimes_k F$. F has two embeddings $1 \otimes_k F$ and $F \otimes_k 1$ into $\widetilde{F}$. We write $F_* = 1 \otimes_k F$ and identify $F = F \otimes_k 1$. For $a = a' \otimes_k 1$ in F we denote the corresponding element $1 \otimes_k a'$ in F_* by a_*. Now $\widetilde{F}$ is a function field over the exact constant field F_* and the generic place P_F of F is the place of degree one of $\widetilde{F}/F_*$ whose residue class map restricted to F is given by $a \mapsto a_*$. The function field $\widetilde{F}/F_*$ is the constant field extension of F by F_* and P_F is equivalent to the generic point on a curve having F/k as a function field. Since P_F is of degree one we have an embedding $\widetilde{F} \longrightarrow F_*((t))$ where $t \in \widetilde{F}$ is a local uniformizer of P_F. We have that $x \in F$ is separating precisely if $x - x_*$ is a local uniformizer for P_F. The restriction of this embedding to F yields an embedding $\phi_t : F \longrightarrow F_*[[t]]$.

Theorem 10. *Upon identifying F and F_* we have $D_x = \phi_{x-x_*}$ for any separating element $x \in F$.*

Proof. Identifying F and F_* the map ϕ_{x-x_*} clearly defines a differentiation of F. Restricted to $k[x]$ it is given by $x \mapsto x_* + t$ since this expresses a polynomial in the local uniformizer $t = x - x_*$. On $k[x]$ it is hence equal to D_x. Since x is separating the equality of D_x and ϕ_{x-x_*} on F follows from the uniqueness of extensions of differentiations.

We can extend D_x to $\widetilde{F}$ by setting $D_x(a) = a$ for $a \in F_*$, or by F_*-linearity in other words. This equals the differentiation D_x obtained by x viewed as separating element of $\widetilde{F}$ and we also have $D_x = \phi_{x-x_*}$ on $\widetilde{F}$.

By Theorem 10 the change of separating element for a differentation has the effect of changing the local uniformizer. More precisely, let $x, y \in F$ be separating. Then $\phi_{y-y_*}(x - x_*) = D_y(x - x_*) = D_y(x) = dx/dy \cdot t + O(t^2)$ and this series has to be substituted for t in $D_x(a)$ in order to obtain $D_y(a)$ for $a \in F$. This discussion yields the usual chain rule for (higher) derivatives. The other familiar rules follow from the properties of a differentiation as in Definition 6 and 7.

5 Orders and Ramification Divisors of Linear Systems

Let L be a linear system of F/k. Recall that L is a set of effective divisors $\{\, (a) + E \,|\, a \in V \backslash \{0\} \,\}$ for a divisor E and some k-linear subspace V of $\mathcal{L}(E)$. We say that L is defined by E and V. The complete linear system L defined by E is the linear system defined by E and $\mathcal{L}(E)$. If L is defined by E and V then it is clearly also defined by $E - (a)$ and aV for any $a \in F^\times$. Furthermore, for any $E \in L$ we have that L is defined by E and the k-linear space V generated by $\{\, a \in F^\times \,|\, (a) = D - E \text{ for } D \in L \,\}$. So alternatively one can think of L as an equivalence class of tuples (E, V), where $(E, V) \sim (E - (a), aV)$.

In the following we write $\deg(L) := \deg(E)$ and $\dim(L) := \dim(V)$. Also, let $L(\mu P) := \{\, D \in L \,|\, v_P(D) \geq \mu \,\}$ for $\mu \in \mathbb{Z}^{\geq 0}$.

Definition 11. *Let L be a linear system and P a place of degree one. The integer $\mu \in \mathbb{Z}^{\geq 0}$ is called (Wronskian) order of L at P if $L(\mu P) \neq L((\mu + 1)P)$.*

Any P for which 0 is not an order of L is called a base point of L. Let L be a linear system defined by E and V. We write $V(\mu P) := \{\, a \in V \,|\, v_P(a) \geq \mu \,\}$. Then $V(\mu P) \neq V((\mu + 1)P)$ if and only if μ is an order of L.

Proposition 12. *Let L be the complete linear system defined by $W - D$. Then μ is a D-gap number of P if and only if $\mu - 1$ is an order of L at P.*

Proof. Abbreviate $V = \mathcal{L}(W - D)$. From the theorem of Riemann-Roch we obtain

$$\dim(D + (\mu - 1)P) - \dim(D + \mu P) = i(D + (\mu - 1)P) - i(D + \mu P) - 1,$$

so μ is a D-gap number if and only if $i(D + (\mu - 1)P) > i(D + \mu P)$. We have $\dim V(\mu P) = i(D + \mu P)$, so this is equivalent to $V((\mu - 1)P) \neq V(\mu P)$ and $\mu - 1$ being an order of V at P.

According to Proposition 12, in order to compute gap numbers and Weierstrass places it suffices to investigate the orders of L at various places.

Proposition 14. *Every order μ of L at P satisfies $0 \le \mu \le \deg(L)$. There are $\dim(L)$ orders of L at P.*

Proof. Let L be defined by E and V such that $v_P(E) = 0$. Then $V(\mu P) \subseteq \mathcal{L}(E - \mu P)$. By definition, $\mu \ge 0$. Furthermore, for $\mu > \deg(L) = \deg(E)$ we have $\dim V(\mu P) \le \dim(E - \mu P) = 0$ and hence $V(\mu P) = V((\mu + 1)P) = 0$, so this μ is not an order. Thus $0 \le \mu \le \deg(L)$ for orders. In order to prove that there are $\dim(L) = \dim(V)$ orders we take $\phi : F \longrightarrow k((\pi))$ to be a P-adic expansion map. We have that k-linearly independent elements are mapped to k-linearly independent series. Using a Gaussian elimination process we see that there is a unique basis w_i of V such that $\phi(w_i) = \pi^{\mu_i} + O(\pi^{\mu_i+1})$ and $0 \le \mu_1 < \cdots < \mu_{\dim(L)}$. Thus the μ_i are all $\dim(L)$ orders of L.

We now want to investigate the orders of L defined by E and V for almost all places simultaneously. For this we consider the linear system L as a linear system of $\tilde{F}$ via the conorm map and investigate it at the generic place P_F.

Definition 15. *The orders of L are defined to be the orders of L at P_F.*

Proposition 16. *Let L be defined by E and V and let $v_1, \ldots, v_n$ be a k-basis of V. Let x be a separating element of F/k. The orders of L are the lexicographically smallest integers $0 = \varepsilon_1 < \cdots < \varepsilon_n$ such that $\det(D_x^{(\varepsilon_i)}(v_j))_{i,j} \neq 0$. The following transformation properties hold: If $w_i = \sum_j \lambda_{i,j} v_j$ and $(\lambda_{i,j})_{i,j} \in k^{n \times n}$ then $\det(D_x^{(\varepsilon_i)}(w_j))_{i,j} = \det(\lambda_{i,j})_{i,j} \det(D_x^{(\varepsilon_i)}(v_j))_{i,j}$. Moreover $\det(D_x^{(\varepsilon_i)}(av_j))_{i,j} = a^n \det(D_x^{(\varepsilon_i)}(v_j))_{i,j}$ for $a \in F$. If y is a separating variable then $\det(D_y^{(\varepsilon_i)}(v_j))_{i,j} = (dx/dy)^{\sum_i \varepsilon_i} \det(D_x^{(\varepsilon_i)}(v_j))_{i,j}$. We have $\varepsilon_i \le \mu_i$ if $0 \le \mu_1 < \cdots < \mu_n$ are integers such that $\det(D_x^{(\mu_i)}(v_j))_{i,j} \neq 0$, or if the μ_i are the orders of L at a place P.*

Proof. Using Theorem 10 and its notation we have $D_x = \phi_{x-x_*}$, so Proposition 16 is nothing else but a proposition about P_F-adic expansions. The first transformation property is clear by the F_*-linearity of ϕ_{x-x_*}. Analogous to the proof of Proposition 14 we can consider a basis w_i of $\sum_i F_* v_i$ of the form $\phi_{x-x_*}(w_i) = b_i t^{\varepsilon_i} + O(t^{\varepsilon_i+1})$ with some $b_i \in F_*^\times$, obtained by a transformation of determinant one. Clearly $\det(D_x^{(\varepsilon_i)}(v_j))_{i,j} = \det(D_x^{(\varepsilon_i)}(w_j))_{i,j} = \prod_i b_i$ and the lexicographical minimality of the ε_i and $\varepsilon_1 = 0$ follow immediately. Furthermore, multiplication by a and changing the local uniformizer are F_*-linear operations, so any two bases with equal determinant are mapped to bases with equal determinant by these operations. Because of $\phi(aw_i) = a_* b_i t^{\varepsilon_i} + O(t^{\varepsilon_i+1})$ for $a_* \in F_*$ and $\phi_{y-y_*}(w_i) = b_i(dx_*/dy_*)^{\varepsilon_i} t^{\varepsilon_i} + O(t^{\varepsilon_i+1})$ the two transformation statements follow immediately. Finally, because of the construction, ε_r is the smallest index such that the span of the $(D_x^{(m)}(v_j))_j$ for $0 \le m < \varepsilon_r$ has dimension $r - 1$. This implies the first statement about the μ_i. For the second we may assume that $v_P(E) = 0$. But then $\det(D_x^{(\mu_i)}(v_j))_{i,j} \neq 0$ from the proof of Proposition 14.

Definition 17. *Let L be a linear system defined by E and V. Let $v_1, \ldots, v_n$ be a basis of V, x a separating element of F/k and $\varepsilon_1, \ldots, \varepsilon_n$ be the orders of L. The divisor $R(L) := (\det(D_x^{(\varepsilon_i)}(v_j))_{i,j}) + (\sum_i \varepsilon_i)(dx) + nE$ is called ramification divisor of L.*

By Proposition 16 the ramification divisor depends indeed only on L. The determinants in Proposition 16 are called Wronskian determinants.

Theorem 18. *Let L be a linear system, ε_i the orders of L and μ_i the orders of L at P. Then for the valuation $v_P(R(L)) \geq \sum_{i=1}^{\dim(L)}(\mu_i - \varepsilon_i)$ and equality holds if and only if $\det(\binom{\mu_i}{\varepsilon_j})_{i,j} \neq 0$ in k.*

Proof. [17] Let L be defined by E and V such that $v_P(E) = 0$. We take $\phi : F \longrightarrow k((\pi))$ to be a P-adic expansion map. Let w_i be a basis of V with $\phi(w_i) = \pi^{\mu_i} + O(\pi^{\mu_i+1})$. Using Lemma 9 we obtain

$$\begin{aligned} \det\bigl(\phi(D_x^{(\varepsilon_i)}(w_j))\bigr)_{i,j} &= \det\left(\binom{\mu_j}{\varepsilon_i}\pi^{\mu_j-\varepsilon_i} + O(\pi^{\mu_j-\varepsilon_i+1})\right)_{i,j} \\ &= \det\left(\binom{\mu_j}{\varepsilon_i}\pi^{\mu_j} + O(\pi^{\mu_j+1})\right)_{i,j} \cdot \pi^{-\sum_i \varepsilon_i} \\ &= \det\left(\binom{\mu_j}{\varepsilon_i}\right)_{i,j} \cdot \pi^{\sum_i(\mu_i-\varepsilon_i)} + O(\pi^{1+\sum_i(\mu_i-\varepsilon_i)}). \end{aligned}$$

Lemma 19. *If ε is an order of L and $\mu \in \mathbb{Z}^{\geq 0}$ such that $\binom{\varepsilon}{\mu} \neq 0$ in k then μ is also an order of L. In particular, if $p = 0$ or $p > \deg(L)$ then $0, \ldots, \dim(L) - 1$ are the orders of L.*

Proof. [17] Let $\mu_1, \ldots, \mu_n$ be the orders of L at P and $\varepsilon_1, \ldots, \varepsilon_n$ the orders of L. If $0 \leq \nu_1 < \cdots < \nu_n$ are integers such that $\det(\binom{\mu_i}{\nu_j})_{i,j} \neq 0$ in k then $\varepsilon_i \leq \nu_i$ for $1 \leq i \leq n$. Indeed, as in the proof of Theorem 18, $\det(D_x^{(\nu_i)}(w_j))_{i,j} = \det(\binom{\mu_i}{\nu_j})_{i,j}\pi^{\sum_i(\mu_i-\nu_i)} + \cdots \neq 0$. The assertion now follows from Proposition 16.

Since $\binom{\varepsilon}{\mu} \neq 0$ we have $0 \leq \mu \leq \varepsilon$. Since $\mu = 0$ is an order we may assume $\mu > 0$. Let r be the largest integer such that $\varepsilon_r < \mu$. The matrix consisting of the rows $(\binom{\varepsilon_1}{\varepsilon_i}, \ldots, \binom{\varepsilon_r}{\varepsilon_i}, \binom{\varepsilon}{\varepsilon_i})$ for $1 \leq i \leq r$ and $(\binom{\varepsilon_1}{\mu}, \ldots, \binom{\varepsilon_r}{\mu}, \binom{\varepsilon}{\mu})$ is upper triangular and has determinant $\binom{\varepsilon}{\mu} \neq 0$ in k. By the first paragraph of the proof applied to a suitable linear subsystem of L we have $\varepsilon_{r+1} \leq \mu$ and hence $\mu = \varepsilon_{r+1}$ by the definition of r. The second statement follows from the first and Proposition 14.

We remark that $\binom{\varepsilon}{\mu} \neq 0 \bmod p$ if and only if $\mu \geq 0$ and the p-adic expansion of μ is coefficientwise less than or equal to the p-adic expansion of ε.

Corollary 20. *Let ε_i be the orders of L and μ_i the orders of L at P. The ramification divisor $R(L)$ is effective and of degree $\deg(R(L)) = (2g-2)(\sum_{i=1}^{\dim(L)} \varepsilon_i) + \dim(L)\deg(L)$. We have $\varepsilon_i \leq \mu_i$ and equality holds for all $1 \leq i \leq \dim(L)$ if and only if P is not in the support of $R(L)$.*

Corollary 21. *The D-Weierstrass places are precisely the places of degree one in the support of $R(L)$ where L is the complete linear system defined by $W - D$. The D-gap numbers of F/k are $\varepsilon_1 + 1, \ldots, \varepsilon_{\dim(L)} + 1$ for ε_i the orders of L.*

Proof. For Corollary 20 combine Theorem 18 and Proposition 16. For Corollary 21 combine Corollary 20 and Proposition 12.

The weight of a D-Weierstrass place P is defined to be $v_P(R(L))$. Also, $R(L)$ is called D-ramification divisor and, for $D = 0$, ramification divisor of F/k.

From the preceding discussion it is clear that the D-Weierstrass places are the places of degree one of F/k where the specialization of the generic place is not stable. The use of differentials is just the use of generic P_F-adic expansions.

6 Algorithms for Differentiations and Weierstrass Places

6.1 Differentiations

In characteristic zero we have $D_x^{(j)}(a) = j!^{-1} d^j a/dx^j$ for all $j \in \mathbb{Z}^{\geq 0}$. The computation of $D_x^{(j)}(a)$ can therefore be reduced to iteratively compute the derivation d/dx, which is easily achieved. In characteristic $p > 0$ however $D_x^{(j)}(a)$ cannot be computed in this way. Theorem 10 suggests that $D_x^{(j)}(a)$ be computed as the j-th coefficient of the P_F-adic expansion of a with respect to the local uniformizer $x - x_*$. For this there are well known techniques like Hensel or Newton lifting available. As it turns out we can do computations even more effectively, which will be described now.

Theorem 22. *Assume $p > 0$. Let $l, r, s \in \mathbb{Z}^{\geq 0}$ with $l \geq 1$, $s < p^l$ and let $a \in F$. There are unique $\lambda_i \in F$ such that $a = \sum_{i=0}^{p^l - 1} \lambda_i^{p^l} x^i$ and for these we have*

$$D_x^{(rp^l + s)}(a) = \sum_{i=0}^{p^l - 1} \binom{i}{s} D_x^{(r)}(\lambda_i)^{p^l} x^{i-s}. \tag{23}$$

Proof. The λ_i are obtained by representing a in the basis $1, x, \ldots, x^{p^l - 1}$ of the F^{p^l}-vector space F. Next we note that $\binom{rp^l + s}{rp^l} = 1 \bmod p$. Indeed, if $s > 0$ then $\binom{rp^l + s - 1}{rp^l - 1} = 0 \bmod p$ since $rp^l + s$ is not divisible by p^l. Using the additivity of binomial coefficients we obtain $\binom{rp^l + s}{rp^l} = \binom{rp^l + s - 1}{rp^l} \bmod p$. Hence we may assume $s = 0$. But then $\binom{rp^l}{rp^l} = 1$ and in conclusion $\binom{rp^l + s}{rp^l} = 1 \bmod p$, as claimed. Using the iterativity property we obtain

$$D_x^{(rp^l + s)} = D_x^{(rp^l)} \circ D_x^{(s)}. \tag{24}$$

From Definition 7 we have $D_x(x^j) = D_x(x)^j = (x + t)^j = \sum_{i=0}^{j} \binom{j}{i} x^{j-i} t^i$. This means $D_x^{(i)}(x^j) = \binom{j}{i} x^{j-i}$. Now let $b, c \in F$ be arbitrary. Again from the definition we see that $D_x(b^{p^l}) = D_x(b)^{p^l}$. Reading off coefficients yields

$D_x^{(i)}(b^{p^l}) = 0$ for $i \neq 0 \bmod p^l$. Using $D_x(b^{p^l}c) = D_x(b^{p^l})D_x(c)$ and $s < p^l$ we thus obtain $D_x^{(s)}(b^{p^l}c) = b^{p^l}D_x^{(s)}(c)$, and combining these observations gives

$$D_x^{(s)}(a) = \sum_{i=0}^{p^l-1} \binom{i}{s} \lambda_i^{p^l} x^{i-s}. \tag{25}$$

For $0 \leq j < p^l$ we have $D_x^{(rp^l)}(x^j) = 0$ and $D_x^{(rp^l)}(b^{p^l}) = D_x^{(r)}(b)^{p^l}$. Similarly as above this yields $D_x^{(rp^l)}(b^{p^l}x^j) = D_x^{(r)}(b)^{p^l}x^j$ and applying $D_x^{(rp^l)}$ to both sides of equation (25) proves equation (23).

In order to compute differentiations using Theorem 22 we need to find p-th power representations $a = \sum_i \lambda_i^p x^i$. One way of achieving this is to realize F as an inseparable extension of F^p of degree p. The following algorithm however gives an easy to implement alternative.

Algorithm 26. *(Power representation)*

Input: *A function field F/k with separating element x and $a \in F$.*
Output: *Elements $\lambda_i \in F$ such that $a = \sum_{i=0}^{p-1} \lambda_i^p x^i$*

```
1. Set $a_0 := a$ and $a_j := j^{-1} da_{j-1}/dx$ for $1 \leq j < p$.
2. Set $b_{p-1} := a_{p-1}$. For $j = p-2, \ldots, 0$ set $b_j := a_j - \sum_{i=j+1}^{p-1} \binom{i}{j} b_i x^{i-j}$.
3. Return $\lambda_i := b_i^{1/p}$ for $0 \leq i < p$.
```

Proof. We have $a_j = D_x^{(j)}(a)$ for $0 \leq j < p$ and $D_x^{(j)}(a) = \sum_{i=j}^{p-1} \binom{i}{j} \lambda_i^p x^{i-j}$. This shows that the algorithm indeed computes the λ_i.

Algorithm 27. *(Differentiations I)*

Input: *A function field F/k with separating element x, an integer $j \geq 0$ and an element $a \in F$.*
Output: *The differentiation $D_x^{(j)}(a)$.*

```
1. If $j = 0$ then return $a$.
2. Write $j = rp + s$ with $r, s \in \mathbb{Z}^{\geq 0}$ and $s < p$.
3. Compute $e := D_x^{(s)}(a) = (s!)^{-1} d^s a/dx^s$.
4. If $r = 0$ then return $e$.
5. Write $e = \sum_{i=0}^{p-1} \lambda_i^p x^i$ using algorithm 26.
6. Compute $\mu_i := D_x^{(r)}(\lambda_i)$ using Algorithm 27 recursively.
7. Return $\sum_{i=0}^{p-1} \mu_i^p x^i$.
```

Proof. The correctness of the algorithm follows from Theorem 22, equation (25).

We could use equation (23) directly in Algorithm 27. However, it is more effective to apply step 3 first since in the p-th power representation computation afterwards more of the λ_i will be zero.

Algorithm 27 can be improved in two ways. Firstly, suppose we want to compute the first n differentiations of an element. Applying Algorithm 27 for these values takes $O(n^2)$ derivation computations d/dx altogether. We can however obtain an iterative version using only $O(n\lceil\log_p(n)\rceil)$ derivation computations d/dx as follows. Let $a = \sum_{i=0}^{p-1}\lambda_i^p x^i$ and assume that we have computed $D_x^{(rp+s)}(a)$. If $s < p-1$ we compute $D_x^{(rp+s+1)}(a) = (s+1)^{-1}d(D_x^{(rp+s)}(a))/dx$. If $s = p-1$ we compute $D_x^{(rp+s+1)}(a) = D_x^{((r+1)p)}(a) = \sum_{i=0}^{p-1} D_x^{(r+1)}(\lambda_i)^p x^i$ applying this strategy recursively to the values $D_x^{(r)}(\lambda_i)$ (which have to be stored). In the following let N denote a function on the symbols a, s, b, L which is thought of as a set of symbol-value pairs. The subscript i on a tuple denotes the i-th entry.

Algorithm 28. *(Recursion)*

```
Input:   The function $N$.
Output:  The changed function $N$.

1. If $N(s) < p-1$ then compute $N(s) := N(s)+1$, $N(b) := N(s)^{-1}dN(b)/dx$
   and return $N$. Terminate.
2. If $N(L)$ is undefined then compute $N(a) := \sum_{i=0}^{p-1}\lambda_i^p x^i$ using Algorithm 26
   and define $N(L) := (\{(a,\lambda_i),(s,0),(b,\lambda_i)\} \mid 0 \le i \le p-1)$.
3. Set $N(s) := 0$ and compute $N(L) := (\,Recursion(N(L)_i) \mid 0 \le i \le p-1\,)$,
   $N(b) := \sum_{i=0}^{p-1}(N(L)_i(b))^p x^i$.
4. Return $N$.
```

Algorithm 29. *(Differentiations II)*

```
Input:   The function field $F/k$ with separating element $x$ and an $a \in F$.
Output:  The differentiations $D_x^{(0)}(a), D_x^{(1)}(a), \ldots$.

1. Set $N := \{(a,a),(s,0),(b,a)\}$.
2. Repeat returning $N(a)$ and redefining $N := Recursion(N)$.
```

Proof. The validity of the algorithm follows from the above considerations. For the running time statement we observe that computing $D_x^{(j_0)}(a), \ldots, D_x^{(j_0+p^j)}(a)$ for $j_0 + p^j < p^{j+1}$ takes $\le (j+1)p^j$ derivation computations. This is clearly true for $j = 0$. Computing p times p^j successive differentiations costs $\le p(j+1)p^j + pp^j = (j+2)p^{j+1}$ derivation computations so the assertion follows by induction.

We remark that the number of elements to be stored in Algorithm 27 and 29 is $O(n)$ as opposed to $O(1)$ in characteristic zero.

For the second improvement we observe that the differentiations have (depending on the representation of F/k) certain denominators which can be estimated. Dealing with numerators and denominators separately can save expensive element inversions and gcd computations. To be more explicit, let $F = k(x,y)$ with $f(x,y) = 0$ and $f \in k[x,z]$ irreducible, monic and separable in the second variable z. We denote the derivative of f with respect to y by $f'(x,y)$.

Proposition 30. *We have* $b^{j+1}f'(x,y)^{2j-1}D_x^{(j)}(a/b) \in k[x,y]$ *for* $a, b \in k[x,y]$ *with* $b \neq 0$ *and* $j \geq 1$.

Proof. The $D_x^{(j)}(a)$ are the coefficients of the P_F-adic expansion of $a \in \widetilde{F}$ with respect to the prime element $x - x_*$. The proof follows by investigating the denominators which arise in an univariate Newton lifting. We leave the details to the reader.

If F/k is represented as the field of fractions of the coordinate ring of a non-plane affine curve, multivariate Newton lifting has to be used instead so that $f'(x,y)$ is replaced by the Jacobian determinant in an appropriate manner.

6.2 Weierstrass Places

The algorithm for computing Weierstrass places is now fairly straightforward by the previous discussion.

Algorithm 31. *(Weierstrass places)*

```
Input:  A function field $F/k$ with separating element $x$ and a divisor $D$.
Output: The $D$-gap numbers and $D$-Weierstrass places.

1. Compute the canonical divisor $W := (dx)$.
2. If $\dim(W-D) = 0$ then the ramification divisor of the complete linear system
   defined by $W-D$ is zero and there are no $D$-gap numbers and $D$-Weierstrass
   places. Terminate.
3. Compute a basis $v_1, \ldots, v_n$ of $\mathcal{L}(W-D)$.
4. Set $\varepsilon_1 := 0$, $M := (v_1, \ldots, v_n)$, $i := 1$, $\varepsilon := 0$ and $G := \{\}$.
5. Let $i := i+1$. If $i > n$ then go to step 8.
6. Let $\varepsilon := \min\{h \in \mathbb{Z}^{>\varepsilon} \mid \binom{h}{g} = 0$ in $k$ for all $g \in G\}$.
7. Let $M' \in F^{i \times n}$ be the matrix obtained by appending $(D_x^{(\varepsilon)}(v_1), \ldots, D_x^{(\varepsilon)}(v_n))$
   to $M$. If $\operatorname{rank} M' > \operatorname{rank} M$ then $M := M'$, $\varepsilon_i := \varepsilon$ and go to step 5.
   Otherwise let $G := G \cup \{\varepsilon\}$ and go to step 6.
8. Compute the ramification divisor $R := \det(M) + (\sum_{i=1}^{n} \varepsilon)(dx) + n(W-D)$
   of the complete linear system defined by $W-D$.
9. Return $\varepsilon_1 + 1, \ldots, \varepsilon_n + 1$ and the degree one places in the support of $R$.
```

Proof. The algorithm is correct by Corollary 21, Lemma 19 and Proposition 4.

The most expensive part of Algorithm 31 is the computation of the orders and the Wronskian determinant. The differentiations are best computed using Algorithm 29. In order to check that the rank has increased it is convenient to work with an echelonized version of M instead, in order to save subsequent echelonization work. Additionally, the denominators of the differentiations as in Proposition 30 can be treated separately in the linear algebra.

Let $F' = Fk_1$ be the constant field extension of F by k_1 and $\mathrm{con}_{F'/F}$ the conorm map from F to F'. Since D_x is extended by k_1-linearity to F' we have

$R(\mathrm{con}_{F'/F}(L)) = \mathrm{con}_{F'/F}(R(L))$. We can thus compute $\mathrm{con}_{F'/F}(D)$-Weierstrass places over the larger constant field k_1 without really having to work in k_1. If for example k_1 is the algebraic closure of k then any place P in the support of $R(L)$, L the complete linear system defined by $W - D$, gives rise to $\deg(P)$ many Galois conjugate D-Weierstrass places defined over the splitting field of the residue class field of P. This results in a very effective way of computing Weierstrass places and their fields of definition without extending the constant field.

Finally we remark that Algorithm 31 can clearly also be used to compute ramification divisors and orders of arbitrary linear systems.

7 Special Models of Algebraic Curves

As an application we describe in this section how the preceding sections may be used to compute a special model of the curve such that projection onto one of the variables gives a map to $\mathbb{P}^1$ of low degree. Equivalently, given a function field with some generators, try to find other generators such that one of them generates a rational subfield of small index, and return the equations they satisfy.

More specifically, assume P is a place of degree one of the function field F/k. For the first pole number r of P we have in general $r \leq g + 1$. However, if P is a Weierstrass place we may hope that r is considerably smaller than $g + 1$. For a hyperelliptic function field we would for example have $r = 2$ while in general we cannot expect to be better than roughly $r = g/2$. Now, if we are given $x \in F$ such that its pole divisor satisfies $(x)_\infty = rP$ we know $[F : k(x)] = r$ and thus have a rational subfield of small index. The strategy is to use such places in the following algorithm. Note that in order to obtain a Weierstrass place of degree one it might be necessary to work with a constant field extension.

Algorithm 32. *(Special model)*

Input: *A function field F/k with separating element a_1 and generators a_i such that $F = k(a_1)[a_2, \ldots, a_n]$. A place P of degree one.*

Output: *Return a separating element b_1 and generators $b_2, \ldots, b_r$ such that $F = k(b_1)[b_2, \ldots, b_r]$, together with a non-singular affine model given by the algebraic relations between the b_i. The b_i are expressed in the a_i. The number r is the first pole number of P.*

```
1. Compute the first pole number $r$ of $P$ together with an element $b_1 \in F$ such that $(b_1)_\infty = rP$.
2. Let $i := 1$ and $d_1 := 0$.
3. If $i = r$ goto step 5. Otherwise let $i := i + 1$.
4. Compute the smallest pole number $d_i$ of $P$ such that $d_i \neq d_j \bmod r$ for $1 \leq j < i$. Compute an element $b_i \in F$ such that $(b_i)_\infty = d_i P$. Goto step 3.
5. Using linear algebra over $k$ compute $\lambda_{i,j,\nu} \in k[b_1]$ with $\deg(\lambda_{i,j,\nu}) \leq (d_i + d_j - d_\nu)/r$ such that $\lambda_{i,j,1} + \sum_{\nu=2}^{r} \lambda_{i,j,\nu} b_\nu = b_i b_j$ for $2 \leq i, j \leq r$.
6. Return the $b_i$ and the equations computed in the previous step.
```

Proof. See also [7, Section 7]. Considering the degree function $\deg = -v_P$ and using a Gröbner reduction (or saturation) argument one can easily see that the b_i exist and that $1, b_2, \ldots, b_r$ forms a $k[b_1]$-basis of the integral closure $\mathrm{Cl}(k[b_1], F)$. Thus $b_i b_j$ can be expressed as a $k[b_1]$-linear combination of the basis, and these equations give a full description of $\mathrm{Cl}(k[b_1], F)$. The degree bound for the $\lambda_{i,j,\nu}$ follows because there is no degree cancellation possible since $d_i \neq d_j \bmod r$.

If we additionally apply the inversion algorithm given below we may skip step 5 and obtain the model from the inversion algorithm.

Remark 33. *Homogenizing this affine model yields a non-singular weighted projective model if b_1 and the homogenizing variable are counted with weight* 1 *and b_i with weight $\lceil d_i/r \rceil$ for $2 \leq i \leq r$. Also, one can show $\lceil d_i/r \rceil \leq \lceil (2g-1)/r \rceil + 1$ which gives the bound $2\lceil (2g-1)/r \rceil + 2$ for the degrees of the models. We further note that r and the d_i are not in general a minimal set of generators of the Weierstrass semigroup at P. Accordingly, there can be relations of the form $b_j = \prod_{i=1}^{j-1} b_i^{m_i}$ with $m_i \in \mathbb{Z}^{\geq 0}$ leading to the elimination of variables from the model. Further improvements in this direction are possible.*

For $g = 0$ one could ask whether the function field F/k is rational. There are no Weierstrass places available but the canonical class contains a divisor W of degree -2. Then $\dim(-W) = 3$ and $D := (a) - W$ for non constant $a \in \mathcal{L}(-W)\backslash k$ is an effective divisor of degree 2. There is hence a place P of degree one or two in D which we can compute. After a possible quadratic constant field extension by the residue class field of P we can assume $\deg(P) = 1$. Then for $x \in \mathcal{L}(P)\backslash k$ we have $F = k(x)$. If we want to avoid a constant field extension when $\deg(P) = 2$ we can compute a conic as the algebraic relation between the two non constant elements in $\mathcal{L}(P)$. On the conic we could then try to find a rational point [3]. For a further discussion see [9].

For $g = 1$ one could ask whether the function field F/k is elliptic. Again, there are no Weierstrass places available but if we are given a place of degree one, Algorithm 32 can be applied to obtain a Weierstrass model (the trace term should additionally be eliminated in characteristic $\neq 2$). For a further discussion see [8].

For $g \geq 2$ one could ask whether the function field F/k is hyperelliptic. In this case there exist Weierstrass places which can be used as input for Algorithm 32 to obtain a hyperelliptic model, after a possible constant field extension. However, there is a generally better method available which is able to work with any place of degree one, see [10].

Inversion

Algorithm 32 represents the b_i in the generators a_i of the function field. It is desirable to also have expressions for the generators a_i in terms of the b_i. We consider the following general problem: Let $k(a_1)[a_2, \ldots, a_n]$ and $k(b_1)[b_2, \ldots, b_m]$ be two representations of the same function field F/k with a_1 and b_1 separating. Assume $k(a_1)[a_2, \ldots, a_n] = k(a_1)[x_2, \ldots, x_n]/I$ for some prime ideal I of dimension zero and $b_i = f_i(a_2, \ldots, a_n)$ with $f_i \in k(a_1)[x_2, \ldots, x_n]$. The problem is to compute J with $k(b_1)[b_2, \ldots, b_m] = k(b_1)[y_2, \ldots, y_m]/J$, and $g_j \in k(b_1)[y_2, \ldots, y_m]$ such that $a_j = g_j(b_2, \ldots, b_m)$. In other words, the problem is to compute the algebraic relations between the other generators and invert the isomorphism given by the expression of the b_i in the a_j. To achieve this let T_a be the ideal of $k(a_1)[x_2, \ldots, x_n, y_1, \ldots, y_m]$ generated by I and $y_i - f_i(x_2, \ldots, x_n)$ for $1 \leq i \leq m$. We have that T_a is a prime ideal because of the linearity of the added expressions and since I is prime. Furthermore, a Gröbner basis of T_a consists of a Gröbner basis of I together with the elements $y_i - f_i(x_2, \ldots, x_n)$ for $1 \leq i \leq m$. The elimination ideal $T_a \cap k(a_1)[y_1]$ is then also prime and contains a monic irreducible generator m_a. Clearly m_a is the minimal polynomial of b_1 over $k(a_1)$. By substituting b_1 for y_1 in T_a we obtain a prime ideal T'_a such that $k(a_1)[b_1][x_2, \ldots, x_n, y_2, \ldots, y_m]/T'_a \cong F$. From m_a we obtain the minimal polynomial m_b of a_1 over $k(b_1)$ and $k(a_1)[b_1] \cong k(b_1)[a_1]$. According to this isomorphism we can rewrite T'_a into T'_b such that $k(b_1)[a_1][x_2, \ldots, x_n, y_2, \ldots, y_m]/T'_b \cong F$. Reversing the above construction symmetrically we first obtain T_b by substituting x_1 for a_1 and then $J = k(b_1)[y_2, \ldots, y_m] \cap T_b$. Furthermore, finding the normal forms of the variables x_i for $1 \leq i \leq n$ mod T'_b with respect to the lexicographical term order gives the g_i. The above intersections and the last reduction step can be carried out by Gröbner basis computations.

8 Examples

8.1 Weierstrass Places

We consider the function field F/k defined by $y^7 + y = x^4$ over $\mathbb{F}_{49}$. Its genus is 9 and it has 176 places of degree one, the maximal number possible for this finite field and genus. Using the algorithms in section 6 we compute the following data. The gap numbers of F/k are $1, 2, 3, 4, 5, 8, 9, 10, 15$. All 176 places of degree one are Weierstrass places. There are 8 Weierstrass places of weight 9 with gap numbers $1, 2, 3, 5, 6, 9, 10, 13, 17$ and 168 Weierstrass places of weight 5 with gap numbers $1, 2, 3, 4, 5, 9, 10, 11, 17$. The ramification divisor has degree 912. The whole computation takes about 30s on a 600MHz computer, using Magma [1,2].

8.2 Special Models

We consider the function field F/k defined by $y^{10} + 4y^7 + xy^6 + (4x^5 + x^2)y^5 + 3x^5y^2 + 2x^6y + 4x^{10} + x^7 = 0$ over $\mathbb{F}_5$. Its genus is 6 and the ramification divisor contains four places of degree 1 and weights $1, 10, 11, 13$, two places of degree 2 and weights $1, 13$, one place of degree 3 and weight 1, and 14 places of degree 6 and weights $1, \ldots, 1, 11$. In Algorithm 32 we take the Weierstrass place of degree 1 and weight 10 which has 3 as its first pole number. We obtain the affine model with Gröbner basis $x^7 - yz + 1, y^2 - z$, hence the plane model $y^3 = x^7 + 1$. We further obtain $b_1 = a_1/(2a_1 + a_2)$ and $b_2 = 2a_1 + a_2$, and for the inverse representation $a_1 = b_1/(b_1^7 + 1)b_2^4$ and $a_2 = (3b_1 + 1)/(b_1^7 + 1)b_2^4$. The internal integral basis computation takes about $2.6s$. The ramification divisor is then computed and factorized in about $10s$. The rest of the computation takes a further 3s, again on a 600MHz computer using Magma [1,2].

Acknowledgements

I would like to thank H. Stichtenoth for suggesting that the computation of Weierstrass places be an interesting problem. I also thank M. van Hoeij for valuable comments on how to compute hyperelliptic models of curves. This work was supported by NaFöG and EPSRC grants and a stay with the Magma group.

References

1. W. Bosma, J. Cannon, and C. Playoust. The Magma algebra system I: The user language. *J. Symbolic Comp.*, 24, 3/4:235–265, 1997.
2. Comp. algebra group. Magma. `http://www.maths.usyd.edu.au:8000/u/magma/`, 2001.
3. J. E. Cremona and D. Rusin. Efficient solution of rational conics. Preprint available under `http://www.maths.nott.ac.uk/personal/jec/conics.ps.gz`, 2002.
4. H. Hasse. Theorie der Differentiale in algebraischen Funktionenkörpern mit vollkommenem Konstantenkörper. *J. Reine angew. Math.*, 172:55–64, 1934.
5. H. Hasse. Theorie der höheren Differentiale in einem algebraischen Funktionenkörper mit vollkommenem Konstantenkörper bei beliebiger Charakteristik. *J. Reine angew. Math.*, 175:50–54, 1936.
6. H. Hasse and F. K. Schmidt. Noch eine Begründung der Theorie der höheren Differentialquotienten in einem algebraischen Funktionenkörper einer Unbestimmten. *J. Reine angew. Math.*, 177:215–237, 1937.
7. F. Hess. Computing Riemann-Roch spaces in algebraic function fields and related topics. *J. Symbolic Comp.*, 33(4):425–445, 2002.
8. M. van Hoeij. An algorithm for computing the Weierstrass normal form. In A. H. M. Levelt, editor, *Proceedings of the International Symposium on Symbolic and Algebraic Computation, ISSAC '95*, pages 90–95, Montreal, Canada, 1995. ACM Press, New York.
9. M. van Hoeij. Rational parametrizations of algebraic curves using a canonical divisor. *J. Symbolic Comp.*, 23, 2-3:209–227, 1997.
10. M. van Hoeij. An algorithm for computing the Weierstrass normal form of hyperelliptic curves. Preprint available under `http://arXiv.org/`, 2002.
11. Kant group. Kash. `http://www.math.tu-berlin.de/~kant`, 2001.
12. H. Matzat. *Ein Vortrag über Weierstraß Punkte.* Universität Karlsruhe, 1975.
13. F. K. Schmidt. Die Wronskische Determinante in beliebigen differenzierbaren Funktionenkörpern. *Math. Z.*, 45:62–74, 1939.
14. F. K. Schmidt. Zur arithmetischen Theorie der algebraischen Funktionen. II: Allgemeine Theorie der Weierstraßpunkte. *Math. Z.*, 45:75–96, 1939.
15. H. Stichtenoth. *Algebraische Funktionenkörper einer Variablen.* Vorlesungen aus dem Fachbereich Mathematik der Universität Essen, 1978.
16. H. Stichtenoth. *Algebraic Function Fields and Codes.* Springer-Verlag, Berlin-Heidelberg-New York, 1993.
17. K.-O. Stöhr and J. F. Voloch. Weierstrass points and curves over finite fields. *Proc. London Math. Soc. (3)*, 52(1):1–19, 1986.
18. O. Teichmüller. Differentialrechnung bei Charakteristik p. *J. Reine angew. Math.*, 175:89–99, 1936.

New Optimal Tame Towers of Function Fields over Small Finite Fields

Wen-Ching W. Li[1,⋆], Hiren Maharaj[2],
Henning Stichtenoth[3], and Noam D. Elkies[4,⋆⋆]

[1] Department of Mathematics, Pennsylvania State University,
University Park, PA 16802, U.S.A
wli@math.psu.edu

[2] Department of Mathematical Sciences, Clemson University,
Clemson, SC 29634, U.S.A
hmahara@clemson.edu

[3] Mathematik und Informatik, Universität GH Essen,
Fachbereich, Germany,
stichtenoth@uni-essen.de

[4] Department of Mathematics, Harvard University,
Cambridge, MA 02138, U.S.A
elkies@math.harvard.edu

1 Introduction

Ihara [11] introduced the quantity $A(q) = \limsup_{g\to\infty} N_q(g)/g$ where $N_q(g)$ is the maximum number of rational places of a function field with genus g and with the finite field $\mathbb{F}_q$ as the full field of constants. Drinfeld and Vladut [2] showed that $A(q) \le \sqrt{q} - 1$. It was also shown by Ihara [11], and Tsfasman, Vladut and Zink [17] in special cases, that $A(q) = \sqrt{q} - 1$ when q is a square. When q is not a square, the exact value of $A(q)$ is currently unknown. While the problem of finding $A(q)$ in this case is an interesting problem in its own right, much motivation comes from implications in asymptotic results in coding theory. Essentially there are three approaches to finding lower bounds for $A(q)$: class field towers [15], modular curves [11], [17], [3], [4] and explicit towers (that is, given explicitly in terms of generators and relations) of function fields. For applications to coding theory though, explicit towers are needed. In [6], a *tower of function fields* over $\mathbb{F}_q$ is defined to be a sequence $\mathcal{F} = (F_1, F_2, F_3, \ldots)$ of function fields F_i, having the following properties:
(i) $F_1 \subseteq F_2 \subseteq F_3 \subseteq \ldots$.
(ii) For each $n \ge 1$, the extension F_{n+1}/F_n is separable of degree $[F_{n+1} : F_n] > 1$.
(iii) the genus $g(F_j) > 1$ for some $j \ge 1$.
(iv) $\mathbb{F}_q$ is the full field of constants of each F_n.

As noted in [6], (ii), (iii) and the Hurwitz genus formula imply that $g(F_n) \to \infty$ as $n \to \infty$.

⋆ Research of this author is supported in part by the NSF grant no. DMS-9970651

⋆⋆ Author of the Appendix

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 372–389, 2002.

For any tower $\mathcal{F} = (F_1, F_2, F_3, \ldots)$ of function fields F_i, let

$$\lambda(\mathcal{F}) := \lim_{i \to \infty} N(F_i)/g(F_i)$$

where $N(F_i)$ is the number of rational places of F_i. It is shown in [6] that this limit is well defined. A tower $\mathcal{F}$ is said to be *asymptotically good* (respectively *asymptotically bad*) if $\lambda(\mathcal{F}) > 0$ (respectively $\lambda(\mathcal{F}) = 0$). It is clear that if $\mathcal{F}$ is a tower over $\mathbb{F}_q$ then $A(q) \geq \lambda(\mathcal{F})$. We say that the tower $\mathcal{F}$ over $\mathbb{F}_q$ is *optimal* if $A(q) = \lambda(\mathcal{F})$.

In the case that q is a square, Garcia and Stichtenoth [5] discovered the first explicit optimal tower over $\mathbb{F}_q$. In [5], the towers are wildly ramified. Subsequently in [10], [3] and [9] explicit tame towers were found. Tame towers have the advantage that the genus computation is simpler. In this paper we exhibit new optimal tame towers found by computer search using the powerful algebraic number theory package KASH [1]. In section 3, we explain in some detail how we do this. We also discuss some results related to towers over prime fields. Optimal tame towers over $\mathbb{F}_4$,$\mathbb{F}_9$, $\mathbb{F}_{25}$ and $\mathbb{F}_{49}$ are presented in section 4. In section 5 it is shown that these towers are new in the sense that they are not subtowers of any of the known towers over these finite fields.

Elkies [3], [4] has shown that every currently known explicit optimal tower over $\mathbb{F}_{q^2}$ is either elliptic modular or Drinfeld modular. He further conjectures that all the optimal towers over $\mathbb{F}_{q^2}$ constructed recursively should be modular. In the Appendix of this paper, he proves that the four new towers described in this paper are again elliptic modular. Since our search is fairly extensive for polynomials of low degree over small finite fields, this gives a strong numerical evidence of his conjecture.

2 Preliminaries

Given $f(x, y) \in \mathbb{F}_q(x, y)$, the tower $\mathcal{F} = (F_1,\ F_2,\ \ldots)$ over $\mathbb{F}_q$ is said to be defined by $f(x, y)$ if $F_1 = \mathbb{F}_q(x_1)$ is the rational function field and for each $n > 1$, $F_n = \mathbb{F}_q(x_1, x_2, \ldots, x_n)$, where $f(x_i, x_{i+1}) = 0$ for $1 \leq i < n$.

The following notation will be used throughout the paper: suppose that the function field $F := \mathbb{F}_q(x, y)$ is defined by some equation $f(x, y) = 0$ where x and y are transcendental over $\mathbb{F}_q$. Let P be a place in $\mathbb{F}_q(y)$ (respectively, in $\mathbb{F}_q(x)$), let P_1, ..., P_ℓ be the places of F which lie above P and let $Q_1, Q_2, \ldots, Q_\ell$ denote the restrictions of each of the places P_1, ..., P_ℓ to $\mathbb{F}_q(x)$ (respectively, to $\mathbb{F}_q(y)$). Then we write

$$Q_1, \ldots, Q_\ell \leftarrow P$$

and, respectively

$$P \rightarrow Q_1, \ldots, Q_\ell.$$

Note that repetitions of the same place may occur among the Q_i's.

Unless otherwise mentioned, we will use the same notation as in [16], for example, we denote the set of places of a function field F by $\mathbb{P}(F)$. Moreover, Kummer's Theorem (Theorem III.3.7 of [16]) will be used many times in the proofs below without any indication. We will use the following result which is proved in [10].

Theorem 1. *Let $\mathcal{F} = (F_1, F_2, F_3, \ldots)$ be a tower of function fields over $\mathbb{F}_q$ satisfying the following conditions:*
(i) All extensions F_{n+1}/F_n are tame.
(ii) The set $R = \{P \in \mathbb{P}(F_1) | P$ is ramified in F_n/F_1 for some $n \geq 2\}$ is finite.
(iii) The set $S = \{P \in \mathbb{P}(F_1) | \deg P = 1$, and P splits completely in all extensions $F_n/F_1\}$ is non-empty.

Then $\mathcal{F}$ is an asymptotically good tower, and one has the following estimate

$$\lambda(\mathcal{F}) \geq \frac{2s}{2g(F_1) - 2 + r} \tag{1}$$

where $s := \#S$ and $r := \sum_{P \in R} \deg P$.

3 The KASH Implementation

Given $f(x, y) \in \mathbb{F}_q(x, y)$, it is in general a time consuming exercise to determine if the corresponding tower satisfies the conditions (*ii*) and (*iii*) of Theorem 1. The main idea of the computer implementation comes from the proof of Theorem 3.1 in [6]. We explain only how we determine if the ramification set R in Theorem 1 is finite - it is obvious how to find the set S. Suppose P is a place of F_n $(n > 1)$ which is ramified in the extension F_n/F_{n-1}. We wish to determine the possibilities for the restriction of P to F_1. Let P_i be restriction of P to $\mathbb{F}_q(x_i)$ for each $1 \leq i < n$. By viewing the tower as a pyramid as in [6], we see, by Abhyankar's Lemma (Proposition III.8.9 in [16]), that the place P_{n-1} must be ramified in the extension $\mathbb{F}_q(x_n, x_{n-1})/\mathbb{F}_q(x_{n-1})$. Since this extension is given by a *known* equation, namely $f(x_{n-1}, x_n) = 0$, we easily determine all possibilities for the place P_{n-1}. To determine the candidates for P_{n-2} we use built-in features of KASH. Continuing in this way, we finally get the set R we want. We impose upper bounds on the degrees of the possible places P_i to ensure that the program terminates in reasonable time. So, if any P_i has degree too large, we simply discard the $f(x, y)$ and try another equation. Thus, the algorithm used is not deterministic - it may well happen that a discarded $f(x, y)$ yields a finite ramification set. Observe that the condition (*ii*) of Theorem 1 is actually weaker than what we check for - it may happen that the set of possible P_1's may be infinite while the set R of Theorem 1 is finite. However, if the set of possible P_1's is finite then R *is* finite.

After checking that conditions (*ii*) and (*iii*) of Theorem 1 are satisfied, the next step is to determine if the tower is infinite. In order to do this, we choose only those towers where there is ramfication in each step - it is an easy matter to get KASH to automatically check this while searching for the set R.

Some general comments on the output of the program are in order. While we did recover all known towers over small finite fields using this approach, it is disappointing that no towers over prime fields were found. The most extensive computations were carried out with degree 2 and 3 polynomials $f(x, y)$ (note that in [8] it is shown that a necessary condition for $f(x, y) \in \mathbb{F}_q[x, y]$ to give rise to an asymptotically good tower is that $\deg_x f = \deg_y f$). It is tempting to conjecture at least that there are no degree 2 polynomials that satisfy the conditions of Theorem 1 over a prime field. For $f(x, y) = y^2 + ax^2 + bx$ ($a, b \in \mathbb{F}_p$, p a prime number), this is a special case of a result proved by Lenstra [12]. In [12], Lenstra gives an elegant proof that a construction of Garcia, Stichtenoth and Thomas presented in [10] (for every finite field which is not prime) cannot work over prime fields. Inspired by Lenstra's work, the following result can be proved [14]:

Proposition 1. *Let p be the characteristic of $\mathbb{F}_q$.*
(1) *The tower over $\mathbb{F}_q$ defined by the polynomial $f(x, y) = y^2 + ax^2 + bx \in \mathbb{F}_p[x, y]$ satisfies the conditions of Theorem 1 if and only if $p = 3$, $a = 1$, $b = \pm 1$ and q is a square.*
(2) *The tower over $\mathbb{F}_q$ defined by the polynomial $f(x, y) = y^3 + ax^3 + bx^2 + cx \in \mathbb{F}_p[x, y]$ satisfies the conditions of Theorem 1 if and only if $p = 2$, $a = b = c = 1$ and q is a square.*

In general it was found that good towers are rare. For example, upon trying out all degree two polynomials over $\mathbb{F}_3$, less than 1000 were found to satisfy condition (*ii*) of Theorem 1 over $\mathbb{F}_9$, and fewer than 300 of these were found to be infinite (using the aforementioned criterion) and satisfy condition (*iii*). As expected, in general, condition (*iii*) was found to be more restrictive than condition (*ii*). With the current computational evidence it seems that more interesting and general theorems of the above type can be proved.

4 The Towers

In this section we prove the main result of this paper.

Theorem 2. *Each of the polynomials below defines an optimal tower over the indicated finite field:*

- $2xy^2 + (x^2 + x + 1)y + x^2 + x + 2$ *over* $\mathbb{F}_9$;
- $(4x + 1)y^2 + (x^2 + x + 2)y + x + 3$ *over* $\mathbb{F}_{25}$;
- $(x^2 + 6)y^2 + xy + x^2 + 4$ *over* $\mathbb{F}_{49}$;
- $x^2y^3 + (x^3 + x^2 + x)y^2 + (x + 1)y + x^3 + x$ *over* $\mathbb{F}_4$.

We shall present the towers over $\mathbb{F}_9$ and $\mathbb{F}_4$ with detailed proofs. The computations for the towers over $\mathbb{F}_{25}$ and $\mathbb{F}_{49}$ are omitted because of the similarity to the tower over $\mathbb{F}_9$.

4.1 Tower over $\mathbb{F}_9$

Let q be a power of 3 and consider the function field $F^{(q)} := \mathbb{F}_q(x, y)$ defined by

$$f(x, y) := 2xy^2 + (x^2 + x + 1)y + x^2 + x + 2 = 0. \tag{2}$$

Lemma 1. *Let $q = 3$. Then $f(x, y)$ is absolutely irreducible and we have the following:*
(o) The polynomials $T^4 + T^3 + T^2 + 2T + 2$ and $T^4 + T^2 + T + 1$ are irreducible over $\mathbb{F}_3$ and so correspond to places of the rational function field $\mathbb{F}_3(T)$.
(i) The place $y^4 + y^3 + y^2 + 2y + 2$ of $\mathbb{F}_3(y)$ is the only place ramified in the extension $F^{(3)}/\mathbb{F}_3(y)$ and

$$x^4 + x^3 + x^2 + 2x + 2 \leftarrow y^4 + y^3 + y^2 + 2y + 2.$$

(ii) The place y^4+y^2+y+1 of $\mathbb{F}_3(y)$ splits completely in the extension $F^{(3)}/\mathbb{F}_3(y)$ and

$$x^4 + x^2 + x + 1,\ x^4 + x^3 + x^2 + 2x + 2 \leftarrow y^4 + y^2 + y + 1.$$

(iii) The place x^4+x^2+x+1 of $\mathbb{F}_3(x)$ is the only place ramified in the extension $F^{(3)}/\mathbb{F}_3(x)$ and

$$x^4 + x^2 + x + 1 \rightarrow y^4 + y^2 + y + 1.$$

Proof: (o). It is easily checked.

One can complete the square in equation (2) to obtain the following two equations:

$$\left(y + \frac{x^2 + x + 1}{x}\right)^2 = \frac{x^4 + x^2 + x + 1}{x^2} \tag{3}$$

and

$$\left(x + \frac{y^2 + 2y + 2}{y + 1}\right)^2 = \frac{y^4 + y^3 + y^2 + 2y + 2}{(y + 1)^2}. \tag{4}$$

Then both the extensions $F^{(3)}/\mathbb{F}_3(x)$ and $F^{(3)}/\mathbb{F}_3(y)$ are degree 2 Kummer extensions and the only place ramified in the extension $F^{(3)}/\mathbb{F}_3(x)$ is $x^4 + x^2 + x+1$ and the only place ramified in the extension $F^{(3)}/\mathbb{F}_3(y)$ is $y^4+y^3+y^2+2y+2$. The irreducibility of $f(x, y)$ follows from (3) and (4). We have thus proved the first parts of (i) and (iii).
(i). Suppose $y^4+y^3+y^2+2y+2 = 0$. Then from equation (4) we have $x = -(y^2+2y+2)/(y+1)$. Using $y^3+y+1 = -1/(y+1)$, we have $y^3 = -(y^2+2y+2)(y+1)$ so that $x = y^3$, a solution to $x^4 + x^3 + x^2 + 2x + 2 = 0$.
(ii). Suppose that $y^4 + y^2 + y + 1 = 0$. Observe that

$$(2y + 2)^4 + (2y + 2)^3 + (2y + 2)^2 + 2(2y + 2) + 2 = 0. \tag{5}$$

Using $1/(y+1) = y^3+2y^2+2y+2$, equation (4) becomes $(x+y^3+2y^2)^2 = 2y^2 = (y+y^3+2)^2$; so that $x = (y^3+y+2) - (y^3+2y^2) = y^2+y+2 = y^{27}$ (a solution of x^4+x^2+x+1) or $x = -(y^3+y+2) - (y^3+2y^2) = y^3+y^2+2y+1 = (2y+2)^9$ (a solution of $x^4 + x^3 + x^2 + 2x + 2$ from equation (5)).
(iii). It follows from (i) and (ii). □

Let w be a (primitive) element of $\mathbb{F}_9$ which satisfies $w^2 + 2w + 2 = 0$.

Lemma 2. *Put $q = 9$. Then in $F^{(9)}$ we have*

$$\begin{aligned}1/x &\to 1/y, y+1\\ x &\to 1/y, y+2\\ x+2 &\to y+1, y+2\\ x+w^3 &\to y, y+w\\ x+w &\to y, y+w^3\\ x+1 &\to y+w, y+w^3\end{aligned}$$

Proof: We show only that $1/x \to 1/y, y+1$. The remaining results are proved in the same way. One can write equation (2) as

$$\frac{2x}{x^2+x+2} + \frac{x^2+x+1}{x^2+x+2}Y + Y^2 = 0$$

where $Y := 1/y$. Taking this equation modulo $1/x$ we get $Y(Y+1) = 0$ so that the place $1/x$ splits completely in the extension $F^{(9)}/\mathbb{F}_9(x)$ giving rise to a zero of Y (hence a pole of y) and a zero of $Y+1$ (hence a zero of $y+1$). □

Define the sequence $\mathcal{T}^{(q)} := \left(F_1^{(q)}, F_2^{(q)}, F_3^{(q)}, \ldots\right)$ by $F_n^{(q)} := \mathbb{F}_q(x_1, x_2, \ldots, x_n)$ where $f(x_i, x_{i+1}) = 0$ for $1 \le i < n$ where f is as defined in equation (2). Then from Lemma 1 (i) and (iii), it follows that the place $x_1^4 + x_1^2 + x_1 + 1$ of $F_1^{(3)}$is totally ramified in each extension $F_n^{(3)}/F_1^{(3)}$. Thus $\mathcal{T}^{(3)}$ is a tower over $\mathbb{F}_3$ and hence over $\mathbb{F}_9$. From Lemma 1 (i) and (ii), it follows that the only places of $F_1^{(3)}$ that ramify in the tower $\mathcal{T}^{(3)}$ are $x_1^4+x_1^2+x_1+1$ and $x_1^4+x_1^3+x_1^2+2x_1+2$. Therefore $R_{\mathcal{T}^{(9)}} = \{$zeroes of $x_1^4+x_1^2+x_1+1$, zeroes of $x_1^4+x_1^3+x_1^2+2x_1+2\}$.

Now put $q = 9$ and let $S = \{1/x_1, x_1, x_1+2, x_1+1, x_1+w, x_1+w^3\}$. Then each place in the set S splits completely in each extension F_n/F_1 by Lemma 2. Using Theorem 1, we obtain $\lambda(\mathcal{T}^{(9)}) \ge 2 \cdot 6/(-2+8) = 2$. Since $A(9) = 2$, it follows that the tower $\mathcal{T}^{(9)}$ is optimal over $\mathbb{F}_9$ with $\lambda(\mathcal{T}^{(9)}) = 2$.

4.2 Tower over $\mathbb{F}_{25}$

The polynomial

$$(4x+1)y^2 + (x^2+x+2)y + x + 3 \tag{6}$$

gives rise to a tower $\mathcal{T}^{(25)}$ over $\mathbb{F}_{25}$ with ramification set given by $R_{\mathcal{T}^{(25)}} = \{$zeroes of $x_1^4+4x_1^3+x_1^2+1$, zeroes of $x_1^4+2x_1^3+4x_1^2+2x_1+2$, zeroes of $x_1^2+4x_1+2\}$. It can be shown that the zeroes of $x_1^4+2x_1^3+4x_1^2+2x_1+2$ are totally ramified in the tower. Let w be a (primitive) element of $\mathbb{F}_{25}$ which satisfies $w^2+4w+2 = 0$. Then $S = \{1/x, x, x+w^j$ for $j = 0, 3, 4, 6, 7, 8, 11, 12, 14, 15, 16, 18, 20, 22\}$ is a set of 16 places of F_1 which split completely in the tower. We conclude from Theorem 2.1 that $\lambda(\mathcal{T}^{(25)}) \ge 2 \cdot 16/(-2+10) = 4$. Since $A(25) = 4$ it follows that the tower $\mathcal{T}^{(25)}$ is optimal over $\mathbb{F}_{25}$ with $\lambda(\mathcal{T}^{(25)}) = 4$.

4.3 Tower over $\mathbb{F}_{49}$

The polynomial

$$(x^2+6)y^2+xy+x^2+4 \tag{7}$$

gives rise to a tower $\mathcal{T}^{(49)}$ over $\mathbb{F}_{49}$ with ramification set given by $R_{\mathcal{T}^{(49)}} = \{$ zeroes of $x_1^4+x_1^2+3$, zeroes of $x_1^4+6x_1^2+3\}$. It can be shown that the zeroes of $x_1^4+x_1^2+3$ are totally ramified in the tower. Let w be a (primitive) element of $\mathbb{F}_{49}$ which satisfies $w^2+6w+3=0$. Then $S=\{1/x_1, x_1, x_1+w^j$ for $j=4$, 5, 11, 12, 16, 17, 23, 24, 28, 29, 35, 36, 40, 41, 47, 48 $\}$ is a set of 18 places of F_1 which split completely in the tower. Using Theorem 2.1, we obtain $\lambda(\mathcal{T}^{(49)}) \geq 2\cdot 18/(-2+8)=6$. Since $A(49)=6$, it follows that the tower $\mathcal{T}^{(49)}$ is optimal over $\mathbb{F}_{49}$ with $\lambda(\mathcal{T}^{(49)})=6$.

4.4 Tower over $\mathbb{F}_4$

Let q be a power of 2. Define the function field $L^{(q)} := \mathbb{F}_q(x,y)$ by the equation

$$g(x,y)=x^2y^3+(x^3+x^2+x)y^2+(x+1)y+x^3+x=0. \tag{8}$$

One easily checks that $g(x,y)$ is absolutely irreducible. In this section we will make free use of results from [13] without indication.

Lemma 3. *Let $q=2$. The place x^3+x+1 is the only place ramified in the extension $L^{(2)}/\mathbb{F}_2(x)$. Moreover x^3+x+1 is totally ramified and*

$$x^3+x+1 \to y^3+y+1.$$

We also have that

$$x^2+x+1, x^3+x+1 \leftarrow y^3+y+1, \tag{9}$$
$$x^3+x^2+1 \leftarrow y^2+y+1, \tag{10}$$
$$x^3+x^2+1 \leftarrow y^3+y^2+1 \tag{11}$$

and the place y^3+y^2+1 is the only place totally ramified in the extension $L^{(2)}/\mathbb{F}_2(y)$.

Proof: Observe that equation (8) can be written as

$$G(x,Y) := x^2Y^3+Y^2+(x^2+x+1)Y+x+1=0 \tag{12}$$

where $Y := \frac{x+1}{xy}$ so that $\mathbb{F}_2(x,y)=\mathbb{F}_2(x,Y)$. It is easily checked that G is smooth. Let $\mathcal{O}_1$ denote the discrete valuation ring in $\mathbb{F}_2(x)$ which corresponds to the place x^3+x+1. Then the ring $\mathcal{O}_1[Y]/(g(x,Y))$ is a Dedekind domain. Now the discriminant of $G(x,Y)$ is $(x^3+x+1)^2$ so that the only place which ramifies in the extension $L^{(2)}/\mathbb{F}_2(x)$ is x^3+x+1. Suppose that $a^3+a+1=0$. Then putting $x=a$ in equation (12), we have $G(a,Y)=a^2Y^3+Y^2+(a^2+a+1)Y+a+1=a^2(Y+a^5)^3$. Thus the place x^3+x+1 is totally ramified in the extension $L^{(2)}/\mathbb{F}_2(x)$. Putting

$x = a$ in equation (8), one obtains $g(a, y) = a^2(y + a^4)^3 = 0$ whence $y = a^4$ which is another solution of $a^3 + a + 1 = 0$. Thus $x^3 + x + 1 \to y^3 + y + 1$.

Next we show (9), (10) and (11). Again suppose that $a^3 + a + 1 = 0$ and put $y = a$ in equation (8). We get $g(x, a) = (a^2+1)x^3+(a^3+a^2)x^2+(a^2+a+1)x+a = a^6(x + a^2)(x^2 + x + 1) = 0$ so that $x^2 + x + 1 = 0$ or $x = a^2$ which is another solution of $a^3 + a + 1 = 0$. Thus (9) follows.

Suppose that $b^2+b+1 = 0$. Putting $y = b$ in equation (8), one gets $g(x, b) = (b^2+1)x^3+(b^3+b^2)x^2+(b^2+b+1)x+b = b(x^3+x^2+1) = 0$ which implies (10).

Observe that equation (8) can be written as

$$h(Z, y) := 1 + (y^2 + 1)Z + (y^2 + y)Z^2 + y^2Z^3 = 0 \tag{13}$$

where $Z := 1/(x(y + 1) + 1)$. Note that $\mathbb{F}_2(y, Z) = \mathbb{F}_2(y, x)$. It is easily checked that h is smooth. Thus the ring $\mathcal{O}_2[Z]/(h(Z, y))$ is a Dedekind domain where $\mathcal{O}_2$ is the discrete valuation ring in $\mathbb{F}_2(y)$ corresponding to the place $y^3 + y^2 + 1$. Suppose that $c^3 + c^2 + 1 = 0$. Then putting $y = c$ in equation (13), we get $h(Z, c) = 1+(c^2+1)Z+(c^2+c)Z^2+c^2Z^3 = c^2(Z+c^4)^3 = 0$. Thus y^3+y^2+1 is totally ramified in the extension $L^{(2)}/\mathbb{F}_2(y)$. Putting $y = c$ in equation (8) we get $g(x, c) = c^3(x + c^4)^3$ so that $x = c^4$ which is another solution to $c^3 + c^2 + 1 = 0$. Thus $x^3 + x^2 + 1 \leftarrow y^3 + y^2 + 1$ as required. □

Choose $w \in \mathbb{F}_4$ so that $w^2 + w + 1 = 0$.

Lemma 4. *Put $q = 4$. Then we have:*

$$x + 1 \to y, y, y + 1 \tag{14}$$

$$1/x, 1/x, x + 1 \leftarrow y + 1 \tag{15}$$

$$x \to 1/y, 1/y, y \tag{16}$$

$$x + 1, x + 1, x \leftarrow y \tag{17}$$

$$x, x, 1/x \leftarrow 1/y \tag{18}$$

$$1/x \to 1/y, y + 1, y + 1 \tag{19}$$

Proof: (14). Putting $x = 1$ in equation (12) we get $Y^3+Y+1 = Y(Y+w)(Y+w^2) = 0$ so that the place $x + 1$ splits completely in the extension $L^{(4)}/\mathbb{F}_4(x)$ giving rise to places P_1, P_2 and P_3 in $L^{(4)}$ which are the respective zeroes of Y, $Y + w$ and $Y + w^2$. It can be checked that P_2 and P_3 must both be zeroes of y. Putting $x = 1$ in equation (8) we see that at least one of P_1, P_2 and P_3 must be a zero of $y + 1$. Since P_2 and P_3 are zeroes of the function y, the remaining place P_1 must be the zero of $y + 1$.

(15). Observe that equation (8) can be written as

$$A^3 + y^2A^2 + (y^2 + y + 1)A + y(y + 1) = 0 \tag{20}$$

where $A := x(y+1)$ so that $\mathbb{F}_4(x, y) = \mathbb{F}_4(A, y)$. Putting $y = 1$ in equation (20), we get $A^3 + A^2 + A = A(A + w)(A + w^2) = 0$ so that the place $y + 1$ splits completely in the extension $L^{(4)}/\mathbb{F}_4(y)$ giving rise to places P_1, P_2 and P_3, say,

which are the respective zeroes of A, $A+w$ and $A+w^2$. It can be checked that P_2 and P_3 are poles of x. From (14), P_1 must be a zero of $x+1$.
(16). Observe that equation (8) can be written

$$Z^3+(x^2+x+1)Z^2+(x+1)Z+x^2(x^2+1)=0 \tag{21}$$

where $Z:=xy$. Note that $\mathbb{F}_4(x,Z)=\mathbb{F}_4(x,y)$. Putting $x=0$ in equation (21), we get $Z^3+Z^2+Z=Z(Z^2+Z+1)=Z(Z+w)(Z+w^2)=0$. Thus the place x of $\mathbb{F}_4(x)$ splits completely in the extension $L^{(4)}/\mathbb{F}_4(x)$ giving rise to places P_1, P_2 and P_3 in $L^{(4)}$ which are the respective zeroes of Z, $Z+w$ and $Z+w^2$. It can be checked that P_2 and P_3 are both poles of y.

Now, putting $y=0$ in equation (20) so that $A=x$ and $x^3+x=x(x+1)^2=0$ so that there are at least two places in $L^{(4)}$ which lie above the place y of $\mathbb{F}_4(y)$, namely, a zero of x and a zero of $x+1$. It follows that P_3 must be a zero of y.
(17). It follows from (14) and (16).
(18). Observe that equation (8) can be written as

$$\frac{1}{y^2+y}+\frac{y^2}{y^2+y}B+\frac{y^2+y+1}{y^2+y}B^2+B^3=0, \tag{22}$$

where $B=1/(x(y+1))$ so that $\mathbb{F}_4(x,y)=\mathbb{F}_4(B,y)$. Taking equation (22) modulo $1/x$ we get $B^3+B^2+B=B(B+w)(B+w^2)=0$ so that $1/y$ splits completely in the extension $L^{(4)}/\mathbb{F}_4(y)$ giving rise to places P_1, P_2 and P_3, say, in $L^{(4)}$. From (16), two of these places, say P_1 and P_2 are zeroes of x. It follows from this and (16), P_3 cannot be a zero of x. From (14), (15) and (17), P_3 is not a zero of $x+1$. Thus P_3 must be a pole of x.
(19). It follows from (15) and (18). □

Define the sequence $\mathcal{T}^{(q)}:=\left(L_1^{(q)},L_2^{(q)},L_3^{(q)},\ldots\right)$ by $L_n^{(q)}:=\mathbb{F}_q(x_1,x_2,\ldots,x_n)$, where $g(x_i,x_{i+1})=0$ for $1\le i<n$ and g is as given by equation (8). Then from Lemma 3 it follows that the place $x_1^3+x_1+1$ of $L_1^{(2)}$ is totally ramified in each extension $L_{n+1}^{(2)}/L_1^{(2)}$. Thus $\mathcal{T}^{(2)}$ is a tower over $\mathbb{F}_2$. Hence $\mathcal{T}^{(4)}$ is a tower over $\mathbb{F}_4$. It also follows from Lemma 3 that the only places of $L_1^{(4)}$ that ramify in the tower are $x_1^3+x_1+1$,$x_1^3+x_1^2+1$ and the zeroes of $x_1^2+x_1+1$. In other words, $R_{\mathcal{T}^{(4)}}=\{\ x_1^3+x_1+1,\ x_1^3+x_1^2+1,$ zeroes of $x_1^2+x_1+1\}$.

Now put $q=4$ and $S=\{1/x_1,x_1,x_1+1\}$. From (14), (16) and (19) it follows that each place in S splits completely in each extension $L_n^{(4)}/L_1^{(4)}$. By Theorem 1 we obtain the result $\lambda(\mathcal{T}^{(4)})\ge 2\cdot 3/(-2+8)=1$. Since $A(4)=1$, it follows that the tower $\mathcal{T}^{(4)}$ is optimal over $\mathbb{F}_4$ with $\lambda(\mathcal{T}^{(4)})=1$. This completes the proof of Theorem 4.1.

The next result will be used in the next section to show that the tower $\mathcal{T}^{(4)}$ is new.

Lemma 5. *The places of* $L_2^{(4)}$ *which ramify in some extension* $L_n^{(4)}/L_2^{(4)}$ $(n>2)$ *all have degree divisible by* 3.

Proof: Observe that $g(w,y) = w^2(y^3+y+1)$ and $g(w^2,y) = w(y^3+y+1)$. Hence there is exactly one place of $L_2^{(4)}$ above each zero of x^2+x+1 and this place has degree 3 since y^3+y+1 is irreducible. □

5 The above Towers Are New

We will show that the towers over $\mathbb{F}_4$ and $\mathbb{F}_9$ are new. The proof that the tower $\mathcal{T}^{(25)}$ ($\mathcal{T}^{(49)}$) is new is similar to the proof that $\mathcal{T}^{(4)}$ (respectively, $\mathcal{T}^{(9)}$) is new. Given two towers $\mathcal{F} = (F_1, F_2, F_3, \ldots)$ and $\mathcal{E} = (E_1, E_2, E_3, \ldots)$ over $\mathbb{F}_q$, the tower $\mathcal{E}$ is said to be a *subtower* [6] of $\mathcal{F}$ or that $\mathcal{F}$ is a *supertower* of $\mathcal{E}$ if there exists an embedding $\iota : \bigcup_{i\geq 1} E_i \to \bigcup_{i\geq 1} F_i$ over $\mathbb{F}_q$. Hence for each $i \geq 1$ there is an index $m := m(i) \geq 1$ such that $\iota(E_i) \subseteq F_m$. A subtower of an optimal tower is optimal [6]. We call a tower *new* if it is not a subtower of a previously known tower. We make one more definition: we say a place P of F_1 has infinite ramification index in the tower $\mathcal{F}$ if for each $j > 1$, there is a place P_j of F_j which lies above P and such that $e(P_j|P) \to \infty$ as $j \to \infty$. We will use the following result to show that the above towers are new.

Theorem 3. *Let $\mathcal{E} := (E_1, E_2, \ldots)$ be a subtower of $\mathcal{F} := (F_1, F_2, \ldots)$ over $\mathbb{F}_q$. Then we have the following:*
(i) Suppose there is a place P of F_1 which is totally ramified in each extension F_n/F_1. Then there is a sequence $i_1 := 1 < i_2 < i_3 < \ldots$ of numbers such that
- for each $j > 1$ there is a place P_{i_j} of E_{i_j} which ramifies in the extension E_{i_j+1}/E_{i_j}
- $P_{i_{j+1}}|P_{i_j}$ for each j
- $\deg P_{i_j}$ divides $\deg P$ for each $j \geq 1$.
(ii) Suppose there is a place P' of E_1 which has infinite ramification index in the tower $\mathcal{E}$. Then there is a place P of F_1 of infinite ramification index in the tower $\mathcal{F}$.

Proof: By omitting some of the F_i's and renumbering if necessary, we may assume that $E_i \subseteq F_i$ for each i.

(i). Let P be a place of F_1 which is totally ramified in each extension F_n/F_1. For each $j \geq 1$ let P_j be the place of F_j which lies above P. Now fix any $j > 1$ and choose i as large as possible such that $E_i \subseteq F_j$ but $E_{i+1} \nsubseteq F_j$. This is possible since $E_1 \subseteq F_1$ and $g(E_n) \to \infty$ as $n \to \infty$. Let M denote the compositum $E_{i+1}F_j$. Then $M \subseteq F_{i+1}$ and since $E_{i+1} \nsubseteq F_j$ we have $[M : F_j] > 1$. Let Q and R be the respective restrictions of the place P_{i+1} to M and F_j. Since P_{i+1} is totally ramified over P, we must have that $e(Q|R) = [M : F_j] > 1$. Also let $Q' := Q \cap E_{i+1}$ and $R' := R \cap E_i$. Then we must have that $e(Q'|R') > 1$, otherwise by Abhyankar's Lemma, we would have $e(Q|R) = 1$. Thus $P' := P \cap E_1$ is ramified in the extension E_{i+1}/E_i. Moreover, $\deg R'$ divides $\deg P_{i+1} = \deg P$. The result is now clear.

(ii). By assumption, for each $j \geq 1$, there is place P'_j in E_j which lies above the place P' such that $e(P'_j|P') \to \infty$ as $j \to \infty$. For each j, choose a place P_j of F_j which lies above the place P'_j. Now, $e(P_j|P') = e(P_j|P'_j)e(P'_j|P'_1) \to \infty$ as

$j \to \infty$. On the other hand we also have $e(P_j|P') = e(P_j|P_j \cap F_1)e(P_j \cap F_1|P')$. Since $e(P_j \cap F_1|P') \leq [F_1 : E_1]$ for all j, it follows that $e(P_j|P_j \cap F_1) \to \infty$ as $j \to \infty$. Since there are only finitely many possibilities for the places $P_j \cap F_1$ all of which lie above the place P', there must be an index j_0 such that the place $P := P_{j_0} \cap F_1$ appears infinitely often in this sequence. The place P has infinite ramification index in the tower $\mathcal{F}$. □

Now we are ready to show that the towers $\mathcal{T}^{(9)}$ and $\mathcal{T}^{(4)}$ are new. First we show that the tower $\mathcal{T}^{(9)}$ is new. In order to use the above results, we first give a listing of the currently known tame towers together with their ramification properties. Note that the list of tame towers over $\mathbb{F}_{25}$ and $\mathbb{F}_{49}$ can be found in [3].

(a) In [10], it is shown that the tower $\mathcal{M}_1$ over $\mathbb{F}_9$ defined by $y^2 + x^2 + x$ is optimal and $R_{\mathcal{M}_1} = \{x_1, x_1 + 1, x_1 + 2\}$ and the places $1/x_1$ and $x_1 + 1$ are totally ramified.

(b) In [9], it is shown that the tower $\mathcal{M}_2$ over $\mathbb{F}_9$ defined by $y^2 - x^2/(x-1)$ is optimal and $R_{\mathcal{M}_2} = \{1/x_1, x_1 + 1, x_1 + 2\}$ with the places $1/x_1$ and $x_1 + 2$ totally ramified in the tower.

(c) In [9], it is shown that the tower $\mathcal{M}_3$ over $\mathbb{F}_9$ defined by

$$y^2 - \frac{x^2+1}{2x} \tag{23}$$

is optimal. We have $R_{\mathcal{M}_3} = \{1/x_1, x_1, x_1 + 1, x_1 + 2, \text{zeroes of } x_1^2 + 1\}$ with the places $1/x_1$ and x_1 totally ramified in the tower.

(d) In [9], it is shown that the tower $\mathcal{M}_4$ over $\mathbb{F}_9$ defined by

$$y^2 - \frac{(x+1)^2}{4x} \tag{24}$$

and the tower $\mathcal{M}_5$ over $\mathbb{F}_9$ defined by

$$y^2 - \frac{(x+3)^2}{8(x+1)} \tag{25}$$

are both isomorphic to the tower $\mathcal{M}_3$.

In each of the above towers $\mathcal{M}_i$ over $\mathbb{F}_9$, there is a degree one place which is totally ramified whereas in the tower $\mathcal{T}^{(9)}$ only degree two places ramify. Thus by Theorem 3 (i), the tower $\mathcal{T}^{(9)}$ cannot be a subtower of any $\mathcal{M}_i$, $i = 1, 2, 3, 4, 5$.

Next we show that the tower $\mathcal{T}^{(4)}$ is new. In [10], it is shown that the tower $\mathcal{N}_1$ over $\mathbb{F}_4$ defined by $f(x, y) = y^3 + x^3 + x^2 + x$ is optimal. For this tower we have $R_{\mathcal{N}_1} = \{x_1, x_1 + 1, \text{zeroes of } x_1^2 + x_1 + 1\}$ and all the (rational) places in $R_{\mathcal{N}_1}$, except the zero of $x_1 + 1$, are totally ramified in the tower. From lemma 5, we have that the places of $L_2^{(4)}$ which ramify in some extension $L_n^{(4)}/L_2^{(4)}$ $(n > 2)$ all have degree divisible by 3. Thus by Theorem 3 (i) $\mathcal{T}^{(4)}$ cannot be a subtower of $\mathcal{N}_1$.

Remarks: Note that the rational functions (23), (24) and (25) define optimal towers over $\mathbb{F}_{p^2}$ for any odd prime p (see [3] and [9]). Elkies [3], [4] has shown that every currently known explicit optimal tower over $\mathbb{F}_{q^2}$ is either modular or Drinfeld modular. In the Appendix, Elkies proves that the four optimal new towers are again elliptic modular.

Acknowledgement

Many thanks to Arnaldo Garcia for helpful discussions and for pointing out an error in the initial version of Theorem 3.

References

1. M. Daberkow, C. Fieker, J. Klüners, M. Pohst, K. Roegner and K. Wildanger, *KANT V4*, in J. Symbolic Comp. **24** (1997), 267-283.
2. V. G. Drinfel'd and S. G. Vladut, Number of points of an algebraic curve, *Funct. Anal.* **17** (1983), 53-54.
3. N. D. Elkies, Explicit modular towers, *Proceedings of the Thirty-Fifth Annual Allerton Conference on Communication, Control and Computing, T. Basar and A. Vardy, eds.* (1997), 23-32.
4. N. D. Elkies, Explicit towers of Drinfeld modular curves. Proceedings of the 3rd European Congress of Mathematics, Barcelona, 7/2000.
5. A. Garcia and H. Stichtenoth, A tower of Artin-Schreier extensions of function fields attaining the Drinfeld-Vladut bound, *Invent. Math.* **121** (1995), 211-222.
6. A. Garcia and H. Stichtenoth, On the asymptotic behaviour of some towers of function fields over finite fields, *J. Number Theory* **61**, (1996), 248-273.
7. A. Garcia and H. Stichtenoth, Asymptotically good towers of function fields over finite fields, *C. R. Acad. Sci. Paris I* **322** (1996), 1067-1070.
8. A. Garcia and H. Stichtenoth, Skew pyramids of function fields are asymptotically bad, in "Coding Theory, Cryptography and Related Topics", Proceedings of a Conference in Guanajuato, 1998 (J. Buchmann et al, Eds.) 111-113, Springer-Verlag, Berlin 2000.
9. A. Garcia and H. Stichtenoth, On tame towers over finite fields, preprint (2001).
10. A. Garcia, H. Stichtenoth and M. Thomas, On towers and composita of towers of function fields over finite fields. *Finite Fields Appl.* **3** (1997), no. 3, 257-274.
11. Y. Ihara, Some remarks on the number of rational points of algebraic curves over finite fields, *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* **28** (1981), 721-724.
12. H. W. Lenstra, Jr., On a Problem of Garcia, Stichtenoth, and Thomas, To appear in *Finite Fields Appl.*
13. D. Lorenzini, An invitation to arithmetic geometry. Graduate Studies in Mathematics, 9. American Mathematical Society, Providence, RI, 1996.
14. H. Maharaj, H. Stichtenoth and J. Wulftange, On a Problem of Garcia, Stichtenoth, and Thomas II, *In preparation.*
15. H. Niederreiter and C. P. Xing, Rational Points on Curves over Finite Fields: Theory and Applications, Cambridge University Press, Cambridge, 2001.
16. H. Stichtenoth. Algebraic Function Fields and Codes, Springer Universitext. Berlin, Heidelberg, New York, 1993.
17. M.A. Tsfasman, S.G. Vladut and T. Zink, Modular curves, Shimura curves and Goppa codes better than the Varshamov-Gilbert bound. *Math. Nachr.* **109** (1982), 21-28.

Appendix: The polynomials of Theorem 2 Define New Modular Towers

Statement of Results

We identify each of the four recursive towers of Theorem 2 with towers of elliptic modular curves. Specifically, we show:

Theorem A. *The n-th curve in each tower of Theorem 2 is isomorphic with the modular elliptic modular curve associated with the following congruence subgroup G of* $\mathrm{PSL}_2(\mathbb{Z})$*:*

- $G = \varGamma_1(5) \cap \varGamma_0(2^n)$ *for* $f(x,y) = 2xy^2 + (x^2+x+1)y + x^2 + x + 2$ *over* $\mathbb{F}_9$*;*
- $G = \varGamma_1(12) \cap \varGamma_0(2^{n+1})$ $f(x,y) = (4x+1)y^2 + (x^2+x+2)y + x + 3$ *over* $\mathbb{F}_{25}$*;*
- $G = \varGamma_1(5) \cap \varGamma_0(2^n)$ *again for* $f(x,y) = (x^2+6)y^2 + xy + x^2 + 4$ *over* $\mathbb{F}_{49}$*;*
- $G = \varGamma_1(9) \cap \varGamma_0(3^{n+1})$ *for* $x^2y^3 + (x^3+x^2+x)y^2 + (x+1)y + x^3 + x$ *over* $\mathbb{F}_4$*.*

Thus, while these towers are indeed not "subtowers" of previously exhibited optimal towers, they are "supertowers" of towers that are either known already ($\mathrm{X}_0(3^{n+1})$ and $\mathrm{X}_0(3 \cdot 2^{n+1})$, see [3]) or easily obtained by known methods ($\mathrm{X}_0(5 \cdot 2^n)$, see below). Still, they have two new features. First, while every previous recursive tower of elliptic modular curves is either of the form $\{\mathrm{X}_0(l^n N_0) : n = 1, 2, 3, \ldots\}$ or a subtower of some $\{\mathrm{X}_0(l^n N_0)\}$, these new towers require intersecting $\varGamma_0(l^n N_0)$ with $\varGamma_1(N_0)$. As a result, one cannot use the usual models of these curves (in which rational functions have rational q-expansions at infinity): as Ihara observed, to obtain asymptotically optimal towers, one must use Igusa's model of the modular curves, which is a twist of the usual one. The second novelty concerns the identification of the coordinates $(x_1, \ldots, x_n)$ with modular functions. In each modular tower exhibited so far, we found a modular function $x_1(\cdot)$ on the upper half-plane satisfying the identity $f(x_1(\tau), x_1(l\tau)) = 0$, leading to the parametrization of $(x_1, x_2, x_3, \ldots, x_n)$ by modular functions $(x_1(\tau), x_1(l\tau), x_1(l^2\tau), \ldots, x_1(l^{n-1}\tau))$. In each of the four new towers, the identity takes the form $f\big(x_1(\tau), \epsilon(x_1(l\tau))\big) = 0$, where ϵ is a fractional linear transformation such that

$$f(x,y) = 0 \Longleftrightarrow f(\epsilon(x), \epsilon(y)) = 0. \tag{26}$$

Thus the coordinate x_{i+1} $(0 < i < n)$ is $\epsilon^i(x_1(l^i\tau))$ instead of the familiar $(x_1(l^i\tau))$. In each case the cyclic group generated by ϵ gives the action of $\varGamma_0(N_0)/\varGamma_1(N_0)$ on the x_1-line $\mathrm{X}_1(N_0)$.

The transformations ϵ were also a key tool in recognizing the modular towers. In each case, we first found ϵ satisfying (26), then identified the quotient subtower with a modular tower of curves $\{\mathrm{X}_0(l^n N_0)\}$. This then suggested what the original tower must be.

The Quadratic Towers

The three quadratic towers are similar enough that, as for Theorem 2, we treat only one of them fully, and indicate how to modify the formulas to obtain the other two.

Consider first the tower $f(x,y) = 2xy^2 + (x^2+x+1)y + x^2 + x + 2$ over $\mathbb{F}_9$. We construct the directed graph with vertex-set S and with edge-set $\{P \to Q : P, Q \in S\}$. (See again Section 2 for the definitions of S and "$P \to Q$", and the end of 4.1 for the set S associated to the tower over $\mathbb{F}_9$; the edges are exhibited in Lemma 2.) We find that this graph has an involution that fixes the places x_1 and $x_1 + 1$ and switches $1/x$ with $x_1 + 2$ and $x + w$ with $x + w^3$. We guess that this involution is a fractional linear transformation ϵ satisfying (26), and readily find that $\epsilon(x) = x/(x-1)$ and verify (26). To form the quotient subtower we introduce variables $X = x + \epsilon(x)$, $Y = y + \epsilon(y)$, and eliminate x, y from $f(x,y) = 0$ to obtain $F(X,Y) = 0$ where

$$F(X,Y) = XY^2 - X^2Y + (X+1)^2. \tag{27}$$

This again has an involution, which we call $w : t \leftrightarrow (1-t)/(1+t)$. We form the quotient subtower in the same way: let $\xi = X + w(X)$, $\eta = Y + w(Y)$, and eliminate X, Y from $F(X,Y) = 0$ to obtain $\phi(\xi, \eta) = 0$ where

$$\phi(\xi,\eta) = (\xi - 1)\eta^2 + (\xi - \xi^2)\eta + \xi^2 + \xi. \tag{28}$$

The size and structure of our graph, and the action of ϵ on it and of w on its quotient by $\{1, \epsilon\}$, suggest that the F and ϕ towers are isomorphic with the modular towers $\{\mathrm{X}_0(5{\cdot}2^n)\}$ and $\{\mathrm{X}_0(5{\cdot}2^n)/w_5\}$. We prove this next by obtaining these towers explicitly in characteristic zero.

Let H, h be the Hauptmoduln for the rational curves $\mathrm{X}_0(10)$, $\mathrm{X}_0(10)/w_5$, that are defined by the eta products

$$H(\tau) = q^{-1}\prod_{n=1}^{\infty}\frac{1+q^n}{(1+q^{5n})^5} = q^{-1} + 1 + q + 2q^2 + 2q^3 - 2q^4 - q^5 \cdots,$$

$$h(\tau) = q^{-1}\prod_{n=1}^{\infty}\Big((1+q^n)(1+q^{5n})\Big)^{-4} = q^{-1} - 4 + 6q - 8q^2 \cdots = \frac{H^2 - 4H}{H+1}$$

(where as usual $q = e^{2\pi i\tau}$). For $i = 1, 2, 3, \ldots$, define

$$H_i := H(2^i\tau), \qquad h_i := h(2^i\tau).$$

Then each pair H_i, H_{i+1} satisfies the same polynomial relation, quadratic in each variable; by comparing q-expansions we find the relation

$$H_{i+1}^2 = H_i(H_iH_{i+1} - 2H_{i+1} - 4). \tag{29}$$

These equations in $H_1, \ldots, H_n$ give the modular curve $\mathrm{X}_0(5 \cdot 2^n)$. Likewise

$$h_{i+1}^2 = h_i(h_ih_{i+1} + 8h_{i+1} + 16). \tag{30}$$

yields $\mathrm{X}_0(5 \cdot 2^n)/w_5$. To compare these modular towers in characteristic 3 with the recursive towers defined by F and ϕ, we consider the "fixed points" of the recursions: the solutions of $F(X,X) = 0$ and $\phi(\xi,\xi) = 0$, as against those of

$H_i = H_{i+1}$ and $h_i = h_{i+1}$. In the case of $\mathrm{X}_0(5 \cdot 2^n)/w_5$ and ϕ, we find that the fractional linear transformation $(h_i, h_{i+1}) = (\xi/(\xi+1), \eta/(\eta+1))$ takes (30) to (28) and thus identifies the two towers. For $\mathrm{X}_0(5 \cdot 2^n)$ a new twist arises: the equation $H_i = H_{i+1}$ has four simple roots, but $F(X, X) = 0$ has two double roots at $X = -1$ and $X = \infty$. We must instead use the equivalent form of the F tower obtained by applying the involution w to only one variable. This yields

$$(1 - X^2)Y^2 - (X^2 + X + 1)Y + 1 = 0.$$

The tower can now be identified with $\{\mathrm{X}_0(5 \cdot 2^n)\}$ by taking $(H_i, H_{i+1}) = (\alpha(X), \alpha(Y))$, where

$$\alpha(t) := (t - I)/\big((I - 1)t - 1\big)$$

and $I^2 = -1$ in $\mathbb{F}_9$.

It remains to identify the tower defined by $f(x_i, x_{i+1}) = 0$ with the tower of curves obtained from $\{\mathrm{X}_0(5 \cdot 2^n)\}$ by taking fiber products with $\mathrm{X}_1(5)$ over $\mathrm{X}_0(5)$. The bottom curve $\mathrm{X}_1(10)$ of this tower is rational, and has a Hauptmodul with a product formula

$$H'(\tau) = q^{-1} \prod_{n=1}^{\infty} (1 - q^n)^{c_n}, \quad \text{where } c_n = \begin{cases} -1, & \text{if } n \equiv \pm 1 \text{ or } \pm 2 \bmod 10; \\ +1, & \text{if } n \equiv \pm 3 \text{ or } \pm 4 \bmod 10; \\ 0, & \text{if } 5|n. \end{cases}$$

By comparing q-expansions we find that $H = H' - 1/H'$. Thus the double cover $\mathrm{X}_1(10)/\mathrm{X}_0(10)$ is ramified at the roots of $H^2 + 4 = 0$. Reducing these roots to $\mathbb{F}_9$ yields $I = \alpha(0)$ and $-I = \alpha(1)$. But $X = 0$ and $X = 1$ are the branch points of the double cover of the X-line by the x-line given by $X = x + \epsilon(x)$. Hence these double covers yield isomorphic supertowers of $\{\mathrm{X}_0(5 \cdot 2^n)\}$, and we are done.

We readily adapt this analysis to the characteristic-7 tower, with $f(x, y) = (x^2 + 6)y^2 + xy + x^2 + 4$. Here the first involution is visible: $\epsilon(t) = -t$. We let $(X, Y) = (x^2, y^2)$, and find the new involution $w : t \leftrightarrow (3/t)$. Then $(\xi, \eta) = (X + 3/X, Y + 3/Y)$ satisfy a quadratic equation that we identify with (30) by taking $(h_i, h_{i+1}) = (3(\xi + 1)/(\xi - 1), 3(\eta + 1)/(\eta - 1))$. Thus the ξ tower is isomorphic with $\{\mathrm{X}_0(5 \cdot 2^n)/w_5\}$. To treat the X tower we apply w to only one of the variables, and identify the resulting equation with the $\{\mathrm{X}_0(5 \cdot 2^n)\}$ recursion (29) by taking $(H_i, H_{i+1}) = (\alpha(X), \alpha(Y))$, where

$$\alpha(t) := -(2t + 2I + 1)/\big(It + 3I + 1\big)$$

and $I^2 = -1$ in $\mathbb{F}_{49}$. Applying α to the branch points $0, \infty$ of the double cover $X = x^2$ recovers the branch points $H = -2I$, $H = 2I$ of the double cover $X_1(10)/X_0(10)$, and again shows that the resulting supertowers are isomorphic.

Finally we outline our analysis of the tower in characteristic 5 with $f(x, y) = (4x + 1)y^2 + (x^2 + x + 2)y + x + 3$. Here ϵ is the involution $t \leftrightarrow (t + 1)/(2t - 1)$. We take $X = x + \epsilon(x) - 1$, $Y = y + \epsilon(y) - 1$ to obtain $(X - 2)Y^2 = (X^2 - 2X + 2)Y + 2(X^2 - X)$. This has an involution $w : t \leftrightarrow 2/t$, so we take $\xi = X + w(X)$, $\eta = Y + w(Y)$ and find $\phi(\xi, \eta) = 0$ where $\phi(\xi, \eta) = (\xi + 2)\eta^2 - (\xi^2 + \xi - 1)\eta +$

$2\xi^2 - \xi + 2$. Now let $\xi = (1 - h_i)/(h_i + 1)$ and $\eta = (1 - h_{i+1})/(1 + h_{i+1})$ to reach $h_i^2 = h_{i+1}(h_{i+1} - 1)(h_i - 1)$. This is the reduction mod 5 of the relation $h_i^2 = h_{i+1}(h_{i+1} + 4)(h_i + 4)$ satified by $h_i := h(2^i\tau)$, where h is the Hauptmodul for $\mathrm{X}_0(12)/w_3$ defined by the eta product

$$h(\tau) = q^{-1}\prod_{n=1}^{\infty}\left(\frac{(1-q^n)(1-q^{3n})}{(1-q^{4n})(1-q^{12n})}\right)^2 = q^{-1} - 2 - q + 7q^3 - 9q^5 + 10q^7 - 23q^9 \cdots.$$

Hence ϕ generates a recursive tower of curves isomorphic with $\mathrm{X}_0(3 \cdot 2^{n+1})/w_3$. For $\mathrm{X}_0(12)$ we use the Hauptmodul

$$H(\tau) = q^{-1}\prod_{n=1}^{\infty}\frac{(1-q^{4n})(1-q^{3n})^3}{(1-q^n)(1-q^{12n})^3} = q^{-1} + 1 + 2q + q^3 - 2q^7 - 2q^9 + 2q^{11} \cdots,$$

related with h by $h = (H^2 - 4H)/(H - 1)$. Then $H_i := H(2^i\tau)$ satisfy $H_i^2 = H_{i+1}(H_{i+1} - 2)(H_i - 2)$. We recover the quadratic relation between X and Y by setting $H_i = \alpha(X)$ and $H_{i+1} = \alpha(w(Y))$ where $\alpha(t) := (t + R + 2)/((R+1)t - R)$ for $R \in \mathbb{F}_{25}$ such that $R^2 = 2$. This confirms that $(X - 2)Y^2 = (X^2 - 2X + 2)Y + 2(X^2 - X)$ generates the modular tower $\{\mathrm{X}_0(3 \cdot 2^{n+1})\}$. Finally, $\mathrm{X}_1(12)$ is rational with Hauptmodul

$$H'(\tau) = q^{-1}\prod_{n=1}^{\infty}(1 - q^n)^{c_n}, \quad \text{where } c_n = \begin{cases} -1, & \text{if } n \equiv \pm 1 \bmod 12; \\ +1, & \text{if } n \equiv \pm 5 \bmod 12; \\ 0, & \text{if } (n, 12) \neq 1. \end{cases}$$

The double cover $\mathrm{X}_1(12)/\mathrm{X}_0(12)$ is given by $H = H' + 1/H'$ and is therefore ramified at $H = \pm 2 = \alpha(\mp 2R)$. Since these values $\mp 2R$ are also taken by $X = x + \epsilon(x) - 1$ at the fixed points $x = \pm R - 2$ of ϵ, they are the branch points of our double cover of the X-line by the x-line given by $X = x + \epsilon(x) - 1$. Again we have completed the identification of the tower $f(x_i, x_{i+1}) = 0$ with a tower of modular curves as claimed in the statement of Theorem A.

The Cubic Tower

We now consider the tower $x^2y^3 + (x^3 + x^2 + x)y^2 + (x+1)y + x^3 + x$ over $\mathbb{F}_4$. This time (26) holds for $\epsilon(x) = 1/(x+1)$, a fractional linear transformation of order 3. Let $X = x + \epsilon(x) + \epsilon^2(x) = (x^3 + x + 1)/(x^2 + x)$ and $Y = (y^3 + y + 1)/(y^2 + y)$. We then eliminate x, y from $f(x, y) = 0$ to obtain $Y^3 = X^3 + X^2 + X$. This yields a known optimal tower over $\mathbb{F}_4$, discovered by Garcia and Stichtenoth [7] and identified with the modular tower $\{\mathrm{X}_0(3^{n+1})\}$ in [3]. Here the modular parametrization is $X_i = 1 + 1/H(3^{i+1}\tau)$ where

$$H(\tau) = 3 + q^{-1}\prod_{n=1}^{\infty}\left(\frac{1-q^n}{1-q^{9n}}\right)^3 = q^{-1} + 5q^2 - 7q^5 + 3q^8 + 15q^{11} - 32q^{14} \cdots.$$

Now $\mathit{\Gamma}_1(9)$ is a rational curve with Hauptmodul

$$H'(\tau) = q^{-1}\prod_{n=1}^{\infty}(1-q^n)^{c_n}, \quad \text{where } c_n = \begin{cases} -1, & \text{if } n \equiv \pm 1 \text{ or } \pm 2 \bmod 9; \\ +2, & \text{if } n \equiv \pm 4 \bmod 9; \\ 0, & \text{if } 3|n. \end{cases}$$

Since $X_1(9)/X_0(9)$ is a cyclic cubic cover, H must be a rational function of H' of degree 3 with cyclic Galois group; by comparing q-expansions we find

$$H = \frac{H'^3 - 3H' - 1}{H'^2 - H} = H' + \varepsilon(H') + \varepsilon^2(H')$$

where $\varepsilon(t) = -(t+1)/t$. Setting $H'_i = H'(3^i\tau)$ we find the cubic recursion

$$H'^3_i = H'^3_{i+1} + (-H'^3_i + 3H'^2_i + 3H'_i + 1)H'^2_{i+1} + (-H'^3_i + 6H'^2_i + 6H'_i + 1)H'_{i+1}.$$

Of necessity this is invariant under the substitution

$$(H'_i, H'_{i+1}) \leftarrow (\varepsilon(H'_i), \varepsilon(H'_{i+1})).$$

Replacing only H'_{i+1} by $\varepsilon(H'_{i+1})$ and reducing mod 2, we obtain the equivalent recursion

$$(H'^3_i + H'^2_i + H'_i + 1)H'^3_{i+1} + H'^3_i H'^2_{i+1} + (H'^3_i + H'^2_i + H'_i)H'_{i+1} + 1 = 0$$

for the tower of modular curves over $\mathbb{F}_4$ corresponding to $\mathit{\Gamma}_1(9) \cap \mathit{\Gamma}_0(3^{n+1})$. We next find a fractional linear transformation α such that

$$(H_i, H_{i+1}) = (\alpha(X), \alpha(Y))$$

identifies this tower with the one obtained from $f(x,y)$. Since the latter tower is optimal, this requires a cubic twist of the cover $X_1(9)/X_0(9)$, forcing α to have coefficients outside the field of definition of the tower. We find that $\alpha(t) = (Ct+1)/(t+C+1)$ works for $C \in \mathbb{F}_8$ such that $C^3 = C+1$. This completes the proof of the last part of Theorem A.

Concluding Remarks

We noted that the cubic tower can be identified with the modular tower specified in Theorem A only over a cubic extension of $\mathbb{F}_4$. This arises because Igusa's models of the curves in the modular tower are cubic twists over $\mathrm{X}_0(3^{n+1})$ of their usual models. The quadratic towers require twists as well: they can be identified with the usual models of modular towers only over quadratic extensions of $\mathbb{F}_{p^2}$. We avoided exhibiting fractional linear transformations that realize this identification over $\mathbb{F}_{p^4}$ (and the lifts of w to fractional linear involutions of x, y) by checking that the branch points over the base curves of the degree-2 subtowers agree with those of $\mathrm{X}_1(10)/X_0(10)$ and $\mathrm{X}_1(12)/X_0(12)$.

Theorem A, together with the computations of Li, Maharaj and Stichenoth reported in the body of the paper, may be regarded as further computational evidence of the modularity conjecture for optimal recursive towers that we proposed in [3]. One might reasonably ask whether this conjecture is falsifiable: how could one prove that a potential counterexample is *not* modular? But modularity imposes stringent conditions on a tower of curves. For instance, the Galois group of its closure over the function field F_1 of the base curve must be of GL_2 type. A tower that failed this necessary condition would automatically be a counterexample. Conversely, if the conjecture is true, it may be possible to demonstrate that the Galois condition holds for every optimal recursive tower, as a step towards proving the conjecture.

Acknowledgements

I thank Ihara for alerting me to the use of Igusa's model of modular curves not dominated by $\mathrm{X}_0(N)$ to obtain asymptotically optimal curves. I thank the Packard Foundation for partial financial support during the preparation of this work. The symbolic computations reported above were greatly facilitated by the computer packages PARI and MACSYMA.

Periodic Continued Fractions in Elliptic Function Fields

Alfred J. van der Poorten and Xuan Chuong Tran

ceNTRe for Number Theory Research,
Macquarie University, Sydney,
NSW 2109, Australia
alf@mpce.mq.edu.au
xctran@hotmail.com

Abstract. We construct all families of quartic polynomials over $\mathbb{Q}$ whose square root has a periodic continued fraction expansion, and detail those expansions. In particular we prove that, contrary to expectation, the cases of period length nine and eleven do not occur. We conclude by providing a list of examples of pseudo-elliptic integrals involving square roots of polynomials of degree four. The primary issue is of course the existence of units in elliptic function fields over $\mathbb{Q}$. That, and related issues are surveyed in the paper's introduction.

1 Introduction

We provide the expansion of all families of quartic polynomials defined over $\mathbb{Q}$ and with periodic continued fraction expansion, and derive from that a list of examples of each family of pseudo-elliptic integrals involving square roots of such polynomials of degree four.

2 Units in Quadratic Function Fields

Let $D(X)$ be a polynomial, not a square, defined over a field $\mathbb{F}$ of characteristic zero, and suppose there are polynomials $x(X)$, $y(X)$ defined over $\mathbb{F}$, with $y \neq 0$, so that $x^2 - Dy^2$ is a constant $-\kappa$, of course in $\mathbb{F}$.

Example 1. Suppose we are given the pseudo-elliptic integral

$$\begin{aligned}&\int^u \frac{4t-1}{\sqrt{t^4-2t^3+3t^2+2t+1}}\,dt\\ &\quad = \log\bigl((u^4-3u^3+5u^2-2u)+(u^2-2u+2)\sqrt{u^4-2u^3+3u^2+2u+1}\,\bigr).\end{aligned}$$

Set $D(u) = u^4-2u^3+3u^2+2u+1$, $x(u) = u^4-3u^3+5u^2-2u$, $y(u) = u^2-2u+2$. We may save ourselves an annoying verification. Add to the given claim the corresponding allegation with $\sqrt{D}$ replaced by $-\sqrt{D}$. On the left we integrate zero, and on the right we obtain $\log(x^2-Dy^2)$; that is, $x^2-Dy^2=-\kappa$ must be a constant.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 390–404, 2002.

Example 2. Because $D \neq \square$, it is plain that $\kappa \neq 0$. Just so, D must be of even degree, $2g+2$ say, and with leading coefficient a square in $\mathbb{F}$. It follows that $\delta(X) = \sqrt{D(X)}$ is represented by a Laurent series in $\mathbb{F}((X^{-1}))$, say $\sum_{h=-g-1}^{\infty} d_h X^{-h}$.

Take $\delta(X) = \sqrt{X^4 - 2X^3 + 3X^2 + 2X + 1}$. Then

$$\begin{aligned}\delta(X) = X^2 - X + 1 + 2X^{-1} + 2X^{-2} - 4X^{-4} - 8X^{-5} - 6X^{-6} + 10X^{-7} \\ + 40X^{-8} + 58X^{-9} + 2X^{-10} - 188X^{-11} - 442X^{-12} - 382X^{-13} + \cdots \quad .\end{aligned}$$

Plainly, the element $u = x - \delta y$ of the function field $\mathbb{K} = \mathbb{F}(X, \delta)$ of the curve $\mathcal{C} : Y^2 = D(X)$ is a non-trivial unit in $\mathbb{K}$. Indeed, it divides a trivial unit $\kappa \in \mathbb{F} \subset \mathbb{K}$. Hence the divisor of u on the Jacobian $\operatorname{Jac}(\mathcal{C})$ of $\mathcal{C}$ is supported only at infinity, thus at just two points, which we may conveniently call ∞_+ and ∞_-. Because it is the divisor of a function it has degree zero and thus there is some integer m — in fact, the regulator of $\mathbb{K}$ — so that $m(\infty_+ - \infty_-)$ is the divisor of a function. That is, $\infty_+ - \infty_-$ is a torsion point of order m on $\operatorname{Jac}(\mathcal{C})$.

It is well understood that the existence of a non-trivial unit in $\mathbb{K}$ guarantees that δ has a periodic continued fraction expansion. In [11] we also explain why, unlike the case of real quadratic irrationals where the continued fraction of the square root $\sqrt{D}$ of any positive nonsquare integer is always periodic, the continued fraction of the square root $\delta(X)$ of a polynomial D is not always periodic. The point is that, by the box principle, Pell's equation $x^2 - Dy^2 = 1$ always has a solution in the number case, but — because there are infinitely many polynomials of bounded degree if the base field $\mathbb{F}$ is infinite — Pell's equation does not necessarily have a solution in the function case. Assisted by ideas of Tom Berry [2], we also detail the structure of the period of the continued fraction expansion of $\sqrt{D(x)}$ when D is a polynomial over a field $\mathbb{F}$ and the expansion of $\sqrt{D(x)}$ happens to be periodic. In particular, we notice that, given the existence of unit $x - \delta y$ with norm $x^2 - Dy^2 = -\kappa$, then $\big((x^2 + Dy^2) - 2\delta xy\big)/\kappa$ is a unit of norm 1, given by a period of the continued fraction expansion of δ. For $\kappa \neq -1$, the unit $x - \delta y$ is said to be given by a *quasi*-period.

We recall that a continued fraction expansion

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{a_4 + \cfrac{1}{a_5 + \ddots}}}}}$$

plainly needs a less wasteful notation, say $[\,a_0\,,\,a_1\,,\,a_2\,,\ldots]$, to represent it.

Example 3. We have

$$\begin{aligned}&\sqrt{X^4 - 2X^3 + 3X^2 + 2X + 1} = \\ &[X^2 - X + 1, \\ &\quad \overline{\tfrac{1}{2}X - \tfrac{1}{2},\, 2X - 2,\, \tfrac{1}{2}X^2 - \tfrac{1}{2}X + \tfrac{1}{2},\, 2X - 2,\, \tfrac{1}{2}X - \tfrac{1}{2},\, 2X^2 - 2X + 2}],\end{aligned}$$

displaying the full period, whereas

$$[X^2 - X + 1\,,\, \tfrac12 X - \tfrac12\,,\, 2X - 2] = x(X)/y(X),$$

with $x(X) = X^4 - 3X^3 + 5X^2 - 2X$ and $y(X) = X^2 - 2X + 2$, already provides a unit, $x(X) - y(X)\sqrt{X^4 - 2X^3 + 3X^2 + 2X + 1}$, of norm -4. One also notices

$$4 \cdot [\tfrac12 X - \tfrac12\,,\, 2X - 2\,,\, \tfrac12 X^2 - \tfrac12 X + \tfrac12\,] = [2X - 2\,,\, \tfrac12 X - \tfrac12\,,\, 2X^2 - 2X + 2].$$

Here, we recall Wolfgang Schmidt's felicitous formulation [15] of a well known fact:

Proposition 1 (Multiplication of continued fractions by a constant).

$$B[\,Ca_0\,,\, Ba_1\,,\, Ca_2\,,\, Ba_3\,,\, Ca_4\,,\, \ldots] = C[\,Ba_0\,,\, Ca_1\,,\, Ba_2\,,\, Ca_3\,,\, Ba_4\,,\, \ldots].$$

Example 4. It is a consequence of the various symmetries and twisted symmetries possessed by the period of the the square root of a quadratic irrational with polynomial trace (such symmetries are instanced by the preceding example), that a quasi-period must be of odd length; that is, if it is of even length then it is in fact a period.

Look carefully at the period $a_1\,,\, a_2\,,\, \ldots\,,\, a_{2r}$:

$$\tfrac12 X - \tfrac12\,,\, 2X - 2\,,\, \tfrac12 X^2 - \tfrac12 X + \tfrac12\,,\, 2X - 2\,,\, \tfrac12 X - \tfrac12\,,\, 2X^2 - 2X + 2\,.$$

Other than for $2r = 6$ and $\kappa = 4$, the following features are not particular to the example. First, the word $a_1 a_2 \cdots a_{2r-1}$ is symmetric. Second, as 'also noticed' above, the second half repeats the first half of the period, up to a twist by κ. In the example, r is too small fully to illustrate that the half period $a_1 \cdots a_{r-1}$ is twisted symmetric: in that $\kappa \cdot [\,a_1\,,\, \ldots\,,\, a_{(r-1)/2}\,] = [\,a_{r-1}\,,\, \ldots\,,\, a_{(r+1)/2}\,]$. Whatever, these observations force r indeed to be odd.

The non-periodic case is considered in [12]. There, the point is that it is easy enough to notice periodicity, but not at all obvious how to prove non-periodicity. Aided by remarks of Jin Yu in [17], the paper [12] instances a simple criterion (based on reduction modulo different primes) that readily allows the detection of non-periodicity from inspection of just several initial partial quotients of the continued fraction expansion.

Below, we apply the results alluded to above to compute all quartic polynomials $D(x)$ over $\mathbb{Q}$ so that $\sqrt{D(x)}$ does have a periodic continued fraction expansion. In the case $\deg D = 4$, the curve $\mathcal{C} : Y^2 = D(X)$ is of genus $g = 1$, and may be considered to coincide with its Jacobian. Thus it suffices to list the various possibilities for the order of torsion points on an elliptic curve, as we may by a celebrated result of Mazur [7], and, following the algorithm given by Adams and Razar [1], to obtain the model $\mathcal{C}$ so as to have located the relevant torsion point at infinity. Specifically, given an elliptic curve $E/\mathbb{Q} : v^2 = u^3 + Au + B$ and a rational point $P = P(a, b)$ on E, the transformation

$$u = \tfrac12(X^2 + Y - a), \qquad v = \tfrac12(X^3 + XY - 3aX - 2b) \tag{1}$$

maps P and the point at infinity O on E to the two points at infinity on

$$E_P : Y^2 = X^4 - 6aX^2 - 8bX + c\,,$$

where $c = -4A - 3a^2$ and $B = b^2 - a^3 - Aa$.

Conversely the formulas

$$X = (v+b)/(u-a)\,, \qquad Y = 2u + a - \big((v+b)/(u-a)\big)^2$$

transform the quartic model E_P back to E; thus (1) is a birational transformation.

The elliptic case $g = 1$ is congenial for reasons additional to Mazur's theorem. For general genus g, it is easy to see that the complete quotients $\delta_h(X)$ of δ are all of the shape

$$\delta_h = (P_h + \sqrt{D})/Q_h\,,$$

with $Q_h \big| D - P_h^2$ and, this remark is in part just setting the notation, the generic step in the continued fraction algorithm for $\delta = \sqrt{D}$ is

$$\delta_h = (P_h + \sqrt{D})/Q_h = a_h - (P_{h+1} - \delta)/Q_h\,. \tag{2}$$

Here the sequences of polynomials (P_h) and (Q_h) are given sequentially by

$$P_{h+1} + P_h = a_h Q_h, \quad \text{and} \quad Q_{h+1}Q_h = D - P_{h+1}^2\,.$$

Proposition 2. *The polynomials* Q *and* P *satisfy* $\deg Q_h \le g = \frac{1}{2}\deg D - 1$ *and* $\deg P_{h+1} = g + 1 = \frac{1}{2}\deg D$ *for all* $h = 0,\ 1,\ \ldots$.

Proof. Given $\deg Q_h \le g$ it follows from $-(P_{h+1} - \sqrt{D})/Q_h$ being a remainder, so that it is of negative degree, that $\deg P_{h+1} = g + 1$ and $\deg(P_{h+1} - \sqrt{D})$ is less than $\deg Q_h$. Thus $Q_{h+1}Q_h = D - P_{h+1}^2$ entails that $\deg Q_{h+1} \le g$. Finally, $\delta_0 = \sqrt{D}$ displays that $Q_0 = 1$, so $\deg Q_0$ is no more than g.

Now notice that $P_{h+1} + P_h = a_h Q_h$ entails that $\deg Q_h = 0$ is equivalent to $\deg a_h = g + 1$. However, $\deg Q_l = 0$ signals that δ has a quasi-period comprising the partial quotients a_1, a_2, …, a_l. Moreover, if this is a primitive such period then, other than for a_l, all those partial quotients have degree at most g. Thus, in the elliptic case, the quasi-period length l implies that the regulator m — the degree of the fundamental unit or, equivalently, the sum of the degrees of the partial quotients comprising the quasi-period — is given by $m = l + 1$. For larger g, the corresponding argument typically does no better than $m \ge l + g$.

Back to the case $\deg D = 4$ and base field $\mathbb{Q}$, we know from [7] that the possible values for m are 2, 3, …, 10, and 12; because those are the possible torsion orders of the 'divisor at infinity' on $\mathcal{C}$.

We recall that a quasi-period of even length is in fact a period, whereas a quasi-period of odd length r might be a period, or it yields a primitive period of length $2r$. It follows from the first reason that we will find primitive periods of length 2, 4, 6, and 8, and for the second reason that there surely will be primitive periods of length 1 and 2, 3 and 6, 5 and 10, 7 and 14, 9 and 18, and 11 and 22. Here one expects the periods of odd length to occur because the norm of the fundamental unit may surely happen to be -1.

However, as it happens, we see below that the periods 9 and 11 do *not* occur.

Example 5. Set $D(X) = X^4 - 2X^3 + 3X^2 + 2X + 1$, and consider the continued fraction expansions of the numbers $\sqrt{D(n)}$ for $n = 1, 2, \ldots$. Of course these expansions are periodic, of respective period lengths $\ell\bigl(D(n)\bigr) = \ell_n$, say. It is notorious that, given an arbitrary positive integer k, not a square, it is in general extraordinarily difficult to predict the period length $\ell(k)$ of the expansion of $\sqrt{k}$. Yet here $\ell_{2n-1} = 17$ and $\ell_{2n} = 7$ for $n = 2, 3, \ldots$. By the way, all the $D(n)$ are 1 modulo 4 so that, in decency, we should have considered the quantities $(\sqrt{D(n)} + 1)/2$ in place of $\sqrt{D(n)}$. Indeed, their periods all have length 5, for $n = 2, 3, \ldots$.

This last remark is apropos, given a theorem of Schinzel [16] to the following effect. Suppose $f(X)$ is a polynomial, not a square, taking positive integer values at $X = 1, 2, \ldots$. Denote by ℓ_n the length of the period of the continued fraction expansion of $\sqrt{f(n)}$. Then $\limsup_{n\to\infty} \ell_n$ is finite if and only if there is a nontrivial unit in the function field $\mathbb{Q}\bigl(X, \sqrt{f(X)}\,\bigr)$, which moreover has *integer* coefficients, that is there is a unit defined over $\mathbb{Z}\bigl(X, \sqrt{f(X)}\,\bigr)$.

In this context, Schinzel speculates on the possible period lengths for quartic polynomials f ([16, p297]) reporting 1 and the even lengths "and possibly also 5, 7, 9, 11 (I have not verified this) … ". Of course, in 1962 the result of Mazur was as yet no more than a conjecture (of Nagell). Related remarks of Schinzel include essentially everything observed above and make clear moreover that these things were mostly already known to Abel and Tchebicheff. For details and references see [16, II §4]. The continued fractions in the easier genus zero case are given by [16, I] and are discussed by van der Poorten and Hugh Williams in [13].

Pseudo-elliptic integrals, as instanced at Example 1, are the subject of [11]; with one change. In [11] we write about *quasi*-elliptic integrals as if these integrals are 'sort of' elliptic, in the sense that a quasi-period certainly kind of is a period (quasi: resembling; as it were …). The qualifier *quasi* was incorrect. It would have been more to the point to speak of *pseudo*-elliptic integrals (pseudo: a word element meaning false, pretended…), emphasising that these integrals have elliptic appearance but are not elliptic at all.

3 Continued Fractions of Quadratic Irrationals

Anyone attempting to compute the truncations $[\,a_0\,,\,a_1\,,\,\ldots\,,\,a_h\,] = x_h/y_h$ of a continued fraction will be delighted to notice that the definition

$$[\,a_0\,,\,a_1\,,\,\ldots\,,\,a_h\,] = a_0 + 1/[\,a_1\,,\,\ldots\,,\,a_h\,]$$

immediately implies by induction on h that there is a correspondence

$$\begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix}\cdots\begin{pmatrix} a_h & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} x_h & x_{h-1} \\ y_h & y_{h-1} \end{pmatrix} \longleftrightarrow [\, a_0 \,,\, a_1 \,,\, \ldots \,,\, a_h \,] = x_h/y_h$$

between products of certain two by two matrices and the convergents of continued fractions. Notice, incidentally, that if a product of matrices corresponds to x_h/y_h then so does any nonzero polynomial multiple of that product of matrices.

Proposition 3 below is discussed in [11].

Proposition 3. *Let δ be a quadratic irrational function with trace t and norm n both polynomials; that is, $\delta^2 - t\delta + n = 0$. Suppose x and y are polynomials so that the matrix*

$$M = \begin{pmatrix} x & -ny \\ y & x - ty \end{pmatrix}$$

has determinant $(x-\delta y)(x-\overline{\delta} y) = (-1)^r\kappa$, with κ a nonzero constant. Then M has a unique decomposition

$$M = \begin{pmatrix} a & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix}\cdots\begin{pmatrix} a_{r-1} & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} (a-t)/\kappa & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & \kappa \end{pmatrix},$$

where a, a_1, $\ldots$, a_{r-1} are polynomials of degree at least one satisfying $a_1 = \kappa a_{r-1}$, $a_2 = a_{r-2}/\kappa$, $a_3 = \kappa a_{r-3}$, $\ldots$. Hence, if r is even then $\kappa = 1$. Moreover

$$\delta = [\, a \,,\, \overline{a_1 \,,\, \ldots \,,\, a_{r-1} \,,\, (2a-t)/\kappa \,,\, \kappa a_1 \,,\, \ldots \,,\, a_{r-1}/\kappa \,,\, 2a-t} \,] \qquad (3)$$

provides the periodic continued fraction expansion of δ.

Of course, if $\kappa = 1$ then δ has period length r rather than $2r$.

Proposition 4. *If δ has quasi-period length r, but period length $2r$ — thus $\kappa \neq 1$ and r is odd — then $\mu\delta$ has period length r if and only if $\mu^2 = 1/\kappa$.*

Proof. Take δ as in (3). By Proposition 1 we see that

$$\mu\delta = [\, \mu a \,,\, \overline{a_1/\mu \,,\, \ldots \,,\, \mu a_{r-1} \,,\, (2a-t)/\mu\kappa \,,\, \kappa a_1/\mu \,,\, \ldots \,,\, a_{r-1}/\mu\kappa \,,\, \mu(2a-t)} \,],$$

so indeed $\mu = 1/\mu\kappa$ is of the essence.

4 Elliptic Curves with Torsion at Infinity

We recall Mazur's theorem limiting the possible rational torsion on a elliptic curve defined over $\mathbb{Q}$.

Proposition 5 (Mazur). *If E is an elliptic curve defined over $\mathbb{Q}$, then the torsion subgroup $E(\mathbb{Q})_{tors}$ of $E(\mathbb{Q})$ is isomorphic to either*

$$\begin{array}{lll} & \mathbb{Z}_m & \textit{for} \quad m = 1,\, 2,\, 3,\, \ldots,\, 10,\, 12 \\ \textit{or} & \mathbb{Z}_2 \times \mathbb{Z}_m & \textit{for} \quad m = 2,\, 4,\, 6,\, 8. \end{array}$$

Thus for each $m \in \{2, 3, \ldots, 10, 12\}$ we need all curves $\mathcal{C}_m : Y^2 = D_m(X)$ with D_m a polynomial of degree 4 and defined over $\mathbb{Q}$ and so that $\mathcal{C}_m$ has a torsion point of order m at infinity, equivalently — see page 393 — so that the continued fraction expansion of $\sqrt{D}$ is periodic with quasi-period length $m-1$. Naturally we lose no generality in normalising so that D_m is monic and has zero trace.

4.1 Tabulations

The first tabulation of rational elliptic curves with given torsion group[1] probably is given by Kubert [6]. Table 3 of [5], copied below, provides a congenial version of Kubert's table, listing in Tate normal form all elliptic curves

$$E : y^2 + (1-c)xy - by = x^3 - bx^2 \tag{4}$$

with $E(\mathbb{Q})_{\text{tors}} \cong \mathbb{Z}_m$ ($m = 4,\ 5,\ \dots,\ 10,\ 12$) and $E(\mathbb{Q})_{\text{tors}} \cong \mathbb{Z}_2 \times \mathbb{Z}_{2m}$ ($m = 2$, 3, 4); in each case the point $(0,0)$ is a torsion point of maximal order.

$E(\mathbb{Q})_{\text{tors}}$	b	c
$\mathbb{Z}_4$	t	0
$\mathbb{Z}_5$	t	t
$\mathbb{Z}_6$	$t(t+1)$	t
$\mathbb{Z}_7$	$t^2(t-1)$	$t(t-1)$
$\mathbb{Z}_8$	$(t-1)(2t-1)$	$\dfrac{(t-1)(2t-1)}{t}$
$\mathbb{Z}_9$	$t^2(t-1)(t^2-t+1)$	$t^2(t-1)$
$\mathbb{Z}_{10}$	$\dfrac{t^3(t-1)(2t-1)}{(t^2-3t+1)^2}$	$-\dfrac{t(t-1)(2t-1)}{t^2-3t+1}$
$\mathbb{Z}_{12}$	$\dfrac{t(2t-1)(2t^2-2t+1)(3t^2-3t+1)}{(t-1)^4}$	$-\dfrac{t(2t-1)(3t^2-3t+1)}{(t-1)^3}$
$\mathbb{Z}_2 \times \mathbb{Z}_4$	$\frac{1}{16}(4t-1)(4t+1)$	0
$\mathbb{Z}_2 \times \mathbb{Z}_6$	$-\dfrac{2(t-1)^2(t-5)}{(t^2-9)^2}$	$-\dfrac{2(t-5)}{t^2-9}$
$\mathbb{Z}_2 \times \mathbb{Z}_8$	$\dfrac{(2t+1)(8t^2+4t+1)}{(8t^2-1)^2}$	$\dfrac{(2t+1)(8t^2+4t+1)}{2t(4t+1)(8t^2-1)^2}$

One notices that the cases $\mathbb{Z}_2 \times \mathbb{Z}_{2m}$ are just special cases of torsion order $2m$; thus, in the sequel, we will not need the last three lines of the table.

With the change of variables $x = u^2x' + r$, $y = u^3y' + su^2x' + t$, where

$$\begin{cases} u = 1, & r = -(c^2 - 2c - 4b + 1)/12, \\ s = (c-1)/2, & t = -(c^3 - 3c^2 - (4b-3)c - (8b+1))/24, \end{cases}$$

we see that the elliptic curve (4) is isomorphic to

$$E:\ y^2 = x^3 - (c_4/48)x - c_6/864\,,$$

[1] At the time, the fact that [6] provided a complete list was of course only conjectural.

where c_4 and c_6 is standard notation for the invariants of the curve (4); see for example [4], or [3]. The point $(0,0)$ is transformed to

$$P = \big((c^2 - 2c - 4b + 1)/12,\ -b/2\big)$$

and, by isomorphism, is a torsion point P on E with maximal order.

The reader interested in constructing a table of rational torsion types such as the one above will find valuable instruction in the papers [8] and [9] of Nitaj.

4.2 Quartic Coverings of Elliptic Surfaces

Finally, the transformation (1) recommended by Adams and Razar [1], see page 392, provides a list of quartic covers

$$\mathcal{C}_m(s) : Y^2 = D_m(X,s) = \big(X^2 + u_m(s)\big)^2 + v_m(s)\big(X + w_m(s)\big) \qquad (5)$$

defined over $\mathbb{Q}(s)$ so that the divisor at infinity on the Jacobian of the curve $\mathcal{C}_m$ is torsion of order m (in brief, so that the point at infinity on the curve is torsion of order m). Here s (which replaces the t of the table for elegant variation) is a parameter ranging over $\mathbb{Q}$ omitting only several isolated values. One checks readily that the continued fraction expansion of Y begins $[\,X^2 + u\ ,\ 2(X - w)/v\ ,\ \ldots\,]$.

We use just brute force to notice that if $m = 2$ then the continued fraction expansion is

$$[\,X^2 + s\ ,\ \overline{2(X^2 + s)/t\ ,\ 2(X^2 + s)}\,]$$

and necessarily

$$\mathcal{C}_2(s,t) : Y^2 = D(X,s,t) = \big(X^2 + s\big)^2 + t, \qquad s \in \mathbb{Q},\ t \in \mathbb{Q} \setminus \{0\}. \qquad (6)$$

The special case $t = 1$ gives period length $r = 1$.

Similary, if $m = 3$ then the continued fraction expansion must be

$$[\,X^2 - s^2\ ,\ \overline{2(X + s)/t\ ,\ 2(X^2 - s^2)}\,]$$

and so

$$\mathcal{C}_3(s,t) : Y^2 = D(X,s,t) = \big(X^2 - s^2\big)^2 + t(X - s), \qquad s \in \mathbb{Q},\ t \in \mathbb{Q} \setminus \{0\}. \ (7)$$

In all other cases we obtain an elliptic surface $D_m(X,s)$ thus with just one rational parameter.

Here and below, we detail only the continued fraction expansions, seemingly breaking the cardinal rule that when dealing with quadratic irrationals one must mind one's P's and Q's. That is, the critical information is contained in the complete quotients $(Y_m + P_h)/Q_h$, rather than in the partial quotients a_h. However, here we lose no information to speak of. The reader can readily confirm that a partial quotient $2(X - c_h)/b_h$ entails that $Q_h = b_h(X + c_h)$, and if $P_h = X^2 + u_m + 2e_h$ then $e_{h+1} = -(e_h + u_m + c_h^2)$. Of course, the partial quotient $2(X^2 + u_m)/k_m$ implies $Q = k_m$ and $e = u_m$. We take $P_0 = 0$ and $Q_0 = 1$ but, in decency, we ought to be expanding $Y_m + (X^2 + u_m)$, thus with $P_0 = X^2 + u_m$. Note that, in any case, $P_1 = X^2 + u_m$, that is, $e_1 = 0$.

4.3 Periods of Even Length

We summarised the case $m = 3$ at (7) on page 397. The case $m = 5$ is

$$\begin{aligned}\mathcal{C}_5(t) : Y_5^2(X,t) &= D(X,t)\\ &= \bigl(X^2 - \tfrac14(t^2-6t+1)\bigr)^2 + 4t\bigl(X - \tfrac12(t-1)\bigr)\,, \quad t \in \mathbb{Q}\setminus\{0\},\end{aligned} \tag{8}$$

with continued fraction expansion

$$\begin{aligned} Y_5(s) = [\,X^2 - \tfrac14(t^2-6t+1)\,,\ &\overline{\bigl(X+\tfrac12(t-1)\bigr)/2t\,,}\\ &\overline{2\bigl(X-\tfrac12(t+1)\bigr)\,,\ \bigl(X+\tfrac12(t-1)\bigr)/2t\,,\ 2\bigl(X^2-\tfrac14(t^2-6t+1)\bigr)}\,].\end{aligned}$$

Just so, $\mathcal{C}_7(t)$ is defined by

$$u_7(t) = -\tfrac14(t^4-6t^3+3t^2+2t+1), \quad v_7(t) = 4t^2(t-1), \quad w_7(t) = -\tfrac12(t^2-t-1),$$

and $Y_7(X,t)$ has continued fraction expansion,

$$\begin{aligned}&[\,X^2+u_7(t)\,,\ \overline{\tfrac12\bigl(X+\tfrac12(t^2-t-1)\bigr)/t^2(t-1)\,,\ 2\bigl(X-\tfrac12(t^2-t+1)\bigr)\,,}\\ &\qquad\qquad \overline{\tfrac12\bigl(X+\tfrac12(t^2-3t+1)\bigr)/t(t-1)\,,}\\ &\overline{2\bigl(X-\tfrac12(t^2-t+1)\bigr)\,,\ \tfrac12\bigl(X+\tfrac12(t^2-t-1)\bigr)/t^2(t-1)\,,\ 2\bigl(X^2+u_7(t)\bigr)}\,].\end{aligned} \tag{9}$$

Finally, for this is the last case with m odd, for $m = 9$ we have

$$\begin{aligned} u_9(t) &= -\tfrac14(t^6-6t^5+9t^4-10t^3+6t^2+1),\\ &\qquad v_9(t) = 4t^2(t-1)(t^2-t+1), \qquad w_9(t) = -\tfrac12(t^3-t^2-1),\end{aligned} \tag{10}$$

with continued fraction expansion

$$\begin{aligned}&[\,X^2+u_9(t)\,,\ \overline{\tfrac12\bigl(X+\tfrac12(t^3-t^2-1)\bigr)/t^2(t-1)(t^2-t-1)\,,}\\ &\quad\overline{2\bigl(X-\tfrac12(t^3-t^2+1)\bigr)\,,\ \tfrac12\bigl(X-\tfrac12(t^3-3t^2+2t-1)\bigr)/t^2(t-1)\,,}\\ &\qquad\qquad \overline{2t\bigl(X-\tfrac12(t^3-3t^2+4t-1)\bigr)/(t^2-t+1)\,,}\\ &\quad\overline{\tfrac12\bigl(X-\tfrac12(t^3-3t^2+2t-1)\bigr)/t^2(t-1)\,,\ 2\bigl(X-\tfrac12(t^3-t^2+1)\bigr)\,,}\\ &\qquad\quad \overline{\tfrac12\bigl(X+\tfrac12(t^3-t^2-1)\bigr)/t^2(t-1)(t^2-t-1)\,,\ 2\bigl(X^2+u_9(t)\bigr)}\,].\end{aligned}$$

4.4 Periods of Odd Length

We have dealt with the case $m = 2$ at page 397. When $m = 4$ we find that

$$\mathcal{C}_4(t) : Y_4(X,t)^2 = D(X) = \bigl(X^2+\tfrac14(4t-1)\bigr)^2 + 4t(X+\tfrac12)\,, \quad t\in\mathbb{Q}\setminus\{0\}, \tag{11}$$

and

$$\begin{aligned} Y_4(X,t) = [\,X^2+\tfrac14(4t-1)\,,\ &\overline{2(X-\tfrac12)/4t\,,\ 2(X-\tfrac12)\,,\ 2\bigl(X^2+\tfrac14(4t-1)\bigr)/4t\,,}\\ &\qquad\qquad \overline{2(X-\tfrac12)\,,\ 2(X-\tfrac12)/4t\,,\ 2\bigl(X^2+\tfrac14(4t-1)\bigr)}\,].\end{aligned}$$

Thus $\kappa_4(t)=4t$. This entails that $Y_4(X,\frac{1}{4}s^2)/s$ has the periodic continued fraction expansion of period length $r=3$:

$$[\,\big(X^2+\tfrac{1}{4}(s^2-1)\big)/s\;,\;\overline{2(X-\tfrac{1}{2})/s\;,\;2s(X-\tfrac{1}{2})\;,\;2\big(X^2+\tfrac{1}{4}(s^2-1)\big)/s}\,].$$

For $m=6$, and $t\in\mathbb{Q}\setminus\{0,-1\}$, the surface $\mathcal{C}_6(t)$ is given by

$$u_6(t)=\tfrac{1}{4}(3t^2+6t-1),\quad v_6(t)=4t(t+1),\qquad w_6(t)=-\tfrac{1}{2}(t-1). \tag{12}$$

and its continued fraction is detailed by

$$\begin{aligned}[\,X^2+\tfrac{1}{4}(3t^2+6t-1)\;,\;\big(X+\tfrac{1}{2}(t-1)\big)/2t(t+1)\;,\;2\big(X-\tfrac{1}{2}(t+1)\big)\;,&\\ \big(X-\tfrac{1}{2}(t+1)\big)/2t\;,\;2\big(X+\tfrac{1}{2}(t-1)\big)/(t+1)\;,&\\ 2\big(X^2+\tfrac{1}{4}(3t^2+6t-1)\big)/4t\;,\;\ldots\,].&\end{aligned}$$

Thus $\kappa_6(t)=4t$. It follows that $Y_6(X,s^2)/2s$ has the periodic continued fraction expansion of period length $r=5$:

$$\begin{aligned}[\,\big(X^2+\tfrac{1}{4}(3s^4+6s^2-1)\big)/2s\;,\;\overline{\big(X+\tfrac{1}{2}(s^2-1)\big)/s(s^2+1)}\;,&\\ \overline{\big(X-\tfrac{1}{2}(s^2+1)\big)/s\;,\;\big(X-\tfrac{1}{2}(s^2+1)\big)/s\;,}&\\ \overline{\big(X+\tfrac{1}{2}(s^2-1)\big)/s(s^2+1)\;,\;2\big(X^2+\tfrac{1}{4}(3t^2+6t-1)\big)/2s}\,].&\end{aligned}$$

Finally, because this provides the last of the cases with odd period length, the elliptic surface $\mathcal{C}_8(t):Y_8^2(X,t)=D_8(X,t)$ is defined by

$$\begin{aligned}u_8(t)=(4t^4+4t^3-16t^2+8t-1)/4t^2,&\\ v_8(t)=4(t-1)(2t-1),\qquad w_8(t)=-(2t^2-4t+1)/2t,&\end{aligned}\tag{13}$$

and, if $t\in\mathbb{Q}\setminus\{0,\frac{1}{2},1\}$, then $Y_8(X,t)$ has the continued fraction expansion

$$\begin{aligned}[\,X^2+u_8(t)\;,\;\tfrac{1}{2}\big(X+(2t^2-4t+1)/2t\big)/(t-1)(2t-1)\;,&\\ 2\big(X-(2t^2-4t+1)/2t\big)\;,\;\tfrac{1}{2}t\big(X-(2t-1)/2t\big)/(t-1)(2t-1)\;,&\\ 2(2t-1)\big(X-(2t-1)/2t\big)/t^2\;,\;\tfrac{1}{2}t^3\big(X-(2t^2-4t+1)/2t\big)/(t-1)(2t-1)^2\;,&\\ 2(2t-1)\big(X+(2t^2-4t+1)/2t\big)/t^3\;,\;\tfrac{1}{2}t^3\big(X^2+u_8(t)\big)/(t-1)(2t-1)^2\;,\;\ldots].&\end{aligned}$$

Thus $\kappa_8(t)=4(t-1)(2t-1)^2/t^3$. It follows that $Y_8\big(X,1/(1-s^2)\big)/2s(1+s^2)$ has a continued fraction expansion with period $r=7$ for $s\in\mathbb{Q}\setminus\{0,\pm1\}$. For example

$$\begin{aligned}\tfrac{1}{20}Y_8(\tfrac{1}{6}X,-\tfrac{1}{3})&=\tfrac{1}{720}\sqrt{X^4-898X^2+1920X+245761}\\ &=[\,(X^2-449)/720\;,\;\overline{3(X-23)/4\;,\;(X+17)/60\;,\;-(X-15)/4\;,}\\ &\qquad\overline{-(X-15)/4\;,\;(X+17)/60\;,\;3(X-23)/4\;,\;2(X^2-449)/720}\,].\end{aligned}$$

Theorem. *There are no rational quartic polynomials $Y^2 = D(X)$ so that the continued fraction expansion of Y has period length nine, or eleven.*

Proof. For $t \in \mathbb{Q} \setminus \{0, \frac{1}{2}, 1\}$, the elliptic surface $\mathcal{C}_{10}(t)$ is given by

$$\begin{aligned} u_{10}(t) &= -\frac{4t^6 - 16t^5 + 8t^4 + 8t^3 - 4t + 1}{4(t^2 - 3t + 1)^2}, \\ v_{10}(t) &= \frac{4t^3(t-1)(2t-1)}{(t^2-3t+1)^2}, \qquad w_{10}(t) = \frac{2t^3 - 2t^2 - 2t + 1}{2(t^2 - 3t + 1)}, \end{aligned} \tag{14}$$

with $\kappa_{10}(t) = -4t(t-1)(t^2-3t+1)$;

The continued fraction expansion of $Y_{10}(X,t)$ is

$$\begin{aligned} Y = \Big[X^2 - \tfrac{4t^6-16t^5+8t^4+8t^3-4t+1}{4(t^2-3t+1)^2}\,,\ \tfrac{(t^2-3t+1)^2}{2t^3(t-1)(2t-1)}\big(X - \tfrac{2t^3-2t^2-2t+1}{2(t^2-3t+1)}\big)\,, \\ 2\big(X + \tfrac{2t^3-4t^2+4t-1}{2(t^2-3t+1)}\big)\,, -\tfrac{t^2-3t+1}{2t(t-1)(2t-1)}\big(X - \tfrac{2t^3-6t^2+4t-1}{2(t^2-3t+1)}\big)\,, \\ -\tfrac{2(t^2-3t+1)}{t}\big(X + \tfrac{2t-1}{2(t^2-3t+1)}\big)\,,\ \tfrac{1}{2t^2(t-1)}\big(X + \tfrac{2t-1}{2(t^2-3t+1)}\big)\,, \\ \tfrac{2(t^2-3t+1)^2}{2t-1}\big(X - \tfrac{2t^3-6t^2+4t-1}{2(t^2-3t+1)}\big)\,, -\tfrac{1}{2t(t-1)(t^2-3t+1)}\big(X + \tfrac{2t^3-4t^2+4t-1}{2(t^2-3t+1)}\big)\,, \\ -\tfrac{2(t^2-3t+1)^3}{t^2(2t-1)}\big(X - \tfrac{2t^3-2t^2-2t+1}{2(t^2-3t+1)}\big)\,, \\ -\tfrac{1}{2t(t-1)(t^2-3t+1)}\big(X^2 - \tfrac{4t^6-16t^5+8t^4+8t^3-4t+1}{4(t^2-3t+1)^2}\big)\,, \ldots\Big]. \end{aligned}$$

It follows from Proposition 4 that there is such an expansion with period length nine if and only if the equation $w^2 = \kappa_{10}(t)$ has a nontrivial solution in rationals t and w, that is, with $w \neq 0$. But there is no such solution.

We transform the equation by $t \mapsto 1/(t-1)$, $w \mapsto w/(t-1)^2$, yielding $w^2 = t^3 - 7t^2 + 15t - 10$, and note that the global minimal model of the cubic curve is $y^2 = x^3 - x^2 - x$. That is curve 80B2(A) in Cremona's tables [4], and we there read that the curve has rank 0 and its only rational point is the torsion point $(0,0)$ of order 2. Hence there is no t such that $\kappa(t)$ is a square, except $t = 0$, 1, but those values give singular curves.

Just so, for $t \in \mathbb{Q} \setminus \{0, \frac{1}{2}, 1\}$, the elliptic surface $\mathcal{C}_{12}(t)$ is given by

$$\begin{aligned} u_{12}(t) &= \frac{12t^8 - 120t^7 + 336t^6 - 468t^5 + 372t^4 - 168t^3 + 36t^2 - 1}{4(t-1)^6}, \\ v_{12}(t) &= \frac{4t(2t-1)(2t^2-2t+1)(3t^2-3t+1)}{(t-1)^4}, \\ w_{12}(t) &= \frac{6t^4 - 8t^3 + 2t^2 + 2t - 1}{2(t-1)^3}, \end{aligned} \tag{15}$$

with $\kappa_{12}(t) = 4t(2t-1)^2(3t^2-3t+1)^3/(t-1)^{11}$; and

$$\begin{aligned}Y_{12}(X,t) = [\,X^2 + u_{12}(t)\,,\ &\tfrac{(t-1)^4}{2t(2t-1)(2t^2-2t+1)(3t^2-3t+1)}\big(X - \tfrac{6t^4-8t^3+2t^2+2t-1}{2(t-1)^3}\big)\,,\\
2\big(X + \tfrac{6t^4-10t^3+8t^2-4t+1}{2(t-1)^3}\big)\,,\ &-\tfrac{(t-1)^3}{2t(2t-1)(3t^2-3t+1)}\big(X - \tfrac{2t^4+2t^3-6t^2+4t-1}{2(t-1)^3}\big)\,,\\
&-\tfrac{2(3t^2-3t+1)}{(t-1)(2t^2-2t+1)}\big(X + \tfrac{2t^4-4t^3+6t^2-4t+1}{2(t-1)^3}\big)\,,\\
&\tfrac{(t-1)^6}{2t(2t-1)(3t^2-3t+1)^2}\big(X - \tfrac{(2t-1)(2t^2-2t+1)}{2(t-1)^3}\big)\,,\\
&\tfrac{2(2t-1)(3t^2-3t+1)}{(t-1)^5}\big(X - \tfrac{(2t-1)(2t^2-2t+1)}{2(t-1)^3}\big)\,,\\
&-\tfrac{(t-1)^{10}}{2t(2t-1)^2(2t^2-2t+1)(3t^2-3t+1)^2}\big(X + \tfrac{2t^4-4t^3+6t^2-4t+1}{2(t-1)^3}\big)\,,\\
&-\tfrac{2(2t-1)(3t^2-3t+1)^2}{(t-1)^8}\big(X - \tfrac{2t^4+2t^3-6t^2+4t-1}{2(t-1)^3}\big)\,,\\
&\tfrac{2(t-1)^{11}}{2t(2t-1)^2(3t^2-3t+1)^3}\big(X + \tfrac{6t^4-10t^3+8t^2-4t+1}{2(t-1)^3}\big)\,,\\
&\tfrac{2(2t-1)(3t^2-3t+1)^2}{(t-1)^7(2t^2-2t+1)}\big(X - \tfrac{6t^4-8t^3+2t^2+2t-1}{2(t-1)^3}\big)\,,\\
&\tfrac{(t-1)^{11}}{2t(2t-1)^2(3t^2-3t+1)^3}\big(X^2 + u_{12}(t)\big)\,,\ \ldots].\end{aligned}$$

Much as before, when $m = 12$ we consider $t(t-1)(3t^2-3t+1) = w^2$ with $w \in \mathbb{Q}$, which expands to $w^2 = 3t^4-6t^3+4t^2-t$. This quartic has a rational point $(1,0)$. The transformation $t \mapsto -1/(t-1)$ and $w \mapsto w/(t-1)^2$ transforms the equation to $w^2 = t^3+7t^2+17t+14$. Its global minimal model is $y^2 = x^3+x^2+x$, which is curve 48A4(A) of Cremona's tables [4]. That curve has rank 0 and its only rational point is the torsion point $(0,0)$ of order 2. Hence there is no t such that $\kappa_{12}(t)$ is a square, except if $t = 0$, 1, which give singular curves. □

5 Pseudo-elliptic Integrals

Listing the fundamental unit in each of the function fields $\mathbb{Q}\big(Y_m(X,t)\big)$ is mere teratology[4] (teratology: the science or study of monstrosities ...), so we provide only examples. Note that to compute a unit $x(X)+y(X)Y$ one either computes the relevant convergent $x_{m-2}(X)/y_{m-2}(X)$ of the cited expansions or, more elegantly, one recalls that the unit is the product of the complete quotients $(Y+P_h)/Q_h$ for $h = 1, \ldots\ m-1$.

The following is a list of example pseudo-elliptic integrals, see [11],

$$\int \frac{f_m(z)\,dz}{\sqrt{D_m(z)}} = \log\big(x_m(z) + y_m(z)\sqrt{D_m(z)}\,\big).$$

In each case the reader might verify that indeed $x' = fy$ and x^2-Dy^2 is constant.

[4] The truly interested reader will learn more from computing them for herself than from studying a list — in any case, the length and complexity of such a list would have forced me to exceed my page limit.

$$\begin{aligned}x_{12}(z) &= z^{12} - 118z^{11} + 16028z^{10} - 1069154z^{9} + 72544053z^{8} - 2910120156z^{7}\\ &+115293384192z^{6} - 2435904763524z^{5} + 49959577428123z^{4} - 3156443198606z^{3}\\ &-6523744685908252z^{2} + 264671040329753798z - 1519185098148240209;\\ y_{12}(z) &= z^{10} - 118z^{9} + 14517z^{8} - 944616z^{7} + 57651426z^{6} - 2264475780z^{5}\\ &+79914037266z^{4} - 1800781684584z^{3} + 34360879041117z^{2}\\ &-338671088037302z + 2242974918048761;\end{aligned}$$

$$f_{12}(z) = 12z + 118; \quad D_{12}(z) = (z^2 + 1511)^2 + 107520(z - 13).$$

$$\begin{aligned}x_{10}(z) &= z^{10} - 125z^{8} - 1600z^{7} + 7450z^{6} + 128000z^{5} + 457750z^{4}\\ &-4504000z^{3} - 22308875z^{2} + 274924375;\\ y_{10}(z) &= z^{8} - 100z^{6} - 1120z^{5} + 4470z^{4} + 64000z^{3}\\ &+183100z^{2} - 1351200z - 4461775;\end{aligned}$$

$$f_{10}(z) = 10z; \quad D_{10}(z) = (z^2 - 25)^2 - 960(z - 1).$$

$$\begin{aligned}x_{9}(z) &= z^{9} - 9z^{8} - 108z^{7} + 828z^{6} + 5454z^{5} - 29646z^{4} - 131868z^{3}\\ &+467532z^{2} + 1190457z - 3028401;\\ y_{9}(z) &= z^{7} - 9z^{6} - 75z^{5} + 627z^{4} + 2403z^{3} - 15579z^{2} - 28377z + 132273;\end{aligned}$$

$$f_9(z) = 9(z + 1); \quad D_9(z) = (z^2 - 33)^2 - 192(z + 3).$$

$$\begin{aligned}x_{8}(z) &= z^{8} - 10z^{7} - 50z^{6} + 1006z^{5} - 976z^{4} - 34526z^{3}\\ &+108946z^{2} + 413690z - 1829009;\\ y_{8}(z) &= z^{6} - 10z^{5} - 25z^{4} + 660z^{3} - 1313z^{2} - 11306z + 41369;\end{aligned}$$

$$f_8(z) = 8z + 10; \quad D_8(z) = (z^2 - 25)^2 + 192(z + 7).$$

$$\begin{aligned}x_{7}(z) &= z^{7} + z^{6} - 31z^{5} - 103z^{4} + 331z^{3} + 1435z^{2} - 429z - 5557;\\ y_{7}(z) &= z^{5} + z^{4} - 22z^{3} - 62z^{2} + 133z + 429;\end{aligned}$$

$$f_7(z) = 7z - 1; \quad D_7(z) = (z^2 - 9)^2 - 64(z - 1).$$

$$\begin{aligned}x_{6}(z) &= z^{6} - 2z^{5} + 8z^{4} - 4z^{3} + 8z^{2} + 8z;\\ y_{6}(z) &= z^{4} - 2z^{3} + 6z^{2} - 4z + 4;\end{aligned}$$

$$f_6(z) = 6z + 2.; \quad D_6(z) = (z^2 + 2)^2 + 8z.$$

$$x_5(z) = z^5 - z^4 + 3z^3 + z^2 + 2; \quad y_5(z) = z^3 - z^2 + 2z;$$

$$f_5(z) = 5z + 1; \quad D_5(z) = (z^2 + 1)^2 + 4z.$$

$$x_4(z) = z^4 - 2z^3 + 2z^2 + 4z - 4; \quad y_4(z) = z^2 - 2z + 2;$$

$$f_4(z) = 4z + 2; \quad D_4(z) = z^4 + 8(z + 1).$$

$$\int \frac{(3z-s)\,dz}{\sqrt{(z^2-s^2)^2+t(z-s)}}$$
$$= \log\bigl(1+2(z+s)(z^2-s^2)/t+2\bigl((z+s)/t\bigr)\sqrt{(z^2-s^2)^2+t(z-s)}\,\bigr).$$

$$\int \frac{2z\,dz}{\sqrt{(z^2+s)^2+t}} = \log\bigl(z^2+s+\sqrt{(z^2+s)^2+t}\,\bigr).$$

One readily recognises the final, $m=2$, example as an elementary integral by setting $w=z^2+s$. That might make one wonder whether there are rational transformations that nakedly reveal the elementary nature of the integrals in each case. The answer is, of course, yes; a helpful reference is [14], pp38*ff*.

6 Remark

The attentive reader will have noticed an unexpected feature of the detailed continued fraction expansions for m at least 4. In each case the third partial quotient, a_2 , is of the shape $2(X-c)$, moreover with $c=-w+1$. Of course, such an observation may well be no more than an artefact of Kubert's parametrisations on which ours are based. Indeed, the curves on page 396 depend on just two[5] parameters, there called b and c, so our three parameters cannot be independent. Specifically, they happen all to satisfy the identity $4(u_m+w_m^2)=v_m$. Although u and w^2 do have the same weight, that weight is different from the weight of v, so that coefficient 2 *is* artificial. A normalisation ($xX \mapsto X$, $x^2Y \mapsto Y$, so $u'=u/x^2$, $v'=v/x^3$, $w'=w/x$) changes the identity to $4(u'^2+w')=xv'$, and the 2 to $2/x$.

References

1. William W. Adams and Michael J. Razar, 'Multiples of points on elliptic curves and continued fractions', *Proc. London Math. Soc.* **41** (1980), 481–498.
2. T. G. Berry, 'On periodicity of continued fractions in hyperelliptic function fields', *Arch. Math.* **55** (1990), 259–266.
3. Henri Cohen, *A Course in Computational Algebraic Number Theory*, Graduate Texts in Mathematics **138** (New York: Springer–Verlag, 1993).
4. J. E. Cremona, *Algorithms for Modular Elliptic Curves*, 2nd edition, Cambridge University Press, 1997.
5. Everett W. Howe, Franck Leprévost, and Bjorn Poonen, 'Large torsion subgroups of split Jacobians of curves of genus two or three', *Forum Math.* **12**.3 (2000), 315–364 (MR2001e:11071).
6. Daniel Sion Kubert, 'Universal bounds on the torsion of elliptic curves', *Proc. London Math. Soc.* **33**.3 (1976), 193–237.
7. B. Mazur, 'Modular curves and the Eisenstein ideal', *Inst. Hautes Études Sci. Publ. Math.* **47** (1977), 33–186.

[5] *Of course* an elliptic curve depends on just two parameters, say the *two* Eisenstein series G_4 and G_6 .

8. Abderrahmane Nitaj, ‘Détermination de courbes elliptiques pour la conjecture de Szpiro’, *Acta Arith.* **85**.4 (1998), 351–376.
9. Abderrahmane Nitaj, ‘Isogènes des courbes elliptiques définies sur les rationnels’, to appear in *J. Combinatorial Math.*
10. Oskar Perron, *Die Lehre von den Kettenbrüchen*, 2nd edition, 1929 (Chelsea Publishing Company, New York, N Y).
11. Alfred J. van der Poorten and Xuan Chuong Tran, ‘Quasi-elliptic integrals and periodic continued fractions’, *Monatshefte Math.*, 131 (2000), 155-169.
12. Alfred J. van der Poorten, ‘Non-periodic continued fractions in hyperelliptic function fields’, (Dedicated to George Szekeres on his 90th birthday), *Bull. Austral. Math. Soc.* **64** (2001), 331–343.
13. A. J. van der Poorten and H. C. Williams, ‘On certain continued fraction expansions of fixed period length’, *Acta Arith.* **89**.1 (1999), 23–35 (MR2000m:11010).
14. Viktor Prasolov and Yuri Solovyev, *Elliptic functions and elliptic integrals*, translated from the Russian manuscript by D. Leites, Translations of Mathematical Monographs, 170. American Mathematical Society, Providence, RI, 1997; x+185 pp.
15. A. Schinzel, ‘On some problems of the arithmetical theory of continued fractions’, *Acta Arith.* **6** (1961), 393–413; and *ibid.* **7** (1962), 287–298.
16. Wolfgang M. Schmidt, ’On continued fractions and diophantine approximation in power series fields’, *Acta Arith.* **95** (2000), 139–166.
17. Jing Yu, ‘Arithmetic of hyperelliptic curves’, manuscript marked *Aspects of Mathematics*, Hong Kong University, 1999; see pp4–6.

Fixed Points and Two-Cycles of the Discrete Logarithm

Joshua Holden

Department of Mathematics, Rose-Hulman Institute of Technology, Terre Haute, IN, 47803-3999, USA, holden@rose-hulman.edu

Abstract. We explore some questions related to one of Brizolis: does every prime p have a pair (g, h) such that h is a fixed point for the discrete logarithm with base g? We extend this question to ask about not only fixed points but also two-cycles. Campbell and Pomerance have not only answered the fixed point question for sufficiently large p but have also rigorously estimated the number of such pairs given certain conditions on g and h. We attempt to give heuristics for similar estimates given other conditions on g and h and also in the case of two-cycles. These heuristics are well-supported by the data we have collected, and seem suitable for conversion into rigorous estimates in the future.

1 Introduction, Previous Work, and Data on Fixed Points

In [4], paragraph F9 includes the following problem, due to Brizolis: given a prime $p > 3$, is there always a pair (g, h) such that g is a primitive root of p, $1 \leq h \leq p - 1$, and

$$g^h \equiv h \mod p \ ? \tag{1}$$

In other words, is there always a primitive root g such that the discrete logarithm $\log_g$ has a fixed point? It has been proved that the number $N(p)$ of such pairs is greater than $\phi(p-1)^2/(p-1) + O(p^{1/2+\epsilon})$, thereby showing that the answer to Brizolis' question is yes at least for sufficiently large p. This result seems to have been first proved by Zhang in [7] and later, independently, by Cobeli and Zaharescu in [2]. Campbell and Pomerance ([6]) have again rediscovered the result and made the value of "sufficiently large" small enough that they expect to be able to use a direct search to finish the problem.

This paper attempts to start a similar project for the two-cycles of $\log_g$, that is the pairs (g, h) such that there is some a between 1 and $p - 1$ such that

$$g^h \equiv a \mod p \quad \text{and} \quad g^a \equiv h \mod p \ . \tag{2}$$

Using the work of Campbell and Pomerance as a starting point we give heuristics for estimating the number of such pairs with and without various side conditions, and provide computational evidence to support them. We expect that the methods used by Campbell and Pomerance would also be useful in turning these heuristics into asymptotic theorems.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 405–415, 2002.

The first observation that Campbell and Pomerance make is that if h is a primitive root modulo p which is also relatively prime to $p-1$, then there is a unique primitive root g satisfying (1), namely $g = h^{\overline{h}}$ reduced modulo p, where $\overline{h}$ denotes the inverse of h modulo $p-1$ throughout this paper. (Note that if h is relatively prime to $p-1$ then h is a primitive root if and only if g is. Likewise, g and h have the same order modulo p if and only if h is relatively prime to $p-1$.) Their technique for estimating $N(p)$ is thus to count the number of such h. One possible underlying heuristic for this is to observe that there are $\phi(p-1)$ possibilities for h which are relatively prime to $p-1$, and we would expect each of them to be a primitive root with probability $\phi(p-1)/(p-1)$. (There are $\phi(p-1)$ primitive roots for p among the numbers between 1 and $p-1$.) This by itself gives a very accurate estimate of the number of solutions to (1) with g a primitive root and h relatively prime to $p-1$, as is shown for some sample p in Table 1. (See Section 3 for details on how the tables were computed.)

Table 1. Solutions to (1) with g PR, h RPPR

p	predicted	observed
10007	2500.5	2539
10009	1096.1	1103
10037	2115.7	2111
10039	812.6	781
10061	1603.2	1605

Campbell and Pomerance also observe that the solutions to (1) with g a primitive root and h relatively prime to $p-1$ make up a positive proportion of the solutions with g a primitive root but no restrictions on h. To obtain a heuristic for this problem which may prove useful we look at a simpler version of Brizolis' problem where g is not necessarily a primitive root. To reduce the amount of excess verbiage, in the rest of this paper we will refer to an integer which is a primitive root modulo p as PR and an integer which is relatively prime to $p-1$ as RP. An integer which is both will be referred to as RPPR and one which has no restrictions will be referred to as ANY. All integers will be taken to be between 1 and $p-1$, inclusive, unless stated otherwise. If $N(p)$ is, as above, the number of solutions to (1) such that g is a primitive root and h is a primitive root which is relatively prime to $p-1$ then we will say $N(p) = N_{(1),g\,\mathsf{PR},h\,\mathsf{RPPR}}(p)$.

With this notation, we now look at $N_{(1),g\,\mathsf{ANY},h\,\mathsf{RP}}(p)$ and $N_{(1),g\,\mathsf{ANY},h\,\mathsf{ANY}}(p)$. In the first case, h has an inverse modulo $p-1$, so as before there is a unique g for each h such that (g,h) satisfies (1). Thus $N_{(1),g\,\mathsf{ANY},h\,\mathsf{RP}}(p) = \phi(p-1)$ with no error term.

On the other hand if h is ANY then there are two possibilities. Let $d = \gcd(h, p-1)$. If h is a d-th power residue modulo p then there are d solutions g to (1), since d divides $p-1$. If h is not a a d-th power residue then there are no solutions to (1). The number of d-th power residues modulo p is $(p-1)/d$, so the chance that h is a residue is $1/d$. Thus we expect on the average 1 pair (g,h)

for every h, or $p-1$ pairs in all, giving us $N_{(1),g\,\mathsf{ANY},h\,\mathsf{ANY}}(p) \approx p-1$. Table 2 gives evidence that this is correct.

Table 2. Solutions to (1) with g ANY, h ANY

p	predicted	observed
10007	10006	10082
10009	10008	9820
10037	10036	10249
10039	10038	10058
10061	10060	9923

Now suppose g is PR, h is ANY. The analysis is the same as in the previous case, except that now each solution g has an estimated chance of $\phi(p-1)/(p-1)$ of being a primitive root modulo p. Thus $N_{(1),g\,\mathsf{PR},h\,\mathsf{ANY}}(p) \approx \phi(p-1)$, as suggested by Table 3.

Table 3. Solutions to (1) with g PR, h ANY

p	predicted	observed
10007	5002	5079
10009	3312	3295
10037	4608	4643
10039	2856	2812
10061	4016	3987

We have not yet mentioned all of the (sixteen) possible combinations of conditions on g and h. By observations made above,

$$N_{(1),g\,\mathsf{PR},h\,\mathsf{RPPR}}(p) = N_{(1),g\,\mathsf{PR},h\,\mathsf{RP}}(p) = N_{(1),g\,\mathsf{PR},h\,\mathsf{PR}}(p) = N_{(1),g\,\mathsf{ANY},h\,\mathsf{RPPR}}(p).$$

We have not yet collected data for $N_{(1),g\,\mathsf{ANY},h\,\mathsf{PR}}(p)$ but there is every reason to believe that it is approximately $\phi(p-1)/(p-1)N_{(1),g\,\mathsf{ANY},h\,\mathsf{ANY}}(p) \approx \phi(p-1)$ since the extra condition on h is independent in our heuristics. Likewise in the cases where g is RP or RPPR we would expect the values to be approximately $\phi(p-1)/(p-1)$ times the corresponding values where g is ANY or PR. (The case where g is RPPR is also mentioned in [4].)

In summary, we have the following:

Theorem 1 (Zhang, independently by others).

$$N_{(1),g\,\mathsf{PR},h\,\mathsf{RPPR}}(p) \approx \phi(p-1)^2/(p-1)$$

Conjecture 1.

(a) $N_{(1),g\,\mathsf{ANY},h\,\mathsf{ANY}}(p) \approx p-1$

(b) $N_{(1),g\,\mathsf{PR},h\,\mathsf{ANY}}(p) \approx \phi(p-1)$
(c) $N_{(1),g\,\mathsf{ANY},h\,\mathsf{PR}}(p) \approx \phi(p-1)$
(d) $N_{(1),g\,\mathsf{RP},h\bullet}(p) \approx \phi(p-1)/(p-1)N_{(1),g\,\mathsf{ANY},h\bullet}(p)$
(e) $N_{(1),g\,\mathsf{RPPR},h\bullet}(p) \approx \phi(p-1)/(p-1)N_{(1),g\,\mathsf{PR},h\bullet}(p)$

2 Two-Cycles: Heuristics

Attacking (2) directly requires the simultaneous solution of two modular equations, presenting both computational and theoretical difficulties. In the fixed point case we started with the situations where h was RP and we could solve (1) immediately. Similarly, in the two-cycle case we will use similar conditions to reduce the solution of two equations to the solution of one equation.

(As an aside, it should be noted that (2) is already in some sense only one equation, as a is in fact explicitly defined. Thus we could write (2) in the form

$$g^{g^h \bmod p} \equiv h \mod p \ .$$

However, this has the serious drawback of an unnatural reduction modulo p in the exponent. There does not seems to be any added insight gained from writing the equation this way which would make up for this problem.)

Consider the modular equation

$$h^h \equiv a^a \mod p \tag{3}$$

Given g, h, and a as in (2), then (3) is clearly satisfied and the common value is g^{ah} modulo p. Conditions on g and h in (2) can (sometimes) be translated into conditions on h and a in (3). On the other hand, given a pair (h, a) which satisfies (3), we can attempt to solve for g such that (g, h) satisfies (2) and translate conditions on (h, a) into conditions on (g, h). We will start with the situations where the equivalence is relatively straightforward.

If h is RP and a is ANY in (3) then we can let $g \equiv a^{\overline{h}}$ modulo p; then it is straightforward to show that we have a two-cycle with h RP and no particular condition on g. (In fact given h there is a one-to-one correspondence between instances of g which are ANY and instances of a which are ANY.) Conversely, given a two-cycle with h RP and g ANY, we have (3) with h RP and a ANY. Thus $N_{(2),g\,\mathsf{ANY},h\,\mathsf{RP}}(p) = N_{(3),h\,\mathsf{RP},a\,\mathsf{ANY}}(p)$. Computationally, the second of these is much easier to compute; instead of looping through both g and h we only need to loop through a and record the value of each a^a modulo p and whether a was RP.

For a heuristic estimate of $N_{(2),g\,\mathsf{ANY},h\,\mathsf{RP}}(p) = N_{(3),h\,\mathsf{RP},a\,\mathsf{ANY}}(p)$, it turns to be useful to make a distinction between two-cycles which are fixed points and "proper" two-cycles. The former correspond to the trivial solutions $h = a$ of (3). (Indeed, we saw already in the case of fixed points that we should set $g \equiv h^{\overline{h}} = a^{\overline{h}}$.) We estimated that there are approximately $\phi(p-1)$ fixed points in this case. The proper two-cycles correspond to pairs (h, a) with $h \neq a$; the values of h^h and a^a modulo p are distributed according to no obvious pattern, so given h we suppose a chance of $1/(p-1)$ that $h^h \equiv a^a$. There are $\phi(p-1)$ values of h which are RP and $p-2$ values of $a \neq h$ for an expected number of nontrivial pairs equal to $(p-1)\phi(p-1)/(p-2) \approx \phi(p-1)$. (We will ignore the $o(1)$ terms coming from $a \neq h$ in the future.)

Conjecture 2. $N_{(2),g\,\mathsf{ANY},h\,\mathsf{RP}}(p) = N_{(3),h\,\mathsf{RP},a\,\mathsf{ANY}}(p) \approx 2\phi(p-1)$.

Table 4 in Section 3 gives values of $N_{(3),h\,\mathsf{RP},a\,\mathsf{ANY}}(p)$ determined by experiment which agree quite well with the estimated ones.

Adding conditions to a does not significantly complicate the analysis. If h and a are both RP in a solution to (3) then it is easy to see that this is equivalent to a solution to (2) with h RP and $\mathrm{ord}_p(g) = \mathrm{ord}_p(h)$, but no other conditions on g. We will say that $N_{(3),h\,\mathsf{RP},a\,\mathsf{RP}}(p) = N_{(2),h\,\mathsf{RP},g\,\mathsf{ORD}\,h}(p)$. We estimate this by separating the trivial and nontrivial pairs (h,a) once again. There are approximately $\phi(p-1)$ of the former and approximately $\phi(p-1)^2/(p-1)$ of the latter, since there are only $\phi(p-1)$ values of a which are RP.

Conjecture 3. $N_{(3),h\,\mathsf{RP},a\,\mathsf{RP}}(p) = N_{(2),h\,\mathsf{RP},g\,\mathsf{ORD}\,h}(p) \approx \phi(p-1) + \phi(p-1)^2/(p-1)$.

If h is RP and a is PR in a solution to (3), then this is equivalent to a solution to (2) with g PR and h RP, so $N_{(2),g\,\mathsf{PR},h\,\mathsf{RP}}(p) = N_{(3),h\,\mathsf{RP},a\,\mathsf{PR}}(p)$. In separating the trivial and nontrivial pairs it is necessary to observe that if $h = a$ then h is RPPR, so the trivial pairs contribute $\approx \phi(p-1)^2/(p-1)$. The nontrivial pairs also contribute $\approx \phi(p-1)^2/(p-1)$.

Conjecture 4. $N_{(2),g\,\mathsf{PR},h\,\mathsf{RP}}(p) = N_{(3),h\,\mathsf{RP},a\,\mathsf{PR}}(p) \approx 2\phi(p-1)^2/(p-1)$.

If either h or a is required to be RPPR in a solution to (3), then both must be. This is equivalent to a solution of (2) with g PR and h RPPR; i.e., $N_{(2),g\,\mathsf{PR},h\,\mathsf{RPPR}}(p) = N_{(3),h\,\mathsf{RPPR},a\,\mathsf{RPPR}}(p)$. The trivial pairs (h,a) contribute $\approx \phi(p-1)^2/(p-1)$. The nontrivial pairs contribute $\approx \phi(p-1)^3/(p-1)^2$, since there are $\approx \phi(p-1)^2/(p-1)$ values each of a and h which are RPPR, but the values of h^h and a^a are now constrained to be PR so there are only $\phi(p-1)$ possibilities.

Conjecture 5. $N_{(2),g\,\mathsf{PR},h\,\mathsf{RPPR}}(p) = N_{(3),h\,\mathsf{RPPR},a\,\mathsf{RPPR}}(p) \approx \phi(p-1)^2/(p-1) + \phi(p-1)^3/(p-1)^2$.

If a is RP but h is not necessarily so, then we may proceed similarly, letting $g \equiv h^{\overline{a}}$ modulo p. If a is RP and h is ANY, this is equivalent to a solution to (2) with h ANY and $\mathrm{ord}_p(g) = \mathrm{ord}_p(h)$. Thus $N_{(2),h\,\mathsf{ANY},g\,\mathsf{ORD}\,h}(p) = N_{(3),h\,\mathsf{ANY},a\,\mathsf{RP}}(p)$. This of course is the same as $N_{(3),h\,\mathsf{RP},a\,\mathsf{ANY}}(p) \approx 2\phi(p-1)$. Similarly, if a is RP and h is PR then this is equivalent to a solution to (2) with g and h both PR, so $N_{(2),h\,\mathsf{PR},g\,\mathsf{PR}}(p) = N_{(3),h\,\mathsf{PR},a\,\mathsf{RP}}(p)$. This is the same as $N_{(3),h\,\mathsf{RP},a\,\mathsf{PR}}(p) \approx 2\phi(p-1)^2/(p-1)$.

Conjecture 6.

(a) $N_{(2),h\,\mathsf{ANY},g\,\mathsf{ORD}\,h}(p) = N_{(3),h\,\mathsf{ANY},a\,\mathsf{RP}}(p) \approx 2\phi(p-1)$.
(b) $N_{(2),h\,\mathsf{PR},g\,\mathsf{PR}}(p) = N_{(3),h\,\mathsf{PR},a\,\mathsf{RP}}(p) \approx 2\phi(p-1)^2/(p-1)$.

The heuristics for (3) so far seem to be well supported by the data (see Section 3), are easy to convert to heuristics for (2), and seem to be suitable for a rigorous approach along the lines of [6]. The situation when neither h nor a is RP is less convenient.

We will first discuss the solutions to (3), and afterwards their relationship to (2). The expected chance that a number is PR is the same as the chance that a number is RP, so we would expect $N_{(3),h\,\mathsf{PR},a\,\mathsf{ANY}}(p) \approx N_{(3),h\,\mathsf{RP},a\,\mathsf{ANY}}(p) \approx 2\phi(p-1)$, and of course the same for $N_{(3),h\,\mathsf{ANY},a\,\mathsf{PR}}(p)$. This appears to be the case. Similarly we expect $N_{(3),h\,\mathsf{PR},a\,\mathsf{PR}}(p) \approx N_{(3),h\,\mathsf{RP},a\,\mathsf{RP}}(p) \approx \phi(p-1) + \phi(p-1)^2/(p-1)$. Finally, the same heuristics predict that $N_{(3),h\,\mathsf{ANY},a\,\mathsf{ANY}}(p) \approx 2(p-1)$. This does not seem to fit well with the data, however. (See Section 3.)

A finer analysis in this case is in order. (The following argument was suggested by an anonymous referee.) Fix the prime p, and let S_m be the set of h which are ANY such that $\operatorname{ord}_p(h^h) = m$. Let T_m be the set of h which are ANY such that $\operatorname{ord}_p(h) = m$. Then the estimated chance that h^h modulo p is a particular number in T_m is $|S_m|/|T_m|$ and the estimated chance that h^h and a^a are the same number modulo p is $|S_m|^2/|T_m|^2$. The number of solutions to (3) with $\operatorname{ord}_p(h^h) = \operatorname{ord}_p(a^a) = m$ is thus $\approx |S_m|^2/|T_m|$, and the total number of nontrivial solutions to (3) is $\approx \sum_{m|p-1} |S_m|^2/|T_m|$.

Now it's not hard to see that h^h has order m if and only if h has order dm for some d dividing $(p-1)/m$ and also $\gcd(h, \operatorname{ord}_p(h)) = d$. So

$$S_m = \bigcup_{d|(p-1)/m} \left(\{\operatorname{ord}_p(a) = dm\} \cap \{\gcd(a, dm) = d\}\right).$$

Supposing as we have been that conditions on order are independent of conditions on greatest common divisors, we have

$$|S_m| \approx \sum_{d|(p-1)/m} \frac{\phi(dm)\phi(m)}{dm} = \frac{\phi(m)}{m}\left(\sum_{d|(p-1)/m} \frac{\phi(dm)}{d}\right)$$

and

$$\sum_{m|p-1} |S_m|^2/|T_m| \approx \sum_{m|p-1} \frac{\phi(m)}{m}\left(\sum_{d|(p-1)/m} \frac{\phi(dm)}{d}\right).$$

Thus we estimate

$$N_{(3),h\,\mathsf{ANY},a\,\mathsf{ANY}}(p) \approx (p-1) + \sum_{m|p-1} \frac{\phi(m)}{m}\left(\sum_{d|(p-1)/m} \frac{\phi(dm)}{d}\right)$$

which gives much better agreement with the data.

In the case where $p-1$ is squarefree, ϕ can be treated as completely multiplicative and this can be simplified to

$$\sum_{m|p-1} \frac{\phi(m)^2}{m}\left(\sum_{d|(p-1)/m} \frac{\phi(d)}{d}\right) = \sum_{m|p-1}\left(\prod_{q|m} \frac{\phi(q)^2}{q}\right)\left(\prod_{q|(p-1)/m}\left(1 + \frac{\phi(q)}{q}\right)\right)$$

$$= \prod_{q|p-1} \left(\frac{\phi(q)^2}{q} + 1 + \frac{\phi(q)}{q} \right) = \prod_{q|p-1} \left(q + 1 - \frac{1}{q} \right).$$

Thus

$$N_{(3),h\,\mathsf{ANY},a\,\mathsf{ANY}}(p) \approx (p-1) + \prod_{q|p-1} \left(q + 1 - \frac{1}{q} \right).$$

In all cases the product is taken over primes q.

A similar analysis can be done in the general case; let $p - 1 = \prod q^{\alpha}$ and let $m = \prod q^{\beta}$, then

$$\begin{aligned} |S_m| \approx \sum_{d|(p-1)/m} \frac{\phi(dm)\phi(m)}{dm} &= \prod_q \phi(q^\beta) \left(\sum_{0 \le \gamma \le \alpha-\beta} \frac{\phi(q^{\alpha+\beta})}{q^{\alpha+\beta}} \right) \\ &= \prod_q \phi(q^\beta) \left(\left(1 - \frac{1}{q} \right) (\alpha - \beta) + \frac{\phi(q^\beta)}{q^\beta} \right) \end{aligned}$$

and

$$\begin{aligned} &\sum_{m|p-1} |S_m|^2/|T_m| \approx \prod_q \left(\sum_{\beta=0}^{\alpha} \phi(q^\beta) \left[\left(1 - \frac{1}{q}\right)(\alpha - \beta) + \frac{\phi(q^\beta)}{q^\beta} \right]^2 \right) \\ &= \prod_q \left(\left(1 - \frac{1}{q}\right)^2 \alpha^2 + \sum_{\beta=1}^{\alpha} q^\beta \left(1 - \frac{1}{q}\right) \left[\left(1 - \frac{1}{q}\right)(\alpha - \beta + 1) \right]^2 \right) \\ &= \prod_q \left(\left(1 - \frac{1}{q}\right)^2 \alpha^2 \right. \\ &\quad + \left(1 - \frac{1}{q}\right)^3 \left[(\alpha+1)^2 \frac{q^{\alpha+1} - q}{q-1} - 2(\alpha+1) \frac{\alpha q^{\alpha+2} - (\alpha+1)q^{\alpha+1} + q}{(q-1)^2} \right. \\ &\quad \left. \left. + \frac{\alpha^2 q^{\alpha+3} - (2\alpha^2 + 2\alpha - 1)q^{\alpha+2} + (\alpha^2 + 2\alpha + 1)q^{\alpha+1} + q^2 + q}{(q-1)^3} \right] \right) \end{aligned} \tag{4}$$

To summarize:

Conjecture 7.

(a) $N_{(3),h\,\mathsf{ANY},a\,\mathsf{ANY}}(p) \approx (p-1) + \sum_{m|p-1} \frac{\phi(m)}{m} \left(\sum_{d|(p-1)/m} \frac{\phi(dm)}{d} \right)$.

(b) If $p-1$ is squarefree then $N_{(3),h\,\mathsf{ANY},a\,\mathsf{ANY}}(p) \approx (p-1) + \prod_{q|p-1} \left(q + 1 - \frac{1}{q} \right)$, where the product is taken over primes q dividing $p-1$.

(c) In general, $N_{(3),h\,\mathsf{ANY},a\,\mathsf{ANY}}(p) \approx (p-1)$ plus the formula given in (4).

(This finer analysis can also be carried out for the other sets of conditions on h and a that we have investigated. The reader will find that the heuristic estimates produced in these cases are the same as those that result from the coarser analyses above.)

We now to the the implications for $N_{(2),g\,\mathsf{ANY},h\,\mathsf{ANY}}(p)$. A solution to (2) certainly gives us a solution to (3) by letting $a \equiv g^h$ modulo p. Thus, for instance, we expect $N_{(2),g\,\mathsf{ANY},h\,\mathsf{ANY}}(p) \lessapprox N_{(3),h\,\mathsf{ANY},a\,\mathsf{ANY}}(p)$. In the other direction, given a solution to (3) we can try to solve $g^{ha} \equiv h^h$ modulo p; this will succeed $1/d$ of the time where $d = \gcd(ha, p-1)$. If there is a solution, then there are d such solutions, which look like $g\xi$ where $\xi^d \equiv 1$ modulo p. Now $(g^a)^h \equiv h^h$, so $h \equiv g^a\zeta$ for some $\zeta^{h'} \equiv 1$, $h' = \gcd(h, p-1)$. Likewise $a \equiv g^h\zeta'$ for some $\zeta'^{a'} \equiv 1$, $a' = \gcd(h, p-1)$. We want to find ξ such that $(g\xi)^a \equiv h \equiv g^a\zeta$ and $(g\xi)^h \equiv a \equiv g^h\zeta'$, or $\xi^a \equiv \zeta$ and $\xi^h \equiv \zeta'$. We would expect that the chance of this happening for a particular ξ would be $a'h'/d^2$. There are d values of ξ such that $(g\xi)^{ha} \equiv h^h$ if there are any, but g only exists $1/d$ of the time. Thus given a pair (h, a) which is a solution to (3) we expect on the average $a'h'/d^2 = \gcd(a, p-1)\gcd(h, p-1)/\gcd(ha, p-1)^2$ pairs (g, h) which are solutions to (2). If h and a are both RP then this number is 1; in general it will be less. This seems to be born out by the data as far as it goes.

3 Two-Cycles: Data

Tables 4 through 7 give the number of solutions to (3) for all of the conditions on h and a discussed above, keeping in mind that conditions on h and a are symmetric. Each table was calculated in a few minutes on a home computer using Maple. Almost all of the observed data points are within a few percent of their predicted values.

Table 4. Solutions to (3) with h RP

p	$N_{a\,\mathsf{ANY}}$ predicted	$N_{a\,\mathsf{ANY}}$ observed	$N_{a\,\mathsf{PR}}$ predicted	$N_{a\,\mathsf{PR}}$ observed
10007	10004	9947	5001.0	5050
10009	6624	6569	2192.1	2186
10037	9216	9092	4231.5	4174
10039	5712	5724	1625.2	1611
10061	8032	8008	3206.4	3176

Table 5. More solutions to (3) with h RP

p	$N_{a\,\mathsf{RP}}$ predicted	$N_{a\,\mathsf{RP}}$ observed	$N_{a,h\,\mathsf{RPPR}}$ predicted	$N_{a,h\,\mathsf{RPPR}}$ observed
10007	7502.5	7516	3750.5	3853
10009	4408.1	4454	1458.8	1449
10037	6723.7	6578	3087.2	3019
10039	3668.6	3690	1043.8	999
10061	5619.2	5572	2243.2	2205

Table 6. Solutions to (3) with h PR

p	$N_{a\,\mathsf{ANY}}$ predicted	$N_{a\,\mathsf{ANY}}$ observed	$N_{a\,\mathsf{PR}}$ predicted	$N_{a\,\mathsf{PR}}$ observed
10007	10004	10001	7502.5	7520
10009	6624	6491	4408.1	4356
10037	9216	9207	6723.7	6668
10039	5712	5857	3668.6	3732
10061	8032	8046	5619.2	5634

Table 7. Solutions to (3) with h ANY, a ANY

p	N predicted	N observed
10007	22516.0	22428
10009	28790.4	28434
10037	24891.5	24638
10039	27323.4	27238
10061	26137.5	26328

Tables 8 and 9 give the number of solutions to (2) for some representative conditions on g and h. Table 8 was computed on a SPARC-station in 7.2 hours, using Maple. (Tables 1 and 3 were computed at the same time.) Table 9 was computed on a Pentium III running Linux in 3.5 hours, using Maple. (Table 2 was computed at the same time.) No particular attempts were made to optimize the code. The numbers for h RP are identical with the corresponding numbers for (3) given above.

The predicted numbers for h ANY were not calculated using the heuristics for (3) discussed above. Instead, we observed that non-trivial solutions to (2) are also equivalent to non-trivial solutions of the equation

$$g^h \equiv \log_g h$$

where the left-hand side is taken to be reduced modulo p and the right-hand side is taken as a number between 0 and $p-2$ if it exists. We assume that the left-hand and right-hand sides are distributed independently. If g is PR, there are $\phi(p-1)$ choices for g. For each g there are $p-1$ choices for h and for each one a $1/(p-1)$ chance that the left-hand and right-hand sides will coincide, for an expected total of $\phi(p-1)$ non-trivial choices. Combined with our predictions for fixed points, this gives $N_{(2),g\,\mathsf{PR},h\,\mathsf{ANY}}(p) \approx 2\phi(p-1)$. If g is ANY, then there are $p-1$ choices for g. The right-hand side only exists if h is a power of g, but the left-hand side can only take on as many values as there are powers of g, so these factors balance out for an expected total of $p-1$. Combining this with fixed point results gives:

Conjecture 8. $N_{(2),g\,\mathsf{ANY},h\,\mathsf{ANY}}(p) \approx 2(p-1)$.

These results agree with the observed numbers within a few percentage points. (The drawback of these heuristics compared to those derived from (3) is that they do not seem as suitable for a rigorous approach.)

Table 8. Solutions to (2) with g PR

p	$N_{h\,\mathsf{ANY}}$ predicted	$N_{h\,\mathsf{ANY}}$ observed	$N_{h\,\mathsf{RP}}$ predicted	$N_{h\,\mathsf{RP}}$ observed
10007	10004	10061	5001.0	5050
10009	6624	6479	2192.1	2186
10037	9216	9125	4231.5	4174
10039	5712	5730	1625.2	1611
10061	8032	7923	3206.4	3176

Table 9. Solutions to (2) with g ANY

p	$N_{h\,\mathsf{ANY}}$ predicted	$N_{h\,\mathsf{ANY}}$ observed	$N_{h\,\mathsf{RP}}$ predicted	$N_{h\,\mathsf{RP}}$ observed
10007	20012	20006	10004	9947
10009	20018	19628	6624	6569
10037	20072	20107	9216	9092
10039	20076	20084	5712	5724
10061	20120	19853	8032	8008

4 Applications, Conclusion, and Future Work

The idea of repeatedly applying the function $x \mapsto g^x \bmod p$ is used in the famous cryptographically secure pseudorandom bit generator of Blum and Micali. ([1]; see also [5] and [3], among others, for further developments.) If one could predict that a pseudorandom generator was going to fall into a fixed point or cycle of small length, this would obviously be detrimental to cryptographic security. Our data suggests, however, that the chance that a pair (g, h) is a non-trivial two-cycle is $1/(p-1)$ for all of the conditions on choosing g and h that we have investigated. Likewise the chance that a pair (g, h) is a fixed point is $1/(p-1)$ except in the case where g is chosen PR and h is chosen RPPR, in which case the chance is $1/\phi(p-1)$ due to the redundancy of the conditions. This might perhaps be taken as an indication that the seed of one of these pseudorandom generators should be chosen not to be RPPR if this is feasible. (In these protocols g is often taken to be PR as a given.)

Most of the results of this paper are perhaps not surprising. We hope, however, that the heuristics introduced will lead to rigorous bounds on the error terms for our estimates. A likely consequence of these bounds would be proofs that every prime has a pair (g, h) which is a non-trivial two-cycle given various conditions on g and h. One area which we are not able to fully develop is the relationship between $N_{(3),h\,\mathsf{ANY},a\,\mathsf{ANY}}$ and $N_{(2),g\,\mathsf{ANY},h\,\mathsf{ANY}}$. Also, it may be possible to clean up the general formula for $N_{(3),h\,\mathsf{ANY},a\,\mathsf{ANY}}$. More work is definitely needed in these areas. Another obvious direction for further work would be to extend our analysis to three-cycles and more generally k-cycles for small values of k.

Acknowledgments

The author would like to thank the anonymous referees for their comments, for pointing out further references, and especially for suggesting the finer analysis for $N_{(3),h\,\mathsf{ANY},a\,\mathsf{ANY}}(p)$ carried out above, including the formula for the squarefree case. Thanks also go to John Rickert for his help with formulas, to Igor Shparlinski for his encouragement and for pointing out references, and to Mariana Campbell and Carl Pomerance for their interest in this project.

References

1. Manuel Blum and Silvio Micali. How to generate cryptographically strong sequences of pseudorandom bits. *SIAM J. Comput.*, 13(4):850–864, 1984.
2. Cristian Cobeli and Alexandru Zaharescu. An exponential congruence with solutions in primitive roots. *Rev. Roumaine Math. Pures Appl.*, 44(1):15–22, 1999.
3. Rosario Gennaro. An improved pseudo-random generator based on discrete log. In M. Bellare, editor, *Advances in Cryptology — CRYPTO 2000*, pages 469–481. Springer, 2000.
4. Richard K. Guy. *Unsolved Problems in Number Theory.* Springer-Verlag, 1981.
5. Sarvar Patel and Ganapathy S. Sundaram. An efficient discrete log pseudo-random generator. In H. Krawczyk, editor, *Advances in Cryptology — CRYPTO '98*, pages 304–317. Springer, 1998.
6. Carl Pomerance. On fixed points for discrete logarithms. Talk given at the Central Section meeting of the AMS, Columbus, OH, September 22, 2001. Joint work with Mariana Campbell.
7. Wen Peng Zhang. On a problem of Brizolis. *Pure Appl. Math.*, 11(suppl.):1–3, 1995.

Random Cayley Digraphs and the Discrete Logarithm

Extended Abstract

Jeremy Horwitz[1] and Ramarathnam Venkatesan[2]

[1] Stanford University, Stanford, CA 94305, USA
horwitz@cs.stanford.edu
[2] Microsoft Research, Redmond, WA 98052, USA
venkie@microsoft.com

Abstract. We formally show that there is an algorithm for DLOG over all abelian groups that runs in expected optimal time (up to logarithmic factors) and uses only a small amount of space. To our knowledge, this is the first such analysis. Our algorithm is a modification of the classic Pollard rho, introducing explicit randomization of the parameters for the updating steps of the algorithm, and is analyzed using random walks with limited independence over abelian groups (a study which is of its own interest). Our analysis shows that finding cycles in such large graphs over groups that can be efficiently locally navigated is as hard as DLOG.

1 Introduction

The Discrete Logarithm Problem (DLOG) defined over abelian groups plays a fundamental role in cryptography as a basis for many primitives (*e.g.*, Diffie-Hellman key exchange, DSS, and ElGamal signatures). The algorithms to find DLOG fall into two types: the generic, black-box, exponential-time algorithms that use only the group structure (*e.g.*, baby-step giant-step and Pollard rho) and the domain-specific subexponential algorithms (*e.g.*, index calculus methods), which are not yet known to exist for groups over elliptic curves. Because of its generality and that it uses a very small amount of space, Pollard rho [8] is practically and theoretically important.

Surprisingly, there is no formal analysis of the classic Pollard rho without random-oracle assumptions. The standard analysis is heuristic: it approximates the rho walk with a totally random walk (*i.e.*, a walk which at every step randomly and independently jumps to another group element) and then infers the existence of a cycle of length $\sqrt{p}$ using the birthday paradox. But, in reality, the walk is far from random: the algorithm only makes a *deterministic walk* (which is crucial for Floyd's algorithm to find a cycle using only a small amount of space) on a 3-regular directed graph over $\mathbb{Z}_p^\times$ that is constructed semi-randomly. By using a random oracle for the moves to the *neighboring* nodes, Teske [11,12] has analyzed both the original Pollard rho as well as more general k-regular graphs (for $k \geq 3$); for $k \geq 6$ she derives an $O(\sqrt{|G|})$ bound for finite abelian groups

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 416–430, 2002.

using a result of Hildebrand. Lack of independence between moves creates difficulty in analysis, especially since the move from a node z depends on (the label of) the node z. Earlier, Bach [2] studied Pollard rho for factoring and showed that the probability of success for the rho method is $c(\log^2 p)/p$ (for some $c > 0$), which is only slightly better than the obvious bound of $1/p$.

We explicitly introduce randomness by slightly modifying the algorithm and then base our treatment on random walks on Cayley graphs over abelian groups. Recall that a k-regular Cayley digraph (directed graph) on a group G has a set S of k generators. Its set of nodes is G and its edges are formed by connecting every α in G by a directed edge to αg_i, for every $g_i \in S$. To solve for $y = g^x$ in G, we construct S with equal number of random powers of y and g, and construct a navigation function $h_E(\alpha)$ (for $\alpha \in G$) which maps into $\{1, 2, \ldots, k\}$ (for $k = O(\log p)$), by picking a random polynomial over a suitable extension of $\mathbb{F}_2$ and truncating its output. We start at some $z_0 \in G$ and move from z_i to z_{i+1} by multiplying z_i by the generator in S with index $h_E(z_i)$. Finally, we look for a collision in the z_is.

Firstly, we show that our modified algorithm, which is a *random* walk *with limited independence* on a *random* Cayley graph (*i.e.*, S is a random k-subset of the group), finds the DLOG in optimal time (up to logarithmic factors). We note that a random choice of generators is important for two reasons: first, to show that the rho algorithm produces a nontrivial relationship (Theorem 1). Second, to guarantee the existence of Cayley graphs over *any* abelian group with an underlying Markov chain that rapidly mixes (without randomization, no such universal construction is known); the rapid-mixing property in turn is crucial for removing the dependence on a random-oracle assumption. This complements the result of Shoup [10] who showed that *generic* algorithms for DLOG must take at least $\sqrt{|G|}$ steps. It would be interesting to know if random walks exploiting specific group properties yield faster algorithms.

This analysis also allows us to show that finding nontrivial cycles (*i.e.*, smaller than the group order) in random Cayley graphs over an abelian group G of order p is as hard as solving DLOG over G. These graphs are *succinctly presented* in the sense that they are defined by simple rules for moving from a node to its neighbors; they are, however, too huge to be explicitly stored. Our succinct graphs have girth (*i.e.*, the length of shortest cycle) $O(\log p)$; however, to computationally efficient algorithms, the girth appears to be exponential in $\log p$. This allows for the construction of secure hash functions. A significantly longer version of this paper (including experimental results which exhibit practical run times and parallel our theoretical results) will be available from the authors.

2 Preliminaries and Statement of Results

In this section we present relevant definitions, motivation, and statements of our results. Our study is from the point of view of path finding or navigating in exponentially large graphs that have simple rules for moving from one node to another. We assume the constraint that one has a limited amount of memory.

2.1 Cayley Digraphs

In view of the Pohlig-Hellman result on DLOG [7], we consider only prime-order groups; we denote the order of the group discussed in this paper by p. In such a group, every element except the identity is a generator. For notions related to graph theory and random walks, we refer the reader to [4].

Let G be a multiplicative abelian group of order p and $S = \{g_1, g_2, \ldots, g_{2n}\} \subseteq G$ (we write $2n$ since $|S|$ will always be even). A *Cayley digraph generated by* S is denoted by $\mathcal{G}(G, S) = \mathcal{G} = (V, E)$ and has the set of nodes $V = G$ and the set of edges $E = \{(g, gg_i) \ : \ g \in G, g_i \in S\}$. (Most papers study undirected versions where, if $g \in S$, then $g^{-1} \in S$, and may additionally assume that the unit $1 \in S$ (*i.e.*, all nodes have self loops); we cannot assume either of these conditions.) A *path* of length t is a sequence $(v_0, v_1, \ldots, v_t)$ with every $(v_i, v_{i+1}) \in E$. A path is a *cycle* if it also satisfies $v_t = v_0$. In this paper, our main parameter is $2n = O(\log p)$, where p is large enough to make DLOG hard, while path lengths t can be exponentially large in $2n$. Since G is abelian, paths (and cycles) of length t admit succinct representations of size $O(2n \log t)$ as: given a path (or cycle), we write it as $X = (x_1, x_2, \ldots, x_{2n}) \in \mathbb{N}^{2n}$ where x_i is the number of the edges of the form (g, gg_i) in the path. Since $g^p = 1$ for any $g \in G$, cycles occur in $\mathcal{G}$ trivially; we will be interested only in *nontrivial* cycles having length $t < p$. We assume that all our paths and cycles are nontrivial and have length $t \le \Lambda$ for a fixed constant $\Lambda = (\log^{O(1)} p)\sqrt{p} = o(p)$. Having $t = o(p)$ avoids wraparound problems even when we add the lengths of a constant number of paths.

Succinct Graphs. We say that $\mathcal{G} = \mathcal{G}(G, S)$ is a *random Cayley digraph over* G if the elements of S are picked from G randomly and independently. By a navigation algorithm for a graph (V, E), we mean some algorithm to compute $f(u, i) = v$, where v is the ith ordered neighbor (under some predefined ordering) of the node u. If the graph is d-regular, then it can be edge colored, for example, with d colors, and we set $f(u, i) = v$ if the edge (u, v) has the ith color. A graph is *succinctly presented* (or *succinct*) if there is a navigation algorithm $f(u, i)$ that runs in time $|u|^{O(1)}$, where $|u|$ is the length of its label. We note that Cayley graphs over $\mathbb{Z}_p^\times$ are succinct because one can take the standard binary representation of integers as the label and compute $f(\alpha, i)$ as αg_i, where g_i is the ith generator in S. Another example is the k-dimensional hypercube with vertex set $\mathbb{Z}_2^k$ with vertices connected if and only if they differ in exactly one co-ordinate.

2.2 Limited Independence

A sequence of random variables $z_0, z_1, \ldots, z_t$ is called *m-wise independent* if any subsequence of at most m variables is independent; in our case they will be uniformly distributed. A 2-wise independent sequence is also called a *pairwise independent* sequence. A function $f(x)$ is *m-wise independent* if, for any sequence of inputs $\{z_i\}_{i=0}^t$, the sequence $\{f(z_i)\}_{i=0}^t$ is m-wise independent. We will randomly choose polynomials of degree $m - 1$ defined over an extension field of $\mathbb{F}_2$; notice that these polynomials are m-wise independent. Indeed, given $\boldsymbol{z} = (z_0, z_1, \ldots, z_{m-1})$ with distinct x_i and $\boldsymbol{y} = (y_0, y_1, \ldots, y_{m-1})$, one can find a polynomial with $f(z_i) = y_i$: solve the equation $\boldsymbol{y} = V\boldsymbol{f}$ (where $V = (z_i^j)_{0 \leq i,j < m}$ is a Vandermonde (and, hence, invertible) matrix) for $\boldsymbol{f} = (f_0, f_1, \ldots, f_{m-1})$ and set $f(x) = \sum_{i=0}^{m-1} f_i x^i$. We note that if we truncate each of the outputs of $f(z_i)$ (to some number of least-significant bits), we will still have an m-wise independent sequence. To see this, note that in this case we are given only the truncated bits of entries in y and we may arbitrarily extend them to fully specify a vector y and proceed as before. By incrementing if necessary, we shall assume that m is even. We now recall the following tail inequality [3] for the sum Z of a sequence of m-wise independent variables taking values in $[0, 1]$. Set $\mu := \mathrm{E}[Z]$ and let $a > 0$. Then, we have $\Pr[|Z - \mu| \geq a] \leq 8((m\mu + m^2)/(a^2))^{m/2}$.

2.3 Finding Cycles in Succinct Graphs and DLOG

While finding paths and cycles efficiently in the usual graphs is well-understood, finding paths and cycles in succinct graphs using only small space may be hard (though, in some cases, such as hypercubes, this is trivial). Indeed, one may view the classic Pollard rho for solving $y = g^x$ as a method both to define (using y and g) a succinctly presented graph together with its navigation algorithm h and to find a cycle in the succinct graph (then solve a linear equation to find DLOG). Our modification to Pollard rho differs only in the definition step and is aimed at bounding the run time and the success probability in the cycle-finding step.

Pollard Rho Algorithm. Let $g \neq 1$ be fixed. Given $y \in G = \langle g \rangle$, the task is to find x such that $y = g^x$. The algorithm (in some simple way) partitions G into three approximately equal-sized sets T_1, T_2, and T_3 (taking care that $1 \notin T_3$). Now, define the navigation algorithm h_ρ: $G \to G$ as: $h_\rho(z) = zg$ for $z \in T_1$, $h_\rho(z) = zy$ for $z \in T_2$, and $h_\rho(z) = z^2$ for $z \in T_3$.

Starting with some fixed $z_0 = g^r$, construct a sequence $\{z_i\}_{i=0}^t$ with $z_{i+1} = h_\rho(z_i)$ until a collision occurs (*i.e.*, $z_u = z_v$ for some $u \neq v$). Then use Floyd's algorithm to find a cycle, which yields a relationship of the form $bx = a + rc \bmod p$.

Remark 1. It is crucial that h above is deterministic if one wants to preserve the main advantages of small space and being able to avoid exhaustive search over the entire group. As noted earlier, in standard analysis for the rho method, one treats the z_is as if they were random and independent (equivalently, one treats the graph as a complete graph and the navigation function h as if it were chosen randomly from the set of all functions from G to G) and uses the birthday paradox to bound $t = O(\sqrt{p})$. Also, we note that there is no formal guarantee for the probability that b^{-1} exists $(\mathrm{mod} p)$, which is required to finally discover x.

Cayley Rho Algorithm. Fix a cyclic group G (of order p) and a generator $g \in G$ with respect to which we will solve DLOG. Where $2n$ is the size of $S \subseteq G$

(the set of generators for the Cayley graph), we, for convenience, assume that $2n$ is a power of 2. (Experiments show that when $2n$ is at least $4 \log_2 p$, the Cayley rho algorithm performs better than the Pollard rho; further details appear in the full version of this paper.) We fix an extension field $E/\mathbb{F}_2$ with $[E : \mathbb{F}_2] = 3\lceil \log p \rceil$ (unless otherwise stated, we always mean the base-2 logarithm) and set $d := \nu \lceil \log p \rceil$, where ν is a small constant. Define H to be the set of all degree-d polynomials from E to E. Let $y = g^x$ be given. We construct an algorithm $\mathcal{C}(y)$ as follows:

1. **Defining the succinct graph**: Randomly choose $r_1, r_2, \ldots, r_n \in \mathbb{Z}_p$ and $s_1, s_2, \ldots, s_n \in \mathbb{Z}_p$. Then let $(g_1, g_2, \ldots, g_{2n})$ be a random permutation of $(g^{r_1}, g^{r_2}, \ldots, g^{r_n}, y^{s_1}, y^{s_2}, \ldots, y^{s_n})$. Let $S := \{g_1, g_2, \ldots, g_{2n}\}$ and $\mathcal{G} = \mathcal{G}(G, S)$ be the *random* Cayley graph generated by S over G.
   Initializing the navigation algorithm: We randomly choose and fix a polynomial $h'\colon E \to E$ from H.
   Computing $h(\alpha)$: Given $\alpha \in G$, we use a standard $\lceil \log p \rceil$-bit binary representation of α and pad it with a suitable prefix of zeros to get $\alpha' \in E$. Define $h_E\colon E \to \{1, 2, \ldots, 2n\}$ so $h_E(\alpha')$ is the $\log_2(2n)$ least-significant bits of binary representation of $h'(\alpha')$. Define $h\colon G \to G$ by $h(\alpha) = \alpha g_c$, where $c := h_E(\alpha')$.
2. As in Pollard rho, we can use a procedure $\mathcal{A}(\mathcal{G})$ which outputs a cycle $X = (x_1, x_2, \ldots, x_{2n})$ in $\mathcal{G}$ (*i.e.*, $\prod g_i^{x_i} = 1$). If the cycle is trivial, we repeat the entire algorithm; else we solve a linear equation (described below). (In case the equation cannot be solved (*i.e.*, it is $0x = 0$), $\mathcal{C}$ must be restarted.)

By abusing notation we may write $h \in \mathsf{H}$ or $h_E \in \mathsf{H}$ (really only $h' \in \mathsf{H}$).

2.4 Notation for Walks

Throughout this paper, we utilize a number of functions (particularly h, h_ρ, h_E, and h_2) to describe our random walks, primarily to simplify our analysis and to make our notation more convenient for both the authors and the readers.

The transition function $h\colon G \to G$ is most similar to a standard transition function for a Markov chain: it takes as input the current state and returns the next state. (The method for its construction is explained in Section 2.3.) The function $h_\rho\colon G \to G$ represents the Pollard rho transition function, which we only mention in a referential context. We use $h_E\colon E \to \{1, 2, \ldots, 2n\}$ (as described in Section 2.3) as an intermediate construction en route to building h. We overload h_E to allow $h_E\colon G \to \{1, 2, \ldots, 2n\}$ (where, in these instances, h_E appropriately pads a natural binary representation of its input with zeroes in order to apply h_E as usual (as described in Section 2.3)). A technical necessity used only in Section 6, $h_2\colon G \times \mathbb{N} \to G$ is constructed from a function h'_2 (which is randomly chosen from a set of bivariate polynomials) just as h is constructed from h'. h_2 is constructed so that when $\gamma_i = 0$, $h_2(z_i, \gamma_i) = h(z_i)$. The probability (over choice of h_2) that there is *no* collision in the walk defined from h_2 is equal to the probability (over choice of h) that there is *no* collision in the walk defined from h. This result is discussed in greater detail in Lemma 13.

2.5 Main Results

Theorem 1 (Near-Optimal Convergence). *Let the Cayley rho algorithm $\mathcal{C}$ take $\widetilde{O}(\sqrt{p})$ (*i.e., *$O(\sqrt{p})$ up to factors of* $\log p$*) moves on the graph. Then, (a) the probability (over the random choices made by $\mathcal{C}$) of a cycle of length $\widetilde{O}(\sqrt{p})$ occurring is a positive constant and (b) when the cycle-finding algorithm $\mathcal{A}$ returns successfully, $\mathcal{C}$ solves* DLOG *with probability at least $(2n^2)^{-1}$; thus the expected number of calls to the cycle-finding algorithm $\mathcal{A}$ is at most $2n^2$.*

Corollary 1 (DLOG $\preceq$ Cycle Finding). *Finding cycles in random Cayley graphs over G is as hard as solving* DLOG *on G.*

The corollary follows from part (b) of the theorem, since it applies to *any* cycle-finding algorithm $\mathcal{A}$. To prove Theorem 1(a), we use the next theorem.

Theorem 2 (Rapid Mixing). *Let $\mathcal{G}$ be a random Cayley digraph over an abelian group G of prime order p and let $z_0 \in G$ be arbitrary. Starting from z_0, let the endpoint of a t-step (totally independent) random walk be z_t. If $t \geq 2\log p$, then, for any $\alpha \in G$, $|\Pr[z_t = \alpha] - 1/p| \leq p^{-2}$.*

Rapid mixing of Cayley graphs is well-studied; however, we could not find a reference for the case of Cayley digraphs with both $O(\log p)$ generators and no self loops that states the required bound ($O(p^{-2})$ rather than $O(1)$) on the deviation from the uniform. However, our proof is simple, and all the required Markov chain properties are derived directly from Lemma 2. Yet, the theorem is insufficient for us to prove results unconditionally; if we assumed that the navigation function is a purely random function, then we would get the result using the above theorem. It is simple to show, using elementary matrix methods, the following: starting at an arbitrary z_i, if a purely random walk on an expander converges to an almost-uniform distribution μ: $G \to [0,1]$ in τ steps (*i.e.*, the node $z_{i+\tau}$ is almost-uniformly distributed), then, for any $t > \tau$, the distribution of z_{i+t} remains almost-uniformly distributed. This need not be true when the walk steps are correlated. However, using that G is abelian, we can show that the walk remains almost-uniformly distributed. This result appears to be the first of its type and is of interest by itself.

3 Proof of Theorem 1(b)

Proof. Let $\mathcal{A}(\mathcal{G})$ find a cycle of length $t = o(p)$. From this cycle, we get an equation of the form $z_0 = z_0 \prod_{i=1}^{2n} g_i^{w_i}$, for some initial node $z_0 \in G$ and $0 \leq w_i \leq t$, where $\sum_{i=1}^{2n} w_i = t$. From the definition of the g_i, we see that $\prod_{i=1}^{n} g^{-r_i w_i} = \prod_{i=1}^{n} y^{s_i w_{n+i}}$. Hence, $-\sum_{i=1}^{n} r_i w_i = x \sum_{i=1}^{n} s_i w_{n+i} \pmod{p}$, which yields x unless $\sum_{i=1}^{n} r_i w_i = \sum_{i=1}^{n} s_i w_{n+i} = 0 \pmod{p}$. The probability that we *cannot* find x (because the aforementioned sums are zero) is bounded above by $1 - \frac{1}{n^2+1}$ (from Lemma 1), so we expect to rerun $\mathcal{A}$ at most $n^2 + 1 \leq 2n^2$ times. □

Lemma 1. *Let $k_1, k_2, \ldots, k_{2n} \in \mathbb{Z}_p$ be such that for $i \neq j$, $k_i \neq \pm k_j \pmod p$. Fix $t = o(p)$ and randomly choose $\sigma \in S_{2n}$. An adversary, given the k_i and t (but not σ), chooses $0 \leq w_i \leq t$ (not all zero) and we say the adversary wins if $\sum_{i=1}^{n} k_{\sigma(i)} w_i = \sum_{j=n+1}^{2n} k_{\sigma(j)} w_j = 0 \pmod p$. Then the probability (over choices of σ) that the adversary wins is at most $1 - \frac{1}{n^2+1}$.*

4 The Markov Chain Induced by G

We define our random walk on $\mathcal{G}$ as follows: starting at an initial node z_0, one picks, uniformly at random, one of the outgoing edges (say, $(z_0, z_0 g_i)$) and moves to the opposite node (*i.e.*, $z_1 := z_0 g_i$). Then we iterate this step, using independent coin flips at each node. The induced Markov chain (which we denote by **MC**) has the transition matrix M with entries $m_{\alpha\beta} = 1/2n$ if there is an edge from the node α to node β (else it is zero); the adjacency matrix $A(\mathcal{G})$ has $a_{\alpha\beta} = 2n m_{\alpha\beta}$. Our graphs are directed and we must work out many of their properties from scratch. We point out that existing literature on rapid mixing cannot be directly used for a variety of reasons: our graphs are directed; we cannot add self loops to guarantee aperiodicity; we need to derive quantitative bounds on the deviation (from the uniform distribution); and, most importantly, our walks are not entirely independent. Here, matrix theory cannot be applied at all, and we utilize a probabilistic argument that capitalizes on the abelian property and shows (in this case) that if a purely random walk is convergent, then so is the related limited-independence random walk.

4.1 Conventions and Markov Chain Preliminaries

Conventions we use include denoting the walk by $z_0, z_1, \ldots, z_t$ and defining a function $\boldsymbol{c}$: $\{0, 1, \ldots, t-1\} \to \{1, 2, \ldots, 2n\}$ so $z_{i+1} = z_i g_{\boldsymbol{c}(i)}$. Notice that the random walk is completely specified by $\boldsymbol{c}$; as such, we often refer to $\boldsymbol{c}$ as a walk.

Let $\Omega_t := \{(x_1, x_2, \ldots, x_{2n}) \in \mathbb{N}^{2n} : \sum_{i=1}^{2n} x_i = t\}$. For $1 \leq j \leq 2n$, set $y_j := \left|\boldsymbol{c}^{-1}(j)\right|$. In other words, the random walk induced by $\boldsymbol{c}$ picks each generator g_i a total y_i times during the t-step random walk. Notice that there is a well-defined map $\boldsymbol{c} \mapsto Y = (y_1, y_2, \ldots, y_{2n})$, which we will write as $\psi(\boldsymbol{c}) = Y$. Let $\lambda(Y) = \Pr_{\boldsymbol{c}}[\psi(\boldsymbol{c}) = Y]$ and $\mu(Y) = |\Omega_t|^{-1} = \binom{t+2n-1}{2n-1}^{-1}$.

The group S_{2n} of permutations of $\{1, 2, \ldots, 2n\}$ acts on Ω_t and we denote its orbits by $T_1, T_2, \ldots, T_N$. We note that $Y = (y_1, y_2, \ldots, y_{2n})$ and Y' both belong to the same orbit T_j if and only if Y is a permutation of Y' (*i.e.*, $Y' = (y_{\sigma(1)}, y_{\sigma(2)}, \ldots, y_{\sigma(2n)})$ for some $\sigma \in S_{2n}$). Clearly this induces an equivalence relation, and we write $Y \sim Y'$ if and only if $Y, Y' \in T_j$ for some j. As usual, we say that T_j is the orbit of Y. An important fact here is that if $Y \sim Y'$, then $\lambda(Y) = \lambda(Y')$, since the sequence $\{\boldsymbol{c}(i)\}_{i=0}^{t-1}$ and $\{\sigma(\boldsymbol{c}(i))\}_{i=0}^{t-1}$ have the same λ probability for any $\sigma \in S_{2n}$.

We first prove a preliminary lemma that is analogous to a result of Erdős and Rényi [6], who showed that random subproducts of the (uniformly-chosen) generators are almost-uniformly distributed. Our method allows one to quantify the dependence of the quality of this distribution in terms of the walk length, as well as to show many properties of the random $\mathcal{G}$.

We fix $g \neq 1$ so that $G = \langle g \rangle$. Recall that we represent a path X of length t_X by a $2n$-tuple of nonnegative integers $X = (x_1, x_2, \ldots, x_{2n})$ such that $\sum_i x_i = t_X$. We say that two distinct nonzero paths X and Y are *linearly correlated* if, as vectors, they are scalar multiples of each other (*i.e.*, $t_X Y = t_Y X \pmod{p}$). Otherwise, they are said to be *linearly uncorrelated.* For example, $X \neq Y$ will be linearly uncorrelated if $t_X = t_Y$ or if they are binary vectors. In addition, if $\max\{t_X^2, t_Y^2\} < p$, it is sufficient that $t_X Y = t_Y X$ holds over $\mathbb{Z}$. Note that if two vectors are linearly *independent* over $\mathbb{F}_p$, they will be linearly *uncorrelated* in our sense, but the converse need not hold.

We consider pairs of linearly uncorrelated paths and conclude that random Ss induce a pairwise-independent function on them. For a given random S, with $g_i := g^{\alpha_i}$, we define a mapping ϕ_S to take a path X to the node $\prod_i g^{\alpha_i x_i}$. Without loss of generality, we may assume that the starting point of the walk is unity. As such, $\phi_S(X)$ is the endpoint of the walk specified by X. We will heavily rely on Corollary 2 below, which is immediate from Lemma 2.

Remark 2. We will assume that S is formed by picking $2n$ elements randomly and independently from G. These need not be distinct, so S can be a multiset. Our main analysis requires only a lower bound on the size of S. By a tiny increase in the number of elements picked, one can be assured that S has $2n$ elements with probability at least $1 - p^{-3}/4$. Constructing S can be viewed as randomly choosing an $S \in \mathsf{S} := \{S \subseteq G \;:\; |S| = 2n\}$.

Also, note that the next lemma allows the case $1 \in S$.

Lemma 2 (Pairwise Independence). *In a Cayley digraph over a group of prime order p, let X and Y be two arbitrary distinct nonzero linearly uncorrelated paths of lengths at most Λ. Then, the mapping $\phi_S(X) := \prod g^{\alpha_i x_i}$ is a pairwise-independent mapping,* i.e., *for any $a, b \in \mathbb{Z}_p$,*

$$\Pr_S\left[\phi_S(X) = g^a \wedge \phi_S(Y) = g^b\right] = \Pr_S\left[\phi_S(X) = g^a\right] \Pr_S\left[\phi_S(Y) = g^b\right] = \frac{1}{p} \cdot \frac{1}{p}\ .$$

Corollary 2. *(a) On $A \subseteq \Omega_t$, for equal-length ($t \leq \Lambda$) paths, the mapping $\phi_S(X)$ is pairwise-independent. (b) The restriction of $\phi_S(X)$ to the set $B := \{(x_1, x_2, \ldots, x_{2n}) \;:\; x_i \in \{0,1\},$ not all zero$\}$ is a pairwise-independent map. In this case, $X \mapsto \phi_S(X)$ is a subset-product map on nonempty sets of generators.*

Corollary 3. *Let $\mathsf{S}_0 \subseteq \mathsf{S}$ be such that $1 - \frac{|\mathsf{S}_0|}{|\mathsf{S}|} \leq \varepsilon$ (with $\varepsilon \leq p^{-2}$). Then,*

$$1 - 2\varepsilon(1 - 1/p) \leq \Pr_S[\phi_S(X) = g^a | S \in \mathsf{S}_0] / \Pr_S[\phi_S(X) = g^a] \leq 1 + 2\varepsilon\ .$$

4.2 Properties of MC

Notice that (unless $S = \{1\}$) the elements of S generate G and, thus, the Cayley digraph $\mathcal{G}$ is strongly connected (*i.e.*, **MC** is irreducible). For any irreducible Markov chain, by the Perron-Frobenius theorem, the adjacency matrix has 1 as the maximal eigenvalue; additionally, this eigenvalue has multiplicity one. To guarantee a stationary distribution of the chain, it must also be aperiodic (stated as Lemma 3). Note that the group structure imposes that the in-degree and the out-degree of any node are the same (both equal to $|S|$), making M doubly stochastic (*i.e.*, every column sums to one, as does every row). Hence, if **MC** has a stationary distribution, it must be the uniform distribution. In addition to allowing us to conclude that **MC** has a unique, uniform stationary distribution, the proof of the following lemma also yields a $\Theta(\log |G|)$ bound for both the diameter and the girth of almost every graph.

Lemma 3. *MC is aperiodic for all but a negligible fraction of choices of S.*

5 Rapid Mixing (Proof of Theorem 2)

We recall standard definitions. The *boundary* of a $D \subseteq V$ is the set $\partial D = \{v \in V \;:\; v \notin D$ and v has incoming edge from some node in $D\}$.

If $U \subseteq V$ and for every subset W of U we have $|\partial W| \geq \varepsilon |W|$, then U is then called *ε-expanding.* We call the subgraph induced by an ε-expanding subset an *ε-expanding graph.* The entire graph $\mathcal{G} = (V, E)$ is called an *ε-expander* if every subset of size at most $\frac{|V|}{2}$ is ε-expanding.

Normally, ε is taken to be a constant as the size of $\mathcal{G}$ grows; one shows that on such expanders a random walk rapidly mixes in the sense that it reaches a distribution exceptionally close to its stationary (uniform) distribution in $O(\log p)$ steps. Cayley graphs and general expanders are the subject of extensive literature and the reader may wish to consult [1,5,9] as well as the short survey in the full version of this paper.

5.1 Outline of the Proof of Theorem 2

First, Lemma 4(a) will allow us to conclude that almost all choices of the set S of generators are *good* in the sense that:

(†) for a sufficiently large walk length t, for $\alpha \in G$ and $A \subseteq \Omega_t$ with $|A| \geq p^5$, $\left| \frac{|\phi_S^{-1}(\alpha) \cap A|}{|A|} - \frac{1}{p} \right| < \frac{1}{p^2}$.

Thus, if we pick a random Y from A, the endpoint $\phi_S(Y)$ will be almost-uniformly distributed. However, a random walk $\boldsymbol{c}$ does not induce a uniform distribution on the tuples $Y \in A$ for arbitrary A, but, if A is an orbit in Ω_t under the action of S_{2n}, the induced distribution $Y \mapsto \Pr_{\boldsymbol{c}}[Y | Y \in A]$ is indeed uniform (within a fixed orbit A). To use (†), we will need $|A| \geq p^5$; however, there are many small orbits (*e.g.*, the orbit of $Y = (s, s, \ldots, s)$). Fortunately, Lemma 5 will help complete the proof by showing the following property:

(‡) with overwhelming probability, a random walk $\boldsymbol{c}$ generates a $\psi(\boldsymbol{c}) = Y$ whose orbit under S_{2n} has, for *every* S, size at least p^5.

5.2 Proof of Theorem 2

Lemma 4. *Fix $\alpha \in G$. (a) If $A \subseteq \Omega_t$ with $|A| \geq p^5$, then, for all but a p^{-2} fraction of $S \in \mathsf{S}$, $|\Pr_{X \in A}[\phi_S(X) = \alpha] - 1/p| < p^{-2}$. (b) If B is as defined in Corollary 2, and if $2n \geq 8 \log p$, then $|\Pr_{X \in B}[\phi_S(X) = \alpha] - 1/p| < p^{-2}$.*

Lemma 4(a) shows that almost all S satisfy (†), so now we address (‡):

Lemma 5. *If $2n$ is a constant multiple of $\log p$, then there is a $t = O(\log p)$ such that, for a t-step random walk $\boldsymbol{c}$, we have*

$$\Pr_{\boldsymbol{c}}[Y := \psi(\boldsymbol{c}) \textit{ has an orbit of size no more than } p^5] = o(p^{-2}) \ .$$

Now we use these lemmata to prove Theorem 2. Notice that, for a random walk $\boldsymbol{c}$, $\Pr_{\boldsymbol{c}}[Y|Y \in T_j] = 1/|T_j|$, since for any two $Y, Y' \in T_j$, we have $Y \sim Y'$ and $\lambda(Y) = \lambda(Y')$. Now arrange the T_j in increasing order by size, and pick the smallest $L \in \mathbb{N}$ so that $|T_L| > p^5$. Now, for *every* "good" (see (†)) S:

$$\begin{aligned}
\Pr_Y[\phi_S(Y) = \alpha] &= \sum_{j=1}^{L} \Pr_Y[\phi_S(Y) = \alpha | Y \in T_j] \Pr_Y[Y \in T_j] \\
&\quad + \sum_{j=L+1}^{N} \Pr_Y[\phi_S(Y) = \alpha | Y \in T_j] \Pr_Y[Y \in T_j] \\
&\leq \sum_{j=1}^{L} \Pr_Y[Y \in T_j] + \sum_{j=L+1}^{N} \left[\frac{1}{p} + \left(\frac{1}{p^2}\right)\right] \Pr_Y[Y \in T_j] \\
&\leq o(p^{-2}) + \left(\frac{1}{p} + \frac{1}{p^2}\right) \sum_{j=L+1}^{N} \Pr_Y[Y \in T_j] \leq o(p^{-2}) + \left(\frac{1}{p} + \frac{1}{p^2}\right) \ .
\end{aligned}$$

We complete the proof of Theorem 2 by noticing that, for every good S, we have a similar lower bound: $\Pr_Y[\phi_S(Y) = \alpha] \geq \sum_{j=L+1}^{N} \Pr_Y[\phi_S(Y) = \alpha | Y \in T_j] \Pr_Y[Y \in T_j] \geq \left(\frac{1}{p} - \frac{1}{p^2}\right) \sum_{j=L+1}^{N} \Pr_Y[Y \in T_j] \geq \left(\frac{1}{p} - \frac{1}{p^2}\right)\left(1 - o(p^{-2})\right)$. □

5.3 Rapid Mixing with Limited Independence

In this section, we will denote by w a lower bound on the local-independence parameter of the hash functions so that $\boldsymbol{c}$ will be w-wise independent (and hence, $d \geq w$). Our analysis is applicable to any $\boldsymbol{c}$ that is w-wise independent. For example, $\boldsymbol{c}(r)$ may depend only on r, $\boldsymbol{c}(r)$ may depend only on z_r, or $\boldsymbol{c}(r)$ may depend on both, possibly with additional parameters. Indeed, we use this fact in Section 6.

We need to compute $\Pr[z_j = \alpha | z_i]$. For convenience, we write $\overline{g_r} = g_{\boldsymbol{c}(r)}$, so that the sequence of generators chosen for the walk is $\overline{g_0}, \overline{g_1}, \ldots, \overline{g_{t-1}}$. We denote the intervals of integers as $[a, b] = \{a, a+1, \ldots, b\}$, $(a, b] = \{a+1, a+2, \ldots, b\}$, etc., and we denote the shift by m of an interval $I = [a, b)$ by $I + m := [a+m, b+$

m). For notational convenience, we define, for an interval I, $\boldsymbol{\pi}(I) := \prod_{r\in I} \overline{g_r}$. Let τ be such that the distribution after τ steps of random walk is within p^{-2} of the uniform. Set $L := \lfloor t/\tau \rfloor - 1$. We will see that, for an overwhelming fraction of hash functions (or, equivalently, $\boldsymbol{c}$), the following *cancellation property* holds for some $A_i \subseteq [t-(i+1)\tau, t-i\tau)$ (for $1 \le i < L$): $\boldsymbol{\pi}(A_i)\boldsymbol{\pi}([t-i\tau, t-(i-1)\tau) \setminus A_{i-1}) = 1$ (where $A_0 := \emptyset$). Hence,

$$\begin{aligned} z_t = z_\tau \boldsymbol{\pi}([\tau, t)) &= z_\tau \boldsymbol{\pi}([\tau, t-2\tau))\boldsymbol{\pi}([t-2\tau, t-\tau) \setminus A_1) \\ &= z_\tau \boldsymbol{\pi}([\tau, t-3\tau))\boldsymbol{\pi}([t-3\tau, t-2\tau) \setminus A_2) \\ &= \cdots = z_\tau \boldsymbol{\pi}([\tau, t-L\tau))\boldsymbol{\pi}([t-L\tau, t-(L-1)\tau) \setminus A_{L-1}) \ . \end{aligned}$$

That is, the walk beyond τ steps repeatedly introduces a multiplicative factor of $1 \in G$ via subproducts over small (*i.e.*, of length at most 2τ) intervals; this does not mean that the z_is repeat, since the terms in the subproducts need not be consecutive (*i.e.*, the A_i need not be intervals). To be precise, τ is defined to be the minimal value so that if $\mu_\tau \colon G \to [0,1]$ is the distribution of the node z_τ, then $|\mu_\tau(\alpha) - 1/p| < p^{-2}$ (for all α and all starting points z_0 for the walk). The exact values for μ_τ may depend on the starting point or the independence parameter of h, but μ_τ is well-defined without knowing these, up to the additive p^{-2} error term. We call μ_τ (up to this error term) the distribution after τ steps.

Our basic parameters will be w, τ and Δ; here Δ (see Lemma 6) is a lower bound on the length of a walk during which every $g_i \in S$ (alternatively, some constant fraction of S) will almost surely be chosen at least once. First we have three simple lemmata:

Lemma 6. *Let $J = [s, s+\Delta) \subseteq [0, \Lambda]$ be given. Then there is a set $\mathsf{H}_{\text{GOOD}} \subseteq \mathsf{H}$ of size at least $|\mathsf{H}|\left(1 - p^{-3}/4\right)$ for whose members the following hold:*

(a) if $\Delta \ge 5(2n)^2$ and $2n\sqrt{6} \ge w \ge 2n \ge 4\log p$, then $\{\overline{g_j} \ : \ j \in J\} = S$ and

(b) if $\Delta \ge \left(\frac{20}{3}\right) w$ and $w > 2n + 3 + 4\log p$, then there exists a $B \subseteq J$ such that $S' := \{\overline{g_j} \ : \ j \in B\}$ has at least $\frac{1}{4}|S|$ elements.

Lemma 7. *Let $\alpha \in G$ be arbitrary. If $2n > 8\log p$, then, for every $J = [s, s+\Delta) \subseteq [0, \Lambda]$ such that $\{\overline{g_j} \ : \ j \in J\} = S$, $\mathsf{S}_{\text{GOOD}} := \{S \ : \ \exists A \subseteq J \text{ s.t. } \boldsymbol{\pi}(A) = \alpha\}$ has probability at least $1 - p^{-3}/4$. Additionally, the conclusion holds under the weaker requirement that $S' := \{\overline{g_j} \ : \ j \in J\}$ has at least $8\log p$ elements.*

Lemma 8. *Let $s \le \Lambda$ and $\alpha \in G$ be arbitrary. Recall that μ_τ is the probability distribution (defined up to $O(p^{-2})$ error terms) after τ steps of a totally independent random walk. Let H_{GOOD} be any set containing at least a $1 - p^{-3}/4$ fraction of H (and assume the degree of the polynomials d is at least $\tau + \Delta$). Set $I := [s, s+\tau)$ and $J := [s+\tau, s+\tau+\Delta)$. Then, for any $A \subseteq J$, there is a ζ such that $|\zeta| \le p^{-2}$, for which*

$$\begin{aligned} \Pr_{\boldsymbol{c}}[\boldsymbol{\pi}(I) = \alpha\boldsymbol{\pi}(J \setminus A)] &= \Pr_{h_E \in \mathsf{H}}[\boldsymbol{\pi}(I) = \alpha\boldsymbol{\pi}(J \setminus A)] \\ &= \mu_\tau(\alpha\boldsymbol{\pi}(J \setminus A)) = \Pr_{h_E \in \mathsf{H}_{\text{GOOD}}}[\boldsymbol{\pi}(I) = \alpha\boldsymbol{\pi}(J \setminus A)] + \zeta \ . \end{aligned}$$

Remark 3. Lemma 6 holds for every interval $J \subseteq [s, \Lambda]$ and Lemma 8 holds for every interval $(I \cup J) \subseteq [s, \Lambda]$, both because we consider paths of every possible length when performing the run-time analysis of the Cayley rho.

Lemma 9 (Rewind). *Let $\frac{3}{20}\Delta' \geq w \geq 2n \geq 32 \log p$ and $d \geq \Delta' + \tau$. Let $i + \tau < j \leq t \leq \Lambda$ and let $\alpha \in G$ be arbitrary. Then there exist $\mathsf{H}_{\text{GOOD}} \subseteq \mathsf{H}$ and $\mathsf{S}_{\text{GOOD}} \subseteq \mathsf{S}$, each of probability at least $1 - p^{-3}/4$, such that the following holds over $h_E \in \mathsf{H}_{\text{GOOD}}$, $S \in \mathsf{S}_{\text{GOOD}}$: $|\Pr_{h,S}[z_j = \alpha] - \Pr_{h,S}[z_{i+\tau} = \alpha]| \leq p^{-2}$.*

Lemma 10. *Put $\Delta' = \Delta + \tau$. Let i, j, k, ℓ be such that $i + \Delta' < j \leq \Lambda$ and $k + \Delta' < \ell \leq \Lambda$, and $[i, j] \cap [k, \ell] \neq \emptyset \Rightarrow (|i - k| > \Delta'$ and $|j - \ell| > \Delta')$. Let $\alpha, \beta \in G$ be arbitrary. If $d \geq 2\Delta'$, then there are sets $\mathsf{H}_{\text{GOOD}} \subseteq \mathsf{H}$ and $\mathsf{S}_{\text{GOOD}} \subseteq \mathsf{S}$, both of probability at least $1 - p^{-3}/4$, such that $|\Pr[z_j = \alpha | z_\ell = \beta; z_i, z_k] - \Pr[z_j = \alpha | z_i]| \leq p^{-2}$ when the probabilities are viewed over $h_E \in \mathsf{H}_{\text{GOOD}}$ and $S \in \mathsf{S}_{\text{GOOD}}$.*

Now we consider the case when one of the walks is too short to guarantee that it mixes to a uniform distribution.

Lemma 11. *Let $\Delta \geq \Delta' + \tau$, with Δ' as in Lemma 9 and let i, j, k, ℓ be such that $\ell < k + \Delta \leq \Lambda$ and $i + \Delta < j \leq \Lambda$, and $[i, j] \cap [k, \ell] \neq \emptyset \Rightarrow (|i - k| > \Delta$ and $|j - \ell| > \Delta)$. Let $\alpha, \beta \in G$ be arbitrary. If $d \geq 2(\tau + \Delta)$ and $|S| \geq 2\Delta$, then there are sets $\mathsf{H}_{\text{GOOD}} \subseteq \mathsf{H}$ and $\mathsf{S}_{\text{GOOD}} \subseteq \mathsf{S}$, both of probability at least $1 - p^{-3}/4$, such that $|\Pr[z_j = \alpha | z_\ell = \beta; z_i, z_k] - \Pr[z_j = \alpha | z_i]| \leq p^{-2}$ when the probabilities are viewed over $h_E \in \mathsf{H}_{\text{GOOD}}$ and $S \in \mathsf{S}_{\text{GOOD}}$.*

6 Run Time of Cayley Rho

Let $z_0, z_1, \ldots, z_t \in G$ denote the sequence produced by the Cayley rho algorithm $\mathcal{C}$. Define the random variables Y_{ij} to be 0 when $z_i \neq z_j$ and 1 otherwise (for $i, j \in \{0, 1, \ldots, t\}$). Then the number of collisions is $Y := \sum_{i<j} Y_{ij}$. Put $\mu := E_{S,h}[Y]$ and $\sigma^2 = E_{S,h}[(Y - \mu)^2]$. We wish to bound $\varrho := \Pr[Y = 0] \leq \Pr[|Y - \mu| \geq \mu] \leq \frac{\sigma^2}{\mu^2} = \frac{E_{S,h}[Y^2]}{\mu^2} - 1$. In this section we prove

Lemma 12. *There is a $v = O(\log p)$ such that if $t \geq 4v\xi\sqrt{p}$, then $\varrho \leq \xi^{-2}$.*

As $\Delta' + \tau = O(\log p)$, we may choose v so that $v \geq 2(\Delta' + \tau)$, where Δ' and τ are as defined in the previous section. We also will assume in this section that $d \geq \chi(2n)$ (for some constant χ) and that $m < 2n$; we try to prove the result in terms of t, optimal up to a constant factor. A path is called *short* if its length is at most v; otherwise the path is called *long*.

In the proof of the lemma, a technical issue stems from the fact that if $Y_{a,a+L} = 1$ corresponds to a cycle, then $Y_{b,b+qL} = 1$ for every $b \geq a$ and every $q \in \mathbb{N}$. Thus, in the equation $\mathrm{E}_{S,h}\left[Y^2\right] = \sum \mathrm{E}_{S,h}\left[Y_{ij} Y_{k\ell}\right]$, if the shortest cycle corresponds to $Y_{a,a+L} = 1$ (*i.e.*, L is minimal), then, for each $q = 1, 2, \ldots, \lfloor t/L \rfloor$ and $b \geq a$, further cycles occur so that we get $Y_{b,b+qL} = 1$. This will make a significant contribution to $E_{S,h}\left[Y^2\right]$, but these correlations are due to the existence of cycles *after* the first cycle occurs, and we need only to find an upper bound for the probability of the absence of cycles.

To this end, we now construct a new walk which coincides with $\boldsymbol{c}$ up to the first collision. As $h(\alpha)$ (for $\alpha \in G$) was constructed from $h_E(y)$ (for $y \in E$) in Section 2.3, so we construct $h_2(\alpha, \gamma)$ (for $\alpha \in G$ and $\gamma \in \{0, 1, \ldots, \Lambda^2\}$) from $h_E(\gamma \circ \alpha)$ ($\circ$ denotes concatenation; γ will be prefixed with the necessary leading zeroes for it to occupy the $2\lceil \log_2 p \rceil$ most-significant bits of the input to h_E). We construct h_2 so $h_2(\alpha, 0) = h(\alpha)$ for all $\alpha \in G$. Given a random walk $\boldsymbol{c}$ and an h_2 constructed from a given h_E, we define a *modified random walk* $\widetilde{\boldsymbol{c}}$ as follows: $\widetilde{\boldsymbol{c}}(0) := h_2(\widetilde{z}_0, \gamma_0)$ and $\widetilde{\boldsymbol{c}}(i) := h_2(\widetilde{z}_i, \gamma_i)$, where $\gamma_i = |\{s \in \{0, 1, \ldots, i\} \ : \ \exists s' < s \ \text{ s.t. } \ \widetilde{z}_s = \widetilde{z}_{s'}\}| < \Lambda^2$ and the $\widetilde{z}_i$ are defined so $\widetilde{z}_{i+1} = \widetilde{z}_i g_{\widetilde{\boldsymbol{c}}(i)}$ and $\widetilde{z}_0 = z_0$ (the z_0 associated with the original walk $\boldsymbol{c}$). Now we have a simple lemma:

Lemma 13. *Let h_E be a random polynomial of degree $d \geq m$. Fix z_0. Then the following hold: (a) for any $t \leq \Lambda$, if $\gamma_t = 0$, then the modified random walk $\widetilde{\boldsymbol{c}}$ agrees with $\boldsymbol{c}$ and they both generate the same sequence $\{z_i\}_{i=0}^{t}$; (b) the following three are m-wise independent functions of their input (α, γ, and α, respectively): $h_\gamma(\alpha) := h_E(\gamma \circ \alpha)$, $h_\alpha(\gamma) := h_E(\gamma \circ \alpha)$ and $h_E(0 \circ \alpha) = h_E(\alpha)$; (c) the modified walk $\{\widetilde{\boldsymbol{c}}_i\}_{i=0}^{t-1}$ is an m-wise independent sequence; and (d) the probability (over h_E) that there is no collision is the same for both $\widetilde{\boldsymbol{c}}$ and $\boldsymbol{c}$.*

We point out that after the first collision (*i.e.*, when $\gamma_i > 0$), the walks $\widetilde{\boldsymbol{c}}$ and $\boldsymbol{c}$ can be markedly different. The modified walk is likely hard to implement in full generality without increasing time or space requirements significantly.

For every fixed $S \in \mathsf{S}$, when the path from z_i to z_j is long, $\Pr_h[Y_{ij} = 1]$ is only approximately $1/p$ (within p^{-2} error). However,

Lemma 14. *If z_i and z_j are endpoints of a short path, then $E_{S,h}[Y_{ij}] = 1/p$.*

Proof. Let $L := j - i \leq v$ be the length of the path $X \in \Omega_L$ from z_i to z_j. Then,

$$\begin{aligned}
\mathrm{E}_{S,h}[Y_{ij}] &= \Pr_{S,\boldsymbol{c}}[Y_{ij} = 1] = \Pr_{S,X=\psi(\boldsymbol{c})}[\phi_S(X) = 1] \\
&= \sum_{\overline{X} \in \Omega_L} \underbrace{\Pr_S[\phi_S(\overline{X}) = 1 | X = \overline{X}]}_{\text{Lemma 2}} \Pr_{\boldsymbol{c}}[X = \overline{X}] = \sum_{\overline{X} \in \Omega_L} \frac{1}{p} \Pr_{\boldsymbol{c}}[X = \overline{X}] = \frac{1}{p} \, .
\end{aligned}$$

Notice that $\Pr_S[\phi_S(\overline{X}) = 1 | X = \overline{X}]$ is well-defined and that Lemma 2 can be applied to compute it. This is consistent with the intuitive observation that on short distances, the Cayley rho walk appears independent. □

Now define $U := \{(i, j, k, \ell) \ : \ 0 \leq i < j \leq t \ \text{ and } \ 0 \leq k < \ell \leq t\}$; each element of U represents a pair of paths: one from i to j and one from k to ℓ. Recall that $v > 2(\Delta' + \tau)$, where Δ' and τ are as in the previous section. Define K to be the set of tuples $(i, j, k, \ell) \in U$ containing entries that satisfy the assumptions of Lemmata 10 and 11; $\overline{K}$ will denote $U \setminus K$. Thus, $\overline{K}$ contains path pairs that are (a) both *short* (*i.e.*, of length at most v) or (b) one of the paths is long and the other one is short but the short one has its end points within a short distance from the endpoints of the long path. Thus $\overline{K}$ contains at most $\binom{t+1}{2}(2v)^2 + \binom{t+1}{2}(2v)^2 = 8v^2\binom{t+1}{2}$ many 4-tuples.

Lemma 15. *If* $(i,j,k,\ell) \in \overline{K}$*, then* $|\Pr[Y_{ij}=1|Y_{k\ell}=1] - 1/p| = O(p^{-2})$.

Proof. We use the basic relations $\Pr[A|B] = \sum_i \Pr[A|BC_i]\Pr[C_i|B]$ (where $\{C_i\}$ is a partition) and $\Pr[A|B] \leq \Pr[A]/\Pr[B]$. Now, let C_1 be the event $(h,S) \in \mathsf{H}_{\text{GOOD}} \times \mathsf{S}_{\text{GOOD}}$ and C_2 its complement. Recall that $\Pr[C_2] = O(p^{-3})$. Then, $\Pr_{S,h}[Y_{k\ell}=1|Y_{ij}=1]$ equals

$$\underbrace{\Pr_{S,h}[Y_{k\ell}=1|Y_{ij}=1;C_1]}_{\text{Lemma 10}}\Pr_{S,h}[C_1|Y_{k\ell}=1] + \Pr_{S,h}[Y_{ij}=1|Y_{k\ell}=1;C_2]\Pr_{S,h}[C_2|Y_{k\ell}=1]\ ,$$

which is no more than $\left(\frac{1}{p} + O(p^{-2})\right)\cdot 1 + 1\cdot\frac{\Pr_{S,h}[C_2]}{\Pr_{S,h}[Y_{k\ell}=1]} \leq \frac{1}{p} + O(p^{-2}) + \frac{O(p^{-3})}{1/p} = \frac{1}{p} + O(p^{-2})$. □

Now we can finish our proof of the lower bound for the probability that a cycle exists (Lemma 12). We notice that $\mathrm{E}_{S,h}[Y^2] = \sum_{\substack{i<j \\ k<\ell}} \mathrm{E}_{S,h}[Y_{ij}Y_{k\ell}] = \sum_{\overline{K}} \mathrm{E}_{S,h}[Y_{ij}Y_{k\ell}] + \sum_K \mathrm{E}_{S,h}[Y_{ij}Y_{k\ell}]$ and proceed to bound each term.

$$\sum_K \mathrm{E}_{S,h}[Y_{ij}Y_{k\ell}] = \sum_K \underbrace{\Pr[Y_{ij}=1|Y_{k\ell}=1]}_{\text{Lemma 15}}\Pr[Y_{k\ell}=1] \leq \sum_K \left(\frac{1}{p} + O(p^{-2})\right)\frac{1}{p}$$
$$= |K|\left(p^{-2} + O(p^{-3})\right) = \left(\binom{t+1}{2}^2 - |\overline{K}|\right)\left(p^{-2} + O(p^{-3})\right)\ .$$

In the complementary range, $\sum_{\overline{K}} \mathrm{E}_{S,h}[Y_{ij}Y_{k\ell}] \leq \sum_{\overline{K}} \mathrm{E}_{S,h}[Y_{ij}] = \sum_{\overline{K}} \frac{1}{p} = |\overline{K}|\frac{1}{p}$. Finally, by Lemma 14, $\mathrm{E}_{S,h}[Y] = \sum Y_{ij} = \sum_{i<j}\frac{1}{p} = \binom{t+1}{2}\frac{1}{p}$. Combining the results, we see that

$$\varrho \leq \frac{E_{S,h}[Y^2]}{E_{S,h}[Y]^2} - 1 \leq \left(\binom{t+1}{2}\frac{1}{p}\right)^{-2}\left[|\overline{K}|\frac{1}{p} + \frac{\binom{t+1}{2}^2 - |\overline{K}|}{p^2}\left(1 + O\left(\frac{1}{p}\right)\right)\right] - 1\ ,$$

which is less than $16v^2p/t^2$. Hence, when $t \geq 4v\xi\sqrt{p}$, we get $\varrho \leq \xi^{-2}$. □

We thank Prasad Tetali for generous discussions related to random walks on directed graphs.

References

1. N. Alon and Y. Roichman, "Random Cayley Graphs and Expanders." *Random Structures and Algorithms*, **5**:271–284, 1994.
2. E. Bach, "Toward a Theory of Pollard's Rho Method." *Information and Computation*, **90**(2):139–155, 1991.

3. M. Bellare and J. Rompel, “Randomness-Efficient Oblivious Sampling.” *Symposium on Foundations of Computer Science (FOCS ’94)*:276–287, 1994.
4. B. Bollobas, *Modern Graph Theory*, Graduate Texts in Mathematics **184**. Springer-Verlag, Berlin, 1998.
5. A. Broder and E. Shamir, “On the Second Eigenvalue of Random Regular Graphs.” *Symposium on the Foundations of Computer Science (FOCS ’87)*:286–294, 1987.
6. P. Erdős and A. Rényi, “Probabilistic Methods in Group Theory.” *Journal d’Analyse Mathématique*, **14**:127–138, 1965.
7. S.C. Pohlig and M.E. Hellman, “An Improved Algorithm for Computing Logarithms over GF(p) and Its Cryptographic Significance.” *IEEE Transactions on Information Theory*, **24**:106–110, 1978.
8. J.M. Pollard, “Monte Carlo Methods for Index Computation (mod p).” *Mathematics of Computation*, **32**(143):918–924, 1978.
9. Y. Roichman, “On Random Random Walks.” *Annals of Probability*, **24**(2):1001–1011, 1996.
10. V. Shoup, “Lower Bounds for Discrete Logarithms and Related Problems.” *Advances in Cryptology: EUROCRYPT ’97 (LNCS 1233)*:256–266, 1997.
11. E. Teske, “Speeding Up Pollard’s Rho Method for Computing Discrete Logarithms.” *Algorithmic Number Theory Symposium III: ANTS-III (LNCS 1423)*:541–554, 1998.
12. E. Teske, “On Random Walks for Pollard’s Rho Method.” *Mathematics of Computation*, **70**:809–825, 2001.

The Function Field Sieve Is Quite Special

Antoine Joux[1] and Reynald Lercier[2]

[1] DCSSI Crypto Lab
51, Bd de Latour Maubourg
F-75700 Paris 07 SP
France
Antoine.Joux@m4x.org
[2] CELAR
Route de Laillé
F-35570 Bruz
France
lercier@celar.fr

Abstract. In this paper, we describe improvements to the function field sieve (FFS) for the discrete logarithm problem in $\mathbb{F}p^n$, when p is small. Our main contribution is a new way to build the algebraic function fields needed in the algorithm. With this new construction, the heuristic complexity is as good as the complexity of the construction proposed by Adleman and Huang [2], i.e $L_{p^n}[1/3, c] = \exp((c + o(1)) \log(p^n)^{\frac{1}{3}} \log(\log(p^n))^{\frac{2}{3}})$ where $c = (32/9)^{\frac{1}{3}}$. With either of these constructions the FFS becomes an equivalent of the special number field sieve used to factor integers of the form $A^N \pm B$. From an asymptotic point of view, this is faster than older algorithm such as Coppersmith's algorithm and Adleman's original FFS. From a practical viewpoint, we argue that our construction has better properties than the construction of Adleman and Huang. We demonstrate the efficiency of the algorithm by successfully computing discrete logarithms in a large finite field of characteristic two, namely $\mathbb{F}2^{521}$.

1 Introduction

Due to their cryptographic significances, the integer factorization problem and the discrete logarithm problem in finite fields have been extensively studied in the last decades. The best methods currently known to solve these problems are index calculus techniques. In the field of integer factorization, the number field sieve (NFS) [15] having surpassed its ancestor, the quadratic sieve [24], is now the fastest of the known factoring algorithms. It exists in two flavors, the general number field sieve which can factor any integer and the special number field sieve which is useful for numbers of a special form: many integers of the form $A^N \pm B$ were factored using the special number field sieve. The latest example is the factorization of $2^{773} + 1$ at CWI [21].

For the computation of discrete logarithms in prime fields, the situation is similar. The quadratic sieve has an analog called the gaussian integer method [6,

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 431–445, 2002.

13]. Similarly a variant of the number field sieve [23] can be used for computing discrete logarithms. Furthermore, as for factorization, we can distinguish between the general number field sieve and the special number field sieve. For example a computation of discrete logarithms modulo $p = (739 \cdot 7^{149} - 736)/3$ can be found in [27].

In this paper, we address the case of discrete logarithm computations in $\mathbb{F}_{p^n}$, when p is small. The best known practical method for the typical case $p = 2$ is due to Coppersmith [5]. It was used in 1992 by Gordon and McCurley [11] to compute discrete logarithms in $\mathbb{F}_{2^{401}}$. In the same paper, the sieving part of the discrete logarithm computation for $\mathbb{F}_{2^{503}}$ was also reported. More recently, the sieving part of a discrete logarithm computation in $\mathbb{F}_{2^{607}}$ using Coppersmith's method has been performed [25]. The linear algebra step of this computation has been finished very recently [26]. From a more theoretical viewpoint, there exists an analog of the general number field sieve called the function field sieve (FFS), which is not restricted to $p = 2$ [1, 2]. From a practical viewpoint, only Coppersmith's algorithm was considered by now in the characteristic two case [7].

Adleman and Huang [2] showed that the asymptotics of the function field sieve can be largely improved. In fact, with these improvements the function field sieve becomes an equivalent of the special number field sieve. In this paper, we propose a different method that achieves the same complexity. Moreover, in most cases, our method is faster. As a consequence, both the asymptotic complexity and the practical implementation turn out to be better than in older works. We finally illustrate this result by a computation in $\mathbb{F}_{2^{521}}$.

2 Algorithmic Considerations

The function field sieve was introduced in [1] for computing discrete logarithms in $\mathbb{F}_{p^n}$ with small values of p. It is quite similar to the number field sieve and it has a complexity of the same order $L_{p^n}[1/3]$. However, there are some crucial differences that allow large improvements. Most notably, a field such as $\mathbb{F}_{p^n}$ can be represented in many different ways. In Coppersmith's algorithm [5], as in the work of Adleman and Huang [2], the key idea was to select a "small representation" of the finite field, more precisely, this is done by selecting a polynomial $\lambda(t)$ of degree as low as possible, such that $t^n - \lambda(t)$ is irreducible. Once $\lambda(t)$ is chosen, it is possible to find two good polynomials having a common root in this field. Clearly, these two constructions focus on a small subset of the possible representations of $\mathbb{F}_{p^n}$. In this paper, we propose a construction that allows a much larger varieties of possible representations. This extra degree of freedom reduces the task of choosing good polynomials at the beginning of the function field sieve algorithm. It turns out that this method keeps the good complexity proved by Adleman and Huang (cf. section 3). Moreover, the selected polynomials are somewhat better. In term of the asymptotic complexity, this is hidden in the $o(1)$, however this yields a significant decrease of the practical run times. In this section, we mostly focus on our new polynomial selection phase.

As a foreword, let us recall that the method we use to compute discrete logarithms in a field $\mathbb{F}_{p^n}$ is derived from the well known Pohlig–Hellman method. The first step is to factor $p^n - 1$. For any small factor ℓ of $p^n - 1$, discrete logarithms modulo ℓ can be found using the Pollard Rho method. For the remaining prime factors ℓ of $p^n - 1$, we use the index-calculus method described here. Finally, we combine the results thanks to the Chinese Remainder Theorem in order to get a result modulo $p^n - 1$.

Therefore, in the sequel, we will assume that we are computing discrete logarithms modulo a large prime factor ℓ of $p^n - 1$. This is not an issue since computing the prime factors of $p^n - 1$ can be done with the special number field sieve with a complexity of the same order $L_{p^n}[1/3]$.

2.1 Representation of $\mathbb{F}_{p^n}$

One classical way to work with $\mathbb{F}_{p^n}$ consists in handling equivalence classes in the quotient of the commutative ring $\mathbb{F}_p[t]$ by one of its proper maximal ideals $f(t)\mathbb{F}_p[t]$ where $f(t)$ is an irreducible element of $\mathbb{F}_p[t]$ of degree n. Each equivalence class is then uniquely determined by a polynomial of $\mathbb{F}_p[t]$ of degree strictly smaller than n. Consequently, any element of $\mathbb{F}_{p^n}$ can be seen as a polynomial of degree smaller than n. With such a representation, adding two elements of $\mathbb{F}_{p^n}$ is the same as adding two elements of $\mathbb{F}_p[t]$. Multiplying two elements of $\mathbb{F}_{p^n}$ is the same as multiplying two elements of $\mathbb{F}_p[t]$ and reducing the result modulo $f(t)$.

Since there are numerous irreducible elements of degree n in $\mathbb{F}_p[t]$, there are numerous ways to represent $\mathbb{F}_{p^n}$. The computation of the map between two representations consists in computing the roots over one representation of $\mathbb{F}_{p^n}$ of a polynomial of degree n whose coefficients are in $\mathbb{F}_p$. From an algorithmic viewpoint, this is known as the "equal degree factorization" problem. This can be done quite efficiently since there exists algorithms for this task whose complexity is polynomial in $\log p^n$ (a good survey can be found in [16]). As a consequence, if some particular representation of $\mathbb{F}_{p^n}$ is well suited to discrete logarithm computations, it is a simple matter to switch from a given representation to the more adapted one. In the sequel, we take that step for granted and forget about the given initial representation.

2.2 General Principle of the FFS

The Function Field Sieve algorithm is an "index-calculus" method. So it can be seen at a high level of abstraction as a two steps algorithm.

Step 1: One fixes a subset $S = \{\gamma_1, \ldots, \gamma_{|S|}\}$ of $\mathbb{F}_{p^n} \simeq \mathbb{F}_p[t]$ called the factor base and tries to collect relations between products of elements of S. So, we have equations of the form

$$\sum_{(\epsilon,\gamma)\in\mathbb{Z}\times S} \epsilon \log_x \gamma = 0 \qquad (1)$$

where x is a generator of the multiplicative subgroup of order ℓ in $\mathbb{F}_{p^n}$. When enough such relations are collected, one obtains the quantities $\log_x \gamma$ via the inversion modulo ℓ of the corresponding linear system.

Step 2: To find the discrete logarithm of an element y which is not in S, one tries random integers ν until $x^\nu y$ is a product of elements of S. Then

$$\log_x y = \left(-\nu + \sum_{(\epsilon,\gamma) \in \mathbb{Z} \times S} \epsilon \log_x \gamma \right) \bmod \ell. \tag{2}$$

The way how the factor base S is chosen is specific to each variation. In the original Function Field Sieve as described by Adleman, the factor base is the image by a morphism ϕ in $\mathbb{F}_{p^n}$ of the generators of two sets S_α and S_β. The set S_α is a set of $\mathbb{F}_p$-rational principal places in the rational function field $\mathbb{F}_p(t)$. The set S_β is a set of $\mathbb{F}_p$-rational principal places defined in an algebraic function field.

Once a random polynomial $\mu(t) \in \mathbb{F}_p[t]$ has been chosen, this function field is defined by an absolutely irreducible bivariate polynomial $H(t, X) = \sum_{i=0}^{d} \sum_{j=0}^{d'} h_{i,j} X^i t^j$ such that $H(t, \mu(t)) = 0 \bmod f(t)$. The mapping ϕ from this algebraic function field to $\mathbb{F}_{p^n}$ is then easily defined by $X \longrightarrow \mu(t)$.

The algorithm consists in finding couples $(r(t), s(t)) \in \mathbb{F}_p[t]^2$, where $r(t)$ and $s(t)$ are relatively prime, such that the polynomial $r(t)\mu(t) + s(t)$ can be written as a product of irreducible polynomials in S_α and such that the divisor associated to the function $r(t)X + s(t)$ can be written as a sum of places in S_β. Following Adleman, such a pair $(r(t), s(t))$ is called "doubly smooth" since $r(t)\mu(t) + s(t)$ is smooth and $r(t)X + s(t)$ is smooth in the sense that the norm over $\mathbb{F}_p[t]$ of $r(t)X + s(t)$ is smooth.

Thanks to eight technical conditions given on $H(t, X)$ by Adleman, these equalities in terms of divisors can be seen as equalities in terms of functions, once raised to the order h of the jacobian power. One can apply the morphism ϕ to get a relation in $\mathbb{F}_{p^n}$. Applying also this morphism on the rational side in the same manner finally yields a relation of the form (1).

2.3 Choice of the Polynomials

In full generality, as explained in [8] for the NFS case, the function field sieve requires two polynomials $f_\alpha(X)$ and $f_\beta(X)$ with a common root μ in $\mathbb{F}_{p^n}$. For the algorithm to be efficient, these polynomials should have small coefficients.

The method suggested in [1] is an adaptation of the base m technique used in NFS [3] to the function field case. The method works as follows: choose a polynomial $m(t)$ and write the definition polynomial $f(t)$ of $\mathbb{F}_{p^n}$ in base $m(t)$ as $\sum h_i(t)\mu(t)^i$. Then $X - \mu(t)$ and $H(t, X) = \sum h_i(t) X^i$ clearly have the common root $\mu(t)$ in $\mathbb{F}_{p^n}$. Thus, we get a rational side (corresponding to the degree one polynomial) and an algebraic side. For the number field sieve, several techniques for the polynomial construction lead to two polynomials of degree greater than one. In this case, we no longer have a rational side, which leads to technical difficulties in the later phases of the number field sieve [23].

The version of FFS suggested by Adleman and Huang in [2] is asymptotically much faster. It works by selecting $f(t)$, the polynomial describing the field representation, to be of the form $f(t) = t^n + \lambda(t)$ where $\lambda(t)$ is of degree as low as possible. Then, they choose a parameter d, let $e = \lceil n/d \rceil$ and construct two polynomials $H(t, X) = X^d + t^{ed-n}\lambda(t)$ and $X - t^e$, with common root $\mu(t) = t^e$. In fact, Coppersmith's algorithm [5] can be seen as a subcase of the algorithm of Adleman and Huang, when $p = 2$ and d is a power of two.

We present here a new technique to build good polynomials in the function field case. As the version of Adleman and Huang, this technique is specific to $\mathbb{F}_{p^n}$ and in general cannot be applied in the number field sieve.

The basic idea is simple, we do the construction backward. Instead of choosing a definition polynomial $f(t)$ beforehand, we only fix p^n. Then we choose a polynomial $H(t, X) = \sum h_i(t)X^i$ of degree d in X (the exact value of d will be determined during the complexity analysis) whose coefficients $h_i(t)$ are polynomials in $\mathbb{F}_p[t]$ with very small degrees in t. Afterward, we choose random polynomials $\mu_1(t)$ and $\mu_2(t)$ of degree at most $\lfloor n/d \rfloor$ in t and check whether $f(t) = \mu_2(t)^d H(t, -\mu_1(t)/\mu_2(t))$ is an irreducible polynomial of degree n over $\mathbb{F}_p$. If the test is successful, we are done, otherwise, we choose another pair $(\mu_1(t), \mu_2(t))$ and restart. Of course, it is essential to correctly choose the coefficients of H to guarantee that f can be of degree n. This implies that the degree of at least one coefficient in H should be the remainder of the division of n by d. Thus the coefficients of H cannot be of arbitrary small degree, however their degrees can be smaller than d in all cases. Moreover, some care should be taken when choosing H. We discuss this point in the next section.

To compare our construction with that of Adleman and Huang, we need to compare the size (degree in t) of the coefficients involved in the two polynomials $H(t, X)$ and $\mu_2(t)X + \mu_1(t)$ (resp. $X - \mu(t)$). A simple way to perform this comparison is to compute the resultant of the two polynomials and compare the respective degrees. With our method, the degree is exactly n as explained above. With the method of Adleman and Huang, the degree is ed and varies from n to $n + d - 1$. Unless d divides n, our construction leads to smaller polynomials and thus to a faster algorithm.

In practice, Coppersmith's algorithm is the only one which has been considered for computing discrete logarithms in large finite fields of small characteristic. When writing its complexity as

$$L_{p^n}[1/3, c] = \exp((c + o(1)) \log(p^n)^{\frac{1}{3}} \log(\log(p^n))^{\frac{2}{3}}),$$

we get a value of c between $(32/9)^{\frac{1}{3}}$ and $c = 4^{\frac{1}{3}}$. More precisely, with Coppersmith's algorithm the value of c is not a constant, since there are good cases and bad cases. At best, we have $c = (32/9)^{\frac{1}{3}}$ and at worst $c = 4^{\frac{1}{3}}$, this is always better than Adleman's FFS where $c = (64/9)^{\frac{1}{3}}$. With our construction or that of Adleman and Huang, we have $c = (32/9)^{\frac{1}{3}}$ in all cases. Thus, from a theoretical viewpoint, our algorithm has a larger scope and is faster than Coppersmith's.

Indeed, it can be used with a characteristic different from 2. In practice, our algorithm is faster in characteristic two than Coppersmith's whenever the optimal choice of degree for H does not turn out to be a power of 2.

2.4 Number Theoretical Conditions on the Chosen Polynomial

When writing down the equation associated to a smooth pair, we must be careful and be sure that these equations really make sense. This involves two technical difficulties.

The first difficulty is that we should not forget any valuation on the algebraic side of the equation. However, when factoring the norm of an ideal, we miss the valuations at infinity. Thus, we need to add these valuations when writing down the equation. In [1], this was done by choosing an algebraic field with several valuations at infinity and by using dehomogenization techniques to compute these valuations. However, this approach is quite cumbersome, specially when writing down the equation. Ideally, we would like to ignore the valuations at infinity. This is possible with the use of so-called "$C_{a,b}$ curves".

Theorem: [18] Let K be a perfect field, $\overline{K}$ the algebraic closure of K, $C_{a,b} \subset \overline{K}$ be a possibly reducible affine algebraic set defined over K, t, X be the coordinates of the affine space, and a, b relatively prime positive integers. Then the following conditions are equivalent.

- $C_{a,b}$ is an absolutely irreducible affine algebraic curve with exactly one K-rational place P_∞ at infinity and the pole divisor of t and X are aP_∞ and bP_∞ respectively.
- $C_{a,b}$ is defined by a bivariate polynomial of the form

$$H(t,X) = h_{a,0}X^a + h_{0,b}t^b + \sum_{ib+ja<ab} h_{i,j}X^it^j \tag{3}$$

  where $h_{i,j} \in K$ for all i, j and $h_{0,b}$, $h_{a,0}$ are nonzero.

As outlined in [18], any bivariate polynomial $H(t,X)$ of the form (3) is absolutely irreducible. So, only two conditions on H among the eight conditions initially given by Adleman must be satisfied.

1. $f(t)$ divides $\mu_2(t)^dH(t,-\mu_1(t)/\mu_2(t))$.
2. The order of the jacobian of the curve defined by $H(t,X)$ is relatively prime to $(p^n-1)/(p-1)$.

The second technical difficulty is the existence of an obstruction group that voids the validity of the equations in certain cases. In the case of the number field sieve, the obstruction group is discussed in [23] and [12]. This obstruction group has two components. The first one comes from the group of units and the second one from the class group of this field. Dealing with the first component is quite difficult and relies on certain maps introduced by Schirokauer. However, dealing with the class group is extremely easy as long as its order is relatively prime with ℓ (the cardinality of the subgroup in which we are computing discrete logarithms). Indeed, in that case, we can simply forget the existence of the class group and everything falls out correctly. For a detailed explanation see [12].

With the FFS, units are handled by Adleman by the valuations at infinity. With $C_{a,b}$ curves, we only have one valuation at infinity, and so, the only units are the elements of $\mathbb{F}_p$. Thus the only obstruction stems from the class group. Since H has a small degree and small coefficients, the class number h is always small. As recalled in [22], $h \leq (\sqrt{p}+1)^{2g}$ where g is the genus of the function field defined by $H(t, X)$. Since in our construction, the polynomial $H(t, X)$ has very small degrees in X and t, g is always small and it is feasible to compute the order of the jacobian to check condition 2. So, since ℓ is supposed to be a large prime because part of the logarithm in small multiplicative subgroup can be determined by other techniques, ℓ and the class number are always relatively prime. Thus, the obstruction group when using this variation of the FFS completely vanishes and the conditions given above can always be considered as satisfied.

2.5 Linear Algebra

The linear algebra consists of two sub-steps, the structured gaussian elimination and an iterative solver based on Lanczos' algorithm.

The way we implement the structured gaussian elimination is completely described in [12]. Lanczos' algorithm is described in [14].

2.6 Computing Individual Logarithms

When computing discrete logarithm with the number field sieve or the function field sieve, finding logarithms of individual numbers is not a negligible task. Indeed, in the theoretical studies of these algorithms, both O. Schirokauer for the number field sieve [23] and L. Adleman for the function field sieve [1] suggest methods where the whole computation essentially needs to be redone for each new logarithm. From a computational viewpoint, this is not acceptable.

However, in [5] a different method was suggested by Coppersmith. A similar method also exists in the large characteristic case [12]. From a theoretical point of view, the complexity of computing an individual logarithm is once again $L_{p^n}[1/3]$, thus it is comparable to Schirokauer's and Adleman's methods. A first attempt at analyzing this approach can be found in [22] (some insights about the complexities are given). In practice, it turns out to be quite efficient.

We now describe Coppersmith's method and adapt it to our construction. The method consists in two steps. In the first step, the individual logarithm computation is split into the logarithm computation of several smaller polynomials, dubbed medium-sized [5]. More precisely, starting from a polynomial $y(t)$, we randomize it by computing $z(t) = x(t)^\nu y(t)$, where $x(t)$ is an element of the factor base. We further write $z(t) = z_1(t)/z_2(t)$, where $z_1(t)$ and $z_2(t)$ have degrees around $n/2$, using the extended Euclidean algorithm. Then we check whether $z_1(t)$ and $z_2(t)$ are smooth with respect to a smoothness bound $L_{p^n}[2/3]$. We now need to compute the logarithm of many polynomials of degree $L_{p^n}[2/3]$.

In the second step, each of these logarithms is further splitted into the logarithm of even smaller polynomials. We stop when all polynomials are below the smoothness bound used in the preprocessing stage.

3 Heuristic Complexity Analysis for Small, Fixed p

In order to give an heuristic analysis of the complexity of an index calculus method, the traditional approach is to assume that the objects we are trying to factor over the chosen factor base in order to find equations behave like random objects. Under such an heuristic assumption, it is quite easy to quantify the probability of smoothness. This leads in turn to a precise evaluation of the number of couples $(r(t), s(t))$ we need to try before getting sufficiently many equations. With our variant of the function field sieve, candidate equations have two sides which need to be factored. In order to simplify the analysis, the classical approach [3, 1] that we now follow is to assume that the factor bases on the left and the right hand sides contain the same irreducible polynomials. In that case, we can group the two sides in a single polynomial by multiplying together the polynomials coming from the left and the right hand sides[1].

In order to evaluate the probability of smoothness, we use the following result, which can be found in [19, 17]. Let $\mathcal{P}(k, m)$ denotes the probability for a random polynomial of degree k to factor into irreducible polynomials of degree lower than or equal to m. Then, when $k^{1/100} \leq m \leq k^{99/100}$, i.e. in all the range of interest in our case, we have:

$$\mathcal{P}(k, m) = \exp\left((1 + o(1))\frac{k}{m}\log\frac{m}{k}\right).$$

Assuming that $r(t)$ and $s(t)$ are polynomials of degree lower or equal than l, we find that the degree of the linear side of a candidate equation is at most $l + \lfloor n/d \rfloor$. Similarly, on the other side, testing the function $r(t)X + s(t)$ for smoothness yields polynomials $r(t)^d H(t, s(t)/r(t))$ the degrees of which are bounded by $dl + d$. Indeed, as seen in section 2.3, $H(t, X)$ is polynomial of degree d in X whose coefficients have degree lower than d in t. When multiplying the two sides together, we get a polynomial of degree $dl + d + l + \lfloor n/d \rfloor$. Going through all the possible pairs $(r(t), s(t))$ such that $\gcd(r(t), s(t)) = 1$, we need to find enough smooth pairs. In fact, the number of pairs $(r(t), s(t))$ such that $\gcd(r(t), s(t)) = 1$ is a constant fraction of all pairs. In the sequel, we estimate this number by p^{2l}. We need as many smooth pairs as the number of elements in the factor base. Each factor base contains about p^{m+1}/m elements, since this number counts all the unitary polynomials of degree up to m. Thus, counting both factor bases, we can give an upper bound of $4p^{m+1}/m$ smooth pairs needed. This can be approximated to p^m in an asymptotic approach. Assuming that m and l are

[1] Because of this approximation, the complexity stated at the end of this section is clearly an upper bound of the heuristic complexity of the algorithm. However, as far as we known, a more precise analysis would not yield a better complexity.

already fixed, we can find the optimal value of d by minimizing the total degree $dl + d + l + \lfloor n/d \rfloor$. Asymptotically, we can forget about rounding to the nearest integer and minimize $dl + d + l + n/d = n/d + (d+1)(l+1) - 1$. Of course, d should be rounded to the nearest integer, which leads to

$$d = \left\lceil \sqrt{\frac{n}{l+1}} \right\rfloor .$$

Replacing d by its value, we find that the total degree is approximately $2\sqrt{n(l+1)}$. Moreover, in order to balance the complexity of the sieving phase, which is quadratic in p^l, and of the linear algebra phase, which is quadratic in p^m when using sparse techniques, we need to choose $l = m$. In order to get enough smooth pairs, the smoothness probability should be of the order of p^{-m}. Taking logarithm, we need to satisfy the following equation:

$$\log \mathcal{P}(2\sqrt{n(l+1)}, m) \approx -m \log p, \text{ i.e,}$$
$$\frac{2\sqrt{n(m+1)}}{m} \log \left(\frac{2\sqrt{n(m+1)}}{m} \right) \approx m \log p.$$

Expressing p^m as $L_{p^n}[1/3, c] = \exp((c+o(1)) \log(p^n)^{\frac{1}{3}} \log(\log(p^n))^{\frac{2}{3}})$, we can write $m = (c + o(1))n^{\frac{1}{3}} \log_p(n)^{\frac{2}{3}}$ and get the following equation on c,

$$\frac{2}{3\sqrt{c}} = c.$$

Thus, we find $c = (4/9)^{\frac{1}{3}}$. Since the complexity of the algorithm is quadratic p^m, it can be written as $L_{p^n}[1/3, 2c] = L_{p^n}[1/3, (32/9)^{\frac{1}{3}}]$.

We conclude that the complexity of discrete logarithm computations in $\mathbb{F}_{p^n}$, when p is small and fixed, is in fact the same as the complexity of factoring special integers with the special number field sieve.

4 Implementation Choices

4.1 Sieving

Sieving is done in $\mathbb{F}_{p^n}$ in a similar way as this is done in $\mathbb{F}_p$ following a now traditional "sieving by vectors" with special–q technique as introduced by Odlyzko for Coppersmith's algorithm [19] or by Pollard for factorizing integers [20]. In fact, we have not implemented any efficient line sieving as proposed in [5] or as generalized in [9].

Using special–q means here that in order to get sufficiently many relations, we sieve many independent sets of values for $(r(t), s(t))$. Each set is defined by an irreducible polynomial $q(t)$ called the special–q, and contains pairs $(r(t), s(t))$ such that $q(t)$ divides $r(t)\mu_1(t) + s(t)\mu_2(t)$. If $\boldsymbol{u}$ and $\boldsymbol{v}$ form a basis of the corresponding lattice, then $(r(t), s(t))$ can be written as $k_u(t)\boldsymbol{u} + k_v(t)\boldsymbol{v}$. Then we can with simple linear algebra send the lattice corresponding to any small prime ideal from the $(r(t), s(t))$ representation to the $(k_u(t), k_v(t))$ representation. Sieving is then done in a big rectangle in the $(k_u(t), k_v(t))$ space by successively marking the points on each of the small prime lattices. A basis for these small prime ideals can be easily obtained from the factorization of $H(t, X)$ modulo the norm of the ideal but the resulting coordinates have size close to the norm of the ideal. We improve this by combining the vectors of this basis in order to get a new basis with coordinates of degree approximately twice as small. This is done in full generality in our implementation by using an adaptation to the ring $\mathbb{F}_2[t]$ of the well-known algorithm of Gauss for reducing lattices in dimension two. We give pseudo-code for this reduction algorithm in figure 1.

```
– Input: A basis of the lattice $(\boldsymbol{u}, \boldsymbol{v})$ with $\boldsymbol{u} = (u_1, u_2)$ and $\boldsymbol{v} = (v_1, v_2)$.
– Output: A reduced basis.
– Step 1: Let $d_u = \max(\deg(u_1), \deg(u_2))$ and $d_v = \max(\deg(v_1), \deg(v_2))$. If $d_u > d_v$, exchange $\boldsymbol{u}$ and $\boldsymbol{v}$.
– Main loop: Do
    • Let $\delta_1 = \deg(v_1) - \deg(u_1)$, $\delta_2 = \deg(v_2) - \deg(u_2)$.
    • Let $\boldsymbol{w^{(1)}} = \boldsymbol{v} - t_1^{\delta} \cdot \boldsymbol{u}$.
    • Let $\boldsymbol{w^{(2)}} = \boldsymbol{v} - t_2^{\delta} \cdot \boldsymbol{u}$.
    • If $\max(\deg(w_1^{(1)}), \deg(w_2^{(1)})) < \max(\deg(w_1^{(2)}), \deg(w_2^{(2)}))$, let $\boldsymbol{w} = \boldsymbol{w^{(1)}}$ else let $\boldsymbol{w} = \boldsymbol{w^{(2)}}$.
    • If $\max(\deg(v_1), \deg(v_2)) > \max(\deg(w_1), \deg(w_2))$, let $\boldsymbol{v} = \boldsymbol{w}$ and declare the loop as active.
    • If $\max(\deg(u_1), \deg(u_2)) \geq \max(\deg(v_1), \deg(v_2))$, exchange $\boldsymbol{u}$ and $\boldsymbol{v}$.
– Until the loop is not declared active for two consecutive executions.
– Output $(\boldsymbol{u}, \boldsymbol{v})$.
```

Fig. 1. Algorithm for reducing lattices in dimension 2 over $\mathbb{F}_2[t]$.

In order to consider pairs $(r(t), s(t))$ such that $\gcd(r(t), s(t)) = 1$, a necessary condition is that $\gcd(k_u(t), k_v(t)) = 1$. So, in the rectangle we allocate for the sieving, positions corresponding to a pair $(k_u(t), k_v(t))$ with two coordinates divisible by t can be omitted [10]. This is a quick shortcut to avoid 25% of the gcd computations. Similarly, we avoid the positions where both coordinates are divisible by $t + 1$.

Depending on the problem we have to handle, it can be computationally interesting to perform such a sieve on the algebraic side too. In this case, each point on the rectangle are marked if they are points of the small prime ideals on the linear side or on the algebraic side. This was for instance the case for the computation described in section 5. This is done by representing the prime ideals on the algebraic side as lattices and handling them as on the linear side.

After selecting good $(k_u(t), k_v(t))$ candidates in this way, we can check efficiently that the corresponding values $r(t)\mu_1(t) + s(t)\mu_2(t)$ are indeed smooth using Berlekamp's algorithm. Then, if for some of the remaining couples $(r(t), s(t))$, the divisor of the function $r(t)X + s(t)$ is smooth too, this produces an algebraic relation between elements of the factor bases.

Remark: Berlekamp's algorithm has got roughly two phases; the construction of a set of "f-reducing polynomials" thanks to the kernel computation of the Berlekamp's matrix and the separation step itself involving the computation of gcds between the list of factors and the reducible polynomials [16]. Let us note that the last phase can be speeded up using the fact that the polynomials that we want to factor have their potential factors stored in the factor bases. Thanks to this table of irreducible polynomials, one can test early in the process whether the partial factors are irreducible. When they are, we remove them from the list of factors. Thus we can spare many of the gcd computations that are necessary when no table of irreducible polynomials is available.

4.2 Linear Algebra

As explained in section 2.5, it is straightforward to apply to the case $\mathbb{F}_{p^n}$, p small, the ideas developed for $\mathbb{F}_p$, p large. Simply, this step is done modulo each large prime factor ℓ of $(p^n-1)/(p-1)$.

The only small improvement we are aware of concerns the characteristic 2 case. When 2^n-1 is a prime, the arithmetic involved in Lanczos' algorithm can be slightly speeded up. This consists in using the classical fact that the reductions modulo 2^n-1 can be done by a single subtraction on the binary representation of the integers involved. When 2^n-1 is not a prime, it is usually better to perform Lanczos' algorithm modulo each large prime factor ℓ of 2^n-1, instead of modulo 2^n-1.

5 Example

Let σ be the mapping defined from the set of integers to $\mathbb{F}_2[t]$ which sends an integer ν (written in an hexadecimal way) to a polynomial $\sigma(\nu)$ such that substituting t by 2 in $\sigma(\nu)$ yields ν (for instance, $\sigma(\mathtt{b}) = t^3+t+1$), we now describe a discrete logarithm computation in $\mathbb{F}_{2^{521}}$.

Precisely, we were able to compute the discrete logarithm of $e(t)$, $\pi(t)$ and $e(t)+\pi(t)$ where

$$\begin{aligned} e(t) = \sigma(\lfloor 2^{519} e \rfloor) &= t^{520}+t^{518}+\ldots+t^6+t^3, \\ \pi(t) = \sigma(\lfloor 2^{519} \pi \rfloor) &= t^{520}+t^{519}+\ldots+t^6+t^3+1. \end{aligned}$$

At first, we fixed a representation of $\mathbb{F}_{2^{521}}$ by choosing a $C_{1,5}$ algebraic curve over $\mathbb{F}_2$ given by

$$H(t,X) = X^5+X+t+1,$$

and checking that the resultant of $H(t,X)$ with the bivariate polynomial

$$\begin{aligned} \mu_2(t)X+\mu_1(t) = \sigma(\mathtt{1b92c17dec4c4cf4f5ab9c1e86f})X+ \\ \sigma(\mathtt{d0e134790925d9e08}) \end{aligned}$$

yields an irreducible polynomial $f(t)$ of degree 521.

Of course, there exists many $C_{a,b}$ curves which could have been used here. However, following an idea developed for factoring integers, we select a polynomial $H(t,X)$ whose number of roots, modulo irreducible polynomials of small degree over $\mathbb{F}_2$, is slightly larger than usual.

The factor base contains the 300 000-th first irreducible polynomials over $\mathbb{F}_2$ once ordered by their σ values and contains the places with norms of degree smaller than 22 in the function field defined by $H(t,X)$. After a three weeks computation on a quadri-processors alpha server 8400 computer, we obtained 472 121 equations in 450 940 unknowns with 9 235 383 nonzero entries.

So we had 300 000 special-q. For each special-q, we marked points in a rectangle $(k_u(t), k_v(t))$ of size $2^{14} \times 2^{14}$ such that the corresponding pairs $(r(t), s(t))$ are candidates for smoothness (cf. section 4.1). This yielded around 2 000 candidates. Testing them with Berlekamp's algorithm, both in the linear and the algebraic side, gave in average 2 equations.

We then applied a structured Gaussian elimination to reduce our system to 197 039 equations in 196 939 unknowns with 12 220 108 nonzero entries [13, 12] (249277 entries were different from ±1, the largest was 29). Time needed for this on only one processor was about one hour.

Then, our parallelized version of Lanczos' algorithm took 10 days over 4 processors to finish the linear inversion modulo $2^{521} - 1$. At the end, we had "logarithms for ideals" of small norms. As a consequence, we had logarithms for small irreducible polynomials:

$$\begin{aligned}\log_t(t+1) = {} & 94681577152122294076175173598650324606 21\\ & 888852201905263910801487998985884345864952201320754968 8251\\ & \qquad 33615526417923163653891428634582550637955161092146219401 59,\end{aligned}$$

$$\begin{aligned}\log_t(t^2+t+1) = {} & 409945320335775766828444393363213401 5543\\ & 756038796071121488062791898236173013002391324856407381 0794\\ & \qquad 9052818943078142206215533143595141990328387727782201876 1891,\end{aligned}$$

$$\vdots$$

Afterwards, we found in few hours, two polynomials,

$$z_1(t) = \sigma(\texttt{17acf35dc9215}) \times \sigma(\texttt{33cab5311}) \times \sigma(\texttt{83b6db37}) \times \sigma(\texttt{88af29f}) \times \sigma(\texttt{4c99eb3}) \times \sigma(\texttt{1a22cdd}) \times \sigma(\texttt{debb79}) \times \sigma(\texttt{6358f}) \times \sigma(\texttt{304f}) \times \sigma(\texttt{6b5}) \times \sigma(\texttt{41b}) \times \sigma(\texttt{75}) \times \sigma(\texttt{2})^4$$

and

$$z_2(t) = \sigma(\texttt{41edc78c5127}) \times \sigma(\texttt{75a6c0fe253}) \times \sigma(\texttt{b66ac13d5}) \times \sigma(\texttt{d422507}) \times \sigma(\texttt{b0b0e11}) \times \sigma(\texttt{d2c45}) \times \sigma(\texttt{81869}) \times \sigma(\texttt{54e1}) \times \sigma(\texttt{a85}) \times \sigma(\texttt{409}) \times \sigma(\texttt{25f}) \times \sigma(\texttt{fd}) \times \sigma(\texttt{3b}) \times \sigma(\texttt{7}) \times \sigma(\texttt{3})$$

such that

$$\sigma(\texttt{3fffcd})^{43} \times e(t) = z_1(t)/z_2(t) \bmod f(t).$$

Then, using at most 6 levels of special–q descents, computing discrete "logarithms for the ideals" of norms larger than $\sigma(\texttt{5df401})$ in the left algebraic field was (thanks to a one hour computation for each ideal, on a unique processor) equivalent to compute discrete "logarithms for ideals" of norms larger than those of the factor base in the algebraic function field. Time needed for computing the corresponding discrete logarithms was at most one hour for each on a unique processor. At the end, we obtained

$$\begin{aligned}\log_t e(t) = {} & 2632477621938341298849947024285383\,6\\ & 028931740709327317719002560095841802532546548170764837586429\,2\\ & 845650245474689082025204387673462677992080095380610945787435\,8.\end{aligned}$$

Similarly, we found

$$\begin{aligned}\log_t \pi(t) = {} & 54752914801211335857888400194404883\\ & 4396041695431692261837322543793760717339868611259553398016090\\ & 470879005113858820909173945556153048761351376719820943349684\,4,\end{aligned}$$

and

$$\begin{aligned}\log_t (e(t) + \pi(t)) = {} & 41592014011202531792054377504019307\\ & 6439975376714991661072042543367168030386736581168078966485150\\ & 627246523930786284689895718995063216522239910056818539851816\,7.\end{aligned}$$

So, as a conclusion, time that we need for computing discrete logarithms in $\mathbb{F}_{2^{521}}$ on a 525 MHz quadri-processor alpha server 8400 computer is approximatively 12 hours for each, once the sieving step (21 days) and the linear algebra step (10 days) is performed.

The software we used is an adaptation to the characteristic two of a $\mathbb{F}_p$ implementation [12] taking advantage of a generic software based on a finite field `C` library called `ZEN` [4].

Remark: Since the current record in this field of research is a computation in $\mathbb{F}_{2^{607}}$ obtained with Coppersmith's algorithm after one year over 100 PCs [26], it is natural to estimate on the basis of our computation over $\mathbb{F}_{2^{521}}$ what would be the time needed by this FFS implementation to handle $\mathbb{F}_{2^{607}}$. This can be easily done by computing $L_{2^{607}}[1/3, (32/9)^{\frac{1}{3}}] \ / \ L_{2^{521}}[1/3, (32/9)^{\frac{1}{3}}]$. This yields a factor of 12 and means a one year computation on a single 525 MHz quadri-processor alpha server 8400 computer. We have performed some experiment in this range, they corroborate this rough estimate.

6 Conclusion

In this paper, we described improvements to the function field sieve for the discrete logarithm problem. With these improvements, we computed discrete logarithms in $\mathbb{F}_{2^{521}}$ and showed that the function field sieve can be considered as an equivalent of the special number field sieve, giving the confirmation that it is faster, both from an asymptotic and from a computational viewpoint, than Coppersmith's algorithm and Adleman's original FFS.

Acknowledgments

The authors would like to thank J.M. Couveignes for some helpful discussions about $C_{a,b}$ curves. They would also like to thank the anonymous referees for valuable comments and for pointing out to them that Adleman and Huang [2] already proved that the Function Field Sieve has the same complexity as the special Number Field Sieve.

References

1. L. M. Adleman. The function field sieve. In *Proceedings of the ANTS-I conference*, volume 877 of *Lecture Notes in Computer Science*, pages 108–121, 1994.
2. L. M. Adleman and M. A. Huang. Function field sieve method for discrete logarithms over finite fields. In *Information and Computation*, volume 151, pages 5–16. Academic Press, 1999.
3. J. P. Buhler, H. W. Lenstra, Jr., and C. Pomerance. Factoring integers with the number field sieve. Pages 50–94 in [15].
4. F. Chabaud and R. Lercier. *ZEN, User Manual.* Available at `http://-www.di.ens.fr/~zen/`.
5. D. Coppersmith. Fast evaluation of logarithms in fields of characteristic two. *IEEE transactions on information theory*, IT-30(4):587–594, July 1984.
6. D. Coppersmith, A. Odlyzko, and R. Schroppel. Discrete logarithms in $\mathbb{F}_p$. *Algorithmica*, 1:1–15, 1986.
7. T. Denny, O. Schirokauer, and D. Weber. Discrete Logarithms: The effectiveness of the Index Calculus Method. In *Proceedings of the ANTS-II conference*, volume 1122 of *Lecture Notes in Computer Science*, pages 337–361, 1996.
8. M. Elkenbracht-Huizing. An implementation of the number field sieve. *Experimental Mathematics*, 5(3):231–253, 1996.
9. S. Gao and J. Howell. A general polynomial sieve. *Designs, Codes and Cryptography*, 18:149–157, 1999.
10. R. Golliver, A. K. Lenstra, and K. McCurley. Lattice sieving and trial division. In *Proceedings of the ANTS-I conference*, volume 877 of *Lecture Notes in Computer Science*, pages 18–27. Springer-Verlag, 1994.
11. D. Gordon and K. McCurley. Massively parallel computation of discrete logarithms. In *Advances in Cryptology — CRYPTO'92*, volume 740 of *Lecture Notes in Computer Science*, pages 312–323. Springer-Verlag, 1993.
12. A. Joux and R. Lercier. Improvements to the general number field sieve for discrete logarithms in prime fields. *Math. Comp.*, 2000. To appear. Preprint available at `http://www.medicis.polytechnique.fr/~lercier/`.
13. B. A. LaMacchia and A. M. Odlyzko. Computation of discrete logarithms in prime fields. *Designs, Codes and Cryptography*, 1:47–62, 1991.
14. B. A. LaMacchia and A. M. Odlyzko. Solving large sparse systems over finite fields. In *Advances in Cryptology — CRYPTO'90*, volume 537 of *Lecture Notes in Computer Science*, pages 109–133. Springer-Verlag, 1991.

15. A. K. Lenstra and H. W. Lenstra, Jr., editors. *The development of the number field sieve*, volume 1554 of *Lecture Notes in Mathematics*. Springer–Verlag, 1993.
16. R. Lidl and H. Niederreiter. *Finite Fields*, volume 20 of *Encyclopedia of Mathematics and its Applications*. Addison–Wesley, 1983.
17. R. Lovorn. *Rigorous Subexponential Algorithms for Discrete Logarithms Over Finite Fields*. PhD thesis, Univ. of Georgia, 1992.
18. R. Matsumoto. Using C_{ab} curves in the function field sieve. *IEICE Trans. Fundamentals*, E82-A(3):551–552, march 1999.
19. A. M. Odlyzko. Discrete logarithms in finite fields and their cryptographic significance. In T. Beth, N. Cot, and I. Ingemarsson, editors, *Advances in Cryptology — EUROCRYP'84*, volume 209 of *Lecture Notes in Computer Science*, pages 224–314. Springer–Verlag, 1985. Available at `http:/www.dtc.umn.edu/~odlyzko`.
20. J.M. Pollard. The lattice sieve. Pages 43–49 in [15].
21. P. Montgomery S. Cavallar and H. te Riele. New record SNFS factorization. Available at `http://listserv.nodak.edu/archives/nmbrthry.html`, november 2000. Factorization of $2^{773} + 1$.
22. O. Schirokauer. The special function field sieve. Preprint.
23. O. Schirokauer. Discrete logarithms and local units. *Phil. Trans. R. Soc. Lond. A 345*, pages 409–423, 1993.
24. R. D. Silverman. The Multiple Polynomial Quadratic Sieve. *Math. Comp.*, 48:329–339, 1987.
25. E. Thomé. Computation of discrete logarithms in $\mathbb{F}_{2^{607}}$. In *Advances in Cryptology — ASIACRYPT'2001*, volume 2248 of *Lecture Notes in Computer Science*, pages 107–124. Springer–Verlag, 2001.
26. E. Thomé. Discrete logarithms in $\mathbb{F}_{2^{607}}$. Available at `http://listserv.nodak.edu/archives/nmbrthry.html`, february 2002.
27. D. Weber and Th. Denny. The solution of McCurley's discrete log challenge. In H. Krawczyk, editor, *Advances in Cryptology — CRYPTO'98*, volume 1462 of *Lecture Notes in Computer Science*, pages 458–471. Springer–Verlag, 1998.

MPQS with Three Large Primes

Paul Leyland[1], Arjen Lenstra[2], Bruce Dodson[3],
Alec Muffett[4], and Sam Wagstaff[5]

[1] Microsoft Research Ltd, 7 JJ Thomson Avenue, Cambridge, CB3 0FB, UK,
pleyland@microsoft.com
[2] Citibank, 1 North Gate Road, Mendham, NJ 07945-3104, USA,
arjen.lenstra@citicorp.com
[3] Lehigh University, Bethlehem, PA 18015-3174, USA,
bad0@Lehigh.edu
[4] Sun Microsystems Professional Services,
Riverside Way, Watchmoor Park, Camberley, GU15 3YL, UK,
alec.muffett@uk.sun.com
[5] CERIAS and Department of Computer Sciences, Purdue University,
West Lafayette, IN 47907-1315, USA,
ssw@cs.purdue.edu

Abstract. We report the factorization of a 135-digit integer by the triple-large-prime variation of the multiple polynomial quadratic sieve. Previous workers [6][10] had suggested that using more than two large primes would be counterproductive, because of the greatly increased number of false reports from the sievers. We provide evidence that, for this number and our implementation, using three large primes is approximately 1.7 times as fast as using only two. The gain in efficiency comes from a sudden growth in the number of cycles arising from relations which contain three large primes. This effect, which more than compensates for the false reports, was not anticipated by the authors of [6][10] but has become quite familiar from factorizations obtained using the number field sieve. We characterize the various types of cycles present, and give a semi-quantitative description of their rather mysterious behaviour.

1 Introduction

The use of large primes in the Multiple Polynomial Quadratic Sieve (MPQS) has been suggested many times. Earlier studies [10][13] suggested that using one large prime is always better than using none, and that the double large prime variation (often called PPMPQS) is more efficient than using fewer large primes when factoring integers with more than about 80 decimal digits. For N near 10^{100} these studies showed that the double-prime version of MPQS is 2 to 2.5 times faster than the single-prime variant. A spectacular example of a PPMPQS factorization was the record-breaking factorization of the 129-digit Scientific American RSA challenge RSA-129 in 1993-4, reported in [1].

In this paper we introduce TMPQS, i.e., MPQS with three large primes. In [6][10] it was suggested that it would be less efficient than PPMPQS because of

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 446–460, 2002.

the very large number of false reports from the sievers, that is quadratic residues which are flagged as possibly being smooth but which subsequently prove not to be. The analysis underlying the suggestion from [10], however, did not take into account an effect that has since then become familiar from number field sieve (NFS) factorizations. In NFS, relations with more than two large primes are found at very little extra cost. Experience with such NFS-relations has shown that, after a slow start, the number of cycles suddenly grows very rapidly [7]. We expected that similar behaviour may also occur with TMPQS, and that it may compensate for the extra false reports. A few preliminary and small-scale computations by the first author suggested that TMPQS was certainly no worse than PPMPQS for N in the 100-digit range and may be slightly better.

To put this hypothesis to the test, we factored the 135-digit cofactor of $2^{803} - 2^{402} + 1$, also known as 2,1606L.c135 in the Cunningham newsletters. It was chosen because it is: somewhat larger than the previous record-breaking MPQS factorization but close enough in size to RSA-129 that it should give an indication of a possible speed-up; comparable with recent large factorizations using the General NFS; not easily factored with the Special NFS. It should be noted, however, that GNFS would have been much faster than the computation we decided to embark upon. We factored 2,1606L.c135 solely to satisfy our curiosity with respect to the relative speed of TMPQS and PPMPQS and to get more insight in the cycle behaviour. Setting the current record for a factorization with the Quadratic Sieve algorithm was entirely incidental...

In Section 2 we describe our implementation of TMPQS. Section 3 provides a detailed description of the growth in the number of cycles and gives a semi-quantitative analysis of the results. Evidence that TMPQS may be expected to outperform PPMPQS for sufficiently large numbers is presented in Section 4.

2 The Multiple Polynomial Quadratic Sieve

We assume that the reader is familiar with MPQS and PPMPQS (see [6] [10]). Compared to PPMPQS, TMPQS requires a different way to process the reports from the siever and different cycle counting and cyle finding software. The implementation of the siever that the first four authors used was almost identical to the 'factoring by email' version described in [9]. Even though other implementations of PPMPQS are readily available, some of them perhaps more efficient than ours, using this old implementation had the advantage that we were able to compare our performance figures directly to those obtained in previous large-scale factorizations, such as that of the 129-digit RSA challenge number [1] which was also based on the software from [9]. The fifth author used an adapted version of his Self-Initializing Quadratic Sieve siever [2], making sure the reported relations were compatible with the ones found by the others. Because of the nature of this version of MPQS, these relations contained markedly more small primes but the large-prime statistics within the relations having large primes, and the relative proportions of the three types of relations, was closely similar.

The remainder of this section contains a brief description of the five stages of TMPQS: parameter selection, sieving, counting and finding cycles, linear algebra, and combination of relations to find the factorization.

2.1 Parameter Selection

In MPQS a **multiplier** $k \in \mathbf{Z}_{>0}$ is chosen such that kN is a quadratic residue modulo relatively many small primes, where N is the number to be factored. For 2,1606L.c135 the value $k = 1$ is optimal.

The **factor base size** is the number of primes with which one sieves. It was chosen as 550 000, so $B_1 = 17\,157\,953$ was the largest prime in the factor base.

The **large prime bound** was chosen as $B_2 = 2^{30}$. Thus, relations consist of integers v such that $v^2 \bmod (2{,}160\mathrm{L.c}135)$ is the product of primes $\leq B_1$ and at most three primes $< B_2$.

The size of the factor base was chosen in a rather ad hoc manner. Sieving experiments were performed for various choices between 400 000 and 750 000. For each experiment a number of sieve-report bounds were tried, attempting to maximize the yield. The value eventually chosen may not be optimal, but it appears not to be too bad. Compared to PPMPQS our choice may be considered to be on the small side, but it is a logical consequence of allowing three as opposed to just two large primes.

The **sieving range**, the number of values sieved per polynomial, was determined experimentally. The siever used by four of us was tested with values lying between 17 million and 100 million, again attempting to maximize the yield. The yield was only slightly dependent on the sieving range but was somewhat higher for the lower values we tried. The bulk of the computation used a value of 17 158 000. The siever employed by the fifth author used a sieving range of only 2 000 000 because that method initializes polynomials so efficiently.

2.2 Sieving

The only difference between the siever used by most of us and that used to factor RSA-129 is that we attempted to factor quadratic residues less than $B_2^3 = 2^{90}$ after primes $\leq B_1$ had been divided out. We used a fast pseudoprimality test to reject candidates greater than B_2 that were not recognized as composites, we used ECM to factor those which were, and rejected those for which ECM failed or found a prime factor larger than B_2. Per report, the pseudoprime test may have to be applied to two different numbers (one $< B_2^3$ and one $< B_2^2$) and ECM may have to be used to factor both numbers (if composite). Thus, in some cases, testing candidate quadratic residues may be more expensive in TMPQS than PPMPQS.

The output of the siever was a series of relations, each consisting of an integer v and the prime factorization of $v^2 \bmod (2{,}1606\mathrm{L.c}135)$. As in [1], we denote a relation as a *ful*, *par*, or *ppr* according to whether it contains zero, one, or two large primes. In addition, we have *tpr* relations which contain three large primes.

As in [1], the sieving process was distributed over many client machines and the relations produced sent to a central machine which checked the correctness of newly arrived data and discarded duplicates. Relations were then added to one of four files, according to whether they were *ful*, *par*, *ppr*, or *tpr* relations.

We began sieving on 10th January 2001 and finished 231 days later on 29th August. By the time we finished sieving, we had accumulated a total of 13 441 627 relations, made up of 62 626 *ful*s, 790 129 *par*s, 4 080 732 *ppr*s and 8 508 140 *tpr*s.

2.3 Counting and Finding Cycles

In principle, once enough relations have been output by the sieving phase, linear dependencies in the matrix of exponent vectors modulo 2 could be found by standard techniques of linear algebra, such as Gaussian elimination or block Lanczos. There are two problems, however. In the first place, one needs to be able to recognize that enough relations have been found. Secondly, the resulting matrix would be far too large to be easily dealt with by these methods.

In [10] 'cycles' were introduced as sets of relations where each large prime occurs an even number of times. It follows that each cycle gives rise to integers w and s such that $(w^2 \bmod (2{,}1606\mathrm{L.c135}))/s$ is a product of primes $\leq B_1$, where w is the product of the v's and s is the product of all large primes in the cycle and thus a square. For our purposes, cycles are therefore equivalent to (less sparse) *ful* relations. Identifying a relation with a vector with bits set for its large primes, cycles are linear dependencies modulo 2 among those vectors.

It follows that in order to recognize if there are enough relations, it suffices to count the number C of independent cycles among the non-*ful* relations, and to check if C plus the number of *ful* relations is larger than 550 000, the size of the factor base. If so, there are enough relations. Because we were interested in the growth of the number of independent cycles, we computed a lower bound for C on a daily basis. This can be done by means of an easy two-step process:

1. From the set of *par*, *ppr*, and *tpr* relations, remove the 'singletons', i.e., relations that have a large prime that does not occur in any other relation. Removing a singleton *ppr* or *tpr* relation may generate further singletons from the other prime(s) present in the relation. So, singleton removal consists of a number of 'pruning passes' that must be carried out iteratively until, during the last pruning pass, no more singletons are removed. Singleton removal can easily be implemented in a variety of ways.
2. Once all singletons have been removed, all remaining *par*, *ppr*, and *tpr* relations belong to a cycle. A close lower bound estimate for C is given by the difference δ of the total number of remaining large primes (not counting multiplicities) and the number of remaining relations, i.e., the oversquareness of the matrix of vectors. (If no *tpr*s are used and all cycles contain a *par* relation, then $\delta = C$; with *tpr*s this is the case if fairly uncommon types of cycles do not occur.)

The last step can conveniently be done using the Union-Find algorithm as in [10]. Union-Find can also directly be applied to the full set of relations, but that leads to a less precise approximation of C. Another reason that we preferred to remove singletons first is that the change in the number of pruning passes provided interesting information about the progress of our computation. This is shown in Section 3, along with a detailed account of the behaviour of the growth of C as relations were added.

Our final collection of relations produced 494 077 independent cycles, only 67 543 of which (14%) did not contain at least one *tpr* relation; there were 62 626 *ful* relations. Linear algebra techniques could have been applied to the collection of *ful* and singleton-less non-*ful* relations but, as mentioned above, this leads to an unattractively large matrix. In NFS this problem is usually dealt with by a partial merging of the complete singleton-less set of relations, removing most of the primes by combining the relations containing them, but keeping some in order to make the linear algebra step as fast as possible [3]. In MPQS, on the other hand, it has been traditional (and certainly not optimal!) to remove all large primes by building a complete set of independent cycles among the non-*ful* relations. This can be achieved using a relatively straightforward adaptation of the method from [10]; we do not elaborate. Of the 494 077 independent cycles, the 487 424 least dense ones were actually used; the longest cycle of these contained 215 relations (whereas the longest present had well over a thousand). With the addition of the 62 626 *ful* relations a matrix with 550 000 rows and 550 050 columns was produced, requiring 616 megabytes of storage (in ASCII format). The average density of the matrix was 411 bits set per row. This is much denser than usual for an MPQS factorization. More sieving would have reduced the density of the matrix, but there was no need to do so as the matrix was tractable with our resources.

2.4 Linear Algebra

Finding dependencies was done with the block Lanczos method of [11] and [12]. We performed this computation twice, once on a conventional uniprocessor machine and once on a cluster of workstations, for two reasons: to be able to make a comparison of the resources used by each implementation; and because having two simultaneous and independent computations increased our chances to meet a tight deadline. In both cases, the Lanczos algorithm used 128–bit vectors and took 4324 iterations to find 57 dependencies.

The uniprocessor machine was fitted with a 1.33GHz AMD Athlon processor and had 768 megabytes of memory. The software was compiled with the GCC compiler; RedHat Linux 7.0 was the operating system. The complete computation took 146 446 seconds, or a little under 40.7 hours. The amount of active virtual memory reached a maximum of 537 megabytes, but fell to 480 megabytes for the main part of the computation. These figures are well below the size of real memory available and so paging was not a problem.

The parallel implementation ran on a cluster of sixteen machines, each of which contained two 300MHz Pentium-II processors and 384 megabytes of memory. (Note that the memory available on a single cluster node would not have been sufficient to hold the entire data set.) We used the Microsoft Visual C++

compiler, together with the MPIPro multi-processor harness communicating via 100Mbps ethernet; the nodes ran the Windows 2000 operating system. We used 12 nodes and only one processor per node as the remainder of the cluster was required by others. Detailed records of the resources used are available only for the master node, which is the node responsible for all the I/O required for the computation. It took 33 808 seconds, or 9.4 hours, for its share of the computation and used 63 megabytes of active memory. The other nodes would have taken about the same cpu time, or very slightly less. One of the slave nodes was observed to be using 53 megabytes of memory. The 10 megabyte difference between these two figures is accounted for by the data structures needed by the master node to co-ordinate the computation as a whole. If we assume the eleven slave nodes each used 53 megabytes, the total memory usage came to 646 megabytes, substantially more than the 480 megabytes used by the uniprocessor.

If we assume that all the nodes took the same 9.4 hours of cpu time, the total computation comes to $12 * 9.4 = 112.8$ hours. Earlier experiments with heavily instrumented versions of the parallel code, admittedly working on much less dense matrices, showed that this is a reasonable assumption. A naive computation of the total number of cpu cycles used by each implementation yields 1.9×10^{14} for the uniprocessor, and 1.2×10^{14} for the cluster. Although it appears at first sight that the parallel implementation is *more* efficient, it must be stressed that this is an over-simplified analysis: the code was compiled with substantially different compilers and run under very different operating systems; and, even when the compilers and operating systems are identical, the runtime can be heavily dependent on memory bandwidth, cache efficiency, and other non-computational effects. In both cases, the linear algebra phase took less than 0.1% as much computation as did the sieving (cf. Section 4).

We warn against extrapolation of our parallel Lanczos result. Nevertheless, based on the apparent feasibility of parallelized Lanczos and the economical feasibility of clusters of small machines, we caution against assuming that fairly small RSA moduli are safe because the matrix problem is said to be too hard.

2.5 Combination of Relations and Production of the Factors

The combination of linearly-dependent relations produced by the linear algebra used exactly the same code as in [9]. Processing each dependency took 18 minutes on a 400MHz machine. The third dependency yielded the factors $p =$ 337 779 774 700 456 816 455 577 092 228 603 627 733 197 301 999 086 530 154 776 370 553 and $q =$ 346 129 173 115 857 975 809 709 331 088 291 920 685 569 205 287 238 835 924 196 565 083 957 of 2,1606L.c135 = 116 915 434 112 329 921 568 236 283 928 181 979 297 762 987 646 390 347 857 868 153 872 054 154 807 376 462 439 621 333 455 331 738 807 075 404 918 922 573 575 454 310 187 518 221. At 66 digits, p is the largest penultimate factor ever found by the MPQS algorithm.

3 Cycle Behaviour

3.1 Introduction

As the number R of relations increases, the number C of independent cycles increases at an ever increasing rate. With one large prime per relation, the growth rate in C is very nearly proportional to R^2. This can be explained using the birthday paradox [9] and was fully analyzed in [10]. For RSA-129, factored with PPMPQS, the quadratic behaviour broke down about half way through the computation [1], and a better approximation for the later stages was that C is proportional to R^4. Obviously, a large fraction of the cycles arose from the *ppr*s.

Our results for the TMPQS factorization of 2,1606L.c135 are summarized in Fig. 1, where $\ln(C)$ is plotted against $\ln(R)$ during the course of the computation. The line is the least squares fit to the data: its slope of 2.7665 shows that the growth is faster than quadratic, but it is clear that a power law is not a good approximation to the behaviour.

3.2 Cycle Types

The observation that the cycle-growth behaviour depends on the type of relations appearing in cycles was made in [1]. As part of this investigation we characterize the cycles more fully to get a better (though far from complete) understanding of the situation. Cycles consisting only of *par* relations we termed S-cycles (S for single large prime). Of the remainder, cycles not containing *tpr* relations we

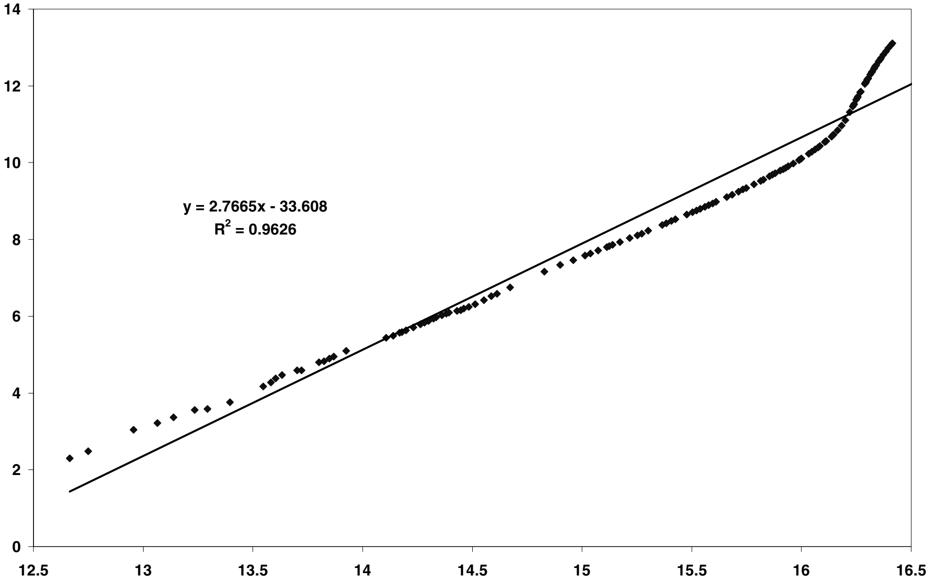


Fig. 1. Behaviour of ln(# cycles) with ln(# relations). Data with $C < 10$ are omitted.

call D-cycles (D for double, since D-cycles contain at least one *ppr* relation and possibly *par* relations), and the remaining cycles are the T-cycles (T for triple). Independent S-cycles may be assumed to consist of two relations. The numbers of independent S, D, and T-cycles are referred to by S, D, and T, respectively. Thus, $C = S + D + T$.

During the course of the computation we calculated C, S, D, and T approximately daily. This allowed us to examine in detail the variation of these quantities as a function of the total number of relations R. The behaviour of each of these quantities as a function of R is described below.

S-cycles. In Fig. 2 we plot $\ln(S)$ against $\ln(R)$. Based on the analysis from [9] and [10] we expect a linear plot with slope close to 2, which is indeed what we find: the least-squares fitted slope is 2.078 with a correlation coefficient squared of 0.9997. At the end of the factorization we had $S = 25\,603$, or 5.18% of the total.

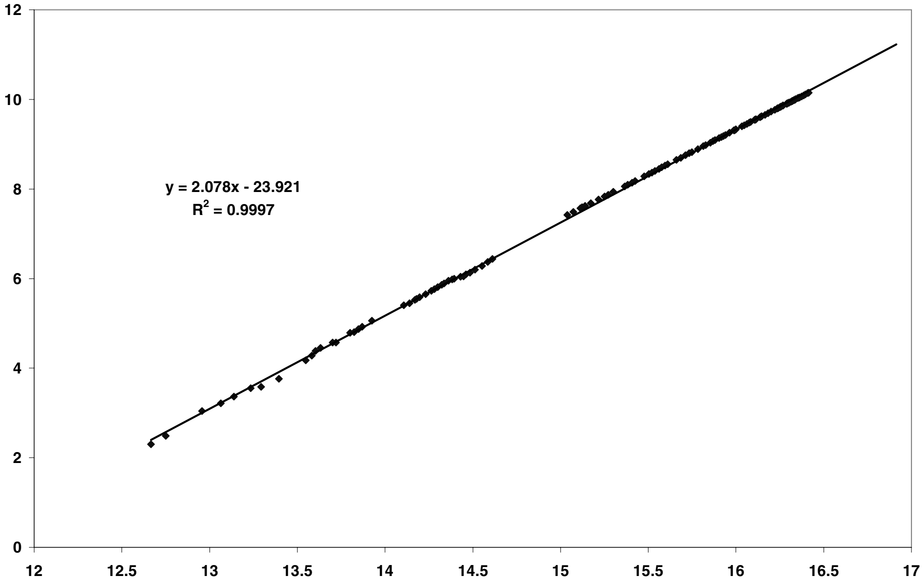


Fig. 2. Behaviour of ln(# S-cycles) with ln(# relations). Data with $S < 10$ are omitted.

D-cycles. In [1] it was observed that $S + D$ grew faster than proportional to R^2, with a suggestion that it may have been proportional to R^4 towards the end of the computation. We are not aware of any other analysis of this quantity. In Fig. 3 we plot $\ln(D)$ against $\ln(R)$. The least square fitted line, with a slope of 3.4969, can be seen to be a remarkably good fit for the data points, though there is a hint that the relationship breaks down at very small values of D. It is tempting to suggest that the slope should be exactly 7/2, but we have *no* theoretical justification for this claim; this would be an interesting subject for further study. By the end of the factorization we had $D = 41\,940$, or 8.49% of the total.

The large gap between $14.7 < \ln(R) < 15.0$ corresponds to a period when counts broken down into S, D, and T were not taken. The edges of the missing-data gap correspond to $D = 100$ and $D = 389$. Prior to this period, T was zero or one — understandably the first T-cycle was eagerly awaited — and D could be calculated from C and the number of large primes occurring in the cycles.

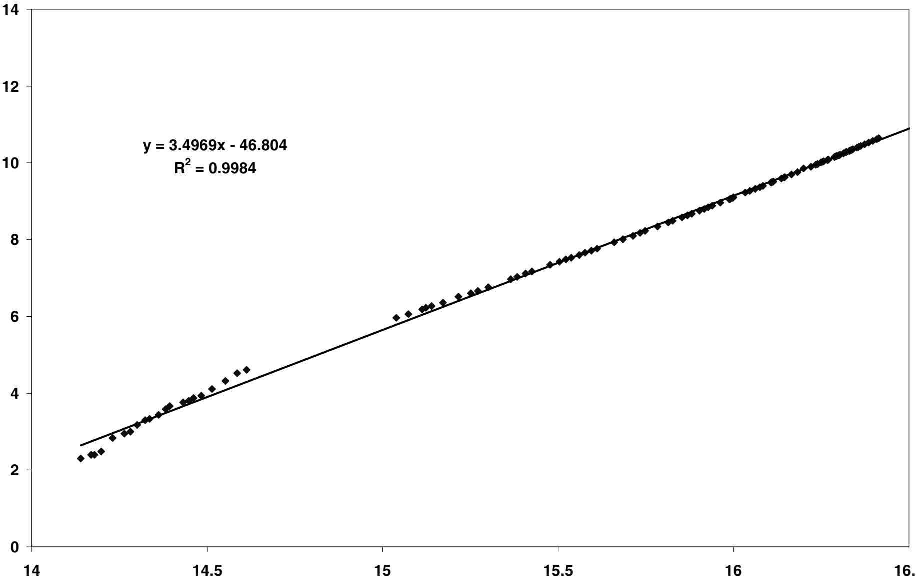


Fig. 3. Behaviour of ln(# D-cycles) with ln(# relations). Data with $D < 10$ are omitted.

T-cycles. When Figs. 2 and 3 are compared with Fig. 1, it is clear that any major non-linearity in the log-log plots must arise from the T-cycles. This is indeed what we find in Fig. 4. The fitted straight line, with a slope of 7.3207, is not a good fit to the data and the data between $16.1 < \ln(R) < 16.3$ shows pronounced curvature. This region corresponds approximately to $1 < R/10^6 < 1.2$. At the end of the factorization we had $T = 426\,534$, or 86.33% of the total.

As can be seen from Fig. 4, the initial growth of T is quite smooth. For $\ln(R) < 15.9$ (i.e., $R < 8\,100\,000$) the data points lie close to a straight line with slope 5.4785 corresponding to a power law with this exponent. Beyond this point, the plot shows an initially rising gradient, reaching a maximum of around 15, and then tailing off again to about 6 near $\ln(R) = 16.35$. In [7] a somewhat similar effect observed during a GNFS factorization was described as explosive growth.

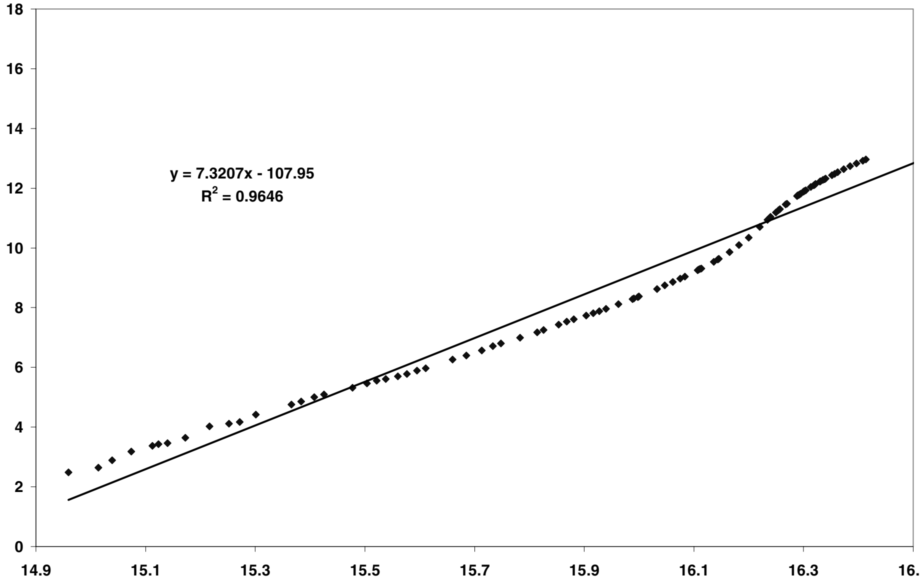


Fig. 4. Behaviour of ln(# T-cycles) with ln(# relations). Data with $T < 10$ are omitted.

In our computation, T showed strongly superpolynomial growth for a short while. The term "explosion" seems too strong for our experience with TMPQS and we prefer "phase transition" as being more appropriate. The two cases, GNFS in [7] and TMPQS, would be expected to show different cycle behaviour. In the former work, relations have four large primes — two algebraic and two rational. The large primes may only be combined with others of the same type, which limits the possible matches. In TMPQS, although there are fewer primes per relation, any of the three primes appearing in a *tpr* relation may combine with any of the (same) primes in other relations. Therefore, it is perhaps not too surprising that we see a phase transition towards the end of the sieving phase.

3.3 Phase Transition in T-Cycle Growth

We present a simple physical model for addition of relations that may justify the phrase "phase transition" to describe the superpolynomial growth in T-cycles. We picture relations as having some of the properties of atoms in chemistry: *par*s are represented as univalent atoms, *ppr*s as divalent, and *tpr*s as trivalent atoms. The silicon - oxygen system shows some of the phenomena we are attempting to model, though care must be taken not to push the analogy too far.

At the beginning of the computation there are no relations, and the model consists of a vacuum without atoms. As relations are added most will not share any primes with relations already present. In our model, these form a monatomic gas. A few will share a prime with another relation; the two atoms will form a chemical bond with one of their valencies. Initially at least, only pairs of *par* relations will have their primes matched (two *par*s with the same large prime). This corresponds in our model to a diatomic molecule formed of two univalent atoms. Other matched relations (diatomic molecules) contain unmatched primes (unsatisfied valencies) and remain available for the addition of further relations (are very reactive in the presence of other atoms and molecules).

As relations are added, the number with matched primes increases. Equivalently, as atoms are added to the gas the density rises and more molecules are formed. Eventually, the density rises high enough that large molecules are produced, some forming rings and chains. Each divalent atom added to a molecule is likely to bind only to one free valency, thereby growing a chain. Initially, at least, each trivalent atom will add an extra free valency to the growing molecule.

The connection between the chemical model and singleton removal is clear: each pruning pass removes those primes which occur only once. In the chemical model, this corresponds to breaking the bonds to those atoms which contain a free valency. When singleton removal has run its course, only cycles remain; when all bonds to reactive sites have been broken, only stable molecules remain.

As the density continues to increase, not only will rings of atoms within a molecule build up, adjacent molecules will connect to each other via a di- or trivalent atom. In real chemical systems, such as the addition of silicon and oxygen atoms to a container, at first the atoms are largely isolated. Then stable O_2 and reactive SiO and Si_2 molecules are formed, and then stable SiO_2 molecules together with a whole raft of highly reactive silicon-oxygen molecules with free valencies. All these atoms and molecules are still in the gas phase. Eventually the density becomes so high that the molecules cross-link in profusion and the system as a whole becomes a liquid, glass, or solid (depending on temperature and pressure) in equilibrium with a vapour. A phase transition has taken place and the properties of the condensed phase are very different from the properties of the earlier gaseous phase.[1] In the TMPQS case, the condensed phase does not appear to be crystalline (we do not find one massive and almost fully interconnected component), but rather more akin to a glass. Many of the relations are connected, but in a number of components and with a large number of lengthy chains which terminate in a free valency and, in all, roughly three-quarters of the relations remain in the gas phase — to mix metaphors badly.

When phase changes occur in physical systems, it is usual for many physical properties to change dramatically over a small variation in a quantity such as temperature, density, or magnetic field strength. Some properties, such as the density increase on condensation or viscosity on polymerization, show themselves as near-step functions with an increased gradient at the phase transition in the phase diagram for the system. Such behaviour is similar to that seen in Fig. 4. Other quantities, such as the specific heat capacity, show relatively flat behaviour well away from the phase transition and rise to a sharp peak at the transition itself. It is normal for the heat capacity of the two phases to be different. To

[1] In a real physical system, bonds are not unbreakable and the condensed phase is constantly exchanging material with the vapour. Matched relations do not split and match with other relations. We did warn against pushing the model too far.

take a physical system almost at random, Doye, Sear and Frenkel [8] describe a phase transition in the molecular configuration of four moderately-sized polymers. Fig. 6a of their paper looks somewhat similar to our Fig. 4 though, it must be admitted, the slopes are very different. When we realise that there is a systematic slope in the $\ln(T)$ versus $\ln(R)$ graph for small values of R and compensate for this feature by subtracting the least squares fitted line to this data ($\ln(T) = 5.4785 \ln(R) - 79.469$) we produce Fig. 5. The resemblance between this plot and Fig. 6a of [8] is remarkable. An even more startling similarity can be seen between Fig. 6b of [8] (the temperature - specific heat capacity plot) and our Fig. 6 which shows a plot of the number of pruning passes during singleton removal against R. The two are almost identical in appearance! In our experiment, the "pruning capacity" rises slowly from about 10 to around 25 in the gaseous phase, grows through a peak of almost 140 at the phase change itself and settles down at about 40 in the condensed phase.

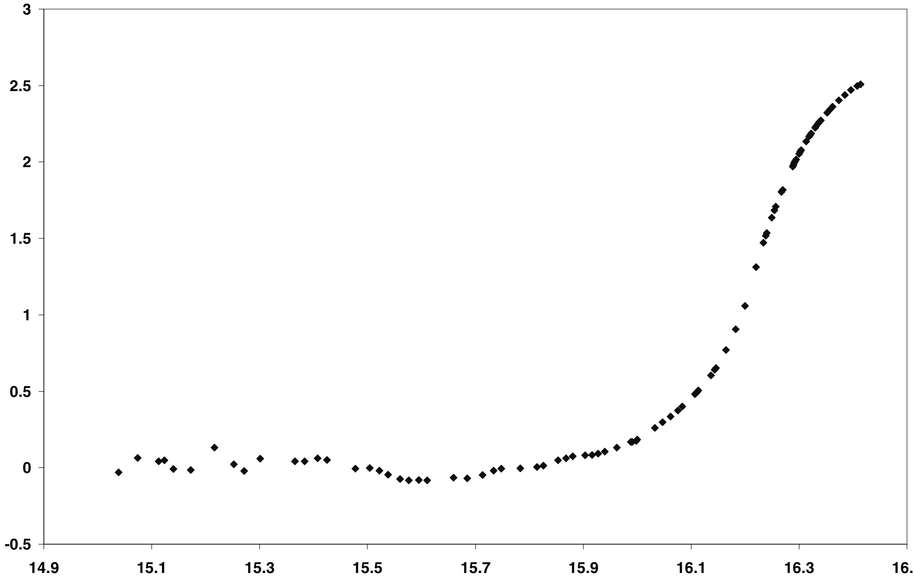


Fig. 5. $\ln(T) - 5.4785 \ln(R) + 79.469$ plotted against $\ln(R)$.

We conclude this section with a repeat of our warning: cycles among relations in a TMPQS factorization are *not* isomorphic to molecules in a chemical mixture, and the analogy should not be pushed too far. Nonetheless, the two systems show remarkably similar behaviour and it would appear that this may be a fruitful field for further study, not least because a similar phenomenon also seems to occur in the GNFS. If the phenomenon were better understood, we may have a hope of selecting sieving parameters which bring forward the onset of the phase transition without unduly slowing down the rate at which relations are produced.

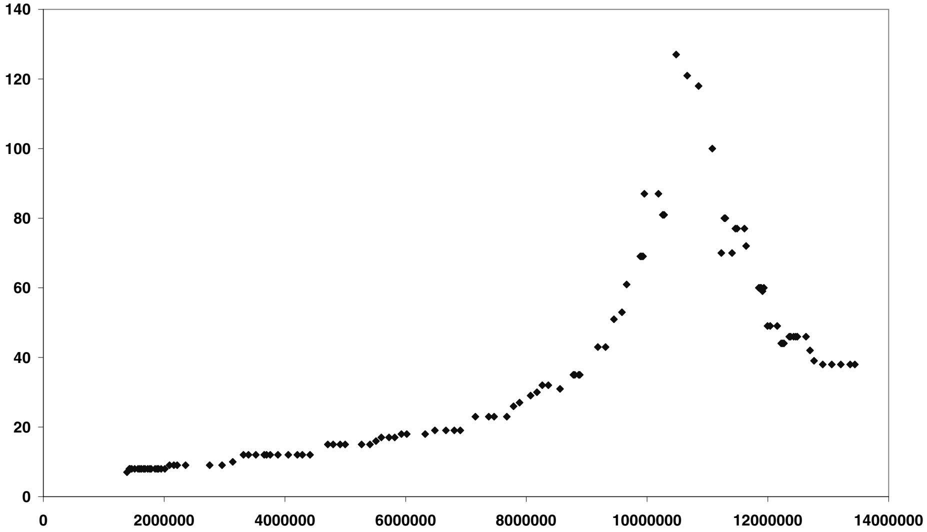


Fig. 6. Behaviour of (# pruning passes) with (# relations).

4 Performance Comparison of TMPQS and PPMPQS

As mentioned in the abstract, one of our objectives was to compare TMPQS and PPMPQS and to test the assertion made in [10] and [6] that TMPQS would be the slower algorithm. Because of the magnitude of the computations required, it is unreasonable to repeat the factorization of a 135-digit integer with PPMPQS purely to make the performance comparison. Fortunately, we can base our comparison on different arguments.

During past PPMPQS factorizations it has often been observed that sieving is approximately half-completed when the number of independent cycles is equal to the number of *ful* relations. As noted in Section 2.3, when we finished we had 62 626 *ful*s and 67 543 cycles not involving *tpr* relations, i.e., the type of cycles generated by PPMPQS. Following this rule of thumb, PPMPQS-sieving would have been about half-completed by the time we were finished. However, this comparison is inaccurate because TMPQS and PPMPQS find *par*s and *ppr*s at different rates: the lower sieve threshold used for TMPQS leads to more false reports (and to *tpr*s) and must therefore affect the *par* and *ppr* yield. For that reason we ran PPMPQS for about a week on the same number, with the same factor base size and large prime bound as used for TMPQS and optimal sieving range. We found that PPMPQS finds non-*tpr*s in 88% of the time of TMPQS (counting only the non-*tpr*s found by TMPQS, but of course TMPQS find *tpr*s too). This would imply that TMPQS does not run about twice as fast than PPMPQS, but only about $0.88 * 2 \approx 1.75$ times faster. An additional small correction may be applied to account for the suboptimality of the PPMPQS factor base size; in our experience this affects the runtime in a very minor way (say, at most 5%).

Based on this analysis, we may conclude that for 2,1606L.c135 and our implementation, TMPQS is more than 1.5 times faster than PPMPQS. This is corroborated by an independent, but admittedly less precise, comparison based on the factorization of RSA-129. During that computation it was found that the complete sieving step for RSA-129 would have taken almost 400 years on a DECStation 5000/25, a fairly common machine in the early 1990s. The first author still has access to such a machine, and ran TMPQS on it for 2,1606L.c135 for 7 117 214 seconds, finding 4901 relations. Given that 13 441 627 relations were needed, this machine would have spent $\frac{13\,441\,627}{4901} \approx 2743$ times 7 million seconds, i.e., about 620 years, to complete the TMPQS sieving. Thus we conclude that this machine would have spent almost 1.6 times as much computation to factor 2,1606L.c135 by TMPQS as it would have done to factor RSA-129 by PPMPQS.

The asymptotic runtime of all variants of the quadratic sieve algorithm to factor N is known to be $L[N] = \exp((1 + o(1))\sqrt{\log N \log\log N})$, for $N \to \infty$. We find that, omitting the $o(1)$'s as usual,

$$\frac{L(2{,}1606\text{L.c135})}{L(5 * \text{RSA-129})} \approx 2.7,$$

(where we use 5 ∗ RSA-129 because there the multiplier was 5) so that we may expect that factoring 2,1606L.c135 by PPMPQS is about 2.7 times harder than factoring RSA-129 by PPMPQS. In 13 years of experience with PPMPQS on many numbers in the 100 to 120 digit range, we have never experienced large deviations from the runtime as 'predicted' by the above method. Thus, in our experience, this estimate is reasonably reliable. Since the actual TMPQS over PPMPQS ratio is 1.6, we conclude that TMPQS leads to a speed-up of $2.7/1.6 \approx 1.7$ over PPMPQS. This estimate is consistent with the one given earlier.

Although TMPQS appears to be an improvement on PPMPQS, it is still not competitive with GNFS when factoring integers of around 135 digits. Results for the GNFS factorizations of RSA-130 and RSA-140 have been published in [5] and in [4] respectively. On a scale where all costs are normalized to RSA-129 = 5000 MIPS-years we used 8000 MIPS-years to factor 2,1606L.c135 whereas RSA-130 required 750 MIPS-years and RSA-140 took 2000 MIPS-years.

5 Conclusions

We have shown that for our implementation and a particular 135 digit number TMPQS outperforms PPMPQS, despite the gloomy prognostications of [6] and [10]. However, for numbers of this size, GNFS is still about six times faster.

The reason that TMPQS performs so much better than expected is due to a sudden growth in the number of cycles if more than two large primes are used. This phenomenon is familiar from the NFS. A full understanding of what happens is still lacking. We have given an interpretation in terms of a phase transition, borrowing terminology from chemical systems. Further work would be valuable, since it may enable us to improve the efficiency of the NFS.

Acknowledgment

We thank Scott Contini and another (anonymous) ANTS reviewer for their very thorough and insightful criticism and comments.

References

1. D. Atkins, M. Graff, A.K. Lenstra, P.C. Leyland, *THE MAGIC WORDS ARE SQUEAMISH OSSIFRAGE*, Proceedings Asiacrypt'94, LNCS 917, Springer-Verlag (1995), 265-277.
2. B. Carrier, S.S. Wagstaff Jr., *Implementing the hypercube quadratic sieve with two large primes*, Technical Report 2001-45, Purdue University CERIAS (2001) URL `http://www.cerias.purdue.edu/papers/archive/2001-45.{ps,pdf}`.
3. S. Cavallar, *Strategies in filtering in the number field sieve*, ANTS-IV, LNCS 1838, Springer-Verlag (2000) 209–231.
4. S. Cavallar, B. Dodson, A.K. Lenstra, P.C. Leyland, W. Lioen, P.L. Montgomery, B. Murphy, H. te Riele, P. Zimmermann, *Factorization of RSA–140 using the number field sieve*, Proceedings Asiacrypt'99, LNCS 1716, Springer-Verlag (1999), 195–207.
5. J. Cowie, B. Dodson, R.-M. Elkenbracht-Huizing, A.K. Lenstra, P.L. Montgomery, J. Zayer, *A world wide number field sieve factoring record: on to 512 bits*, Proceedings Asiacrypt'96, LNCS 1163, Springer-Verlag (1996), 382–394.
6. R. Crandall, C. Pomerance, *Prime Numbers, a computational perspective*, Springer (2001) 237.
7. B. Dodson, A.K. Lenstra, *NFS with four large primes: an explosive experiment*, Proceedings Crypto'95, LNCS 963, Springer-Verlag (1995) 372–385.
8. J.P.K. Doye, R.P. Sear, D. Frenkel, *The effect of chain stiffness on the phase behaviour of isolated homopolymers*, J. Chemical Physics 108, (1998) 2134–2142 URL `http://brian.ch.cam.ac.uk/~jon/papers/homop/homop.html`.
9. A.K. Lenstra, M.S. Manasse, *Factoring by electronic mail*, Proceedings Eurocrypt'89, LNCS 434, Springer-Verlag (1990) 355–371.
10. A.K. Lenstra, M.S. Manasse, *Factoring with two large primes*, Math. Comp. 63 (1994) 785-798.
11. P.L. Montgomery, *A block Lanczos algorithm for finding dependencies over GF(2)*, Proceedings Eurocrypt'95, LNCS 921, Springer-Verlag (1995), 106–120.
12. P.L. Montgomery, *Distributed Linear Algebra*, 4th Workshop on Elliptic Curve Cryptography, Essen (2000), URL `http://www.cacr.math.uwaterloo.ca/conferences/2000/ecc2000/montgomery.{ppt,ps}`.
13. C. Pomerance, *Analysis and comparison of some integer factoring algorithms* in H.W. Lenstra Jr, R. Tijdeman, editors, *Computational methods in number theory, Part I*, Math. Centre Tracts, Math. Centrum (1982) 89–139.

An Improved Baby Step Giant Step Algorithm for Point Counting of Hyperelliptic Curves over Finite Fields

Kazuto Matsuo[1], Jinhui Chao[2], and Shigeo Tsujii[3]

[1] Research and Development Initiative, Chuo University,
42-8 Ichigaya Honmuracho, Shinjuku-ku, Tokyo, 162-8473 Japan
mats@m.ieice.org

[2] Dept. of Electrical, Electronic and Communication Engineering, Chuo University,
1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8851 Japan
jchao@elect.chuo-u.ac.jp

[3] Dept. of Information and System Engineering, Chuo University,
1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8851 Japan
tsujii@ise.chuo-u.ac.jp

Abstract. Counting the number of points of Jacobian varieties of hyperelliptic curves over finite fields is necessary for construction of hyperelliptic curve cryptosystems. Recently Gaudry and Harley proposed a practical algorithm for point counting of hyperelliptic curves. Their algorithm consists of two parts: firstly to compute the residue modulo an integer m of the order of a given Jacobian variety, and then search for the order by a square-root algorithm. In particular, the parallelized Pollard's lambda–method was used as the square-root algorithm, which took 50CPU days to compute an order of 127 bits.

This paper shows a new variation of the baby step giant step algorithm to improve the square–root algorithm part in the Gaudry-Harley algorithm. With knowledge of the residue modulo m of the characteristic polynomial of the Frobenius endomorphism of a Jacobian variety, the proposed algorithm provides a speed up by a factor m, instead of $\sqrt{m}$ in square–root algorithms. Moreover, implementation results of the proposed algorithm is presented including a 135–bit prime order computed in 16 hours on Alpha 21264/667MHz.

1 Introduction

Security of hyperelliptic curve cryptosystems depends in an essential way on the orders of Jacobian varieties of the hyperelliptic curves used in the systems. In particular, it is believed that if a small genus curve is used, and the order of its Jacobian variety contains a prime number with a small cofactor, coprime to the characteristic of the definition field and immune to the Weil/Tate pairing reduction [7], then the hyperelliptic curve cryptosystem can resist all known attacks except maybe the Weil–Descent attack [8].

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 461–474, 2002.

Therefore, the order counting for Jacobian varieties of random hyperelliptic curves is one of the most important problems in construction of hyperelliptic curve cryptosystems.

Recently several researches have been reported on point counting algorithms. In particular, efficient algorithms [15,11,12] have been proposed for curves over small characteristic finite fields. Using these algorithms, it is possible to compute orders of Jacobian varieties in sizes for cryptographic usage (e.g. 160 bits) over such fields,

On the other hand, the situation is quite different for point counting of curves over finite fields with arbitrary characteristics. Although a number of theoretical results such as [21,14,1,13] have been known, it is only until very recent that a practical point counting algorithm for curves over large characteristic finite fields is proposed and implemented by Gaudry and Harley [9,10].

The Gaudry–Harley algorithm consists of two parts: the first part is to compute the residue modulo m, $m \in \mathbb{Z}_{>0}$ of the order of a given Jacobian variety. The second part is to search for the order by an algorithm square root complexity using its modulo m residue. (We will misuse "square-root algorithm" referring to these search algorithms hereafter). This is a natural generalization of the point counting algorithm of elliptic curves proposed in [20] to hyperelliptic curves. In particular, the parallelized Pollard's lambda–method was used as the searching algorithm. It seemed that the square-root algorithm is the most time consuming part in the Gaudry–Harley algorithm. For an example, this part took actually 50CPU days to compute an order of 127 bits.

This paper proposes an improvement of the baby step giant step algorithm, and applies it as the square–root algorithm in the Gaudry-Harley algorithm. It is shown that with knowledge of the residue modulo m of the characteristic polynomial of the Frobenius endomorphism of a Jacobian variety, the proposed search algorithm provides a speed up by a factor m, instead of $\sqrt{m}$ in the usual square–root algorithms. (All complexity estimates hereafter are in terms of operations in the Jacobians). Moreover, implementation results of the proposed algorithm are presented, including a 135–bit order is computed in 16 hours on Alpha 21264/667MHz by using the proposed algorithm. A recently obtained example of 160 bit is also shown in the appendix.

2 Hyperelliptic Curves over Finite Fields and the Orders of Their Jacobian Varieties

Let p be an odd prime number, $\mathbb{F}_q$ a finite field of order q with $\mathrm{char}(\mathbb{F}_q) = p$. Let g be a positive integer. Then a genus g hyperelliptic curve $C/\mathbb{F}_q$ is defined as follows:

$$\begin{aligned} C :& Y^2 = F(X), \\ & F(X) = X^{2g+1} + f_{2g}X^{2g} + \cdots + f_0, \end{aligned} \tag{1}$$

where $f_i \in \mathbb{F}_q$, $\mathrm{disc}\,(F) \neq 0$.

We will restrict ourselves to $g = 2$ cases hereafter.

Let $\mathcal{J}_C$ be the Jacobian variety of C, $\mathcal{J}_C(\mathbb{F}_q)$ its $\mathbb{F}_q$-rational points. It is known that $\mathcal{J}_C(\mathbb{F}_q)$ is a finite Abelian group, so that discrete logarithm based cryptosystems can be constructed on it.

The characteristic polynomial $\chi_q(X)$ of the qth power Frobenius endomorphism of $\mathcal{J}_C$ is given as follows [23,14]:

$$\chi_q(X) = X^4 - s_1X^3 + s_2X^2 - s_1qX + q^2, \tag{2}$$

where $s_i \in \mathbb{Z}$ and

$$|s_1| \le 4\sqrt{q}, \tag{3}$$
$$|s_2| \le 6q. \tag{4}$$

Then the order $\#\mathcal{J}_C(\mathbb{F}_q)$ of $\mathcal{J}_C(\mathbb{F}_q)$ can be obtained as

$$\begin{aligned}\#\mathcal{J}_C(\mathbb{F}_q) &= \chi_q(1) \\ &= q^2 + 1 - s_1(q+1) + s_2 \end{aligned}\tag{5}$$

from $\chi_q(X)$ [23].

From (3), (4) and (5), $\#\mathcal{J}_C(\mathbb{F}_q)$ is bounded within the Hasse–Weil range:

$$L_o := \left\lceil (\sqrt{q}-1)^4 \right\rceil \le \#\mathcal{J}_C(\mathbb{F}_q) \le \left\lfloor (\sqrt{q}+1)^4 \right\rfloor =: H_o. \tag{6}$$

3 The Gaudry–Harley Algorithm

In this section a rough description of the Gaudry–Harley algorithm and its implementation results are given according to [9,10]. (See [9,10] for further details).

In Algorithm 1, we show an outline of the Gaudry–Harley algorithm.

Algorithm 1. Gaudry-Harley point counting algorithm

```
Input: A genus 2 HEC $C/\mathbb{F}_q$
Output: $\#\mathcal{J}_C(\mathbb{F}_q)$
Compute $\#\mathcal{J}_C(\mathbb{F}_q) \bmod 2^e$ by the halving algorithm
for prime numbers $l = 3, 5, \ldots, l_{max}$ do
    Compute $\chi_q(X) \bmod l$ by a Schoof-like algorithm
    Compute $\#\mathcal{J}_C(\mathbb{F}_q) \bmod l$ from $\chi_q(X) \bmod l$j
end for
Compute $\chi_q(X) \bmod p$ by using the Cartier-Manin operator
Compute $\#\mathcal{J}_C(\mathbb{F}_q) \bmod p$ from $\chi_q(X) \bmod p$
Compute $\#\mathcal{J}_C(\mathbb{F}_q) \bmod m, m = 2^e \cdot 3 \cdots l_{max} \cdot p$ by CRT
Compute $\#\mathcal{J}_C(\mathbb{F}_q)$ by a square root algorithm using $\#\mathcal{J}_C(\mathbb{F}_q) \bmod m$
```

The capability of Schoof-like algorithm therefore the largest value of l_{max} which can be computed in Step 2 of Algorithm 1 is subject to many factors, such as the size of $\mathbb{F}_q$. To construct secure curves, $l_{max} = 13$ is a reasonable estimate at the present. Since the computation of the Cartier–Manin operator in Step 6 costs exponential time in $\log p$, it can not be applied to curves over prime fields of large sizes in cryptographic usage. Therefore, Step 6 and 7 will be skipped when the algorithm is applied to curves over prime fields. For the details of the Cartier–Manin operator and its computation, see [17,18,26,9,10].

Gaudry and Harley computed the orders of the Jacobian varieties of hyperelliptic curves over a 64–bit prime field and a degree 3 extension of a 16–bit prime field. The orders are 127 bits and 128 bits respectively, and these results seem to be present records of point counting of hyperelliptic curves over prime fields and over large characteristic fields.

However, construction of secure hyperelliptic curve for cryptographic usage is still out of reach. For an example, it took 50CPU days in Step 9 when the 127–bit order is computed. Besides, point counting algorithms usually have to be repeated before a secure curve is found.

4 A Baby Step Giant Step Algorithm Using $\#\mathcal{J}_C(\mathbb{F}_q)$ mod m

It is known that compared with the baby step giant step algorithm, the parallelized Pollard's lambda–method used in Step 9 of Algorithm 1 has merits such as it can be parallelized and needs only constant amount of memories. Both algorithms have essentially the same computational complexities in CPU time and can be used in Step 9 in Algorithm 1.

In this section, as a preliminary to the following sections, we describe a modified version of the standard baby step giant step algorithm which can be used in Step 9 in Algorithm 1. This algorithm is an extension of the algorithm that has been applied for point counting of elliptic curves in [20].

From Step 8 of Algorithm 1, one knows $N_r \in \mathbb{Z}$ such that

$$\#\mathcal{J}_C(\mathbb{F}_q) = N_r + mN_m, 0 \le N_r < m. \tag{7}$$

Therefore, $\#\mathcal{J}_C(\mathbb{F}_q)$ can be obtained by searching for N_m among

$$\lfloor L_o/m \rfloor \le N_m \le \lfloor H_o/m \rfloor . \tag{8}$$

Now we set $n \approx \sqrt{R_o}$, where

$$R_o \approx (H_o - L_o)/m = 8q^{3/2}/m + O(q/m). \tag{9}$$

Then the candidates of $N_m = i + nj$ can be obtained by finding (i, j) such that

$$(N_r + mi)\mathcal{D} = -mnj\mathcal{D} \tag{10}$$

for all $\mathcal{D} \in \mathcal{J}_C(\mathbb{F}_q)$ by searching for a collision between the lhs and the rhs of (10) among

$$0 \le i < n, \tag{11}$$

$$\left\lfloor \frac{L_o}{mn} \right\rfloor - 1 \le j \le \left\lfloor \frac{H_o}{mn} \right\rfloor . \tag{12}$$

Assuming $\#\mathcal{J}_C(\mathbb{F}_q)$ is a prime order and q is large enough, one can compute $\#\mathcal{J}_C(\mathbb{F}_q)$ from the pair (i,j) obtained by the above computation as follows:

$$\#\mathcal{J}_C(\mathbb{F}_q) = N_r + m(i+nj). \tag{13}$$

When $\#\mathcal{J}_C(\mathbb{F}_q)$ is not a prime number, to compute $\#\mathcal{J}_C(\mathbb{F}_q)$ one may have to calculate $N_r + m(i+nj)$ for different divisors $\mathcal{D} \in \mathcal{J}_C(\mathbb{F}_q)$ and their least common multiple[5,9]. However, it is possible to test whether $\#\mathcal{J}_C(\mathbb{F}_q)$ is a prime order or not once the $N_r + m(i+nj)$ is found for a single $\mathcal{D} \in \mathcal{J}_C(\mathbb{F}_q)$. Thus, in construction of secure curves, one can abandon the curves with non-prime orders and only looking for prime orders.

This algorithm costs $O(q^{3/4}/\sqrt{m})$ according to (9).

5 An Improved Baby Step Giant Step Algorithm

It can be noticed that in Algorithm 1 in Section 3, the residues modulo m of s_i for $\chi_q(X)$ in (2) as well as $\#\mathcal{J}_C(\mathbb{F}_q) \bmod m$ can also be obtained by using either Schoof-like algorithm or the Cartier–Manin operator.

In this section, we propose an improved baby step giant step algorithm which makes effectively use of the residues $s_i \bmod m$. In such a way, this algorithm speeds up the baby step giant step algorithm in Section 4 and other square-root algorithms by a factor $O(\sqrt{m})$.

In fact, the searching range of s_i can be effectively reduced by the following tighter estimates of their boundaries.

Lemma 1. *s_1 is bounded by*

$$s_{1l} := -\lfloor 4\sqrt{q} \rfloor \leq s_1 \leq \lfloor 4\sqrt{q} \rfloor =: s_{1u} \tag{14}$$

and s_2 is bounded by

$$s_{2l} := \lceil 2\sqrt{q}|s_1| - 2q \rceil \leq s_2 \leq \left\lfloor \frac{1}{4}s_1^2 + 2q \right\rfloor =: s_{2u}. \tag{15}$$

Proof. The upper bound in (15) is due to [6]. The lower bound was pointed out to the authors by Fumiyuki Momose. See Appendix A for a proof. □

Now let $s_i' \in \mathbb{Z}$ such that

$$0 \leq s_i' < m, \tag{16}$$
$$s_1 = s_1' + mt_1, t_1 \in \mathbb{Z}, \tag{17}$$
$$s_2 = s_2' + mt_2', t_2' \in \mathbb{Z}. \tag{18}$$

Then t_1 in (17) is bounded by

$$L_1 := \left\lfloor \frac{s_{1l}}{m} \right\rfloor \leq t_1 \leq \left\lfloor \frac{s_{1u}}{m} \right\rfloor =: H_1 \tag{19}$$

due to (14) and t'_2 in (18) is bounded by

$$L'_2 := \left\lfloor \frac{s_{2l}}{m} \right\rfloor \leq t'_2 \leq \left\lfloor \frac{s_{2u}}{m} \right\rfloor =: H'_2 \tag{20}$$

due to (15). Moreover let t_2, t_3 be integers such that

$$t'_2 = t_2 + nt_3, \ t_2, t_3 \in \mathbb{Z}, \tag{21}$$

$$0 \leq t_2 < n \tag{22}$$

for a positive integer n, then t_3 is bounded by

$$L_3 := \left\lfloor \frac{s_{2l}}{mn} \right\rfloor - 1 \leq t_3 \leq \left\lfloor \frac{s_{2u}}{mn} \right\rfloor =: H_3 \tag{23}$$

due to (20).

Consequently we have

$$\#\mathcal{J}_C(\mathbb{F}_q) = q^2 + 1 - s'_1(q+1) + s'_2 - m(q+1)t_1 + mt_2 + mnt_3 \tag{24}$$

by substituting (17), (18), (21) into (5). Hence $\#\mathcal{J}_C(\mathbb{F}_q)$ can be computed by finding (t_1, t_2, t_3) satisfying

$$(q^2 + 1 - s'_1(q+1) + s'_2 - m(q+1)t_1 + mnt_3)\mathcal{D} = -mt_2\mathcal{D} \tag{25}$$

for all $\mathcal{D} \in \mathcal{J}_C(\mathbb{F}_q)$ in the ranges of (19), (22), (23). These computations are carried out by collision searching between the lhs and the rhs of (25).

Next we determine the most effective value of n.

The number of the pairs (s_1, s_2) satisfying (14) and (15) is roughly $32q^{3/2}/3$ because

$$\int \frac{1}{4}s_1^2 + 2q - (2\sqrt{q}|s_1| - 2q)\mathrm{d}s_1 = s_1(\frac{1}{12}s_1^2 - \sqrt{q}|s_1| + 4q). \tag{26}$$

Therefore the number S of the triples (t_1, t_2, t_3) is

$$S \approx \frac{32q^{3/2}}{3m^2}. \tag{27}$$

Now we set n as

$$n_0 \approx \sqrt{S} = \frac{4\sqrt{6}q^{3/4}}{3m} \tag{28}$$

then the number of point additions for all (t_1, t_2, t_3) satisfying (14) and (15) is roughly n on both the lhs and the rhs of (25). The algorithm under this setting works most efficiently. Therefore the computational complexity of the algorithm is $O(q^{3/4}/m)$ according to (28). Thus, using this algorithm in the square–root algorithm part of the Gaudry–Harley algorithm, the computation of $\#\mathcal{J}_C(\mathbb{F}_q)$ is $O(\sqrt{m})$ times faster than by the original baby step giant step algorithm in Section 4 and other square-root algorithms.

A new prime order searching algorithm using the proposed algorithm is shown in Algorithm 2.

Remark 1. s_i obtained in the process of the algorithm shown are not always correct values [6].

Algorithm 2. An improved baby step giant step algorithm for finding prime order curves

Input: A genus 2 HEC $C/\mathbb{F}_q$, $m, s_1', s_2' \in \mathbb{Z}_{>0}$ such that $s_i \equiv s_i' \bmod m$ and $0 \leq s_i' < m$

Output: $\#\mathcal{J}_C(\mathbb{F}_q)$, if it is a prime number and $q \geq 137$

```
$n \leftarrow \left\lfloor 4\sqrt{6}q^{3/4}/(3m) \right\rceil$
$l \leftarrow q^2 + 1 - s_1'(q+1) + s_2'$
Choose a random $\mathcal{D} \in \mathcal{J}_C(\mathbb{F}_q)\backslash\{0\}$
$B \leftarrow \{(b_j = -jm\mathcal{D}, j) \mid 0 \leq j < n\}$
Sort the table $B$ by the entry $b_j$
$\mathcal{D}_1 \leftarrow l\mathcal{D}$
for $i = -\lceil \lfloor 4\sqrt{q} \rfloor / m \rceil \dots \lfloor 4\sqrt{q}/m \rfloor$ do
  $\mathcal{D}_2 \leftarrow \mathcal{D}_1 - im(q+1)\mathcal{D}$
  $s_1 \leftarrow s_1' + im$
  for $k = \lfloor (\lceil 2\sqrt{q}|s_1| \rceil - 2q)/(mn) \rfloor - 1 \dots \lfloor (\lfloor s_1^2/4 \rfloor + 2q)/(mn) \rfloor$ do
    $\mathcal{D}_3 \leftarrow \mathcal{D}_2 + kmn\mathcal{D}$
    if $^{\exists}j$ such that $b_j = \mathcal{D}_3$ then
      $l \leftarrow l + (-i(q+1) + j + kn)m$
      if $l =$ a prime number then
        Output $l$ as $\#\mathcal{J}_C(\mathbb{F}_q)$ and terminate
      else
        $\#\mathcal{J}_C(\mathbb{F}_q)$ is not a prime number and terminate
      end if
    end if
  end for
end for
```

6 Implementation and Construction of Prime Order Curves

Algorithm 2 is implemented to construct genus 2 prime order hyperelliptic curves. We also implemented the computation of both $s_i \bmod 2$ and $s_i \bmod p$ by the Cartier–Manin operator in order to obtain $s_i \bmod m$.

6.1 Computation of s_i mod 2

For construction of prime order curves, the residues of s_i modulo 2 are in fact fixed by $2 \nmid \#\mathcal{J}_C(\mathbb{F}_q)$ according to the following lemma.

Lemma 2.

$$2 \nmid \#\mathcal{J}_C(\mathbb{F}_q) \Leftrightarrow F : \textit{irreducible}/\mathbb{F}_q \Leftrightarrow 2 \nmid s_i \tag{29}$$

Below, we choose irreducible F in the definition equation (1) of curves and set $s_i \equiv 1 \bmod 2$.

6.2 Computation of s_i mod p

We compute $s_i \bmod p$ by using the Cartier–Manin operator [17,18,26,9,10].

When the characteristic p is large, the dominant part of the Cartier–Manin operator computation is to compute

$$U = \sum u_i X^i = F^{(p-1)/2} \tag{30}$$

for F in (1). This computation itself can be efficiently carried out using FFT multiplication. Here we speed it up even further by the following tricks. We firstly compute

$$V = \sum v_i X^i = \begin{cases} F^{(p-1)/4}, & \text{if } 4 \mid p-1 \\ F^{(p-3)/4}, & \text{if } 4 \nmid p-1 \end{cases} \tag{31}$$

using FFT multiplication. Then it is sufficient to compute only u_{p-2}, u_{p-1}, u_{2p-2}, u_{2p-1} to determine $s_i \bmod p$, from v_i, f_i as

$$U = \begin{cases} V^2, & \text{if } 4 \mid p-1. \\ FV^2, & \text{if } 4 \nmid p-1. \end{cases} \tag{32}$$

This technique can reduce both the computational time and the required memory of the original version using FFT to roughly a half.

6.3 Implementation of Algorithm 2

Algorithm 2 is speeded up by using the following techniques in implementation.

1. Both the computational time and the required memory can be reduced by a factor roughly $1/\sqrt{2}$ using of the property that $-\mathcal{D}$ can be obtained easily from a given $\mathcal{D} \in \mathcal{J}_C(\mathbb{F}_q)$. This is done by choosing $n = \sqrt{2}n_0$, the boundaries of t_2 to be $-n \le t_2 \le n-1$, and the condition of j in Step 12 to be $B_j = \pm\mathcal{D}_3$ and so on. See [20,22] for further details.
2. Although Algorithm 2 is designed to minimize the cost of the worst case computation, it is more appropriate to design an algorithm minimizing the average cost for computation of prime order curves. One can minimize the average cost by choosing $n = (1/\sqrt{2})n_0$ [24,2]. This reduces the average time by a factor roughly $2\sqrt{2}/3$ and the required memory by a factor roughly $1/\sqrt{2}$.
3. We use a 32–bit hash value of b_j in the table B and the precomputation table described in [16] to reduce the required memory.
4. Since practical speed of Algorithm 2 depends on the addition speed on $\mathcal{J}_C(\mathbb{F}_q)$, we use an improved Harley addition algorithm shown in [19].
5. The algorithm is terminated once one checked out that $\mathcal{J}_C(\mathbb{F}_q)$ has a non-prime order.

Remark 2. The average time is reduced by a factor roughly 2/3 and the required memory is reduced roughly to a half, by using the techniques of 1. and 2. simultaneously, here $n = n_0$ is used.

6.4 Implementation Results

Algorithm 3 shows an outline of the construction algorithm of prime order curves used in this section.

Algorithm 3. Construction of a prime order genus 2 hyperelliptic curve

```
Input: A finite field $\mathbb{F}_q$ and $p = \mathrm{char}(\mathbb{F}_q)$
Output: A prime order curve $C$ and $\#\mathcal{J}_C(\mathbb{F}_q)$
repeat
    Choose a monic irreducible polynomial $F/\mathbb{F}_q, \deg F = 5$ randomly
    $C : Y^2 = F$
    Compute $s_{CMi} \equiv s_i \bmod p, 0 \le s_{CMi} < p$ by using the Cartier–Manin operator
    $m \leftarrow 2p$, $s'_i \leftarrow s_{CMi}$ if $2 \nmid s_{CMi}$, else $s'_i \leftarrow s_{CMi} + p$
    Compute $\#\mathcal{J}_C(\mathbb{F}_q)$ by Algorithm 2
until $\#\mathcal{J}_C(\mathbb{F}_q) =$ a prime number
Output $C$ and $\#\mathcal{J}_C(\mathbb{F}_q)$
```

This section shows two examples of genus 2 hyperelliptic curves with prime orders constructed by Algorithm 3 and also timings to compute their orders.

These computations are carried out on a Pentium III/866MHz and a Alpha 21264/667MHz respectively. NTL [25] is used for finite field and polynomial operations.

Example 1. A 123–bit prime order Jacobian variety of the following curve

$$\begin{aligned}
&C_1/\mathbb{F}_q : Y^2 = F_1(X),\\
&F_1 = X^5 + (567033\alpha^2 + 322876\alpha + 957805)X^4 +\\
&(1123698\alpha^2 + 933051\alpha + 141410)X^3 + (393269\alpha^2 + 233572\alpha + 708577)X^2 +\\
&\qquad (692270\alpha^2 + 350968\alpha + 788883)X + 968896\alpha^2 + 895453\alpha + 589750
\end{aligned}$$

is obtained by Algorithm 3, where

$$\begin{aligned}
&\mathbb{F}_q = \mathbb{F}_p(\alpha),\\
&\alpha^3 + 1073470\alpha^2 + 34509\alpha + 1223366 = 0,\\
&p = 1342181.
\end{aligned}$$

The order of $\mathcal{J}_{C_1}(\mathbb{F}_q)$ is

$$\#\mathcal{J}_{C_1}(\mathbb{F}_q) = 5846103764014694479322329315740285931.$$

The computation of $\#\mathcal{J}_{C_1}(\mathbb{F}_q)$ took 197 minutes on Pentium III/866MHz and less than 1GB memory. Table 1 shows the timing of main parts of Algorithm 3 and Algorithm 2 for computing $\#\mathcal{J}_{C_1}(\mathbb{F}_q)$.

Table 1. Timing of computing $\#\mathcal{J}_{C_1}(\mathbb{F}_q)$ on Pentium III/866MHz

Algorithm	Step	Time (min.)
Algorithm 3	Step 4	7
Algorithm 2	Step 4	70
	Step 5	1
	Step 6 – Step 21	119
Total		197

Example 2. A 135–bit prime order Jacobian variety of the following curve

$$\begin{aligned}
&C_2/\mathbb{F}_q : Y^2 = F_2(X),\\
&F_2 = X^5 + (2817153\alpha^2 + 3200658\alpha + 1440424)X^4 +\\
&(3310325\alpha^2 + 481396\alpha + 1822351)X^3 + (108275\alpha^2 + 120315\alpha + 469800)X^2 +\\
&\quad (2168383\alpha^2 + 1244383\alpha + 5010679)X + 4682337\alpha^2 + 53865\alpha + 2540378
\end{aligned}$$

is obtained by Algorithm 3, where

$$\begin{aligned}
&\mathbb{F}_q = \mathbb{F}_p(\alpha),\\
&\alpha^3 + 4519302\alpha^2 + 3749080\alpha + 607603 = 0,\\
&p = 5491813.
\end{aligned}$$

The order of $\mathcal{J}_{C_2}(\mathbb{F}_q)$ is

$$\#\mathcal{J}_{C_2}(\mathbb{F}_q) = 2743433545758123404547331161181818187339271.$$

The computation of $\#\mathcal{J}_{C_2}(\mathbb{F}_q)$ took 16 hours on Alpha 21264/667MHz and less than 4GB memory. Table 2 shows the timing of main parts of Algorithm 3 and Algorithm 2 for computing $\#\mathcal{J}_{C_2}(\mathbb{F}_q)$.

Table 2. Timing of computing $\#\mathcal{J}_{C_2}(\mathbb{F}_q)$ on Alpha 21264/667MHz

Algorithm	Step	Time (min.)
Algorithm 3	Step 4	42
Algorithm 2	Step 4	330
	Step 5	20
	Step 6 – Step 21	557
Total		949

Remark 3. In both Example 1 and 2, the giant steps (Step 6–21 in Algorithm 2) were slower than the baby steps (Step 4, 5 in Algorithm 2). However the average cost of the giant steps is the same as of the baby step. Moreover, the cost of the baby steps is fixed for a fixed definition field.

7 Conclusion and Outlook

This paper proposed an improvement of the baby step giant step algorithm for point counting of hyperelliptic curves over finite fields. In construction of secure hyperelliptic curves of genus 2, with knowledge of the residues modulo m of the characteristic polynomials of the Frobenius endomorphisms, the new algorithm can speed up searching by a factor m, instead of $\sqrt{m}$ in original square–root algorithms. Moreover the algorithm is implemented to find a 135–bit prime order curve.

Computation of an order of a hyperelliptic curve in the size of cryptographic usage is possible if both the algorithm proposed in this paper and the Gaudry-Harley's Schoof–like algorithm are used simultaneously. (See Appendix B for an example). However, computation of the residues modulo a prime l by the Schoof–like algorithm becomes impractical for large l. Thus, it seems difficult at present to use these algorithms to find a curve which is cryptographically interesting.

On the other hand, if one could somehow compute the residues modulo l of s_i for l up to 31, then combining with the proposed algorithm, it will be possible to efficiently compute 160–bit orders for curves over prime fields. In fact, the memory required by the proposed algorithm will be less than 150MB.

Acknowledgement

The authors would like to thank Fumiyuki Momose for Lemma 1. The authors would also like to thank Seigo Arita and Koh-ichi Nagao for their interesting discussions. The authors would also like to thank the referees and Andreas Stein for their helpful comments to improve this paper.

A part of this research was supported by Telecommunications Advancement Organization of Japan (TAO).

References

1. Adleman, L.M., Huang, M.-D. Counting rational points on curves and Abelian varieties over finite fields. In Cohen, H., ed. *ANTS-II*, Lecture Notes in Computer Science, **1122** Springer-Verlag (1996) 1–16
2. Blackburn, S.R., Teske, E. Baby–step giant–step algorithms for non–uniform distributions. In Bosma, W., ed. *ANTS-IV*, Lecture Notes in Computer Science, **1838**, Springer-Verlag (2000) 153–168
3. Bosma, W., Cannon, J. *Handbook of Magma functions*, University of Sydney, (2001) `http://magma.maths.usyd.edu.au/`
4. Cassels, J.W.S., Flynn, E.V. *Prolegomena to middlebrow arithmetic of curves of genus 2*, London Mathematical Society Lecture Note Series, **230**, Cambridge University Press, 1996.
5. Cohen, H. *A Course in Computational Algebraic Number Theory*, Graduate Text in Mathematics, **138**, Springer-Verlag, 1993.

6. Elkies, N.D. Elliptic and modular curves over finite fields and related computational issues. In Buell, D.A., Teitlbaum, J.T., eds. *Computational perspectives on number theory*, AMS (1995) 21–76
7. Frey, G., Rück, H.-G. A remark concerning m-divisibility and the discrete logarithm in the divisor class group of curves, *Math. Comp.* **62** (1994) 865–874
8. Galbraith, S.D. Weil descent of Jacobians. preprint (2001)
9. Gaudry, P., Harley, R. Counting points on hyperelliptic curves over finite fields. In Bosma, W., ed. *ANTS-IV*, Lecture Notes in Computer Science, **1838**, Springer-Verlag (2000) 297–312
10. Gaudry, P. *Algorithmique des courbes hyperelliptiques et applications à la cryptologie*, PhD thesis, École polytechnique (2000)
11. Gaudry, P. Algorithms for counting points on curves. Talk at ECC 2001, The Fifth Workshop on Elliptic Curve Cryptography, Waterloo (2001) `http://www.cacr.math.uwaterloo.ca/conferences/2001/ecc/gaudry.ps`
12. Gaudry, P., Gürel, N. An extension of Kedlaya's point–counting algorithm to superelliptic curves. In Boyd, C., ed. *Advances in Cryptology - ASIACRYPT2001*, Lecture Notes in Computer Science, **2248**, Springer-Verlag (2001) 480–494
13. Huang, M.-D., Ierardi, D. Counting rational point on curves over finite fields. *J. Symb. Comp.*, **25**, (1998) 1–21
14. Kampkötter, W. *Explizite Gleichungen für Jacobische Varietäten hyperelliptischer Kurven*, PhD thesis, GH Essen (1991)
15. Kedlaya, K.S. Counting points on hyperelliptic curves using Monsky–Washinitzer cohomology. to appear in the *J. Ramanujan Mathematical Society* (2001)
16. Lehmann, F., Maurer, M., Müller, V., Shoup, V. Counting the number of points on elliptic curves over finite fields of characteristic greater than three. In Adleman, L., M.D.Huang, eds. *ANTS-I*, Lecture Notes in Computer Science, **877**, Springer-Verlag (1994) 60–70
17. Manin, J.I. The theory of commutative formal groups over fields of finite characteristic. Russian Mathematical Surveys **18** (1963) 1–83
18. Manin, J.I. The Hasse–Witt matrix of an algebraic curve. *Trans. AMS* **45** (1965) 245–264
19. Matsuo, K., Chao, J., Tsujii, S. Fast genus two hyperelliptic curve cryptosystems. Technical Report ISEC2001-31, IEICE Japan (2001)
20. Menezes, A., Vanstone, S., Zuccherato, R. Counting points on elliptic curves over $\mathbb{F}_{2^m}$. *Math. Comp.* **60** (1993) 407–420
21. Pila, J. Frobenius maps of Abelian varieties and finding roots of unity in finite fields. *Math. Comp.* **55** (1990) 745–763
22. Stein, A., Teske, E. Optimized baby step–giant step methods and applications to hyperelliptic function fields. Technical Report CORR 2001-62, Department of Combinatorics and Optimization, University of Waterloo (2001)
23. Stichtenoth, H. *Algebraic function fields and codes*, Universitext, Springer-Verlag, 1993.
24. Teske, E. Square–root algorithms for the discrete logarithm problem (A survey), In *Public–Key Cryptography and Computational Number Theory*, Walter de Gruyter, Berlin–New York (2001) 283–301
25. Shoup, V. A tour of NTL, (2001) `http://www.shoup.net/ntl/`
26. Yui, N. On the Jacobian varieties of hyperelliptic curves over fields of characteristic $p > 2$. *J. Algebra* **52** (1978) 378–410

Appendix A. A Proof of the Lower Bound of s_2

This section shows a proof of the lower bound of s_2 in Lemma 1:

$$s_2 \geq \lceil 2\sqrt{q}|s_1| - 2q \rceil \,. \qquad (*)$$

Let $\alpha, \alpha^\rho, \beta, \beta^\rho$ be the eigenvalues of the qth-power Frobenius endomorphism of $\mathcal{J}_C$, where ρ is for complex conjugate. Let $a_1 = \alpha + \alpha^\rho$ and $a_2 = \beta + \beta^\rho$. Then

$$s_1 = a_1 + a_2,$$
$$s_2 = a_1 a_2 + 2q,$$

$a_i \in \mathbb{R}$, and $|a_i| \leq 2\sqrt{q}$, because $|\alpha|$=$|\beta|$= $\sqrt{q}$ [23].

Firstly, we assume $a_i \geq 0$, then

$$s_2 - 2\sqrt{q}s_1 + 2q = (a_1 - 2\sqrt{q})(a_2 - 2\sqrt{q}) \geq 0.$$

This leads to $(*)$.

Next, assume $a_i < 0$, then

$$s_2 + 2\sqrt{q}s_1 + 2q = (a_1 + 2\sqrt{q})(a_2 + 2\sqrt{q}) \geq 0.$$

This leads also to $(*)$.

Finally, assume $a_1 \geq 0$ and $a_2 < 0$. Then

$$-s_2 + 2\sqrt{q}s_1 - 2q = (a_1 - 2\sqrt{q})(-a_2 + 2\sqrt{q}) \leq 0.$$

which leads to $(*)$ again, if $s_1 \geq 0$. Moreover we also have

$$-s_2 - 2\sqrt{q}s_1 - 2q = -(a_1 + 2\sqrt{q})(a_2 + 2\sqrt{q}) \leq 0.$$

which leads to $(*)$, if $s_1 \leq 0$.

Appendix B. A Recent Example

After we submitted an earlier version of this paper, a 160–bit (but non–prime) order was obtained by using both the Gaudry-Harley's Schoof–like algorithm and the algorithm proposed in this paper simultaneously. A 64-bit hash value of b_j is used in the table B of Algorithm 2. For the computation of the Gaudry-Harley's Schoof–like algorithm, we used Magma V.2.8 [3] and its inner package of the Gaudry-Harley's Schoof–like algorithm written by Gaudry.

For the curve

$$\begin{aligned}
&C_3/\mathbb{F}_q : Y^2 = F_3(X), \\
&F_3 = X^5 + (508797\alpha^3 + 672555\alpha^2 + 940125\alpha + 153314)X^3 + \\
&(330843\alpha^3 + 367275\alpha^2 + 910087\alpha + 1002854)X^2 + \\
&(488395\alpha^3 + 873290\alpha^2 + 734350\alpha + 7072)X + \\
&\qquad\qquad 180553\alpha^3 + 25142\alpha^2 + 806296\alpha + 724502,
\end{aligned}$$

the order of the Jacobian variety is obtained as

$$\begin{aligned}\#\mathcal{J}_{C_3}(\mathbb{F}_q) =&1461445886397612447866396786769393107114349704111\\ =&37 \times 79 \times 6055499440163 \times\\ &8256651526520020642310545028 7439,\end{aligned}$$

where

$$\begin{aligned}&p = 2^{20} - 5,\\ &\mathbb{F}_q = \mathbb{F}_p(\alpha),\\ &\alpha^4 + 278680\alpha^3 + 445675\alpha^2 + 218811\alpha + 653340 = 0.\end{aligned}$$

In the process to compute $\#\mathcal{J}_{C_3}(\mathbb{F}_q)$, Pentium III/866MHz is used for the Schoof-like algorithm part and Itanium/800MHz for the other parts and less than 12GB memory is required.

Table 3 shows the timing of main parts to compute $\#\mathcal{J}_{C_3}(\mathbb{F}_q)$.

Table 3. Timing of computing $\#\mathcal{J}_{C_3}(\mathbb{F}_q)$

Algorithm	l / Step	Time
Schoof-like	3	27sec.
	5	14min. 46sec.
	7	3hrs. 10min. 37sec.
	11	20days 20hrs. 23min. 38sec.
Algorithm 3	Step 4	10min. 42sec.
Algorithm 2	Step 4	1day 23hrs. 22min. 22sec.
	Step 5	2hrs. 5min. 15sec.
	Step 6 – Step 21	2days 23hrs. 3min. 34sec.
Total		26days 19hrs. 31min. 21sec.

Factoring $N = pq^2$ with the Elliptic Curve Method

Peter Ebinger[1] and Edlyn Teske[2]

[1] Universität Karlsruhe
Institut für Algorithmen und Kognitive Systeme
Postfach 6980, 76128 Karlsruhe, Germany
peter@ebinger.de
[2] University of Waterloo
Department of Combinatorics & Optimization
Waterloo, Ontario, N2L 3G1 Canada
eteske@uwaterloo.ca

Abstract. Various cryptosystems have been proposed whose security relies on the difficulty of factoring integers of the special form $N = pq^2$. To factor integers of that form, Peralta and Okamoto introduced a variation of Lenstra's Elliptic Curve Method (ECM) of factorization, which is based on the fact that the Jacobi symbols $\left(\frac{a}{N}\right)$ and $\left(\frac{a}{P}\right)$ agree for all integers a coprime with q. We report on an implementation and extensive experiments with that variation, which have been conducted in order to determine the speed-up compared with ECM for numbers of general form.

1 Introduction

In the past years, several cryptographic systems whose security relies on the difficulty of factoring numbers of the form $N = pq^2$ have been proposed. For example, Takagi [Tak98] introduced an RSA-type cryptosystem modulo $p^k q$, where $k \geq 2$. Hühnlein et al. (see [Hue00] and the references given there) consider cryptosystems based on the difficulty of computing discrete logarithms in class groups of non-maximal quadratic orders of discriminant $\Delta_q = \Delta q^2$, where the trapdoor information is the conductor q and hence hidden in the factors of Δ_q. The EPOC [FKM$^+$00] cryptosystems work with a secret key g_p in $(\mathbb{Z}/(p^2q)\mathbb{Z})^*$ whose order modulo p^2 is p. In all these examples, both prime factors are of approximately the same size.

Cryptosystems based on the problem of factoring pq^2 do not *a priori* need to work with a larger modulus than the RSA cryptosystem (see [MvOV96]) whose security relies on the difficulty of factoring numbers of the form pq. The most powerful algorithm to factor pq is the number field sieve whose complexity depends on the size of the number to be factored. For factoring pq^2, also the complexity of ECM has to be considered, which is a function of the size of the least factor. Now, it is estimated [Len01] that factoring a 1024-bit RSA modulus $N = pq$ is computationally equivalent to finding a 341-bit factor of a 3-prime

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 475–490, 2002.

modulus N of the same size. The question is if the same is true if $N = pq^2$ rather than a 3-prime modulus.

In 1996, Okamoto and Peralta [PO96] introduced a variant of ECM [Len87] that was said to speed up the factorization of numbers of the form pq^2 for p sufficiently smaller than q. This algorithm used a concept called Jacobi signatures, and was later improved through the introduction of (pseudo-)random walks based on Jacobi symbols (see [Per01]). Both variants are based on the fact that if $N = pq^2$, then for all $a \not\equiv 0 \pmod q$ we have $\left(\frac{a}{N}\right) = \left(\frac{a}{p}\right)$. (If $a \equiv 0 \pmod q$ and $\left(\frac{a}{N}\right) \neq \left(\frac{a}{p}\right)$, we have $\left(\frac{a}{N}\right) = 0$ and $\gcd(a, N) \in \{q, q^2\}$, and we can factor N.)

Based on theoretical estimates, Peralta [Per01] states that the improved algorithm using Jacobi symbols speeds up ECM by a factor in the order of $\log p$. However, this is compared with the one-stage variant of ECM (referred to as Standard ECM in this paper) rather than the improved versions of ECM consisting of two stages (called continuations). To estimate the impact of the Peralta-Okamoto algorithm on the factoring problem the question is of how much can the use of Jacobi symbols speed up ECM to factor $N = pq^2$ compared with the fastest available variant of ECM to factor numbers of general form?

Using the implementation of ECM in the computer algebra system LiDIA [LiD00], we implemented the Peralta-Okamoto algorithm, to which we refer as the Jacobi Symbol Continuation, and did extensive experiments to optimize its parameters. We determined average running times for factoring pq^2 with $p \approx q$ and p, q between 10^9 and 10^{14} (30 to 47 bits) with this implementation and compared them with timings using the LiDIA implementation of the so-called Improved Standard Continuation of ECM, which is designed to find factors of integers of general form. We found that the Improved Standard Continuation was, on average, about twice as fast as our implementation of the Jacobi Symbol Continuation. Although these findings seem to suggest that factoring a number of the form $N = pq^2$ ($p \approx q$) with ECM is no easier than finding a factor of approximately the same size of a number of general form, further experiments should be conducted for considerably larger factors where the Fast Fourier Transform Continuation is best.

Our paper is organized as follows. In Section 2 we outline the Elliptic Curve Method and some of its continuations for factoring integers of general form. In Section 3 we present the Jacobi Symbol Continuation of ECM for numbers of the form pq^2. This includes a discussion of how to optimally implement and work with the random walks that are needed in this continuation. Section 4 gives theoretical estimates for the running times of the Improved Standard Continuation and the Jacobi Symbol Continuation. In Section 5 we report on various experiments to optimize the choice of parameters in the Jacobi Symbol Continuation. Section 6 contains a performance comparison with the Improved Standard Continuation. We conclude in Section 7.

2 The Elliptic Curve Method (ECM)

In brief, the idea of ECM is the following: one uses the addition formulae for the group of points of an elliptic curve over a finite field to perform arithmetic operations in the set of points of an elliptic curve defined over $(\mathbb{Z}/N\mathbb{Z})$, ignoring the fact that $(\mathbb{Z}/N\mathbb{Z})$ is not a field for N composite. This either works well, or inversion modulo N fails and thus a factor of N is revealed. Crucial for the performance of ECM is to make sure that the latter happens with a sufficiently high probability.

In our following description of ECM and its variants we restrict ourselves to the essentials. We refer to [CP01] for a survey on various tricks to speed up the algorithms.

2.1 Elliptic Curves

Let us first briefly review elliptic curves modulo p (p prime) and modulo N (N composite).

If p is prime, $p \neq 2, 3$ and $\mathbb{F}_p$ is the finite field with p elements, the *elliptic curve* E over $\mathbb{F}_p$ is given by the equation $y^2 = x^3 + ax + b$, where a,$b \in \mathbb{F}_p$ and such that $4a^3 + 27b^2 \neq 0$ (in $\mathbb{F}_p$). The set of points (x, y) satisfying this equation, together with the point at infinity that is denoted by $\mathcal{O}$ and serves as neutral element, forms a finite abelian group, $E(\mathbb{F}_p)$, which is written additively. The inverse of the point $P = (x_P, y_P)$ is $-P = (x_P, -y_P)$. To add the points $P = P(x_P, y_P)$ and $Q = (x_Q, y_Q)$, both $P, Q \neq \mathcal{O}$ and $P \neq -Q$ we do the following: Let

$$\lambda = \frac{y_P - y_Q}{x_P - x_Q} \qquad \text{if } x_P \neq x_Q, \qquad \lambda = \frac{3x_P^2 + a}{2y_P} \qquad \text{if } x_P = x_Q \ . \tag{1}$$

Then $R = P + Q$, where $R = (x_R, y_R)$ with

$$x_R = \lambda^2 - x_P - y_P \quad \text{ and } \quad y_R = -y_P - \lambda(x_R - x_P) \ . \tag{2}$$

Now let N be an odd integer not divisible by three, i.e., $N = \prod_{i=1}^{r} p_i$ with $p_i > 3$ for all prime factors p_i. Let $a, b \in \mathbb{Z}/N\mathbb{Z}$ with $4a^3 + 27b^2 \in (\mathbb{Z}/N\mathbb{Z})^*$. In the projective plane, we define the elliptic curve

$$E(\mathbb{Z}/N\mathbb{Z}) := \{(x : y : z) \in \mathbb{P}^2(\mathbb{Z}/N\mathbb{Z}) : \ y^2 z = x^3 + axz + bz^3\} \ .$$

If the p_i are known and pairwise distinct, the addition law in $E(\mathbb{Z}/N\mathbb{Z})$ can be performed by applying the addition law in each group $E(\mathbb{Z}/p_i\mathbb{Z})$ and then using Chinese Remaindering. This works since $E(\mathbb{Z}/N\mathbb{Z}) \cong E(\mathbb{Z}/p_1\mathbb{Z}) \times \cdots \times E(\mathbb{Z}/p_r\mathbb{Z})$. If the p_i are not known, for addition in $E(\mathbb{Z}/N\mathbb{Z})$ we simply use the same formulae as for addition in $E(\mathbb{Z}/p\mathbb{Z}) = E(\mathbb{F}_p)$ with p prime. Then we also can switch to affine coordinates, and can use the formulae given above. However, the arithmetic in (1), (2) is modulo N, rather than modulo the prime p. Thus, to compute λ we have to compute the inverses modulo N of $x_P - x_Q$ and $2y_P$, respectively. If $x_P - x_Q$, or $2y_P$, is not coprime to N, the Euclidean algorithm to compute the respective inverse fails, and this reveals a non-trivial factor of N. It is easy to see that this happens if $U + V = \mathcal{O}_{p_i}$ for some prime factor p_i of N but $U + V \neq \mathcal{O}_N$. (Here $\mathcal{O}_m$ stands for the point at infinity of $E(\mathbb{Z}/m\mathbb{Z})$.) This motivates the strategy of ECM, which is: compute multiples of kP of a point P on the curve modulo N and hope that $kP = \mathcal{O}_{p_i}$ but $kP \neq \mathcal{O}_N$.

2.2 Standard Version of ECM

We have the following algorithm:

Algorithm: Standard ECM.
Input: N.
Output: A non-trivial factor of N, or "no success".

1. Choose a random curve $E = E_{a,b}$ and a random point $P = (x, y)$ on E:
$$a, x, y \in_R [0, \ldots, N-1] \text{ and } b = y^2 - x^3 - ax \ ,$$
and check that $\gcd(4a^3 + 27b^2, N) = 1$. [Here, and throughout the paper, $\in_R$ indicates chosen at random with respect to the uniform distribution.]
2. Choose *smoothness bounds* B and C. ($B, C \in \mathbb{N}$, to be specified below.)
3. For all primes $q \leq B$ do
$$\begin{aligned} e_q &= \max\{e : q^e \leq C < q^{e+1}\} \ . \\ P &\leftarrow q^{e_q} \cdot P \ . \end{aligned}$$
If the latter computation fails because some inverse modulo N cannot be computed, a factor of N is found. Return that factor.
4. If you want to continue, go back to Step 1. Otherwise, output "no success".

Upon completion of Step 3, Standard ECM has computed the multiple $k_{B,C} \cdot P$ of the point P, where

$$k_{B,C} = \prod_{2 \leq q \leq B} q^{e_q} \ , \qquad q^{e_q} \leq C \ .$$

Then $k_{B,C}P = \mathcal{O}_p$ if and only if all prime factors of $\operatorname{ord} P$ modulo p are less than or equal to B, and all prime powers dividing $\operatorname{ord} P$ are less than or equal to C. Here, $\operatorname{ord} P$ modulo p stands for the element order of the point P in $E(\mathbb{Z}/p\mathbb{Z})$.

Remark 1. In fact, the scalar multiple in Step 3 is calculated using the Montgomery-Chudnowsky representation of an elliptic curve (see [Mon92]). This representation uses projective coordinates and has the advantage that no inversions are needed for doubling and addition of points. A nontrivial factor of N then won't be detected by a failure of an inversion in Step 3. Instead, at the end of Step 3, we compute $\gcd(N, z)$, where z is the z-coordinate of $k_{B,C}P$ in Montgomery-Chudnowsky representation, which is zero modulo p if $k_{B,C}P = \mathcal{O}_p$.

2.3 Smoothness and Semi-smoothness

Definition 1. *Let B be a positive integer. An integer n is called* smooth with respect to B *if all its prime factors are less than or equal to B.*

For a point P on the elliptic curve $E(\mathbb{Z}/N\mathbb{Z})$ and for a prime factor p of N the Hasse bound on $\#E(\mathbb{Z}/p\mathbb{Z})$ implies that $\operatorname{ord} P \pmod{p} \leq p + 1 + 2\sqrt{p}$. Thus, with $C = p + 1 + 2\sqrt{p}$, Standard ECM succeeds for any elliptic curve whose order modulo p is B-smooth. In practice we put $B = C$, and accordingly, we let $k_B = k_{B,C}$.

Remark 2. The smoothness properties of *randomly chosen integers* are well studied and can be expressed in terms of Dickman's rho function. See, for example, [CP01]. Lenstra [Len87] shows that the smoothness properties of group orders of elliptic curves modulo p in the interval $]p+1-\sqrt{p}, p+1+\sqrt{p}[$ resemble those of randomly chosen integers in that interval.

It is obvious that the larger B is, the larger is the probability that an elliptic curve group order is B-smooth. On the other hand, the larger B is, the more expensive is the cost of Step 3 of Standard ECM. An efficient way to increase the size of the largest possible prime factor of $\#E(\mathbb{Z}/p\mathbb{Z})$ so that ECM still works is the *large-prime variant.* It is based on the fact that for a random integer it is very likely that exactly one prime factor is much bigger than all the other factors.

Definition 2. *Let B, D be positive integers. An integer n is* semi-smooth with respect to B and D *if $p \leq B$ for all prime factors p of n with the possible exception of one prime $q \mid n$, for which $q \leq D$ and $q^2 \not| n$.*

For $u > v > 1$ let $\sigma(u, v)$ denote the asymptotic probability that a random integer modulo p is semi-smooth with respect to $p^{1/u}$ and $p^{1/v}$. Precisely, let $\sigma(u, v) = G(1/u, 1/v)$, where, for $0 < \alpha < \beta < 1$, $G(\alpha, \beta) = \lim_{x\to\infty} \Psi(x, x^\beta, x^\alpha)/x$ with $\Psi(x, y, z) =$

$$\#\{n \leq x \,:\, n = n_1 n_2 \cdots n_r, n_1 \geq n_2 \geq \cdots \geq n_r, n_i \text{ prime }, n_1 \leq y, n_2 \leq z\}.$$

Bach and Peralta [BP96] give explicit formulae to efficiently compute $\sigma(u, v)$.

2.4 Continuations

If Standard ECM is not successful in factoring N, we switch from Montgomery-Chudnowsky representation to the Weierstrass representation and proceed with a second stage, called *continuation.* For this, we take the point $W := k_B P$ calculated in Step 3 of Standard ECM and hope that $\operatorname{ord} W$ modulo p is a prime number less than or equal some integer D. This is the case if $\operatorname{ord} P$ is semi-smooth with respect to B and D.

The *Standard Continuation* of ECM works as follows: for all primes $r \in (B, D]$, compute rW. If $rW = \mathcal{O}_p$ for some prime factor p of N but $rW \neq \mathcal{O}_N$, a non-trivial factor of N is revealed. This can be implemented to take $O(D-B)$ group operations.

The *Improved Standard Continuation* (ISC) is: Use the baby-step giant-step method to find a prime $r \in (B, D]$ such that $rW = \mathcal{O}_p$ for some $p \mid N$. If $r \in (B, D]$ and $w = \lceil \sqrt{D} \rceil$, then there exist $0 \leq i < w$ and $1 \leq j \leq w$ such that $r = jw - i$. Then $rW = O_p$ if and only if $jwW = iW$ in $E(\mathbb{Z}/p\mathbb{Z})$. Computing the baby steps iW and the giant steps jwW takes $O(\sqrt{D})$ group operations. Table look-ups for equality checks, as they are usually used in baby-step giant-step applications, are not possible here since p is not known. Instead, for an equality check $iW = jwW$ modulo some prime factor p of N we compute $\gcd(x_{iW} - x_{jwW}, N)$. This is necessary only for i of the form $i = jw - r$ with $r \in (B, D]$ prime, i.e., by the prime number theorem, for about $D/\ln D - B/\ln B$ values of r. By accumulating the differences $(x_{iW} - x_{jwW})$, each such gcd computation can be replaced by a multiplication followed by reduction modulo N.

The optimal choices for the semi-smoothness parameters depend on the size of the prime factor p we want to find. In Table 1 we give sample values for B $(= C)$ and D as used in the LiDIA implementation [LiD00] of the ISC. We also indicate the corresponding u and v such that $B = p^{1/u}$ and $D = p^{1/v}$, and the semi-smoothness probabilities $\sigma(u, v)$. The latter were computed using software that was made available to us by Peralta.

Table 1. Semi-smoothness bounds - Selected LiDIA parameter for Improved Standard Continuation. Semi-smoothness probabilities.

Size of p	B	D	u	v	$\sigma(u,v)$	Size of p	B	D	u	v	$\sigma(u,v)$
10^9	349	4297	3.5	2.5	0.08	10^{12}	997	31489	4.0	2.7	0.04
10^{10}	411	8861	3.8	2.5	0.06	10^{13}	1439	54617	4.1	2.7	0.03
10^{11}	659	17981	3.9	2.6	0.05	10^{14}	2111	89501	4.2	2.8	0.02

There is another continuation, based on the *Fast Fourier Transform* [Mon92]. This continuation is the faster than the ISC only if we look for very big factors p of N, of 20 or more digits [Mue95]. We do not go into the details of this method since in our study, we will be working with smaller factors and thus will be using the ISC as a reference.

Numerous problems in computational number theory that can be solved using baby-step giant-step techniques can also be tackled using Pollard's rho method [Pol75,Pol78]. Let us look at that option. As before, let $W = k_B P$. Let $t_N = \operatorname{ord} W$ modulo N and $t_p = \operatorname{ord} W$ modulo p. We could use Pollard's rho method with an iterating function F that via $Q_{i+1} = F(Q_i)$ defines a (pseudo-) random walk $(Q_i)_{i \in \mathbb{N}}$ in the subgroup generated by W. However, with the factorization of N unknown, in general this is possible only modulo N. In general, we *cannot* define an iterating function such that $Q_{i+1} = F(Q_i)$ modulo p. Thus the expected running time of a Pollard rho application is $O(\sqrt{t_N})$ and not $O(\sqrt{t_p})$, i.e., in terms of the number N to be factored rather than the factor p.

3 Continuation of ECM with Jacobi Symbols

Let p be an odd prime. For an integer a, consider the congruence

$$x^2 \equiv a \pmod{p} \ . \tag{3}$$

We define the *Legendre symbol* $\left(\frac{a}{p}\right)$ as follows: $\left(\frac{a}{p}\right) = 0$ if $p|a$; $\left(\frac{a}{p}\right) = 1$ if $p \not| a$ and (3) has solutions; and $\left(\frac{a}{p}\right) = -1$ if (3) has no solutions. This definition can be extended to composite numbers, as follows (cf. [Coh93]):

Definition 3. *Let $N \in \mathbb{N}$. For an integer a, we define the* Jacobi *symbol* $\left(\frac{a}{N}\right)$*:* $\left(\frac{a}{1}\right) = 1$. $\left(\frac{a}{2}\right) = 0$ *if a is even, and* $\left(\frac{a}{2}\right) = (-1)^{(a^2-1)/8}$ *if a is odd.*
If $N = \prod_{i=1}^{l} p_i$ with the p_i not necessarily distinct, then $\left(\frac{a}{N}\right) = \prod_{i=1}^{l}\left(\frac{a}{p_i}\right)$*, where* $\left(\frac{a}{p}\right)$ *is the Legendre symbol.*

From now on, let N be of the form $N = pq^2$. Then, for all $a \not\equiv 0 \pmod{q}$,

$$\left(\frac{a}{N}\right) = \left(\frac{a}{p}\right) \ . \tag{4}$$

We now explain how this property can be exploited in the *Jacobi Symbol Continuation* (JC) of ECM.

Let $W = k_B P$ be the point computed by ECM's first stage, i.e., Step 3 of Standard ECM, and let t_p and t_N as above. We define a sequence $(Q_i)_{i \in \mathbb{N}} \subseteq E(\mathbb{Z}/N\mathbb{Z})$ by $Q_0 = W$ and $Q_{i+1} = F(Q_i)$, where

$$F(Q_i) = \begin{cases} 2Q_i \ , & \text{if } \left(\frac{x_{Q_i}}{N}\right) = 1 \ , \\ Q_i + Q_0 \ , & \text{otherwise} \ . \end{cases} \tag{5}$$

This sequence eventually becomes periodic. Since $Q = Q' \pmod{p}$ implies that $F(Q) = F(Q') \pmod{p}$, its period depends only on t_p, not on t_N. Experimentally, we find that on average this sequence becomes periodic after $\approx 1.9\sqrt{t_p}$ iterations. This value is based on taking averages over several 10000 runs with various values t_p ranging from 10^3 to 10^8, using the same approach as in [Tes98b].

Remark 3. A sequence $(Q_i)_{i \in \mathbb{N}}$ generated by a random function F becomes periodic after an expected number of $\sqrt{t_p \pi/2} \approx 1.25\sqrt{t_p}$ iterations [Har60]. We know from [Tes01] that wider iterating functions (in the sense that more partitions are involved) better simulate random functions. For example, if for $m = 1, 2, 3$ we let $m_i \in_R [1, D]$, $V_i = m_i W$ and define

$$F(Q_i) = \begin{cases} 2Q_i & \text{if } \left(\frac{x_{Q_i}}{N}\right) = 1 \text{ and } \left(\frac{x_{Q_i}+1}{N}\right) = 1 \ , \\ Q_i + V_1 & \text{if } \left(\frac{x_{Q_i}}{N}\right) = 1 \text{ and } \left(\frac{x_{Q_i}+1}{N}\right) \neq 1 \ , \\ Q_i + V_2 & \text{if } \left(\frac{x_{Q_i}}{N}\right) \neq 1 \text{ and } \left(\frac{x_{Q_i}+1}{N}\right) = 1 \ , \\ Q_i + V_3 & \text{if } \left(\frac{x_{Q_i}}{N}\right) \neq 1 \text{ and } \left(\frac{x_{Q_i}+1}{N}\right) \neq 1 \ , \end{cases} \tag{6}$$

then we find that on average the sequence given by $Q_0 = W$ and $Q_{i+1} = F(Q_i)$ becomes periodic after only $1.5\sqrt{t_p}$ iterations. However, in contrary to the iterating functions considered in [Tes01], the above widening of the iterating function is rather costly: we need to evaluate two Jacobi symbols in each iteration, instead of just one in (5). As a consequence, we find that in terms of CPU-time, the second sequence takes about 1.15 to 1.2 times longer to become periodic. We therefore work with the function in (5).

Once two points Q_i and Q_j have been found for which $Q_i = Q_j \bmod p$ but $Q_i \neq Q_j \bmod N$, we are likely to find a nontrivial factor of N as $\gcd(x_{Q_i} - x_{Q_j}, N)$. To find such a match (Q_i, Q_j), we use an algorithm from a family of cycle-finding algorithms due to Brent [Bre80]. This algorithm is more than 1.5 times faster[1] than Floyd's method, which is usually referred to in this context. It works as follows:

> Brent's cycle-finding algorithm
> We have an auxiliary element R and an auxiliary index r. We initialize $R = Q_0$ and $r = 2$. We compute the next term Q_i. If $Q_i = R$, a match is found and we stop. If $i \geq r$, we replace r by $2r$ and R by Q_i. Then we compute the next term of the sequence, and so forth.

That is, each term Q_i is checked for equality with the term Q_j where $j = 0$ if $i = 1, 2$, and $j =$ the largest power of 2 strictly less than i if $i > 2$. Of course, in our application the equality check is done by computing $\gcd(x_{Q_i} - x_R, N)$. As in the ISC, gcd's can be accumulated: every s-th iteration we compute the gcd of the product of s consecutive differences $(x_{Q_i} - x_R)$ and N, where s is optimized to obtain minimal running times.

Summing up, the JC works as follows: Let $Q_0 = W$. Use the iterating function (5) to define the sequence $(Q_i)_{i \in \mathbb{N}}$ and use Brent's algorithm to find a match $Q_i = Q_j \bmod p$. Such a match reveals a nontrivial factor of N if $Q_i \neq Q_j \bmod N$.

3.1 When to Discard a Curve?

In the ISC, the maximal running time spent with each curve $E(\mathbb{Z}/N\mathbb{Z})$ is predetermined by the semi-smoothness bounds B and D. If after the application of the baby-step giant-step method no factor of N has been found, we conclude

[1] We compare the performance of cycle-finding algorithms in terms of its *expected delay factors*: If E_ρ is the expected number of iterations until a match occurs, and $E_{l(\rho)}$ is the expected number of iterations until a match is found by a certain cycle-finding method, the expected delay factor is defined as $\delta = E_{l(\rho)}/E_\rho$. For a random function $F : S \to S$, we have $E_\rho = \sqrt{\pi/2}\sqrt{|S|}$; then for Floyd's method, we have $\delta_{\mathrm{Fl}} \approx 3.09/\sqrt{\pi/2} \approx 2.47$ (cf. [Pol75]), while for Brent's cycle-finding algorithm we (experimentally) find $\delta_{\mathrm{Br}} \approx 1.97/\sqrt{\pi/2} \approx 1.57$. Cycle-finding algorithms with even smaller delay factors exist (see [Tes98a]), but they require more equality checks per iteration and thus lose their advantage for our application here.

that the order of W modulo p was not semi-smooth with respect to B and D, and the curve is to be discarded. In the JC, however, the probabilistic character of Pollard's rho method results in a spread of the actual running times for various runs of the algorithm. To illustrate this, let us first consider a randomly chosen iterating function $F : S \to S$ (S any finite set). Let ρ denote the number of iterations until a match $Q_i = Q_j$ occurs. Then the expected value of ρ is, of course, $\beta := \sqrt{|S|\pi/2}$ and, from [Har60], the probability that a match occurs after at most $k\beta$ iterations is given as

$$\mathcal{P}(\rho \le k\beta) = \int_0^{k\sqrt{\pi/2}} xe^{-x^2/2}\,dx = 1 - e^{-k^2\pi/4} =: \mathcal{P}(k)\ . \tag{7}$$

Thus, for example, with probability $\mathcal{P}(1/2) = 1 - e^{-\pi/16} \approx 18\%$ it takes only at most half as long as expected until a match occurs, and with probability $1 - \mathcal{P}(2) = e^{-\pi} \approx 4\%$ it takes at least twice as long as expected. Now, in the JC, we expect a match to be detected after approximately $1.9\sqrt{t_p}\cdot\delta_{\mathrm{Br}} \approx 3\sqrt{t_p}$ iterations. Assuming that the corresponding probability distribution is essentially the same as in the random case, we replace β by $\beta' = 3\sqrt{t_p}$ in (7). Then we find that in the JC a match is indeed detected after at most $3\sqrt{t_p}$ iterations with probability only 54%. On the other hand, plugging $k = 4/3, 5/3$ and 2 into (7) we find that with probabilities $25\%, 11\%$ and 4% it takes at least $4\sqrt{t_p}$, $5\sqrt{t_p}$ and $6\sqrt{t_p}$ iterations, respectively, to detect a match. Thus, if after $3\sqrt{D}$ iterations no match has been found, this may be either because $\operatorname{ord} W$ is not semi-smooth with respect to B and D, *or* because we simply have bad luck in that particular execution of the rho method. A decision when a curve should be given up must be made, and we decided for our implementation to stop after $6\sqrt{D}$ iterations. It is worth noting that this decision does not influence the actual performance of the optimized algorithm, since the semi-smoothness bounds will be optimized experimentally.

3.2 A Remark on Semi-smoothness Probabilities

As a consequence of the spread of the actual number of iterations for a given order t_p, the semi-smoothness probabilities $\sigma(u, v)$ do not exactly reflect the probability of success of the JC for given parameters $B = p^{1/u}$ and $D = p^{1/v}$. When executing $6\sqrt{D}$ iterations, we have a 96% chance of success to factor $N = pq^2$ if $\#E(\mathbb{Z}/p\mathbb{Z})$ is semi-smooth with respect to B and D. But at the same time, we *implicitly* work with larger values of D as well. For example, we also have a 54% chance of success to find the factor p should $\#E(\mathbb{Z}/p\mathbb{Z})$ be semi-smooth with respect to B and $4D$. In fact, the probability of success of the JC for given semi-smoothness parameters B, D is bounded below by

$$\sum_{i=1}^{r} \mathcal{P}(2\sqrt{D/D_i}) \cdot \Big(\sigma\big(u, v(D_i)\big) - \sigma\big(u, v(D_{i-1})\big)\Big)\ ,$$

for any choice of integers D_i with $0 \le D_0 < D_1 < \cdots < D_r < \infty$, and where $v(D_i)$ is such that $D_i = p^{1/v(D_i)}$.

4 Estimated Running Times

We now give some rough estimates for the expected running times of the JC and the ISC. All constants in the following notation are machine- and implementation- dependent. They also depend on the size N of the number to be factored; this dependence is polynomial in $\log N$.

- M: The cost of calculating a scalar multiple tP of a point on an elliptic curve $E(\mathbb{Z}/N\mathbb{Z})$ in Montgomery-Chudnowsky representation is expressed as $M \log t$.
- A: The cost of adding two points on $E(\mathbb{Z}/N\mathbb{Z})$ in Weierstrass representation. [For simplicity, we treat doubling a point and adding two points equally.]
- G: The cost of calculating the greatest common divisor of two numbers of size N.
- I: The cost of a multiplication of integers of size N followed by reduction modulo N.
- J: The cost of calculating a Jacobi symbol $\left(\frac{a}{N}\right)$ for an integer $a \pmod N$.

In Table 2, we list sample values for some of these costs, measured on a Sun Ultra 60 Workstation using LiDIA.

Table 2. Sample running times for point addition on $E(\mathbb{Z}/N\mathbb{Z})$, gcd computation, Jacobi symbol computation

# digits of p, q	# digits of $N = pq^2$	A [μs]	G [μs]	J [μs]
9	27	153.2	19.5	115.9
10	30	169.9	22.3	128.4
11	33	193.3	25.9	143.0
12	36	210.0	29.7	159.4
13	39	228.3	33.1	172.6
14	42	255.9	37.9	185.6

Remark 4. The Jacobi symbol computation as implemented in LiDIA is based on [Coh93, Algorithm 1.4.12] (the “binary” method). For comparison, we also implemented routines based on [CP01, Algorithm 2.3.5], the routine found in Victor Shoup’s NTL, and the LiDIA routine stripped off those commands not needed when N is known to be positive and odd. The running time differences were within 4%, with the LiDIA routines being the fastest. But see also Sect. 7.

With p denoting the prime factor to be found, we work with semi-smoothness bounds $B = p^{1/u}$ and $D = p^{1/v}$, where $u > v > 1$. Note that the optimal choices of u and v are not necessarily the same for the ISC and the JC.

We first estimate the cost of Standard ECM. By the prime number theorem, there are $\pi(B) \approx B/\ln B = up^{1/u}/(\ln p)$ primes less or equal to B. Each such prime is bounded by $B = p^{1/u}$, so that the total cost of Standard ECM is about $1.44Mp^{1/u}$. [Here we use that $\ln p/\log p \approx 1.44$.]

In the ISC, we now have to add the expected cost for the baby-step giant-step stage. If no factor is found in this stage, this takes roughly $\sqrt{D}$ baby and $\sqrt{D}$ giant steps. Otherwise, we need an expected number of $\sqrt{D}$ baby steps and $0.5\sqrt{D}$ giant steps. This results in a total cost $2A\sqrt{D} = 2Ap^{1/(2v)}$ if successful and $3/2 \cdot A\sqrt{D} = 3/2 \cdot Ap^{1/(2v)}$ if not successful. Assuming that s differences $(x_{iW} - x_{jwW})$ are accumulated before a greatest common divisor is computed, we have to perform about $D/(s \ln D)$ gcd computations. This is at an approximate total cost $(I + G/s)D/\ln D = (I + G/s)p^{1/v}v/\ln p$. We expect to have to consider about $1/\sigma(u, v)$ elliptic curves modulo N, so that the total cost of the ISC can be roughly estimated as

$$T_{\text{ISC}} = \frac{1.44Mp^{1/u} + 2Ap^{1/(2v)} + (I + G/s)vp^{1/v}/\ln p}{\sigma(u, v)} \ .$$

In the JC, if no factor is found for a given elliptic curve, we execute $6\sqrt{D} = 6p^{1/(2v)}$ iterations of the (pseudo-)random walk, and otherwise an expected number of $3\sqrt{t_p} \leq 3\sqrt{D}$ iterations. Each iteration requires one Jacobi symbol evaluation and one addition of points. We further assume that s differences $(x_R - x_{Q_i})$ are accumulated before a greatest common divisor is computed. Assume that we have to consider about $1/\sigma(u, v)$ elliptic curves modulo N (see, however, Section 3.2). Then the total cost of the JC can be roughly estimated as

$$T_{\text{JC}} = \frac{1.44Mp^{1/u} + 6(J + A + I + G/s)p^{1/(2v)}}{\sigma(u, v)} \ . \tag{8}$$

These running times show that asymptotically, the JC is faster in factoring $N = pq^2$ than the ISC, assuming that D is of the same order of magnitude in both algorithms. However, for the range where the ISC is deployed it cannot be easily predicted from these values which method is better. Also, we still have to optimize the choice of B and D for the JC.

5 Optimizing the Jacobi Symbol Continuation

Due to the doubly probabilistic nature of the JC, a large number of experiments needs to be done to determine the optimal semi-smoothness parameters B and D for finding a prime factor of a certain size. We thus decided to experiment with prime factors p having between 10 and 15 decimal digits. Then performing between 10000 and 2500 runs of the JC for each pair of parameters (B, D) is feasible given our computational resources, and still sufficiently reliable average values should be obtained. For prime factors in that range, the ISC is the best variant of ECM to factor numbers of general form.

For each order of magnitude of p, i.e., for $p \approx 10^k$, where $k =$ 9,10,11,12,13,14, we did a first round of experiments to get an idea in which ranges for B and D to look for optimized parameters. The parameter choice for this first round was also influenced by the optimized LiDIA parameters given in Table 1 and by theoretical estimates using (8). As a result, for $k = 9, 10, 11, 12, 13, 14$ we selected 6 values each of B and D that we combined to 36 pairs (B, D). For each pair (B, D), we conducted the following experiment:

1. Let $m = 100$ if $k = 9, 10$, and $m = 50$ if $k = 11, 12, 13, 14$.
2. For $i = 1$ to m do
3. Select random primes $p, q \in [10^k, 2 \cdot 10^k)$, and let $N = pq^2$.
4. For $j = 1$ to m do
   (a) Choose a random curve $E(\mathbb{Z}/N\mathbb{Z})$ as described in Standard ECM.
   (b) Apply Step 3 of Standard ECM with parameter $B(= C)$. If p is found, go to Step 4(d).
   (c) Apply the JC, with parameter D. If p is found, go to Step 4(d). If p has not been found after $6\sqrt{D}$ iterations, go back to Step 4a.
   (d) Record the running times for Step 4(b) and Step 4(c), and the total running time.
5. Take the average of the total running time over all m^2 rounds of the algorithm.
6. Determine which percentage of the total running time was spent with the JC, on average.

As an example, in Table 3 we show the results of this experiment for $k = 13$. Here, for each pair (B, D), the second value shows the average total running time (in seconds) on a Sun Ultra 60 Workstation, while the first value shows the average percentage of the total running time that was spent on the JC. Given these results, it is hard to tell which choice of (B, D) is best. First, the average of the running times do not seem to be stable, which has to be attributed to the large spread of the running time, and even 2500 runs of the algorithm are not enough to balance that out. Secondly, the differences between the 36 running times are quite small. To pick the best pair (B, D) out of the 36 pairs under consideration, for each value of B, we take the average of the running times for the 6 values of D, and compute the deviation of that average from the average

Table 3. Jacobi Symbol Continuation: Average running times for $p \approx 10^{13}$ (50 different numbers, 50 times each)

D B	50000	55000	60000	65000	70000	75000	Deviation
2200	58%, 8.56	59%, 8.74	60%, 9.00	61%, 9.19	62%, 8.84	62%, 9.13	+0.1%
2300	56%, 8.76	57%, 8.64	59%, 8.74	59%, 8.40	60%, 8.86	61%, 9.02	−1.9%
2400	55%, 8.70	56%, 8.89	57%, 8.59	58%, 9.16	59%, 8.95	60%, 9.04	−0.2%
2500	54%, 9.03	55%, 8.83	56%, 8.82	57%, 9.11	58%, 9.08	59%, 8.80	+0.5%
2600	54%, 8.98	55%, 8.95	56%, 8.90	57%, 8.65	58%, 9.18	58%, 9.06	+0.6%
2700	52%, 8.33	54%, 8.98	54%, 8.69	56%, 9.09	56%, 9.30	57%, 9.49	+0.9%
Deviation	−2.0%	−0.7%	−1.3%	+0.3%	+1.5%	+2.1%	

of all 36 running times. We pick that value of B whose deviation is the smallest, i.e. $B_{\mathrm{JC},13} = 2300$. We do the same with D, and hence pick $D_{\mathrm{JC},13} = 50000$.

Interestingly, the data in Table 3 show that more than half of the total running time, on average, is spent on the continuation. This a higher proportion than in the case of the ISC [Mue95], and does not agree with the rule of thumb [CP01, p.307] that only a fraction of 1/4 to 1/2 should be spent on the continuation stage. However, to reduce the proportion spent on the JC would require a larger value for B, or a smaller value for D. In our initial experiments, both such choices led to worse running times.

In Table 4 we summarize the optimized semi-smoothness bounds for all ranges of p under consideration. As with Table 1, we also indicate the corresponding u and v such that $B = p^{1/u}$ and $D = p^{1/v}$, and the semi-smoothness probabilities $\sigma(u, v)$. Unfortunately, data for $k = 12$ are not available; we realized too late that the corresponding process on our machine had died.

Table 4. Semi-smoothness bounds - Optimized parameter for Jacobi Symbol Continuation. Semi-smoothness probabilities.

Size of p	B	D	u	v	$\sigma(u,v)$	Size of p	B	D	u	v	$\sigma(u,v)$
10^9	325	1750	3.6	2.8	0.05	10^{12}	n/av	n/av			
10^{10}	550	7500	3.6	2.6	0.06	10^{13}	2300	50000	3.9	2.8	0.04
10^{11}	950	10000	3.7	2.8	0.05	10^{14}	3400	80000	4.0	2.9	0.03

6 Comparison with the Improved Standard Continuation

We finally compare the running times for the optimized (as in Section 5) JC with the ISC as implemented in LiDIA. For this, we did the same experiment as in Section 5, only with the JC replaced by the ISC, and with the optimal semi-smoothness bounds as in Table 1. The results are given in Table 5. These data show that the ISC is about twice as fast as the JC. This outcome is rather disappointing, and was to our surprise. After all, the semi-smoothness bound D enters the complexity of the JC only in terms of $\sqrt{D}$. On the other hand,

Table 5. Comparison of running times.

	Jacobi Symbol Cont.			Impr. Standard Cont.			
Size of p	B	D	Time	B	D	Time	$\mathrm{Time_{JC}}/\mathrm{Time_{ISC}}$
10^9	325	1750	0.52s	349	4297	0.27s	1.9
10^{10}	550	7500	1.2s	411	8861	0.56s	2.1
10^{11}	950	10000	2.0s	659	17981	1.1s	1.8
10^{12}	n/av	n/av	n/av	997	31489	1.9s	
10^{13}	2300	50000	8.5s	1439	54617	4.5s	1.9
10^{14}	3400	80000	16.5s	2111	89501	8.1s	2.0

the ISC has to accommodate $D/\ln D$ equality checks, which results in $D/\ln D$ multiplications modulo N when greatest common divisors are accumulated.

However, there is a crucial difference between the ISC and the JC: Applied to factor $N = pq^2$, the ISC succeeds if $\#E(\mathbb{Z}/p\mathbb{Z})$ *or* $\#E(\mathbb{Z}/q\mathbb{Z})$ is semi-smooth. That is, the ISC looks for p and q simultaneously. On the other hand, the JC is successful only in the event of semi-smoothness modulo p. This might explain why, despite all our efforts in optimizing the method, the JC takes twice as long as the ISC.

7 Conclusion and Future Work

Our implementation of the Jacobi Symbol Continuation of the Elliptic Curve Method for factoring numbers of the form pq^2 results in running times that are about twice as high as for the Improved Standard Continuation. This is for p and q of approximately the same size, and between 10^9 and 10^{14}.

However, there still can be applications where the JC is indeed superior, and we hope our study stimulates further work in that direction. First, consider the case that p is much smaller than q. Then both the ISC and the JC look for only one factor, p. We did not consider this setting in our work since we were interested in cryptographic implications of the JC, where we always have $p \approx q$. Second, consider much larger prime factors p and q. Then the good asymptotic properties of the running time of the JC might have a chance to kick in. The reference method would then be the Fast Fourier Transform Continuation (for p, q of 20 and more decimal digits). Extensive such experiments will allow for conclusions for p, q in the cryptographic range, i.e. of 100 and more decimal digits. We had to work with smaller primes for reasons given in Section 5.

Further experiments with the JC should include the following improvements, both of which were pointed out to us by John Pollard [Pol].

First, there are faster algorithms for the Jacobi symbol computation by Meyer Eikenberry and Sorenson [MS98]. For 100-digit N without small prime factors, they are reported to speed up the computation of $\left(\frac{a}{N}\right)$ by a factor of 1.7 to 1.9. The speed-up increases with the size of N: for 1000-digit N we find a speed-up by a factor of up to 2.3. It can readily be seen that also with these faster routines, the iterating function (5) is still preferable over (6). But upon replacing J in (8) by $J/1.9$, we expect that the overall running time of the JC should go down by a factor of 1.25. Adjusting the optimal semi-smoothness bounds to the new situation might slightly increase this factor.

Another, minor, speed-up can be achieved in the cycle-finding algorithm: There, only those differences $x_{Q_i} - x_R$ need to by considered for the accumulated gcd for which $\left(\frac{x_{Q_i}}{N}\right) = \left(\frac{x_R}{N}\right)$. This test can be done without extra cost since both Jacobi symbols need to be computed anyways. As a result, the cost I in (8) can be reduced to $I/2$. It can be further cut down by also considering $\left(\frac{x_{Q_{i+1}}}{N}\right)$ and $\left(\frac{x_{F(R)}}{N}\right)$, $\left(\frac{x_{Q_{i+2}}}{N}\right)$ and $\left(\frac{x_{F^2(R)}}{N}\right)$, etc., which then requires some more sophisticated administration of terms.

References

BP96. E. Bach and R. Peralta. Asymptotic semismoothness probabilities. *Mathematics of Computation*, 65:1701–1715, 1996.

Bre80. R. P. Brent. An improved Monte Carlo factorization algorithm. *BIT*, 20:176–184, 1980.

Coh93. H. Cohen. *A Course in Computational Algebraic Number Theory.* Springer-Verlag, Berlin, 1993.

CP01. R. Crandall and C. Pomerance. *Prime Numbers. A Computational Perspective.* Springer-Verlag New York, 2001.

FKM$^+$00. E. Fujisaki, T. Kobayashi, H. Morita, H. Oguro, T. Okamoto, S. Okazaki, D. Pointcheval, and S. Uchiyama. EPOC — efficient probabilistic public-key encryption. Submission to NESSIE, 2000. https://www.cosic.esat.-kuleuven.ac.be/nessie/workshop/submissions.html.

Har60. B. Harris. Probability distributions related to random mappings. *Annals of Math. Statistics*, 31:1045–1062, 1960.

Hue00. D. Hühnlein. Quadratic orders for NESSIE — overview and parameter sizes of three public key families. Technical Report TI-3/00, TU Darmstadt, Germany, 2000. http://www.informatik.tu-darmstadt.de/TI/-Veroeffentlichung/TR/Welcome.html.

Len87. H. W. Lenstra, Jr. Factoring integers with elliptic curves. *Ann. of Math.*, 126:649–673, 1987.

Len01. A. K. Lenstra. Unbelievable security. Matching AES security using public key cryptosystems. In *Advances in Cryptology - ASIACRYPT 2001*, volume 2248 of *Lecture Notes in Computer Science.* Springer-Verlag, 2001.

LiD00. LiDIA Group, Technische Universität Darmstadt, Darmstadt, Germany. *LiDIA - A library for computational number theory, Version 2.0*, 2000.

MS98. S. Meyer Eikenberry and J. P. Sorenson. Efficient algorithms for computing the Jacobi symbol. *Journal of Symbolic Computation*, 26:509–523, 1998.

Mon92. P. Montgomery. *An FFT extension of the elliptic curve method of factorization.* PhD thesis, University of California, Los Angeles, 1992.

Mue95. A. Müller. Eine FFT-Continuation für die elliptische Kurvenmethode. Master's thesis, Universität des Saarlandes, Saarbrücken, Germany, 1995. Diplomarbeit.

MvOV96. A. Menezes, P. van Oorschot, and S. A. Vanstone. *Handbook of Applied Cryptography.* CRC Press, 1996.

Per01. R. Peralta. Elliptic curve factorization using a "partially oblivious" function. In K.-Y. Lam, I. Shparlinski, H. Wang, and C. Xing, editors, *Cryptography and Computational Number Theory: Workshop in Singapore 1999*, volume 20 of *Progress in Computer Science and Applied Logic*, pages 123–128. Birkhäuser, 2001.

PO96. R. Peralta and E. Okamoto. Faster factoring of integers of a special form. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Science*, E79-A(4), 1996.

Pol. J. M. Pollard. Private communication, February 2002.

Pol75. J. M. Pollard. A Monte Carlo method for factorization. *BIT*, 15(3):331–335, 1975.

Pol78. J. M. Pollard. Monte Carlo methods for index computation (mod p). *Mathematics of Computation*, 32(143):918–924, 1978.

Tak98. T. Takagi. Fast RSA-type cryptosystem modulo $p^k q$. In *Advances in Cryptology - CRYPTO '98*, volume 1462 of *Lecture Notes in Computer Science*, pages 318–326. Springer-Verlag, 1998.

Tes98a. E. Teske. A space efficient algorithm for group structure computation. *Mathematics of Computation*, 67:1637–1663, 1998.

Tes98b. E. Teske. Speeding up Pollard's rho method for computing discrete logarithms. In *Algorithmic Number Theory Seminar ANTS-III*, volume 1423 of *Lecture Notes in Computer Science*, pages 541–554. Springer-Verlag, 1998.

Tes01. E. Teske. On random walks for Pollard's rho method. *Mathematics of Computation*, 70:809–825, 2001.

A New Scheme for Computing with Algebraically Closed Fields

Allan Steel

School of Mathematics and Statistics
University of Sydney
NSW 2006 Australia
allan@maths.usyd.edu.au

Abstract. A new scheme is presented for computing with an algebraic closure of the rational field. It avoids factorization of polynomials over extension fields, but gives the illusion of a genuine field to the user. A technique of modular evaluation into a finite field ensures that a unique genuine field is simulated by the scheme and also provides fast optimizations for some critical operations. Fast modular matrix techniques are also used for several non-trivial operations. The scheme has been successfully implemented within the Magma Computer Algebra System.

1 Introduction

This paper presents a new scheme, implemented within the Magma Computer Algebra System [4], for computing with an algebraic closure of the rational field **Q**. The scheme works by automatically constructing larger and larger algebraic extensions of **Q** as needed during a computation, thus giving the illusion to the user of computing with an algebraic closure of **Q**. The defining polynomials are not necessarily irreducible over the subfields—factorization over algebraic number fields is avoided, and the defining polynomials are automatically modified when factors are found during computations with the field. These factors often arise naturally because of the structure of an algorithm which is computing over the field.

A similar scheme was already proposed before (the D5 system [6]), but in this case an algorithm based on the field must handle the parallelism which occurs when one must compute with several roots of a reducible polynomial, leading to situations where a certain expression evaluated at one root is invertible but evaluated at another root is *not* invertible.

The new scheme presented here has no such difficulty: all the roots of a squarefree polynomial are returned as distinct elements of a genuine field, and any algorithm working over a general field need not be modified in any way to handle the separate roots.

This paper concentrates on the theoretical model underlying the scheme; because of space restrictions, it is impossible to give detailed examples of how the scheme works in practice. For many examples and more information, see the chapter "Algebraically Closed Fields" in the Handbook of Magma Functions [3] or the same chapter in the Online Help of the Magma Homepage [10].

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 491–505, 2002.

2 Definition of an Algebraically Closed Field

2.1 Basic Presentation

The main type of object which we will develop in this paper will be called an ACF, standing for "algebraically closed field". In the implementation, such an object will only be represented at any given moment by a certain finite-degree extension of $\mathbf{Q}$, but since the user will get the illusion that the field is algebraically closed, we will let ACF label the current object.

An ACF A will be represented by the quotient of a rank-n multivariate polynomial ring by a triangular ideal I with n defining polynomials (defined below). In general, I will not be a maximal ideal, so the quotient will not be a field. However, the other key component of A will be a certain sequence Γ of n elements in some finite field which will allow a "modular evaluation" technique, and this will have two separate but critical properties:

1. There will be a unique maximal ideal J containing I which is determined by I and Γ. Thus the quotient by J will define a unique field and the user will get the illusion of working with this field.
2. Some quick tests will be able to be performed in the finite field (via Γ), thus making some fundamental arithmetic operations very fast.

In this paper, we will always have the rational field $\mathbf{Q}$ as the base field, but the scheme can be made to work for any other base field for which one can implement a modular evaluation technique similar to the one described here.

2.2 Triangular Ideals

Throughout the paper, let $\mathbf{Q}_i$ denote $\mathbf{Q}[x_1, \ldots, x_i]$ and $\mathbf{Q}_0 = \mathbf{Q}$ and for $n \geq 1$, let $\mathbf{Q}_n$ have the lexicographical monomial ordering with $x_1 < x_2 < \cdots < x_n$ (see [5, Chap. 2, §2] for details on monomial orderings).

Definition 2.1. *A sequence of n polynomials $(f_1, \ldots, f_n) \in (\mathbf{Q}_n)^n$ is called a* triangular basis *if, for $1 \leq i \leq n$:*

1. *The greatest variable occurring in f_i is x_i.*
2. *f_i is monic, written as a polynomial in x_i.*

The trivial sequence () is defined to be a triangular basis for $\mathbf{Q}_0 = \mathbf{Q}$. An ideal I of $\mathbf{Q}_n$ is called triangular *if it possesses a triangular basis (i.e., if it is generated as an ideal by some triangular basis).*

As an example, let $n = 3$ and $f_1 = x_1^2+1$, $f_2 = x_2^3-x_1$ and $f_3 = x_3^2-x_1x_2+1$. Then (f_1, f_2, f_3) is a triangular basis in $\mathbf{Q}_3 = \mathbf{Q}[x_1, x_2, x_3]$.

It is easy to see that a triangular basis of an ideal I is a Gröbner basis of I (with respect to the above order). This means that we can form the unique normal form modulo I of an element in $\mathbf{Q}_n$. This is the only fact which we will use from the theory of Gröbner bases. See also [1,9] for more discussion concerning triangular ideals.

2.3 Evaluating into a Finite Field

Definition 2.2. *For a prime number p, define $\mathbf{Q}_{n(p)}$ to be the set of all elements $f \in \mathbf{Q}_n$ such that p does not divide the denominator of any coefficient of f. $\mathbf{Q}_{n(p)}$ is clearly a ring.*

Definition 2.3. *Suppose that $\Gamma = (\gamma_1, \ldots, \gamma_n)$ is a sequence of n elements of a finite field E of characteristic p. Define $\phi_\Gamma : \mathbf{Q}_{n(p)} \to E$ to be the natural map which maps a rational a/b to $(a \bmod p) \cdot (b \bmod p)^{-1} \in E$ and which maps x_i to γ_i, for $1 \le i \le n$. ϕ_Γ is clearly a homomorphism (the mod-p homomorphism coupled with an evaluation homomorphism).*

Definition 2.4. *Suppose that $\Gamma = (\gamma_1, \ldots, \gamma_n) \in E^n$ with E a finite field of characteristic p. If a triangular ideal I of $\mathbf{Q}_n$ has the property that $\phi_\Gamma(f) = 0$ for all $f \in I \cap \mathbf{Q}_{n(p)}$, then we say that I is* compatible with *Γ.*

2.4 Determining a Unique Field

Theorem 2.1. *Suppose that I is a triangular ideal of $\mathbf{Q}_n$ with triangular basis $(g_1, \ldots, g_n)$ with $g_i \in \mathbf{Q}_{n(p)}$ for each i and such that I is compatible with $\Gamma = (\gamma_1, \ldots, \gamma_n) \in E^n$, where E is a finite field of characteristic p. Then there is a unique maximal ideal J of $\mathbf{Q}_n$ which is compatible with Γ and such that $J \supseteq I$. Thus $\mathbf{Q}_n/J$ is a field. J is also triangular, and a triangular basis $(q_1, \ldots, q_n)$ of J can be constructed directly with $q_i \in \mathbf{Q}_{n(p)}$ for each i.*

Proof. We perform induction on n. We first handle the case of $n = 0$, where I is the zero ideal of $\mathbf{Q}_0 = \mathbf{Q}$, and $\Gamma = ()$; everything trivially holds by taking J to be the zero ideal (with the empty triangular basis).

Now assume that the theorem is true for $n - 1$. Let I_{n-1} be the triangular ideal of $\mathbf{Q}_{n-1}$ with triangular basis given by the restriction to $\mathbf{Q}_{n-1}$ of $(g_1, \ldots, g_{n-1})$. By assumption, there exists a unique maximal ideal J_{n-1} of $\mathbf{Q}_{n-1}$ with $J_{n-1} \supseteq I_{n-1}$ and triangular basis $(q_1, \ldots, q_{n-1})$. Let g be the element of $(\mathbf{Q}_{n-1}/J_{n-1})[x_n]$ corresponding to g_n. Factorize g over the field $\mathbf{Q}_{n-1}/J_{n-1}$ into powers of monic irreducibles. As $\phi_\Gamma(g_n) = 0$, and E is a field, there must exist exactly one irreducible q of g with $\phi_\Gamma(q_n) = 0$, where q_n is the polynomial of $\mathbf{Q}_n$ corresponding to q (q_n is in $\mathbf{Q}_{n(p)}$, by a variant of Gauss' lemma: the prime p cannot be introduced into a denominator of a monic factor since p does does not divide a denominator of g_n).

So if we let J be the ideal of $\mathbf{Q}_n$ generated by J_{n-1} (lifted to $\mathbf{Q}_n$) and q_n, then $\mathbf{Q}_n/J \cong (\mathbf{Q}_{n-1}/J_{n-1})[x_n]/\langle q \rangle$, with q irreducible, so $\mathbf{Q}_n/J$ is a field and J is maximal. Also, appending q_n to $(q_1, \ldots, q_{n-1})$ (lifted to $\mathbf{Q}_n$) yields a triangular basis of J with $q_i \in \mathbf{Q}_{n(p)}$ for each i, so J is compatible with Γ. The uniqueness of J follows from the uniqueness of J_{n-1} and q. □

Definition 2.5. *Denote the ideal J of Theorem 2.1, uniquely determined from I and Γ, by $\mathcal{J}(I, \Gamma)$.*

Definition 2.6. *Suppose that I is a triangular ideal of $\mathbf{Q}_n$ with triangular basis $(g_1, \ldots, g_n)$ such that I is compatible with $\Gamma = (\gamma_1, \ldots, \gamma_n) \in E^n$, where E is a finite field of characteristic p. Let $J = \mathcal{J}(I, \Gamma)$. Suppose further that for $1 \leq i \leq n$, the resultant in x_i of g_i with the derivative of g_i (with respect to x_i) is not in J. An* ACF *A is defined to be the quotient $(\mathbf{Q}_n/I)/(J/I)$. By the Third Isomorphism Theorem, A is of course isomorphic to $\mathbf{Q}_n/J$ so by Theorem 2.1, A is a field. We call E the* modular evaluation field *of A and n the* rank *of A. We also define the trivial ACF to be $(\mathbf{Q}_n/I)/(J/I)$ where $n = 0$, I and J are the zero ideals of $\mathbf{Q}_0 = \mathbf{Q}$, and $\Gamma = ()$; this is clearly isomorphic to $\mathbf{Q}$.*

The technical condition on the triangular basis of I involving resultants will simply ensure that each of the g_i is squarefree over the appropriate subfield of A (see below).

Since J is uniquely determined from I and Γ, J will be used extensively in our theoretical presentation and analysis and A will present the illusion to the user of acting like $\mathbf{Q}_n/J$. Because of the way A is defined, an element a of A will have the theoretical form $(r + I) + J/I$, where $r \in \mathbf{Q}_n$. Clearly a is the zero element of A if and only if $r \in J$.

In the implementation, however, J will not be constructed explicitly because we wish to avoid factorization, but I and Γ will be the information which is known and used, so that is why we present A as the (theoretically redundant) quotient $(\mathbf{Q}_n/I)/(J/I)$. The element a will be represented by $r \in \mathbf{Q}_n$, kept reduced modulo I. The main difficulty is that we may have $r \in J$, but $r \notin I$, and we need to detect this case without knowing J explicitly.

3 Properties of an ACF

3.1 Evaluation Properties

Lemma 3.1. *Let J be an ideal of $\mathbf{Q}_n$ with triangular basis $(q_1, \ldots, q_n)$ such that $q_i \in \mathbf{Q}_{n(p)}$ for $1 \leq i \leq n$ and such that J is compatible with $\Gamma \in E^n$, where E is a finite field of characteristic p. The cosets of $\mathbf{Q}_n/J$ can each be written uniquely in the form $r + J$, where r is in normal form modulo J, so let $(\mathbf{Q}_n/J)_{(p)}$ be the set of elements of $(\mathbf{Q}_n/J)$ whose unique coset representatives are in $\mathbf{Q}_{n(p)}$. Then $(\mathbf{Q}_n/J)_{(p)}$ is a subring of $\mathbf{Q}_n/J$ and there is a natural well-defined homomorphism $\phi_{\Gamma/J} : (\mathbf{Q}_n/J)_{(p)} \to E$ given by $(r + J) \mapsto \phi_\Gamma(r)$.*

Proof. $(\mathbf{Q}_n/J)_{(p)}$ is clearly a ring, since the sum or product of two of its elements can be reduced to normal form modulo J without introducing a denominator divisible by p (since the q_i form a Gröbner basis of J and are in $\mathbf{Q}_{n(p)}$). Since J is compatible with Γ, $\phi_{\Gamma/J}$ is a well-defined homomorphism. □

Lemma 3.2. *Let $A = (\mathbf{Q}_n/I)/(J/I)$ be an ACF with $\Gamma \in E^n$ and $J = \mathcal{J}(I, \Gamma)$, where E is a finite field of characteristic p. Let $A_{(p)}$ be the set of all elements a in A such if $a = (r + I) + J/I$, where r is in normal form modulo J, then $r \in \mathbf{Q}_{n(p)}$. Then $A_{(p)}$ is a subring of A, and ϕ_Γ can be extended to be a well-defined homomorphism from $A_{(p)}$ to E. Also, for any element $a \in A$ having the* *form $(r+I)+J/I$, where r is reduced modulo I (but not necessarily reduced modulo J), then if $r \in \mathbf{Q}_{n(p)}$, then $a \in A_{(p)}$ and the homomorphism can be evaluated at a by using r (without knowing the normal form of r modulo J).*

Proof. Since $A \cong \mathbf{Q}_n/J$, $A_{(p)}$ is clearly a ring isomorphic to $(\mathbf{Q}_n/J)_{(p)}$ and the extension of ϕ_Γ is well-defined, by Lemma 3.1. Also, just as in the previous proof, if $a = (r+I)+J/I$ with $r \in \mathbf{Q}_{n(p)}$ and with r reduced modulo I, then the normal form of r modulo J must also be in $\mathbf{Q}_{n(p)}$ (because the elements of the triangular (Gröbner) basis of J lie in $\mathbf{Q}_{n(p)}$), so $a \in A_{(p)}$ and since $J \supseteq I$ and both ideals are compatible with Γ, the last statement is correct. □

Corollary 3.1. *Let A, Γ, E and J be as in the last Lemma. Then ϕ_Γ extends to a natural homomorphism from $A_{(p)}[z]$ to $E[z]$ which can be evaluated without using J. So if $q, g \in A_{(p)}[z]$ with q dividing g, then the image of q in $E[z]$ divides the image of g in $E[z]$. Also, a monic factor of a monic polynomial in $A_{(p)}[z]$ must also lie in $A_{(p)}[z]$.*

Proof. The last statement follows from the variant of Gauss' Lemma in the proof of Theorem 2.1. □

3.2 Isomorphisms and Subfields

Definition 3.1. *Suppose that I and I' are both triangular ideals of $\mathbf{Q}_n$ which are compatible with $\Gamma = (\gamma_1, \ldots, \gamma_n) \in E^n$, where E is a finite field. Suppose also that $\mathcal{J}(I,\Gamma) = \mathcal{J}(I',\Gamma) = J$, and that I and I' both satisfy the resultant condition in Definition 2.6. Then we say that the two ACFs $A = (\mathbf{Q}_n/I)/(J/I)$ and $A' = (\mathbf{Q}_n/I')/(J/I')$ are isomorphic. Also, the map $\psi : A \to A'$ defined by $(r+I)+J/I \mapsto (r+I')+J/I'$ is clearly a natural isomorphism (because $I \subseteq J$ and $I' \subseteq J$).*

The key point of this definition is that the one Γ sequence must be common to both ACFs for them to be considered isomorphic in our model (it is insufficient for them to be simply isomorphic as fields).

Lemma 3.3. *Suppose $A = (\mathbf{Q}_n/I)/(J/I)$ is an ACF, where $\Gamma = (\gamma_1, \ldots, \gamma_n)$ and $J = \mathcal{J}(I,\Gamma)$, and let $(g_1, \ldots, g_n)$ be a triangular basis of I. Let I_{n-1} be the triangular ideal of $\mathbf{Q}_{n-1}$ generated by the restriction to $\mathbf{Q}_{n-1}$ of $(g_1, \ldots, g_{n-1})$, and let Γ_{n-1} be $(\gamma_1, \ldots, \gamma_{n-1})$. Then $A_{n-1} = (\mathbf{Q}_{n-1}/I_{n-1})/(J_{n-1}/I_{n-1})$, where $J_{n-1} = \mathcal{J}(I_{n-1}, \Gamma_{n-1})$, is also an ACF which is a subfield of A. Also, $A \cong (A_{n-1}[x_n])/\langle q \rangle$, where q is irreducible, q divides the polynomial g in $A_{n-1}[x_n]$ corresponding to g_n, and the polynomial $q_n \in \mathbf{Q}_n$ corresponding to q is in $\mathbf{Q}_{n(p)}$ and $\phi_\Gamma(q_n) = 0$. Finally, $g \in A_{n-1}[x_n]$ is squarefree.*

Proof. It is clear that I_{n-1} is compatible with Γ_{n-1}, and satisfies the resultant condition in Definition 2.6, so A_{n-1} is a well-defined ACF. All of the claims in the second last sentence follow from the construction of J in the proof of Theorem 2.1. Finally, since the resultant of g_n with its derivative is not in J (by Definition 2.6), the resultant of g with its derivative is not zero, so g is squarefree. □

Definition 3.2. *We will often just say the "modular evaluation of f" for $f \in A_{n-1(p)}[x_n]$ to mean the element of E obtained by mapping the coefficients of f into E via Corollary 3.1 applied to the subfield A_{n-1} and by mapping x_n to γ_n.*

Lemma 3.4. *Suppose $A = (\mathbf{Q}_n/I)/(J/I)$ is an ACF, where $J = \mathcal{J}(I, \Gamma)$. Let $A_{n-1} = (\mathbf{Q}_{n-1}/I_{n-1})/(J_{n-1}/I_{n-1})$ be the subfield of A as in Lemma 3.3. Suppose that $A'_{n-1} = (\mathbf{Q}_{n-1}/I'_{n-1})/(J_{n-1}/I'_{n-1})$ is isomorphic to A_{n-1}, via the isomorphism $\psi_{n-1} : A_{n-1} \to A'_{n-1}$, and that $I'_{n-1} \supseteq I_{n-1}$. Then there is a natural extension $A' = (\mathbf{Q}_n/I')/(J/I')$ of A'_{n-1} which is isomorphic to A via a natural isomorphism ψ, with $I' \supseteq I$.*

Proof. Define I' to be the ideal of $\mathbf{Q}_n$ generated by the embedding of I'_{n-1} in $\mathbf{Q}_n$ and the n-th triangular basis polynomial g_n of I. Clearly $I' \supseteq I$, since $I'_{n-1} \supseteq I_{n-1}$ and $g_n \in I$. As I'_{n-1} is compatible with Γ_{n-1}, we must have $\mathcal{J}(I, \Gamma) = \mathcal{J}(I', \Gamma)$, and I' satisfies the resultant condition in Definition 2.6 since I'_{n-1} does, and g_n and J are unchanged, so A' is a well-defined ACF and is isomorphic to A. Defining $\psi : A \to A'$ by $(r + I) + J/I \mapsto (r + I') + J/I'$ is easily seen to be a well-defined isomorphism, since $I \subseteq J$ and $I' \subseteq J$. □

4 Arithmetic Operations

4.1 Presentation

In this section we show how the key arithmetic operations are performed in an ACF. These operations are: addition, subtraction, multiplication, the testing of whether an element is zero or not, and inversion. These operations clearly suffice to represent a field effectively; other operations are easily derived from them. For example, testing equality of two elements is done by testing whether their difference is zero, and so on.

Let $A = (\mathbf{Q}_n/I)/(J/I)$ be an ACF. We wish to perform arithmetic operations in A without explicitly using the maximal ideal J, as it is not known in the implementation. In our theoretical model, each arithmetic function for a field A will return a new field $A' = (\mathbf{Q}_n/I')/(J/I')$ which is isomorphic to A, and the result(s) will be with respect to A'. From a theoretical point of view, A' is effectively the same field as A, but from an implementation point of view, the ideal I' representing A' will now allow one to perform trivially the desired operation without knowing J.

This theoretical approach of returning isomorphic fields allows us to represent rigorously the way that simplifications (bringing the ideal I closer to J) will occur, without having to worry about the field or its elements changing during an algorithm. In the implementation, however, we always work with *only one* ACF A, and modify it in place whenever there is a simplification: the ideal I is replaced with the new ideal I', and all elements of A are reduced modulo I' (we keep a list of pointers to all hitherto computed elements of A within A itself, so that we can reduce them at this point—this is easily managed).

An element of A is represented as polynomial in $\mathbf{Q}_n$, reduced modulo the current ideal I. The basic operations of addition, subtraction and multiplication are handled easily, since we simply do the operation in $\mathbf{Q}_n$ and reduce modulo I in each case.

4.2 Zero Testing and Inversion

Amazingly, testing whether an element is the zero element of the field is the most difficult operation in the whole scheme! The crucial algorithm ZeroTest below takes an element a of A and returns a new ACF A' which is isomorphic to A but also with the important property that it is trivial in the implementation to test within A' whether a is the zero element of A or not.

Algorithm ZeroTest(a)
Input: a, an element of an ACF $A = (\mathbf{Q}_n/I)/(J/I)$, with $J = \mathcal{J}(I, \Gamma)$.
Output: A new ACF $A' = (\mathbf{Q}_n/I')/(J/I')$ with $I' \supseteq I$ and $\mathcal{J}(I', \Gamma) = J$, and an isomorphism $\psi : A \to A'$, such that if a is the zero element of A then for all $r' \in \mathbf{Q}_n$ such that $\psi(a) = (r' + I') + J/I'$, we have $r' \in I'$.

1. If n is 0, return A and Id_A (the identity map from A to A).
2. Write $a = (r + I) + J/I$ with $r \in \mathbf{Q}_n$ and with r reduced modulo I.
3. Let $A_{n-1} = (\mathbf{Q}_{n-1}/I_{n-1})/(J_{n-1}/I_{n-1})$ be the subfield of A as in Lemma 3.3 and let $f \in A_{n-1}[x_n]$ correspond to r. Let $(g_1, \ldots, g_n)$ be a triangular basis of I and let $g \in A_{n-1}[x_n]$ correspond to g_n.
4. Let E be the modular evaluation field of A of characteristic p. If r is not in $\mathbf{Q}_{n(p)}$, skip to the next step. Otherwise, let $\bar{f}$ be the polynomial in $E[z]$ obtained by mapping the coefficients of f into E using $(\gamma_1, \ldots, \gamma_{n-1})$ (via Corollary 3.1 applied to A_{n-1}) and by mapping x_n to z. Let $\bar{g}$ be the similar polynomial in $E[z]$ corresponding to g. If $\bar{f}$ and $\bar{g}$ are coprime, then return A and Id_A.
5. A_{n-1} is a field, and we assume by induction that we can effectively perform the Euclidean algorithm on f and g, without using J explicitly, to obtain a field $A'_{n-1} = (\mathbf{Q}_{n-1}/I'_{n-1})/(J_{n-1}/I'_{n-1})$ isomorphic to A_{n-1}, the isomorphism $\psi_{n-1} : A_{n-1} \to A'_{n-1}$, and $c \in A'_{n-1}[x_n]$ with c the monic GCD of f and g (moved to $A'_{n-1}[x_n]$ via ψ_{n-1}).
6. If $c = 1$ or $c = g$, construct $A' = (\mathbf{Q}_n/I')/(J/I')$ and $\psi : A \to A'$ from I'_{n-1} and g_n and ψ_{n-1} (using Lemma 3.4) and return A' and $\psi : A \to A'$.
7. If the modular evaluation of c is zero, then let $h = c$; otherwise let $h = g/c$.
8. Let s be the element of $\mathbf{Q}_n$ corresponding to h. Let I' be the triangular ideal of $\mathbf{Q}_n$ with triangular basis $(g'_1, \ldots, g'_{n-1}, s)$. Return $A' = (\mathbf{Q}_n/I')/(J/I')$ and the natural map $\psi : A \to A'$ with $(r + I) + J/I \mapsto (r + I') + J/I'$.

The ACF returned by IsZero makes it possible to compute the inverse of any non-zero element, which is done by the algorithm Inverse.

Algorithm Inverse(a)
Input: a, an element of $A = (\mathbf{Q}_n/I)/(J/I)$ where A is an ACF returned by IsZero applied to (an earlier form of) a, and a is not zero.
Output: A new isomorphic ACF $A' = (\mathbf{Q}_n/I')/(J/I')$, the isomorphism $\psi : A \to A'$, and an element $b' \in A'$ such that $I' \supseteq I$ and b' is the inverse of $a' = \psi(a)$ (i.e, so that $a'b' - 1$ is the zero element of A').

```
1. If $n$ is 0, any representative of $a$ is a constant, so return $A$, $\mathrm{Id}_A$ and the (constant) inverse of $a$.
2. Write $a = (r + I) + J/I$ with $r \in \mathbf{Q}_n$ and with $r$ reduced modulo $I$.
3. Let $A_{n-1} = (\mathbf{Q}_{n-1}/I_{n-1})/(J_{n-1}/I_{n-1})$ be the subfield of $A$ as in Lemma 3.3 and let $f \in A_{n-1}[x_n]$ correspond to $r$. Let $(g_1, \dots, g_n)$ be a triangular basis of $I$ and let $g \in A_{n-1}[x_n]$ correspond to $g_n$.
4. $A_{n-1}$ is a field, and we assume by induction that we can effectively perform the extended Euclidean algorithm on $f$ and $g$, without using $J$ explicitly, to obtain a field $A'_{n-1} = (\mathbf{Q}_{n-1}/I'_{n-1})/(J_{n-1}/I'_{n-1})$ isomorphic to $A_{n-1}$, the isomorphism $\psi_{n-1} : A_{n-1} \to A'_{n-1}$, and $c, u, v \in A'_{n-1}[x_n]$ with $c$ the monic GCD of $f'$ and $g'$ and $c = u \cdot f' + v \cdot g'$ (where $f'$ and $g'$ are $f$ and $g$ moved to $A'_{n-1}[x_n]$ by $\psi_{n-1}$, respectively). Assert that $c$ is one.
5. Construct $A'$ and $\psi : A \to A'$ from $I'_{n-1}$ and $g_n$ and $\psi_{n-1}$ (using Lemma 3.4) and return $A'$ and $\psi : A \to A'$, and the element in $A'$ corresponding to $u$.
```

Theorem 4.1. *Algorithms* ZeroTest *and* Inverse *are correct and do not explicitly use the ideal J in either case.*

Proof. We will prove the algorithms are correct by induction in parallel, since they effectively call each other recursively. We will show that the claims on the outputs of each algorithm are correct in all cases, and that all steps are valid and do not use J explicitly.

First of all, the case where $n = 0$ (so $I = J = 0$ and A is isomorphic to $\mathbf{Q}$) is clearly handled correctly in Step 1 of each algorithm: the coset representative r is always a constant.

Assume now that the algorithms are correct for fields of rank $n - 1$.

In Step 2 of each algorithm, we find a polynomial $r \in \mathbf{Q}_n$ which represents a such that r is reduced modulo I. We do not know J in practice, so r could be non-reduced modulo J, but algorithm IsZero will effectively determine whether r is actually in J or not.

After Step 3 of each algorithm, we have $A \cong A_{n-1}[x_n]/\langle q\rangle$, where q is irreducible and divides g, g is squarefree, and the modular evaluation of each of q and g is zero, by Lemma 3.3. Ignore Step 4 of ZeroTest for the moment. The application in each algorithm of the Euclidean algorithm uses addition, subtraction, multiplication, inversion and zero testing in the subfield A_{n-1}, all of which can be done without explicit use of J, by our induction assumption, and the results will be returned over a subfield A'_{n-1} isomorphic to A_{n-1}, via the isomorphism $\psi_{n-1} : A_{n-1} \to A'_{n-1}$.

First consider IsZero. If c, the GCD of f and g, is one, then since q divides g, q cannot divide f, so $f \bmod q$ is non-zero, so since f corresponds to r, we have $f \notin J$, so a is non-zero. If c is g, then f mod g is zero, so a is the zero element of A, and after constructing A', any r' representing $\psi(a)$ will reduce to zero modulo I' as I' contains g_n corresponding to g. Thus after constructing A' in Step 6, the claim on the output is satisfied in both cases (trivially, in the first case).

Step 4 of ZeroTest is simply a quick test to see whether the GCD of f and g is one. Since the coefficients of f and g can be evaluated into E, then if the GCD of f and g is non-trivial, then the GCD of $\bar{f}$ and $\bar{g}$ must also be non-trivial, by Corollary 3.1. Thus if Step 4 returns, then this is equivalent to the case that c is one in Step 6, so Step 4 is correct.

If c is not one, then q must divide exactly one of c and g/c, since q is irreducible and divides g, which is squarefree. Since c is neither one nor g, c and g/c are both proper factors of g. Now the modular evaluation of q is zero, so the factor of g which q divides will have zero modular evaluation. The other factor is not divisible by q, so it cannot have zero modular evaluation. (These factors can be evaluated since, by Corollary 3.1, they are monic divisors of g which can itself be evaluated.) Thus Step 7 must assign h to the correct factor of g which is divisible by q and which has zero modular evaluation. (As c is monic, no inversions or zero-tests need be done if we compute the quotient g/c, so the subfield will not change.)

In either case, $s \in \mathbf{Q}_n$, which is assigned in Step 8 to correspond to h, must be in J, so $J \supseteq I' \supset I$. Thus $A' = (\mathbf{Q}_n/I')/(J/I')$ is a new well-defined ACF which is isomorphic to A, as h divides g and the modular evaluation of h is zero so I' and Γ define the same J uniquely (and s satisfies the resultant condition since h is squarefree). The returned map ψ, as defined, is easily seen to be a well-defined isomorphism because $I' \supset I$ (and can be implemented without the explicit knowledge of J). Finally, if a is the zero element, then q must divide f and so also c, so $h = c$ and any r' representing $\psi(a)$ will reduce to zero modulo I', as I' contains s which corresponds to c, so the claim on the output is correct. Thus IsZero is correct and J is not used explicitly.

Finally, consider Inverse. This is very similar, and we need only prove that the GCD c is one in Step 4. We can assume that a is non-zero and the input field was returned from IsZero. If that algorithm returned at any step before Step 7, then clearly the GCD c was one then, so will be one again here. If we went through Steps 7 and 8, then since a is non-zero, we must have had the case that q divided g/c (so a factor of g was found even though a is not the zero element: the element of A corresponding to g/c was the zero element). Now c was coprime with g/c, so f was coprime to g/c, and the polynomial corresponding to g/c was included in I'. So the GCD must now be one in Inverse. Thus Inverse is correct and J is not used explicitly. □

Note that the correctness of Inverse depends strongly on the fact that we compute the modular GCD in Step 4 of IsZero instead of just testing whether the modular evaluation of f is zero. If we were to do only the latter, then this would still suffice as a correct quick test for whether a is zero, but it would not detect the case that a is non-zero but the GCD c is non-trivial, so the coprime factor g/c would not be inserted in I, so Inverse would fail.

As an example, suppose an ACF has the defining ideal $I = \langle \alpha^2-2, \beta^2-8 \rangle$ with $\beta > \alpha$, and suppose we call IsZero on $e = \beta - 2\alpha$. Clearly $(\beta - 2\alpha)(\beta + 2\alpha) \in I$, and the GCD c in Step 5 will be set to e. The modular evaluation of c will depend arbitrarily on Γ. If the evaluation is zero, then I' will be $\langle \alpha^2 - 2, \beta - 2\alpha \rangle$ and e will be zero. Otherwise I' will be $\langle \alpha^2 - 2, \beta + 2\alpha \rangle$ and e will be non-zero (and simplified to become -4α). So the simplification occurs in either case, but whether e will be zero or not cannot be predicted beforehand (yet things will stay consistent forever, in all cases).

4.3 Remarks on the Implementation

In the implementation, the key step in the IsZero algorithm to make the whole scheme efficient is the modular GCD test (in Step 4). Since the prime p is normally about the machine word size (of the order of 10^9), it is very rare that Step 4 cannot be applied. Without this quick test, the scheme is simply too naive when the rank of the field becomes non-trivial, because computing the GCDs via the Euclidean algorithm every time a zero test is needed leads to huge coefficient blowup in the recursive calls. So this is one huge advantage our scheme has over simpler schemes.

The other key optimization in the implementation is the use of a fast matrix technique to perform the GCD or XGCD computations (instead of the Euclidean algorithm). This technique uses a very fast p-adic matrix nullspace algorithm, ignoring intermediate coefficient growth over $\mathbf{Q}$. We sketch this very briefly.

The quotient ring $\mathbf{Q}_n/I$ can be considered as a finite-dimensional vector space (since I is a zero-dimensional ideal). Thus we can form a $\mathbf{Q}$-vector space monomial basis $M = (m_1, \ldots, m_d)$ of $\mathbf{Q}_n/I$ (see [5, Chap. 5, §3, Prop. 4] or [2, Prop. 9.4]), which is sorted lexicographically (with the smallest monomial 1 coming first). We then compute the representation matrix B for the multiplication action of r on M: for each i, let the i-th row of B be the vector corresponding to $r \cdot m_i$, using M to index the columns.

For Step 5 of ZeroTest, we compute the (left) nullspace N of B, then we echelonize a basis of N from the right and thus get a v in N (with $v \cdot B = 0$) whose first non-zero entry starting from the right is as left as possible. Then because of the lexicographical order for M, the polynomial corresponding to v is a polynomial which annihilates r modulo I with smallest possible leading monomial (w.r.t. the monomial order) and this polynomial corresponds to the GCD c.

For Step 4 of Inverse, we attempt to solve the linear system $s \cdot B = w$ for the vector $s \in \mathbf{Q}^d$, where w is the vector $(1, 0, \ldots)$ in $\mathbf{Q}^d$ corresponding to the polynomial 1. If there is a solution s, then the polynomial corresponding to s is the inverse of a. If there is no solution s, then there is a non-trivial nullspace N and we do the same steps as in the previous paragraph to cause a simplification of A in IsZero. Then we start again to compute the inverse. Since we know that a is non-zero, there must eventually be a solution s yielding the inverse.

5 Extending an ACF and Finding Roots

We have shown how to perform arithmetic effectively with an already formed ACF, but we must show how to build up an ACF effectively from $\mathbf{Q}$!

5.1 Computing Roots of a Squarefree Polynomial

Let $A = (\mathbf{Q}_n/I)/(J/I)$ be an ACF, and suppose f is a monic squarefree polynomial in $A[x]$ and k an integer with $1 \leq k \leq \operatorname{Deg}(f)$. The following algorithm attempts to construct a new ACF $\tilde{A}$ which contains A as a subfield but also has the crucial property that it contains k *distinct* roots $\beta_1, \ldots \beta_k$ of f. The algorithm may fail, but we address the case of failure below.

Algorithm SquarefreeRoots
Input: An ACF $A = (\mathbf{Q}_n/I)/(J/I)$ (with $J = \mathcal{J}(I, \Gamma)$), a monic squarefree polynomial $f \in A[z]$, and an integer k with $1 \leq k \leq \operatorname{Deg}(f)$.
Output: If successful, a new ACF $\tilde{A} = (\mathbf{Q}_n/\tilde{I})/(\tilde{J}/\tilde{I})$ with a subfield isomorphic to A, the embedding $\chi : A \to \tilde{A}$, and k *distinct* elements $\beta_1, \ldots, \beta_k$ of $\tilde{A}$ which are roots of f, lifted to $\tilde{A}$.

1. Let $\chi : \mathbf{Q}_n \to \mathbf{Q}_{n+k}$ be the natural embedding of $\mathbf{Q}_n$ into $\mathbf{Q}_{n+k}$ (with x_i mapped to x_i for $1 \leq i \leq n$). We can extend χ to a natural embedding of $\mathbf{Q}_n[z]$ into $\mathbf{Q}_{n+k}[z]$. Let $\tilde{I}$ be the ideal of $\mathbf{Q}_{n+k}$ generated by $\chi(I)$ and the k polynomials $(\chi(f))(x_{n+j})$ for $1 \leq j \leq k$. $\tilde{I}$ is clearly a triangular ideal since I is triangular and f is monic.
2. Let E be the evaluation finite field of A, with characteristic p. If $f \notin A_{(p)}[z]$ then FAIL. Otherwise, all the coefficients of f can be evaluated into E, so map f to $e \in \tilde{E}[z]$, using ϕ_Γ. If e is not squarefree, then FAIL.
3. Let $\tilde{E}$ be a minimal-degree splitting field of e over E and let $\tilde{e}$ be e mapped to $\tilde{E}[z]$. As $\tilde{e}$ is squarefree and with degree at least k we can compute k distinct roots $\delta_j \in \tilde{E}$ of $\tilde{e}$ for $1 \leq j \leq k$. Let $\tilde{\gamma}_i$ be γ_i lifted to E' for $1 \leq i \leq n$ and let $\tilde{\gamma}_{n+j} = \delta_j$ for $1 \leq j \leq k$. Let $\tilde{\Gamma} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_{n+k})$. Let $\tilde{A} = (\mathbf{Q}_n/\tilde{I})/(\tilde{J}/\tilde{I})$, where $\tilde{J} = \mathcal{J}(\tilde{I}, \tilde{\Gamma})$.
4. Let $\beta_j = (x_{n+j} + \tilde{I}) + \tilde{J}/\tilde{I}$ for $1 \leq j \leq k$. Return $\tilde{A}$, the natural extension of χ to $\chi : A \to \tilde{A}$, and the k roots $\beta_1, \ldots, \beta_k \in \tilde{A}$.

Theorem 5.1. *Algorithm* SquarefreeRoots *is correct.*

Proof. If the algorithm succeeds, then $\tilde{A}$ is a well-defined ACF since $\tilde{I}$ is clearly compatible with $\tilde{\Gamma}$ by construction and the new polynomials in $\tilde{I}$ satisfy the resultant condition since f is squarefree. As $\tilde{\Gamma}$ equals Γ in its first n entries, and $\tilde{I}$ contains the embedding of I in $\mathbf{Q}_{n+k}$, A is clearly isomorphic to the subfield of $\tilde{A}$ defined by the first n variables and the first n entries of $\tilde{\Gamma}$. As $\tilde{I}$ contains $(\chi(f))(x_{n+j})$ and $\tilde{J} \supseteq \tilde{I}$, clearly $f(\beta_j)$ is zero for $1 \leq j \leq k$. Since the $\tilde{\gamma}_{n+j}$ (for $1 \leq j \leq k$) are distinct by construction, the β_j must be distinct (their differences must be non-zero in $\tilde{A}$). □

If SquarefreeRoots fails, then we have to simplify A fully to obtain the ideal J explicitly (see Section 7 below for how this is done). The explicit computation of J can certainly be very expensive, but fortunately SquarefreeRoots will fail very rarely in practice, since the prime p is normally about the machine word size, so is unlikely to divide the discriminants of typical polynomials.

Once we have the maximal ideal J, with triangular basis $(q_1, \ldots, q_n)$, then we start from scratch with the trivial ACF and a new prime finite field E', and then successively call SquarefreeRoots with q_1, q_2, etc. (and with $k = 1$ each time) to build up to a new ACF A' with a different evaluation sequence Γ', but with the same $I = J$ as before. We then call SquarefreeRoots on f with the original k to obtain $\tilde{A}$ and the desired roots. If any of these calls to SquarefreeRoots fails, then we start again with a new finite field of different characteristic (again, failure is very unlikely).

5.2 Implementation Remarks

In the implementation, when given squarefree $f \in A[z]$ and k, to compute k roots of f we do various optimizations before calling the above algorithm.

1. If f is linear of the form $z - a$, then we can return a as is, of course.
2. For each of the current generators $\alpha_i = (x_i + I) + J/I \in A$, for $1 \leq i \leq n$, we test whether $f(\alpha_i)$ or $f(-\alpha_i)$ is zero (using IsZero). If so, then α_i or $-\alpha_i$ is a root, respectively. This picks up the common case that one asks for the roots of the same polynomial several times, and avoids the creation of many redundant polynomials (but is not strictly necessary for correctness as equality with earlier roots would be handled by the scheme as is). Note that if we use the negative generators $-\alpha_i$ in this test, then we must also check using IsZero that each new root is distinct from the other roots, since it is possible, for example, that $-\alpha_2 = \alpha_1$.
3. If f can be written over $\mathbf{Q}$, we factorize f over $\mathbf{Q}$ (which will be very easy in practice) and then for each irreducible factor, we apply the previous two steps and then SquarefreeRoots if necessary, until k roots are found.

The fact that we allow general k in SquarefreeRoots (not just the degree of the polynomial f) can be quite useful when one wishes, say, only one root of a polynomial and not all the conjugates of the root, as they will cause the field to have higher rank than necessary and this can make the full simplification of the field much more difficult (if that is needed later).

Finally, the roots with multiplicities of a general polynomial in $A[z]$ are also computed easily. Since A is a field, we can simply apply the standard squarefree factorization algorithm (which uses only derivative and GCD computations in $A[z]$) to the polynomial, and then use the above methods for each squarefree factor found; the corresponding multiplicities are simply attached to each root at the end.

5.3 Simple Defining Polynomials

One of the key features of our scheme is that even though the defining polynomials in I defining the roots of f returned by SquarefreeRoots are simply f evaluated in each new variable, the modular evaluation feature ensures that the roots are distinct. In this way, the defining polynomials in I are simple and sparse if f is.

A direct algebraic way to ensure distinct roots does not share this simplicity. Suppose we have $f(x) \in K[x]$, and we wish to have an extension of K in which α and β are roots of f but also distinct. One way is to let the defining polynomials be $f(\alpha)$ and $g(\beta, \alpha) = \frac{f(\beta)-f(\alpha)}{\beta-\alpha}$. But the defining polynomial g for β is more complicated than simply $f(\beta)$. This method can be generalized for computing more roots of f, but each successive defining polynomial becomes worse at each step.

As an example to illustrate the point, say we asked for the roots over an ACF A of $f = x^{10}+x+1$. A splitting field of f has absolute degree $10! = 3628800$. But the ideal I of the ACF simply has 10 “copies” of f for its defining polynomials. Now if we were to invert an expression involving many of the roots, then of course the result could be a huge expression which would be impractical to represent. However, as long as we work with each root separately, it will be just as if we work in $\mathbf{Q}[x]/\langle f\rangle$ in each case, which is quite practical. The point is that the defining polynomials for the roots stay as simple and sparse for as long as possible, and will only become “messy” if the user does something which forces this to happen. Furthermore, for another polynomial f of degree 10, this may never happen anyway, as the absolute splitting field degree may be quite small. In contrast, the scheme in the previous paragraph would just not work for a degree 10 polynomial because the defining polynomials would simply have too many terms!

6 Minimal Polynomial

In this section we show how to compute minimal polynomials of ACF elements.

Algorithm MinimalPolynomial(a)
Input: a, an element of an $A = (\mathbf{Q}_n/I)/(J/I)$ with $J = \mathcal{J}(I, \Gamma)$.
Output: A new ACF $A' = (\mathbf{Q}_n/I')/(J/I')$ isomorphic to A, the isomorphism $\psi : A \to A'$, and an irreducible polynomial $m \in \mathbf{Q}[x]$ such that $m(\psi(a))$ is the zero element of A' (or $m(a)$ is the zero element of A).

```
1. Write $a = (r + I) + J/I$ with $r \in \mathbf{Q}_n$.
2. Compute the minimal polynomial $M \in \mathbf{Q}[x]$ of $(r + I)$ in the quotient ring $\mathbf{Q}_n/I$.
3. Factorize $M$ over $\mathbf{Q}$ as $\prod_{i=1}^{k} p_i^{s_i}$ where each $p_i$ is irreducible over $\mathbf{Q}$.
4. For each $i = 1, \ldots, k$, let $e = p_i(a)$ and call ZeroTest on $e$ to obtain $A'$ and $\psi : A \to A'$ and then test whether $\psi(e)$ is the zero element of $A'$ (without using $J$); if so, return $A'$, $\psi : A \to A'$ and $p_i$ (i.e, return at the first successful $i$).
```

Theorem 6.1. *Algorithm* MinimalPolynomial *is correct, and does not explicitly use J.*

Proof. Since $M(r+I)$ is zero in $\mathbf{Q}_n/I$, $M(a)$ must also be zero in A, since $I \subseteq J$. But since A is a field and $M(a) = 0$, one irreducible factor of M over $\mathbf{Q}$ must be the minimal polynomial of a. Thus for exactly one i in Step 4, p_i must be the minimal polynomial of a so the evaluation $p_i(a)$ will be discovered to be the zero element when moving to A' (without using J explicitly, by Theorem 4.1), and p_i will be correctly returned at that point. □

In the implementation, the minimal polynomial of $r + I$ in $\mathbf{Q}_n/I$ is again computed by a fast matrix technique: either one can compute the minimal polynomial of the representation matrix B of r above, or one can compute the powers of r modulo I and write these as vectors until a dependency is found; in either case a fast p-adic nullspace algorithm avoids intermediate coefficient growth over $\mathbf{Q}$.

Note also that in the implementation, if $m(x)$ is the value returned, then $m(a)$ will be represented exactly as the zero polynomial after the call.

7 Simplifying an ACF and Computing an Absolute Field

Let $A = (\mathbf{Q}_n/I)/(J/I)$ (with $J = \mathcal{J}(I, \Gamma)$) be an ACF, and suppose we wish to simplify A fully; that is, we wish to compute the ideal J explicitly. This may be very expensive of course—the whole point of the scheme is to avoid factorizations and hope that factors are found during the running of an algorithm using the field! But the fully simplified field may still be desired by the user, and is necessary in the (rare) case of failure of Algorithm SquarefreeRoots.

The full simplification algorithm works by fully factoring the successive polynomials in the triangular basis of I over the previous subfield constructed. For each factorization, we select the irreducible factor which evaluates to zero at Γ (there must always be exactly one). The factorization is done by a variant of Trager's algorithm [11], and the univariate factorization over $\mathbf{Z}$ at the base level is done by the standard Berlekamp-Zassenhaus (BZ) algorithm [8, p. 452] coupled with the new fast combination algorithm of van Hoeij [12].

Before we do the full simplification, we use a fast method to obtain a partial simplification: if $\alpha_1, \ldots, \alpha_n$ are the generators of A, then we call the algorithm MinimalPolynomial on each α_i and then also on $\alpha_i + \alpha_j$ for $1 \leq i < j \leq n$. We discard all the results, but elements of J will be found. This often finds most of the necessary simplifications, and takes advantage of the fast matrix techniques so is quite fast.

To compute an absolute number field $K = \mathbf{Q}[z]/\langle f \rangle$ such that A is isomorphic to K, we first fully simplify A to get J, then put J into normal position or shape lemma form (see [2, Sec. 8.6] for example). This gives not only an absolute field K, but also an isomorphism from A to K. A modification of the FGLM algorithm [7] can be used, which again uses fast p-adic matrix techniques in the implementation. Of course, we can only compute an absolute number field in practice when the absolute degree is not too large (say, up to about 1000).

8 An Example

This example briefly illustrates how the scheme performs in its Magma implementation (on a 400MHz Sun Ultrasparc). The Cyclic-6 roots ideal I is generated by 6 polynomials in 6 variables over $\mathbf{Q}$. The Gröbner basis (GB) w.r.t. the lexicographical order has 17 polynomials (computed in 1.2 seconds).

The (affine) variety of I is the set of all solutions to the system of equations implied by I. It is computed over an ACF by successively computing roots of polynomials in the GB. This takes 4.0 seconds and there are 156 elements in the variety. After this, the ACF has rank 28, and the defining polynomials have degrees 1, 2, 4 or 8 (e.g., $a^4 + 4a^3 + 15a^2 + 4a + 1$ and $b^2 + 4b + 1$).

The ACF is fully simplified in 2.9 seconds and this modifies it to have only 3 quadratic defining polynomials, while all the other defining polynomials become linear (so their variables are eliminated from any coordinates of the solutions). It then takes 0.1 seconds to find an absolute polynomial $f = x^8+4x^6-6x^4+4x^2+1$ for the ACF. We thus discover that if we start again with the degree-8 number field $K = \mathbf{Q}[x]/\langle f\rangle$, then the variety of the ideal over an algebraic closure of $\mathbf{Q}$ can in fact be fully constructed over K.

References

1. Philippe Aubry, Daniel Lazard, and Marc Moreno Maza. On the theories of triangular sets. *J. Symbolic Comp.*, 28(1-2):105–124, 1999.
2. Thomas Becker and Volker Weispfenning. *Gröbner Bases.* Graduate Texts in Mathematics. Springer, New York–Berlin–Heidelberg, 1993.
3. Wieb Bosma and John Cannon. *Handbook of Magma Functions.* University of Sydney, 2001.
4. Wieb Bosma, John Cannon, and Catherine Playoust. The Magma Algebra System I: The User Language. *J. Symbolic Comp.*, 24(3):235–265, 1997.
5. David Cox, John Little, and Donal O'Shea. *Ideals, Varieties and Algorithms.* Undergraduate Texts in Mathematics. Springer, New York–Berlin–Heidelberg, 2nd edition, 1996.
6. J. Della Dora, C. Dicrescenzo, and D. Duval. About a new method for computing in algebraic number fields. In B.F. Caviness, editor, *Proc. EUROCAL '85*, volume 204 of *LNCS*, pages 289–290, Linz, 1985. Springer.
7. Jean-Charles Faugère, Patrizia Gianni, Daniel Lazard, and Teo Mora. Efficient computations of zero-dimensional Gröbner bases by change of ordering. *J. Symbolic Comp.*, 16:329–344, 1993.
8. Donald E. Knuth. *The Art of Computer Programming*, volume 2. Addison Wesley, Reading, Massachusetts, 3rd edition, 1997.
9. Daniel Lazard. Solving zero-dimensional algebraic systems. *J. Symbolic Comp.*, 13(2):117–131, 1992.
10. Magma Computational Algebra System: Homepage. University of Sydney. URL:http://magma.maths.usyd.edu.au.
11. Barry M. Trager. Algebraic factoring and rational function integration. In R.D. Jenks, editor, *Proc. SYMSAC '76*, pages 196–208. ACM press, 1976.
12. Mark van Hoeij. Factoring polynomials and the knapsack problem. *J. Number Th.*, to appear 2002. URL:http://www.math.fsu.edu/~hoeij/paper/knapsack.ps.

Additive Complexity and Roots of Polynomials over Number Fields and $\mathfrak{p}$-adic Fields

J. Maurice Rojas*

Department of Mathematics
Texas A&M University
TAMU 3368
College Station, Texas 77843-3368
USA
rojas@math.tamu.edu
http://www.math.tamu.edu/~rojas

Abstract. Consider any nonzero univariate polynomial with rational coefficients, presented as an elementary algebraic expression (using only integer exponents). Letting $\sigma(f)$ denotes the additive complexity of f, we show that the number of rational roots of f is no more than

$$15+\sigma(f)^2(24.01)^{\sigma(f)}\sigma(f)!.$$

This provides a sharper arithmetic analogue of earlier results of Dima Grigoriev and Jean-Jacques Risler, which gave a bound of $C^{\sigma(f)^2}$ for the number of real roots of f, for $\sigma(f)$ sufficiently large and some constant C with $1<C<32$. We extend our new bound to arbitrary finite extensions of the ordinary **or** p-adic rationals, roots of bounded degree over a number field, and geometrically isolated roots of multivariate polynomial systems. We thus extend earlier bounds of Hendrik W. Lenstra, Jr. and the author to encodings more efficient than monomial expansions. We also mention a connection to complexity theory and note that our bounds hold for a broader class of fields.

1 Introduction

This paper presents another step in the author's program [Roj02] of establishing an effective arithmetic analogue of fewnomial theory. (See [Kho91] for the original exposition of fewnomial theory, which until now has always used the real or complex numbers for the underlying field.) Here, we show that the number of **geometrically isolated** roots (cf. section 2) of a polynomial system over any fixed $\mathfrak{p}$-adic field (and thereby any fixed number field) can be bounded from above by a quantity depending solely on the additive complexity of the input equations.

So let us first clarify the univariate case of **additive complexity**: If $\mathcal{L}$ is any field, we say that $f\in\mathcal{L}[x]$ has **additive complexity $\leq$ s (over $\mathcal{L}$)** iff there exist

[*] This research was partially supported by a grant from the Texas A&M College of Science.

C. Fieker and D.R. Kohel (Eds.): ANTS 2002, LNCS 2369, pp. 506–515, 2002.

constants $c_1, d_1, \ldots, c_s, d_s, c_{s+1} \in \mathcal{L}$ and arrays of nonnegative integers $[m_{i,j}]$ and $[m'_{i,j}]$ with $f(x) = c_{s+1} \prod_{i=0}^{s} X_i^{m_{i,s+1}}$, where $X_0 = x$, $X_1 = c_1 X_0^{m_{0,1}} + d_1 X_0^{m'_{0,1}}$, and $X_j = c_j \left(\prod_{i=0}^{j-1} X_i^{m_{i,j}} \right) + d_j \left(\prod_{i=0}^{j-1} X_i^{m'_{i,j}} \right)$ for all $j \in \{2, \ldots, s\}$. We then define the **additive complexity (over $\mathcal{L}$) of f**, $\sigma_{\mathcal{L}}(\mathbf{f})$, to be the least s in such a presentation of f as an algebraic expression. Note in particular that additions or subtractions in repeated sub-expressions are thus not counted, e.g., $9(x-7)^{99}(2x+1)^{43} - 11(x-7)^{999}(2x+1)^3$ has additive complexity ≤ 3.

It has been known since the work of Allan Borodin and Stephen A. Cook around 1974 [BC76] that there is a deep connection between additive complexity over the real numbers $\mathbb{R}$ and the number of real roots of a nonzero polynomial in $\mathbb{R}[x]$. For example, they showed that there is a real constant K such that the number of real roots of f is no more than $2^{2^{\cdot^{\cdot^{2^{K\sigma_{\mathbb{R}}(f)}}}}}$, where the number of exponentiations is $\sigma_{\mathbb{R}}(f) - 1$ [BC76]. Jean-Jacques Risler, using Khovanski's famous Theorem on Real Fewnomials [Kho80,Kho91], then improved this bound to $(\sigma_{\mathbb{R}}(f)+2)^{3\sigma_{\mathbb{R}}(f)+1} 2^{\left(9\sigma_{\mathbb{R}}(f)^2+5\sigma_{\mathbb{R}}(f)+2\right)/2}$ [Ris85, pg. 181, line 6]. (Dima Grigoriev derived a similar bound earlier [Gri82] and both results easily imply a simplified bound of $C^{\sigma_{\mathbb{R}}(f)^2}$ for the number of real roots of f, for $\sigma_{\mathbb{R}}(f)$ sufficiently large and some constant C with $1 < C < 32$.)

Here, based on a recent near-optimal **arithmetic** analogue of Khovanski's Theorem on Real Fewnomials found by the author (cf. section 2 below), we give arithmetic analogues of these additive complexity bounds. Our first main result can be stated as follows:

Theorem 1. *Let p be any rational prime and let $\log_p(\cdot)$ denote the base p logarithm function. Also let $c := \frac{e}{e-1} \leq 1.582$, let $\mathcal{L}$ be any degree d algebraic extension of $\mathbb{Q}_p$, and let $f \in \mathcal{L}[x] \setminus \{0\}$. Then f has no more than $2^{O\left(\sigma_{\mathcal{L}}(f) \log\left(p^d \sigma_{\mathcal{L}}(f)\right)\right)}$ roots in $\mathcal{L}$. More precisely,* $1 + \frac{dp(p^d-1)}{p-1} + \frac{4cdp(p^d-1)^2}{p-1}\left(1 + d\log_p\left(\frac{2d}{\log p}\right)\right)$

$$+\frac{1}{3} \sum_{j=3}^{\sigma_{\mathcal{L}}(f)} j(6c)^j (p^d - 1)^j \left(1 + d\log_p\left(\frac{d}{\log p}\right)\right)\left(1 + d\log_p\left(\frac{2d}{\log p}\right)\right)^{j-1} j! \text{ is a}$$

valid upper bound, and just the first $\sigma_{\mathcal{L}}(f) + 1$ summands suffice if $\sigma_{\mathcal{L}}(f) \leq 2$.

Remark 1. *Our bounds can be improved further and this is detailed in remark 6 of section 3.* ◇

Remark 2. *Note that via the obvious embedding $\mathbb{Q} \subset \mathbb{Q}_2$, theorem 1 easily implies a similar statement for $\mathcal{L}$ a number field. A less trivial extension to number fields appears in theorem 2 below.* ◇

Example 1. *Taking $\mathcal{L} = \mathbb{Q}_2$, we obtain respective upper bounds of* 1, 3, 35, 50195, *and* 6471489 *on the number of roots of f in $\mathbb{Q}_2$, according as $\sigma_{\mathbb{Q}_2}(f)$ is*

0, 1, 2, 3, or 4.[1]

For instance, we see that for any non-negative integers $\alpha, \beta, \gamma, \delta, \varepsilon, \lambda, \mu, \nu$ and constants $c_1, d_1, c_2, d_2, c_3 \in \mathbb{Q}_2$, the polynomial

$$c_3 x^\alpha \left(c_1 x^\beta + d_1 x^\gamma\right)^\delta \left[c_2 \left(c_1 x^\beta + d_1 x^\gamma\right)^\varepsilon + d_2 x^\lambda \left(c_1 x^\beta + d_1 x^\gamma\right)^\mu\right]^\nu$$

has no more than 35 *roots in $\mathbb{Q}_2$ (or $\mathbb{Q}$ obviously). See remark 5 below for improvements of some of these bounds.*

Note that for $\sigma_{\mathbb{R}}(f) \in \{0,1,2,3,4\}$ Risler's bound on the number of real roots respectively specializes to 4, 20736, 274877906944, 5497558138880000000000*, and* 126315281744229461505151771531542528.$\diamond$

The importance of bounds on the number of roots in terms of additive complexity is two-fold: on the one hand, we obtain a new way to bound the number of roots in $\mathcal{L}$ of any univariate polynomial with coefficients in $\mathcal{L}$. Going the opposite way, we can use information about the number of roots in $\mathcal{L}$ of a given univariate polynomial to give a lower bound on the minimal number of additions and subtractions necessary to evaluate it. More to the point, a recent theorem of Shub and Smale establishes a deep connection between the number of integral roots of a univariate polynomial, a variant of additive complexity, and certain fundamental complexity classes.

To make this precise, let us consider another formalization of algebraic expressions. Rather than allowing arbitrary recursive use of integral powers and field operations, let us be more conservative and do the following: Suppose we have $f \in \mathbb{Z}[x_1]$ expressed as a sequence of the form $(1, x_1, f_2, \ldots, f_N)$, where $f_N = f(x_1)$, $f_0 := 1$, $f_1 := x_1$, and for all $i \geq 2$ we have that f_i is a sum, difference, or product of some pair of elements (f_j, f_k) with $j, k < i$. (Such computational sequences are also known as **straight-line programs** or **SLP's**.) Let $\tau(\mathbf{f})$ denote the smallest possible value of $N-1$, i.e., the smallest length for such a computation of f. Clearly, $\tau(f)$ also admits a definition in terms of multivariate polynomial systems much like that of $\sigma_{\mathcal{L}}(f)$. So it is clear that $\tau(f) \geq \sigma_{\mathcal{L}}(f)$ for all $f \in \mathbb{Z}[x_1]$ and $\mathcal{L} \supseteq \mathbb{Z}$, and that $\sigma_{\mathcal{L}}(f)$ is often dramatically smaller than $\tau(f)$.

The Shub–Smale τ Theorem *[BCSS98, theorem 3, pg. 127] Suppose there is an absolute constant κ such that for all nonzero $f \in \mathbb{Z}[x_1]$, the number of distinct roots of f in $\mathbb{Z}$ is no more than $(\tau(f)+1)^\kappa$. Then $\mathbf{P}_{\mathbb{C}} \neq \mathbf{NP}_{\mathbb{C}}$.*

In other words, an analogue (regarding complexity theory over $\mathbb{C}$) of the famous unsolved $\mathbf{P} \stackrel{?}{=} \mathbf{NP}$ question from computer science (regarding complexity theory over the ring $\mathbb{Z}/2\mathbb{Z}$) would be settled. The question of whether $\mathbf{P}_{\mathbb{C}} \stackrel{?}{=} \mathbf{NP}_{\mathbb{C}}$ remains open as well but it is known that $\mathbf{P}_{\mathbb{C}} = \mathbf{NP}_{\mathbb{C}} \Longrightarrow \mathbf{NP} \subseteq \mathbf{BPP}$. (This observation is due to Steve Smale and was first published in [Shu93].) The complexity class **BPP** is central in randomized complexity and cryptology, and the last inclusion (while widely disbelieved) is also an open question. The truth of

[1] All calculations in this paper were done with the assistance of `Maple` and the corresponding `Maple` code can be found on the author's web-page.

the hypothesis of The Shub–Smale τ Theorem, also know as the **τ-conjecture**, is yet another open problem, even for $\kappa = 1$.

Observing that the number of integral roots of f is no more than $\deg f$ (by the fundamental theorem of algebra), and that $\deg f \leq 2^{\tau(f)}$ (since $\deg f_{i+1} \leq 2\max_{j<i} \deg f_j$), we easily obtain the following crude upper bound.

Proposition. *The number of integral roots of $f \in \mathbb{Z}[x_1] \setminus \{0\}$ is at most $2^{\tau(f)}$.*

As of April 2002, no asymptotically sharper bound in terms of $\tau(f)$ appears to be known![2] However, taking a 2-adic approach via theorem 1, we immediately obtain the following improvement.

Corollary. *The number of integral roots of $f \in \mathbb{Z}[x_1] \setminus \{0\}$ is $2^{\mathcal{O}\left(\sigma_{\mathbb{Q}_2}(f) \log \sigma_{\mathbb{Q}_2}(f)\right)}$.*

This bound, while apparently not polynomial in $\tau(f)$, at least has the advantage that it is frequently much smaller than $2^{\tau(f)}$. For instance, our corollary tells us that the polynomial from example 1 has no more than 35 integral roots, while the proposition above would give us a non-constant upper bound of at least α, since this example (if not identically zero) has degree $\geq \alpha$.

Whether our 2-adic approach can be pushed farther to solve the τ-conjecture is an intriguing open question. In particular, it isn't even known if there is a family of f with $2^{\Omega\left(\sigma_{\mathbb{Q}_2}(f)\right)}$ roots in $\mathbb{Q}_2$.

Remark 3. *Curiously, using additive complexity over a different complete field — $\mathbb{R}$ — can* **not** *lead to a solution of the τ-conjecture: there are examples of $f \in \mathbb{Z}[x_1]$ with $\sigma_{\mathbb{R}}(f) = \mathcal{O}(r)$ and over 2^r real (but irrational) roots [Roj00, sec. 3, pg. 13] (see [BC76] for an even bigger lower bound).* ◇

Our main results are proved in section 3, where we in fact prove sharper versions. There we also prove a refined number field analogue of theorem 1, which we now state. Recall that if L is a subfield of $\mathbb{C}$ and $x \in \mathbb{C}$ then we say that **x is of degree $\leq \delta$ over L** iff x lies in an algebraic extension of L of degree $\leq \delta$.

Theorem 2. *Following the notation of theorem 1, let $\delta \in \mathbb{N}$ and suppose instead now that $\mathcal{L}$ is a degree d algebraic extension of $\mathbb{Q}$. Then the number of roots of f in $\mathbb{C}$ of degree $\leq \delta$ over $\mathcal{L}$ is $2^{\mathcal{O}\left(\sigma_{\mathcal{L}}(f)\left(d\delta + \log \sigma_{\mathcal{L}}(f)\right)\right)}$. More precisely,*

$$1 + c(d\delta + 10)2^{d\delta+1}\log_2\left(\frac{d\delta}{\log 2}\right) + c^2(d\delta+10)^2 4^{d\delta+2}\log_2\left(\frac{d\delta}{\log 2}\right)\log_2\left(\frac{2d\delta}{\log 2}\right)$$
$$+\frac{2}{3}\sum_{j=3}^{\sigma_{\mathcal{L}}(f)} j(6c)^j 2^{d\delta j}\left(1 + 2d^2\delta^2\log_2\left(\frac{d^2\delta^2}{\log 2}\right)\right)\left(1 + 2d^2\delta^2\log_2\left(\frac{2d^2\delta^2}{\log 2}\right)\right)^{j-1} j!$$

is a valid upper bound, and just the first $\sigma_{\mathcal{L}}(f) + 1$ summands suffice if $\sigma_{\mathcal{L}}(f) \leq 2$.

This family of bounds can also be sharpened further and this is also detailed in remark 6 of section 3.

[2] Using Descartes' Rule of Signs instead of the fundamental theorem of algebra does not easily yield a sharper bound: the number of monomial terms of f_i grows even faster as a function of $\tau(f)$ than $\deg f_i$.

In summary, Theorems 1 and 2 are the first bounds on the number of roots in a local field or number field which make explicit use of additive complexity. In particular, our results thus extend an earlier result of Lenstra on polynomials with few monomial terms to the setting of an even sharper input encoding. Recall that for any field L we let $L^* := L \setminus \{0\}$.

Lenstra's Theorem. *[Len99, prop. 7.2 and prop. 8.1] Following the notation of Theorems 1 and 2, suppose now that $\mathcal{L}$ is a degree d extension of $\mathbb{Q}_p$ (the* **local** *case)* **or** *$\mathbb{Q}$ (the* **global** *case), and that f has exactly m monomial terms. Then f has no more than $c(q_{\mathcal{L}} - 1)(m-1)^2 \left(1 + e_{\mathcal{L}} \log_p \left(\frac{e_{\mathcal{L}}(m-1)}{\log p}\right)\right)$ roots in $\mathcal{L}^*$ in the local case (counting multiplicities), where $e_{\mathcal{L}}$ and $q_{\mathcal{L}}$ respectively denote the ramification index and residue field cardinality of $\mathcal{L}$. Furthermore, f has no more than $c(m-1)^2(d\delta + 10) \cdot 2^{d\delta+1} \log_2 \left(\frac{d\delta(m-1)}{\log 2}\right)$ roots in $\mathbb{C}^*$ of degree $\leq \delta$ over $\mathcal{L}$ in the global case (counting multiplicities).*

Remark 4. *Recall that $q_{\mathcal{L}}$ is always an integer power of p and $e_{\mathcal{L}} \log_p q_{\mathcal{L}} = d$.* ⋄

Example 2. *Considering the polynomial from example 1 once again, note that Lenstra's Theorem can not even give a constant upper bound for the number of roots in $\mathbb{Q}_2^*$, since the number of monomial terms depends on λ (among other parameters). On the other hand, in the absence of an expression for f more compact than a sum of m monomial terms, Lenstra's bound is quite practical.* ⋄

Remark 5. *Hendrik W. Lenstra has observed that $B(\mathcal{L}, 2, 1)$ is in fact the number of roots of unity in $\mathcal{L}$, which is in turn bounded above by $\frac{e_{\mathcal{L}} p(q_{\mathcal{L}}-1)}{p-1}$ [Len99]. He has also computed $B(\mathbb{Q}_2, 3, 1) = 6$ (giving $3x_1^{10} + x_1^2 - 4$ as a trinomial which realizes the maximum possible number of nonzero roots in $\mathbb{Q}_2$) [Len99, prop. 9.2]. As a consequence (following easily from our proof of theorem 1), the first three summands of our main formula from theorem 1 can be replaced by $1 + \frac{e_{\mathcal{L}} p(q_{\mathcal{L}}-1)}{p-1} + \frac{e_{\mathcal{L}} p(q_{\mathcal{L}}-1) B(\mathcal{L},3,1)}{p-1}$, and our bounds from example 1 can be improved to 3 and 15 in the respective cases $\sigma_{\mathbb{Q}_2}(f) = 1$ and $\sigma_{\mathbb{Q}_2}(f) = 2$. (This is how we derived the bound cited in the abstract.)* ⋄

As mentioned earlier, our main results follow easily from the author's recent arithmetic multivariate analogue of Descartes' Rule [Roj02]. In fact, Arithmetic Multivariate Descates' Rule even allows us to derive multivariate extensions of Theorems 1 and 2 which we state below. So let us precede our proofs by a brief discussion of this important background result.

2 Useful Multivariate Results

Suppose $\mathbf{f_1}, \ldots, \mathbf{f_k} \in \mathcal{L}[x_1^{\pm 1}, \ldots, x_n^{\pm 1}] \setminus \{0\}$, and $\mathbf{m_i}$ is the total number of distinct exponent vectors appearing in f_i (assuming all polynomials are written as sums of monomials). We call $\mathbf{F} := (f_1, \ldots, f_k)$ a $\mathbf{k} \times \mathbf{n}$ polynomial system over $\mathcal{L}$ of **type** $(\mathbf{m_1}, \ldots, \mathbf{m_k})$, and we call a root ζ of F **geometrically isolated** iff ζ is a zero-dimensional component of the underlying scheme over the algebraic closure of $\mathcal{L}$ defined by F. If $\mathcal{L}$ is a finite extension of $\mathbb{Q}_p$ (resp. $\mathbb{Q}$) then we say that we are in the **local** (resp. **global**) case.

Arithmetic Multivariate Descartes' Rule (Special Case). *[Roj02, cor. 1 of sec. 2 and cor. 2 of sec. 3] Let p be any (rational) prime and d, δ positive integers. Suppose $\mathcal{L}$ is any degree d algebraic extension of $\mathbb{Q}_p$ or $\mathbb{Q}$, and let $\mathcal{L}^* := \mathcal{L} \setminus \{0\}$. Also let $m := (m_1, \ldots, m_n) \in \mathbb{N}^n$, $N := (N_1, \ldots, N_n) \in \mathbb{N}^n$, and F an $n \times n$ polynomial system over $\mathcal{L}$ of type m such that the number of variables occuring in f_i is exactly N_i. Define $\mathbf{B}(\mathcal{L}, \mathbf{m}, \mathbf{N})$ to be the maximum number of isolated roots in $(\mathcal{L}^*)^n$ of such an F in the local case, counting multiplicities.*[3] *Then*

$$B(\mathcal{L}, m, N) \leq c^n q_{\mathcal{L}}^n \prod_{i=1}^{n} \left\{ m_i(m_i-1)N_i \left[1 + e_{\mathcal{L}} \log_p \left(\frac{e_{\mathcal{L}}(m_i-1)}{\log p} \right) \right] \right\},$$

where $c := \frac{e}{e-1} \leq 1.582$, and $e_{\mathcal{L}}$ and $q_{\mathcal{L}}$ are respectively the ramfication index and residue field cardinality of $\mathcal{L}$.

Furthermore, moving to the global case, let us say a root $x \in \mathbb{C}^n$ of F is of **degree** $\leq \delta$ **over** $\mathcal{L}$ *iff every coordinate of x is of degree $\leq \delta$ over $\mathcal{L}$, and let us define $\mathbf{A}(\mathcal{L}, \delta, \mathbf{m}, \mathbf{N})$ to be the maximum number of isolated roots of such an F in $(\mathbb{C}^*)^n$ of degree $\leq \delta$ over $\mathcal{L}$, counting multiplicities.*[3] *Then*

$$A(\mathcal{L}, \delta, m, N) \leq 2c^n 2^{d\delta n} \prod_{i=1}^{n} \left\{ m_i(m_i-1)N_i \left[1 + 2d^2\delta^2 \log_2 \left(\frac{d^2\delta^2(m_i-1)}{\log 2} \right) \right] \right\}.$$

Various other improvements of these bounds are detailed in [Roj02]. However, let us at least point out that our bound above is nearly optimal: For **fixed** $\mathcal{L}$, $\log B(\mathcal{L}, (\mu, \ldots, \mu), (n, \ldots, n))$ and $\log A(\mathcal{L}, (\mu, \ldots, \mu), (n, \ldots, n))$ are $\Theta(n \log \mu)$, where the implied constant depends on $\mathcal{L}$ (and d and δ) [Roj02, example 2].

Via our definition of additive complexity we will reduce the proofs of our main results to an application of Arithmetic Multivariate Descartes' Rule. In particular, it appears that any further improvement to our main results will have to come from a different technique. For now, we have the following generalization of Theorems 1 and 2.

Definition 1. *Following the notation above, given any $k \times n$ polynomial system $F = (f_1, \ldots, f_k)$ over $\mathcal{L}$, let us define its* **additive complexity over** $\mathcal{L}$, $\sigma_{\mathcal{L}}(\mathbf{F})$, *to be the smallest s such that*

$$F(x_1, \ldots, x_n) = \left(c_{n+s+1}^{(1)} \prod_{i=1}^{n+s} X_i^{m_{i,n+s+1}^{(1)}}, \ldots, c_{n+s+1}^{(k)} \prod_{i=1}^{n+s} X_i^{m_{i,n+s+1}^{(k)}} \right),$$

[3] The multiplicity of any isolated root here, which we take in the sense of intersection theory for a scheme over the algebraic closure of $\mathcal{L}$ [Ful98], turns out to always be a positive integer when $k = n$ (see, e.g., [Smi97,Roj99]).

where $X_j := x_j$ *for all* $j \in \{1, \ldots, n\}$, $X_j = c_j \left(\prod_{i=1}^{j-1} X_i^{m_{i,j}}\right) + d_j \left(\prod_{i=1}^{j-1} X_i^{m'_{i,j}}\right)$ *for all* $j \in \{n+1, \ldots, n+s\}$, $c_1, d_1, \ldots, c_{n+s}, d_{n+s}, c^{(1)}_{n+s+1}, \ldots, c^{(k)}_{n+s+1} \in \mathcal{L}$, *and* $[m_{i,j}]$, $[m'_{i,j}]$, *and* $[m^{(\ell)}_{i,j}]$ *are arrays of positive integers.* $\diamond$

Theorem 3. *Following the notation above, F has no more than*

$$1 + B(\mathcal{L}, 2, 1) + (1 + B(\mathcal{L}, 2, 1)B(\mathcal{L}, 3, 1)) + \sum_{\ell=3}^{\sigma_{\mathcal{L}}(F)} \binom{n+\ell-1}{n-1}$$
$$B(\mathcal{L}, (\underbrace{2, \ldots, 2}_{n}, \underbrace{3, \ldots, 3}_{\ell-n}), (n+1, n+2, \ldots, n+\ell-1, n+\ell-1))$$

geometrically isolated roots in $\mathcal{L}^n$, *or*

$$1 + A(\mathcal{L}, \delta, 2, 1) + (1 + A(\mathcal{L}, \delta, 2, 1)A(\mathcal{L}, \delta, 3, 1)) + \sum_{\ell=3}^{\sigma_{\mathcal{L}}(F)} \binom{n+\ell-1}{n-1}$$
$$A(\mathcal{L}, \delta, (\underbrace{2, \ldots, 2}_{n}, \underbrace{3, \ldots, 3}_{\ell-n}), (n+1, n+2, \ldots, n+\ell-1, n+\ell-1))$$

geometrically isolated roots in $\mathbb{C}^n$ *of degree* $\leq \delta$ *over* $\mathcal{L}$, *according as we are in the local or global case. In particular, for each bound, the first* $\sigma_{\mathcal{L}}(F) + 1$ *summands suffice if* $\sigma_{\mathcal{L}}(F) \leq 2$.

In closing, let us point out a topological anomaly: Over $\mathbb{R}$, one can go even farther and bound the number of connected components of the zero set of a multivariate polynomial in terms of additive complexity [Gri82,Ris85]. Unfortunately, since $\mathbb{Q}_p$ is totally disconnected as a topological space [Kob84], one can not derive any obvious analogous statement in our arithmetic setting. This is why we consider only geometrically isolated roots in the multivariate case. Nevertheless, it would be quite interesting to know if one could bound the number of higher-dimensional **irreducible** components defined over $\mathcal{L}$ in terms of additive complexity, when $\mathcal{L}$ is a $\mathfrak{p}$-adic field.

3 Proving Theorems 1–3

We will give a proof of Theorem 3 which simultaneously yields Theorems 1 and 2 for free.

Proof of Theorem 3 (and Theorems 1 and 2): First note that by the definition of additive complexity, $(x_1, \ldots, x_n)$ is a geometrically isolated root of $F \Longrightarrow (X_1, \ldots, X_{n+s})$ is a geometrically isolated root of the polynomial system $G = \mathbf{O}$, where the corresponding equations are exactly

$$c^{(1)}_{n+s+1} \prod_{i=1}^{n+s} X_i^{m^{(1)}_{i,n+s+1}} = 0 \quad , \ldots , \quad c^{(k)}_{n+s+1} \prod_{i=1}^{n+s} X_i^{m^{(k)}_{i,n+s+1}} = 0,$$

$$X_{n+1} = c_{n+1}\left(\prod_{i=1}^{n} X_i^{m_{i,n+1}}\right) + d_{n+1}\left(\prod_{i=1}^{n} X_i^{m'_{i,n+1}}\right)$$

$$\vdots$$

$$X_{n+s} = c_{n+s}\left(\prod_{i=1}^{n+s-1} X_i^{m_{i,n+s}}\right) + d_{n+s}\left(\prod_{i=1}^{n+s-1} X_i^{m'_{i,n+s}}\right),$$

where $s := \sigma_{\mathcal{L}}(F)$, $X_i = x_i$ for all $i \in \{1, \ldots, n\}$, and the c_i, d_i, $c_i^{(j)}$, $m_{i,j}$, and $m'_{i,j}$ are suitable constants. This follows easily from the fact that corresponding quotient rings $\mathcal{L}[x_1]/\langle f\rangle$ and $\mathcal{L}[X_0, \ldots, X_s]/\langle G\rangle$ are isomorphic, thus making $\mathbb{C}_p[x_1]/\langle f\rangle$ and $\mathbb{C}_p[X_0, \ldots, X_s]/\langle G\rangle$ isomorphic, where $\mathbb{C}_p$ denotes the completion of the algebraic closure of $\mathbb{Q}_p$. In particular, $k \leq n$ easily implies that F has no geometrically isolated roots in $\mathcal{L}$ at all, so we can assume that $k \geq n$.

So we now need only count the geometrically isolated roots of G in $\mathcal{L}^{n+s}$ (or the geometrically isolated roots of F in $\mathbb{C}^{n+s}$ of degree $\leq \delta$ over $\mathcal{L}$) precisely enough to conclude. Toward this end, note that the first n equations of $G = \mathbf{O}$ imply that at least n distinct X_i must be 0, for otherwise $(X_1, \ldots, X_{n+s})$ would not be an isolated root. Note also that if we have exactly n of the variables $X_1, \ldots, X_{n+\ell}$ set to 0, then the first $n+\ell$ equations of G completely determine $(X_1, \ldots, X_{n+\ell})$. Furthermore, by virtue of the last $s-\ell$ equations of G, the value of $(X_1, \ldots, X_{n+\ell})$ **uniquely** determines the value of $(X_{n+\ell+1}, \ldots, X_{n+s})$. So it in fact suffices to find the total number of geometrically isolated roots (with all coordinates nonzero) of all systems of the form $G' = \mathbf{O}$, where the equations of G' are exactly $(0=0)$ or

$$\varepsilon_1 X_{n+1} = c_{n+1}\left(\prod_{i=1}^{n} X_i^{m_{i,n+1}}\right) + d_{n+1}\left(\prod_{i=1}^{n} X_i^{m'_{i,n+1}}\right)$$

$$\vdots$$

$$\varepsilon_\ell X_{n+\ell} = c_{n+\ell}\left(\prod_{i=1}^{n+\ell-1} X_i^{m_{i,n+\ell}}\right) + d_{n+s}\left(\prod_{i=1}^{n+\ell-1} X_i^{m'_{i,n+\ell}}\right),$$

where $\varepsilon_i \in \{0,1\}$ for all i, $X_{n+\ell} = \varepsilon_\ell = 0$, exactly $n-1$ of the variables $X_1, \ldots, X_{n+\ell-1}$ have been set to 0, and ℓ ranges over $\{1, \ldots, n\}$. Note in particular that the $j^{\underline{\text{th}}}$ equation involves no more than $n+j$ variables for all $j \in \{1, \ldots, \ell-1\}$, and that the $\ell^{\underline{\text{th}}}$ equation involves no more than $n+\ell-1$ variables.

To conclude, we thus see that G has no more than 1, $1 + B(\mathcal{L}, 2, 1)$,

$$\rho(\mathcal{L}) := 1 + B(\mathcal{L}, 2, 1) + (r_n + B(\mathcal{L}, 2, 1)B(\mathcal{L}, 3, 1)), \text{ or}$$

$$\rho(\mathcal{L}) + \sum_{\ell=3}^{s} \binom{n+\ell-1}{n-1} B(\mathcal{L}, (\underbrace{2, \ldots, 2}_{n}, \underbrace{3, \ldots, 3}_{\ell-n}), (n+1, n+2, \ldots, n+\ell-1, n+\ell-1))$$

geometrically isolated roots in $\mathcal{L}^{n+s}$ in the local case, according as s is 0, 1, 2, or ≥ 3, where r_n is 0 or 1 according as $n=1$ or $n \geq 2$. The corresponding statement for the global case, where we replace $B(\mathcal{L}, m, N)$ by $A(\mathcal{L}, \delta, m, N)$ throughout and count geometrically isolated roots in $\mathbb{C}^{n+s}$ of degree $\leq \delta$ over $\mathcal{L}$ instead, is also clearly true. This proves theorem 3.

Theorems 1 and 2 then follow immediately by specializing the above formulae to $n = 1$, applying Arithmetic Multivariate Descartes' Rule, and performing an elementary calculation. ■

Remark 6. *It follows immediately from our proof that we can restate Theorems 1 and 2 in sharper intrinsic terms. That is, the bounds from our proof above can immediately incorporate any new upper bounds for the quantities $B(\mathcal{L}, m, N)$ and $A(\mathcal{L}, \delta, m, N)$.* $\diamond$

Remark 7. *Note that the same proof will essentially work verbatim if we replace $\mathcal{L}$ throughout by* **any** *field admitting a multivariate analogue of Descartes' Rule.* $\diamond$[4]

Acknowledgement

The author thanks the two anonymous referees for their astute comments.

References

BCSS98. Blum, Lenore; Cucker, Felipe; Shub, Mike; and Smale, Steve, *Complexity and Real Computation,* Springer-Verlag, 1998.

BC76. Borodin, Allan and Cook, Stephen A., *"On the Number of Additions to Compute Specific Polynomials,"* SIAM J. Comput. **5** (1976), no. 1, pp. 146–157.

Ful98. Fulton, William, *Intersection Theory,* $2^{\underline{\text{nd}}}$ ed., Ergebnisse der Mathematik und ihrer Grenzgebiete 3, **2**, Springer-Verlag, 1998.

Gri82. Grigor'ev, Dima Yu., *"Lower Bounds in the Algebraic Complexity of Computations,"* The Theory of the Complexity of Computations, I; Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov (LOMI) **118** (1982), pp. 25–82, 214.

Kho80. Khovanski, Askold G., *"On a Class of Systems of Transcendental Equations,"* Dokl. Akad. Nauk SSSR **255** (1980), no. 4, pp. 804–807; English transl. in Soviet Math. Dokl. **22** (1980), no. 3.

Kho91. __________________, *Fewnomials,* AMS Press, Providence, Rhode Island, 1991.

Kob84. Koblitz, Neal I., *p-adic Numbers, p-adic Analysis, and Zeta-Functions,* $2^{\underline{\text{nd}}}$ ed., Graduate Texts in Mathematics, 58, Springer-Verlag, New York-Berlin, 1984.

Len99. Lenstra, Hendrik W., Jr., *"On the Factorization of Lacunary Polynomials,"* Number Theory in Progress, Vol. 1 (Zakopane-Kóscielisko, 1997), pp. 277–291, de Gruyter, Berlin, 1999.

Ris85. Risler, Jean-Jacques, *"Additive Complexity and Zeros of Real Polynomials,"* SIAM J. Comput. **14** (1985), no. 1, pp. 178–183.

[4] In particular, via the approach of our proofs, it is possible to improve slightly the bounds of [Gri82,Ris85] over $\mathbb{R}$. We leave this as an exercise for the interested reader.

Roj99. Rojas, J. Maurice, *"Toric Intersection Theory for Affine Root Counting,"* Journal of Pure and Applied Algebra, vol. 136, no. 1, March, 1999, pp. 67–100.

Roj00. ________________, *"Algebraic Geometry Over Four Rings and the Frontier to Tractability,"* Contemporary Mathematics, vol. 270, Proceedings of a Conference on Hilbert's Tenth Problem and Related Subjects (University of Gent, November 1–5, 1999), edited by Jan Denef, Leonard Lipschitz, Thanases Pheidas, and Jan Van Geel, pp. 275–321, AMS Press (2000).

Roj02. ________________, *"Arithmetic Multivariate Descartes' Rule,"* Math ArXiV preprint `math.NT/0110327`, submitted for publication.

Shu93. Shub, Mike, *"Some Remarks on Bézout's Theorem and Complexity Theory,"* From Topology to Computation: Proceedings of the Smalefest (Berkeley, 1990), pp. 443–455, Springer-Verlag, 1993.

Smi97. Smirnov, Andrei L., *"Torus Schemes Over a Discrete Valuation Ring,"* St. Petersburg Math. J. **8** (1997), no. 4, pp. 651–659.

Author Index