Abelian Surfaces over Finite Fields as Jacobians

Daniel Maisner and Enric Nart with an Appendix by Everett W. Howe

CONTENTS

For any finite field $k = \mathbb{F}_q$, we explicitly describe the k-isogeny classes of abelian surfaces defined over k and their behavior under finite field extension. In particular, we determine the absolutely simple abelian surfaces. Then, we analyze numerically what surfaces are k-isogenous to the Jacobian of a smooth projective curve of genus 2 defined over k. We prove some partial results suggested by these numerical data. For instance, we show that every absolutely simple abelian surface is k-isogenous to a Jacobian. Other facts suggested by these numerical computations are that the polynomials $t^4 + (1-2q)t^2 + q^2$ (for all q) and $t^4 + (2-2q)t^2 + q^2$ (for q odd) are never the characteristic polynomial of Frobenius of a Jacobian. These statements have been proved by E. Howe. The proof for the first polynomial is attached in an appendix.

1. INTRODUCTION

Let C be a projective smooth curve of genus 2 defined over a finite field $\mathbb{F}_q$. If $N_m := \sharp C(\mathbb{F}_{q^m})$ denotes the number of points of C over the m-th degree extension of $\mathbb{F}_q$, the zeta function of C can be written as:

$$Z(C/\mathbb{F}_q, t) = \exp\left(\sum_{m\geq 1} N_m \frac{t^m}{m}\right) = \frac{1 + a_1 t + a_2 t^2 + q a_1 t^3 + q^2 t^4}{(1-t)(1-qt)}, \quad (1\text{–}1)$$

for certain integers a_1, a_2, related to N_1, N_2 by:

$$N_1 = a_1 + q + 1, \qquad N_2 = 2a_2 - a_1^2 + q^2 + 1. \quad (1\text{–}2)$$

In this paper, we are interested in determining which rational functions appear as the zeta function of a curve C of genus 2, or equivalently, what pairs of integers a_1, a_2 satisfy (1–1) for a certain curve C, or equivalently, for what pairs of nonnegative integers (N_1, N_2) there exists a curve of genus 2 having N_1 points over $\mathbb{F}_q$ and N_2 points over $\mathbb{F}_{q^2}$.

To any curve C, as above, we can attach a more feasible object: its Jacobian, $J(C)$, which is an abelian surface over $\mathbb{F}_q$. The isogeny class of $J(C)$ is determined

2000 AMS Subject Classification: Primary 11G20, 14G15; Secondary 11G10

Keywords: Abelian surface, zeta function, finite field, Jacobian variety

1058-6458/2001 $0.50 per page
Experimental Mathematics **11**:3, page 321

by the characteristic polynomial of its Frobenius endomorphism, which is easily described in terms of a_1 and a_2:

$$f_{J(C)}(t) = t^4 + a_1 t^3 + a_2 t^2 + q a_1 t + q^2.$$

Thus, we can split the characterization of the zeta functions of curves of genus 2 in two steps: first characterize the pairs (a_1, a_2) arising from characteristic polynomials of abelian surfaces over $\mathbb{F}_q$ and, afterwards, determine what abelian surfaces are $\mathbb{F}_q$-isogenous to the Jacobian of a smooth projective curve defined over $\mathbb{F}_q$.

The first step is no mystery. The roots of the characteristic polynomial $f_A(t)$ of the Frobenius endomorphism of an abelian surface A are q-Weil numbers. This leads to bounds on a_1 and a_2 which determine a finite subset of $\mathbb{Z}[t]$ containing all possible polynomials of the form $f_A(t)$. Moreover, by results of Honda and Tate, the $\mathbb{F}_q$-isogeny classes of simple abelian varieties A defined over $\mathbb{F}_q$ are ruled by the q-Weil numbers, classified under the action of the absolute galois group. Combined with results of Tate [Waterhouse and Milne 69] computing the dimension of A in terms of the minimal polynomial of the corresponding q-Weil number, this makes it possible to determine explicitly all pairs (a_1, a_2) for which the polynomial $t^4 + a_1 t^3 + a_2 t^2 + q a_1 t + q^2$ is of the form $f_A(t)$ for a certain abelian surface A defined over $\mathbb{F}_q$. These conditions were described in [Rück 90] and [Xing 94]. In Section 2 we review these results, and we present them in a more explicit form (Theorem 2.9). In particular, we obtain a list of all simple abelian supersingular surfaces, that completes that of [Xing 96], where some cases are missing. Also, we include an exhaustive study of the behavior of the simple abelian surfaces under finite field extension, obtaining an explicit description of the absolutely simple varieties in terms of the pair (a_1, a_2) (Theorem 2.15).

The second step, to determine the Jacobians among all abelian surfaces, seems to be a very difficult question; we present a numerical analysis. In Section 3, we develop an algorithm computing all curves of genus 2 up to k-isomorphism and quadratic twist. The algorithm has been implemented in MATHEMATICA using the package FF designed by Guàrdia to work over finite fields of arbitrary degree over the prime field [Guàrdia 98]. As a by-product, we obtain the complete list of all curves of genus 2 without rational points (Theorem 3.2).

In Sections 4 and 5, we display, for any $q \leq 16$, the numerical results obtained by counting for each isogeny class of abelian surfaces over $\mathbb{F}_q$ how many nonisomorphic curves have a Jacobian belonging to the class. This is achieved by running the algorithm of Section 3 and by computing for each curve the corresponding pair (a_1, a_2). These numerical results present some regular behavior which has led us to prove some partial results, both in the positive and negative direction. For instance, we show that every absolutely simple abelian surface is k-isogenous to the Jacobian of a smooth projective curve of genus 2 (Theorem 4.3). The ordinary case has been proved in [Howe 95] and the nonordinary case is a consequence of the work [Howe 96] and our characterization of the absolutely simple surfaces.

Other facts suggested by our numerical computations are that the polynomials $t^4+(1-2q)t^2+q^2$ (for all q) and $t^4 + (2-2q)t^2 + q^2$ (for q odd) are never the characteristic polynomial of Frobenius of the Jacobian of a smooth projective curve of genus 2 defined over $\mathbb{F}_q$. These statements have been proved by E. Howe. The proof for the first polynomial is attached in an appendix and the proof for the second polynomial appears in [Howe 02]

2. ISOGENY CLASSES OF ABELIAN SURFACES OVER FINITE FIELDS

2.1 Characteristic Polynomials of Abelian Surfaces

Let A be an abelian surface defined over the finite field $\mathbb{F}_q$, where $q = p^a$, $a \geq 1$ for a certain prime number p. We denote by

$$f_A(t) = t^4 + a_1 t^3 + a_2 t^2 + q a_1 t + q^2 \in \mathbb{Z}[t] \qquad (2\text{–}1)$$

the characteristic polynomial of the Frobenius endomorphism of A. By abuse of language, we shall sometimes refer to $f_A(t)$ as the characteristic polynomial of A. This polynomial determines A up to $\mathbb{F}_q$-isogeny and the four roots of $f_A(t)$ in $\bar{\mathbb{Q}}$ (counting multiplicites) are q-Weil numbers; more precisely,

$$f_A(t) = (t - \pi_1)(t - \frac{q}{\pi_1})(t - \pi_2)(t - \frac{q}{\pi_2}),$$

with π_1, π_2 q-Weil numbers, not necessarily different. We recall that a q-Weil number is an algebraic integer such that its image under every complex embedding has absolute value $\sqrt{q}$. If A is simple, then $f_A(t) = h_A(t)^e$ for some irreducible polynomial $h_A(t) \in \mathbb{Z}[t]$. By results of Honda and Tate, the mapping

$$A \mapsto \pi \quad \text{root of } h_A(t),$$

is a bijection between $\mathbb{F}_q$-isogeny classes of simple abelian varieties (of any dimension) and conjugation classes of q-Weil numbers (of any degree) [Tate 69].

We review results of Rück and Xing finding necessary and sufficient conditions for a polynomial of the type (2–1) to be the characteristic polynomial of an abelian surface over $\mathbb{F}_q$. We start with well-known bounds on the size of a_1, a_2:

Lemma 2.1. *Let $f(t) \in \mathbb{Z}[t]$ be a monic polynomial of degree 4. The following conditions are equivalent:*

(i) $f(t) = (t-\pi_1)(t-\frac{q}{\pi_1})(t-\pi_2)(t-\frac{q}{\pi_2})$, *with π_1, π_2 q-Weil numbers.*

(ii) $f(t) = (t^2-\beta_1 t+q)(t^2-\beta_2 t+q)$, $\beta_i \in \mathbb{R}$, $|\beta_i| \le 2\sqrt{q}$, $i = 1, 2$.

(iii) $f(t) = t^4 + a_1 t^3 + a_2 t^2 + q a_1 t + q^2$, *with*

$$|a_1| \le 4\sqrt{q}, \quad 2|a_1|\sqrt{q} - 2q \le a_2 \le \frac{a_1^2}{4} + 2q. \quad (2\text{–}2)$$

Proof: The relationship $\beta_i = \pi_i + \frac{q}{\pi_i}$ shows immediately that (i) is equivalent to (ii) (cf. [Waterhouse and Milne 69, p. 59]). Analogously, (ii) is equivalent to (iii) since we can relate pairs β_1, β_2 satisfying (ii) with pairs a_1, a_2 satisfying (iii) by:

$$(x-\beta_1)(x-\beta_2) = x^2 + a_1 x + a_2 - 2q. \qquad \square$$

Definition 2.2. A polynomial $f(t) \in \mathbb{Z}[t]$ satisfying the conditions of Lemma 2.1 will be called a Weil polynomial.

Remark 2.3. The bound on $|a_1|$ can be refined to $a_1 \le 2[2\sqrt{q}]$, which is much better than $4\sqrt{q}$ for q nonsquare and large. In fact,

$$|a_1| > 2[2\sqrt{q}] \Longrightarrow 2|a_1|\sqrt{q} - 2q > \left[\frac{a_1^2}{4} + 2q\right],$$

so that in this case there is no integer a_2 satisfying (2–2).

It is easy to characterize when a Weil polynomial is irreducible:

Lemma 2.4. *Let $f(t) = t^4 + a_1 t^3 + a_2 t^2 + q a_1 t + q^2 \in \mathbb{Z}[t]$ be a Weil polynomial and let $\Delta = a_1^2 - 4a_2 + 8q$. Then, the following conditions are equivalent:*

(i) $f(t)$ is irreducible in $\mathbb{Z}[t]$.

(ii) Δ is not a square in $\mathbb{Z}$ and $|a_1| < 4\sqrt{q}$, $2|a_1|\sqrt{q} - 2q < a_2 < \frac{a_1^2}{4} + 2q$.

(iii) Δ is not a square in $\mathbb{Z}$ and $(a_1, a_2) \ne (0, -2q)$.

Proof: The equalities $|a_1| = 4\sqrt{q}$ or $a_2 = 2q + a_1^2/4$ lead to $\Delta = 0$, whereas the equality $2|a_1|\sqrt{q} - 2q = a_2$ leads to either $a_1 = 0, a_2 = -2q$ or to q a square and $\Delta = (|a_1| + 4\sqrt{q})^2$. Thus, (ii) and (iii) are equivalent.

With the notation of Lemma 2.1, Δ is the discriminant of $(x-\beta_1)(x-\beta_2)$; hence, if Δ is a square, then $\beta_1, \beta_2 \in \mathbb{Z}$ and $f(t)$ decomposes in $\mathbb{Z}[t]$. Thus, (i) implies (iii). Conversely, if $f(t)$ is not irreducible in $\mathbb{Z}[t]$, then either some π_i belongs to $\mathbb{Z}$, or some π_i is a quadratic integer with conjugate q/π_i, or π_1, π_2 are conjugate quadratic integers, $\pi_2 \ne q/\pi_1$. In the first two cases, some β_i belongs to $\mathbb{Z}$ and Δ is a square, whereas in the third case, π_1, π_2 are real and $f(t) = (t^2 - q)^2$. Thus, (iii) implies (i). $\square$

If A is a simple abelian surface defined over $\mathbb{F}_q$ whose characteristic polynomial decomposes in $\mathbb{Z}[t]$, then $f_A(t)$ has to be the square of a quadratic irreducible polynomial. The only real quadratic q-Weil numbers are $\pm\sqrt{q}$ (for a odd) and the corresponding simple abelian variety has dimension 2. We compute the dimension of the simple abelian variety associated with a pair of complex conjugate quadratic q-Weil numbers.

Proposition 2.5. *Let $\beta \in \mathbb{Z}$, with $|\beta| < 2\sqrt{q}$ and let $b = v_p(\beta)$ (taking $b = \infty$ if $\beta = 0$). Let $F(t) = t^2 - \beta t + q$ and let $d = \beta^2 - 4q$ be the discriminant of $F(t)$. Let B be the simple abelian variety defined over $\mathbb{F}_q$ with $h_B(t) = F(t)$. Then:*

$$dim(B) = \begin{cases} \frac{a}{(a,b)}, & \text{if } b < \frac{a}{2}, \\ 2, & \text{if } b \ge \frac{a}{2}, \quad d \in \mathbb{Q}_p^{*2}, \\ 1, & \text{if } b \ge \frac{a}{2}, \quad d \notin \mathbb{Q}_p^{*2}. \end{cases}$$

Proof: By [Waterhouse and Milne 69, pp. 58–59], we have $f_B(t) = h_B(t)^e$ and

$$\dim(B) = e = \text{least common denominator of } \frac{v_p(F_\nu(0))}{a},$$

where ν runs among the finite places of $\mathbb{Q}(\sqrt{d})$ lying above p and $F_\nu(t)$ denotes the corresponding factor of $F(t)$ in $\mathbb{Q}_p[t]$. If d is not a square in $\mathbb{Q}_p$, then $F(t)$ is irreducible in $\mathbb{Q}_p[t]$, $v_p(F(0)) = a$, and $e = 1$. If d is a square in $\mathbb{Q}_p$ then $F(t) = F_1(t)F_2(t)$ in $\mathbb{Q}_p[t]$ and denoting $b_i = v_p(F_i(0))$, an easy manipulation of Newton polygons shows that

$$b \ge \frac{a}{2} \Longrightarrow b_1 = b_2 = \frac{a}{2} \Longrightarrow e = 2,$$

$$b < \frac{a}{2} \Longrightarrow b_1 = b, \quad b_2 = a - b \Longrightarrow e = \frac{a}{(a,b)}.$$

$\square$

Corollary 2.6. *By adequate choice of q and β, we can find simple abelian varieties of arbitrarily large dimension with $h_B(t) = t^2 - \beta t + q$.*

Definition 2.7. We say that an integer $\beta \in \mathbb{Z}$, $|\beta| \leq 2\sqrt{q}$, is a q-Waterhouse number if there is an elliptic curve E defined over $\mathbb{F}_q$ such that $f_E(t) = t^2 - \beta t + q$. Equivalently, $\beta \in \mathbb{Z}$ is a q-Waterhouse number if either $|\beta| = 2\sqrt{q}$ or $|\beta| < 2\sqrt{q}$ and the simple abelian variety B associated to the polynomial $t^2 - \beta t + q$ has dimension 1.

Waterhouse found in [Waterhouse 69] very explicit conditions determining the q-Waterhouse numbers. We list below similar explicit conditions for the 2-dimensional case:

Corollary 2.8. *Let $\beta \in \mathbb{Z}$, $|\beta| < 2\sqrt{q}$. There exists a simple abelian surface B defined over $\mathbb{F}_q$ with $h_B(t) = t^2 - \beta t + q$ if and only if a is even and*

$$\beta = \pm\sqrt{q}, \quad p \equiv 1 \pmod 3, \text{ or } \quad \beta = 0, \quad p \equiv 1 \pmod 4.$$

Proof: Straightforward by Proposition 2.5. □

Now we can resume the explicit determination of the Weil polynomials corresponding to abelian surfaces defined over $\mathbb{F}_q$:

Theorem 2.9. *Let $f(t) = t^4 + a_1 t^3 + a_2 t^2 + q a_1 t + q^2 \in \mathbb{Z}[t]$ be a Weil polynomial and let*

$$\Delta = a_1^2 - 4a_2 + 8q, \quad \delta = (a_2 + 2q)^2 - 4q a_1^2.$$

Then, $f(t)$ is the characteristic polynomial of a simple abelian surface defined over $\mathbb{F}_q$ if and only if one of the following conditions holds:

(M) *Δ is not a square in $\mathbb{Z}$, $v_p(a_1) = 0$, $v_p(a_2) \geq \frac{a}{2}$ and δ is not a square in $\mathbb{Z}_p$.*

(O) *Δ is not a square in $\mathbb{Z}$ and $v_p(a_2) = 0$.*

(SS1) *(a_1, a_2) belongs to the following list:*

$(0,0)$, a odd, $p \neq 2$, or: a even, $p \not\equiv 1$ (mod 8),

$(0,q)$, a odd,

$(0,-q)$, a odd, $p \neq 3$, or: a even, $p \not\equiv 1$ (mod 12),

$(\pm\sqrt{q}, q)$, a even, $p \not\equiv 1$ (mod 5),

$(\pm\sqrt{5q}, 3q)$, a odd, $p = 5$,

$(\pm\sqrt{2q}, q)$, a odd, $p = 2$.

(SS2) *(a_1, a_2) belongs to the following list:*

$(0,-2q)$, a odd,

$(0,2q)$, a even, $p \equiv 1$ (mod 4),

$(\pm 2\sqrt{q}, 3q)$, a even, $p \equiv 1$ (mod 3).

Moreover, let β_1, β_2 be the roots of the quadratic polynomial $x^2 + a_1 x + (a_2 - 2q)$, with discriminant Δ. Then, $f(t) = f_A(t)$ for an abelian surface $A \sim E_1 \times E_2$ if and only if Δ is a square in $\mathbb{Z}$ and β_1, β_2 are q-Waterhouse numbers. In this case, the elliptic curves E_1, E_2 are $\mathbb{F}_q$-isogenous if and only if $\Delta = 0$.

Proof: By Lemma 2.4 in the cases (M), (O), (SS1), $f(t)$ is irreducible and the conditions determining when $f(t) = f_A(t)$ for some surface A were found in [Rück 90]. Actually, Rück wrote condition (SS1) as

$$v_p(a_1) \geq \frac{a}{2}, \quad v_p(a_2) \geq a, \quad f(t) \text{ has no roots in } \mathbb{Z}_p,$$

but it is easy to check that the irreducible Weil polynomials with (a_1, a_2) satisfying this last condition are precisely those listed in condition (SS1) above.

The case where $f(t)$ is reducible (SS2) is a consequence of Proposition 2.5 and it was first described in [Xing 94]. □

Corollary 2.10. *If $\mathbb{F}_q$ is the prime field $\mathbb{F}_p$ (that is $q = p$), then every Weil polynomial is the characteristic polynomial of an abelian surface defined over $\mathbb{F}_q$.*

Proof: Assume that $(a_1, a_2) \in \mathbb{Z}^2$ satisfies the inequalities (2–2). If $\Delta = a_1^2 - 4a_2 + 8q$ is a square in $\mathbb{Z}$, then the integers $\beta = (-a_1 \pm \sqrt{\Delta})/2$ satisfy: $|\beta| < 2\sqrt{p}$ and any integer satisfying this inequality is a p-Waterhouse number. If Δ is not a square in $\mathbb{Z}$ and $(a_1, a_2) \neq (0, -2q)$, then (a_1, a_2) falls in one of the cases (M), (O), (SS1). □

Corollary 2.11. *A Weil polynomial is the characteristic polynomial of a simple supersingular abelian surface defined over $\mathbb{F}_q$ if and only if it appears in the list (SS1) or (SS2).*

Proof: The supersingular condition is equivalent to $v_p(a_1) \geq a/2$, $v_p(a_2) \geq a$. In the list of simple abelian surfaces given in Theorem 2.9, only those of (SS1) and (SS2) satisfy this condition. □

This result completes the list given in [Xing 96], where some cases are missing.

(a_1, a_2)	L
(0,0), (a odd, $p \neq 2$) or (a even, $p \not\equiv 1 \pmod 8$), $p \not\equiv 1 \pmod 4$	$\mathbb{F}_{q^2}$
(0,0), (a odd, $p \neq 2$) or (a even, $p \not\equiv 1 \pmod 8$), $p \equiv 1 \pmod 4$	$\mathbb{F}_{q^4}$
$(0, q)$, a odd, $p \not\equiv 1 \pmod 3$	$\mathbb{F}_{q^2}$
$(0, q)$, a odd, $p \equiv 1 \pmod 3$	$\mathbb{F}_{q^6}$
$(0, -q)$, (a odd, $p \neq 3$) or (a even, $p \not\equiv 1 \pmod{12}$), $p \not\equiv 1 \pmod 3$	$\mathbb{F}_{q^2}$
$(0, -q)$, (a odd, $p \neq 3$) or (a even, $p \not\equiv 1 \pmod{12}$), $p \equiv 1 \pmod 3$	$\mathbb{F}_{q^3}$
$(\pm\sqrt{q}, q)$, a even, $p \not\equiv 1 \pmod 5$	$\mathbb{F}_{q^5}$
$(\pm\sqrt{5q}, 3q)$, a odd, $p = 5$	$\mathbb{F}_{q^5}$
$(\pm\sqrt{2q}, q)$, a odd, $p = 2$	$\mathbb{F}_{q^4}$
$(0, -2q)$, a odd	$\mathbb{F}_{q^2}$
$(0, 2q)$, a even, $p \equiv 1 \pmod 4$	$\mathbb{F}_{q^2}$
$(\pm 2\sqrt{q}, 3q)$, a even, $p \equiv 1 \pmod 3$	$\mathbb{F}_{q^3}$

TABLE 1. The minimum field L of decomposition of the supersingular surfaces.

2.2 Absolutely Simple Abelian Surfaces

We now characterize in terms of the pair (a_1, a_2) when an abelian surface A defined over $\mathbb{F}_q$ is absolutely simple. By abuse of language, we denote simply by $A = (a_1, a_2)$ the (isogeny class of an) abelian surface determined by a pair (a_1, a_2) satisfying the conditions of Theorem 2.9.

We have classified the simple abelian surfaces in three groups: (M) for mixed, (O) for ordinary and (SS1), (SS2) for supersingular. They can be distinguished by the Newton polygon of their characteristic polynomial, which has 3, 2, 1 sides, respectively. The number of sides of the Newton polygon is invariant under scalar extension; thus, attending to the particular shape of the polygon, we see that after scalar extension, a simple surface of type (M), (O), (SS) either remains simple of the same type or decomposes as the product of two elliptic curves, which are, respectively, ordinary×supersingular, ordinary×ordinary, and supersingular×supersingular. Actually, we shall prove that all simple surfaces of type (M) are absolutely simple.

Since the invariant Δ can be a square in $\mathbb{Z}$, or the characteristic polynomial can be reducible only for supersingular simple surfaces, we have

Lemma 2.12. *Let A be a nonsupersingular simple abelian surface defined over $\mathbb{F}_q$. The following conditions are equivalent:*

(i) A remains simple over $\mathbb{F}_{q^n}$.

(ii) The invariant $\Delta(\mathbb{F}_{q^n})$ is not a square in $\mathbb{Z}$.

(iii) The characteristic polynomial $f_{A_{|\mathbb{F}_{q^n}}}(t)$ is irreducible.

The proof of the following observation is straightforward.

Lemma 2.13. *Let $A = (a_1, a_2)$ be an abelian surface defined over $\mathbb{F}_q$ and let $A_{|\mathbb{F}_{q^2}} = (b_1, b_2)$, $A_{|\mathbb{F}_{q^3}} = (c_1, c_2)$. Then*

$$b_1 = 2a_2 - a_1^2, \quad b_2 = a_2^2 - 2qa_1^2 + 2q^2;$$

$$c_1 = a_1(a_1^2 - 3a_2 + 3q), \quad c_2 = a_2^3 + 6q^2a_1^2 - 3q^2a_2 - 3qa_1^2a_2.$$

Moreover,

$$\Delta(\mathbb{F}_{q^2}) = a_1^2\Delta, \quad \Delta(\mathbb{F}_{q^3}) = (q - a_1^2 + a_2)^2\Delta.$$

We can tell the minimum field L of decomposition of the supersingular surfaces just by checking Lemma 2.13 and Theorem 2.9. (See Table 1.)

In the nonsupersingular case, it is easy to analyze, using Lemmas 2.12 and 2.13, the decomposition in $\mathbb{F}_{q^n}$, for $n = 2, 3, 4, 6$:

Proposition 2.14. *Let $A = (a_1, a_2)$ be a simple abelian surface defined over $\mathbb{F}_q$, which is not supersingular. Then*

(i) A decomposes over $\mathbb{F}_{q^2}$ iff $a_1 = 0$.

(ii) A decomposes over $\mathbb{F}_{q^3}$ iff $q = a_1^2 - a_2$.

(iii) A is simple over $\mathbb{F}_{q^2}$ and decomposes over $\mathbb{F}_{q^4}$ iff $a_1^2 = 2a_2$.

(iv) A is simple over $\mathbb{F}_{q^2}$ and $\mathbb{F}_{q^3}$ but decomposes over $\mathbb{F}_{q^6}$ iff $a_1^2 = 3(a_2 - q)$.

(v) If A is simple over $\mathbb{F}_{q^4}$ then it is simple over $\mathbb{F}_{q^8}$.

(vi) If A is simple over $\mathbb{F}_{q^4}$ and $\mathbb{F}_{q^6}$ then it is simple over $\mathbb{F}_{q^{12}}$.

Proof: Suppose that A decomposes over $\mathbb{F}_{q^n}$ and $n = 2$ or 3. Then $\Delta(\mathbb{F}_{q^n})$ is a square in $\mathbb{Z}$, but since $\Delta(\mathbb{F}_{q^2}) = a_1^2\Delta$ (respectively, $\Delta(\mathbb{F}_{q^3}) = (q - a_1^2 + a_2)^2\Delta$) and Δ is not a square in $\mathbb{Z}$, this implies $a_1 = 0$ (respectively, $q - a_1^2 + a_2 = 0$). Conversely, if $a_1 = 0$ (respectively, $q - a_1^2 + a_2 = 0$), then $\Delta(\mathbb{F}_{q^n}) = 0$ and A decomposes over $\mathbb{F}_{q^2}$ by Lemma 2.12. This proves (i) and (ii).

By (i), A decomposes over $\mathbb{F}_{q^4}$ and not before if and only if $a_1 \neq 0$ and $b_1 = 0$. This is equivalent to $b_1 = 0$ since the condition $a_1 = 0 = b_1$ is satisfied only by the supersingular surface $A = (0, 0)$. This proves (iii).

By (i) and (ii), A decomposes over $\mathbb{F}_{q^6}$ and not before if and only if $c_1 = 0$, $a_1 \neq 0$, $q \neq a_1^2 - a_2$. The two first conditions are equivalent to $a_1^2 - 3a_2 + 3q = 0$ and this latter condition already implies that $q \neq a_1^2 - a_2$. In fact, $a_1^2 - 3a_2 + 3q = 0 = q - a_1^2 + a_2$ leads to $(a_1, a_2) = (\sqrt{3q}, 2q)$ which is either impossible or satisfied only by a supersingular surface. This proves (iv).

Suppose that A is simple over $\mathbb{F}_{q^4}$ and decomposes over $\mathbb{F}_{q^8}$. By (iii), applied to the surface $A_{|\mathbb{F}_{q^2}}$, we have $b_1^2 = 2b_2$, but this equation is impossible. In fact, it leads to

$$a_1^4 + 2a_2^2 - 4a_2a_1^2 + 4qa_1^2 - 4q^2 = 0.$$

This relation implies a_1, a_2 both even and $4q^2 \equiv 0 \pmod 8$, which is possible only for $p = 2$. But then, A would be supersingular. This proves (v).

Suppose that A is simple over $\mathbb{F}_{q^4}$ and $\mathbb{F}_{q^6}$, but it decomposes over $\mathbb{F}_{q^{12}}$. By (iv), applied to the surface $A_{|\mathbb{F}_{q^2}}$, we have $b_1^2 = 3(b_2 - q^2)$, which is impossible. In fact, it leads to

$$a_1^4 + (6q - 4a_2)a_1^2 + a_2^2 - 3q^2.$$

The discriminant of this quadratic equation in a_1^2 is $12(a_2 - 2q)^2$, which is a square in $\mathbb{Z}$ only if $a_2 = 2q$; but then $a_1^2 = q$ and (for a even) A would be supersingular. This proves (vi). □

Actually, Proposition 2.14 collects all possible cases in which a non-supersingular simple abelian surface is not absolutely simple.

Theorem 2.15. *Let $f(t) = t^4 + a_1t^3 + a_2t^2 + qa_1t + q^2 \in \mathbb{Z}[t]$ be a Weil polynomial and let*

$$\Delta = a_1^2 - 4a_2 + 8q, \quad \delta = (a_2 + 2q)^2 - 4qa_1^2.$$

Then there exists an absolutely simple abelian surface A defined over $\mathbb{F}_q$ with $f(t) = f_A(t)$ if and only if Δ is not a square in $\mathbb{Z}$ and either

(a) $v_p(a_1) = 0$, $v_p(a_2) \geq a/2$, δ is not a square in $\mathbb{Z}_p$, or

(b) $v_p(a_2) = 0$, $a_1^2 \notin \{0, q + a_2, 2a_2, 3(a_2 - q)\}$.

Proof: We have already checked that all surfaces other than those listed above are not absolutely simple. We prove now that if $A = (a_1, a_2)$ is a nonsupersingular simple abelian surface which is not absolutely simple, then $a_1^2 \in \{0, q + a_2, 2a_2, 3(a_2 - q)\}$. For such a surface, the characteristic polynomial $f_A(t)$ is irreducible. Let π be one of its roots in $\bar{\mathbb{Q}}$ and let $K = \mathbb{Q}(\pi)$ be the quartic field generated by π. The quadratic algebraic integer $\beta = \pi + q/\pi$ belongs to K, hence, the discriminant Δ of its minimal polynomial over $\mathbb{Q}$ is a square in K, so that K contains $\mathbb{Q}(\sqrt{\Delta})$ as a quadratic subfield.

By Lemma 2.12, A decomposes over $\mathbb{F}_{q^n}$ if and only if the characteristic polynomial of $A_{|\mathbb{F}_{q^n}}$ reduces and this is equivalent to $\mathbb{Q}(\pi^n) \subsetneq K$. Take n minimum with this property and let L be a quadratic subfield of K containing π^n. If $\mathrm{Gal}(K/L) = \{1, \sigma\}$, we have

$$\pi^n \in L \Longleftrightarrow (\pi^n)^\sigma = \pi^n \Longleftrightarrow \pi^\sigma = \epsilon\pi,$$

where $\epsilon \in K$ is a primitive n-th root of 1, by the minimality of n. Thus, n belongs to the set $\{2, 3, 4, 5, 6, 8, 10, 12\}$. By Proposition 2.14, the cases $n = 8, 12$ are not possible and in the cases $n = 2, 3, 4, 6$, we have $a_1^2 \in \{0, q + a_2, 2a_2, 3(a_2 - q)\}$.

Finally, assume that $n = 5$ or 10. Then $K = \mathbb{Q}(\mu_5)$ is a cyclic extension of $\mathbb{Q}$ whose only quadratic subfield is $\mathbb{Q}(\sqrt{5})$. In this case, $\pi^n = \pm\sqrt{q^n}$, since π^n is a real Weil number. Thus, A would be supersingular. □

The ordinary case (b) has already been settled in [Howe and Zhu 02].

Corollary 2.16. *The minimum positive integer n for which a not absolutely simple abelian surface over $\mathbb{F}_q$ decomposes over $\mathbb{F}_{q^n}$ belongs to $\{1, 2, 3, 4, 5, 6\}$.*

Corollary 2.17. *If an abelian surface A defined over $\mathbb{F}_q$ decomposes over $\overline{\mathbb{F}}_q$ as the product of two elliptic curves, one supersingular, the other ordinary, then A decomposes already over $\mathbb{F}_q$.*

3. CURVES OF GENUS 2 OVER FINITE FIELDS

3.1 Generalities on Curves of Genus 2

Let k be a perfect field. Any smooth projective curve C defined over k of genus 2 is hyperelliptic; that is, it admits

a k-morphism, $x\colon C \longrightarrow \mathbb{P}^1$, of degree 2. In particular, the function field $k(C)$ is a separable quadratic extension of $k(\mathbb{P}^1)$. The k-automorphism, $\iota\colon C \longrightarrow C$, corresponding to the nontrivial element of $\mathrm{Gal}(k(C)/k(\mathbb{P}^1))$ is called the *hyperelliptic involution* of C. Any two k-morphisms of degree 2 from C to $\mathbb{P}^1$ differ by a k-automorphism of $\mathbb{P}^1$. Thus, the hyperelliptic involution does not depend on the particular choice of the morphism x. The fixed points of ι are the ramification points of any such morphism and they coincide also with the Weierstrass points of C. By the Hurwitz genus formula, the different of $k(C)/k(\mathbb{P}^1)$ is a divisor D of degree 6. If $\mathrm{char}(k) \neq 2$, D consists of six different points, but if $\mathrm{char}(k) = 2$, there are three different possibilities for the structure of this divisor [Igusa 60],[Lachaud 91]:

(a) $D = 5P_\infty$,
(b) $D = 3P_\infty + P_0$,
(c) $D = P_\infty + P_0 + P_1$.

Since the divisor D is defined over k, the points P_∞ and P_0 in cases (a) and (b) are defined over k too. However, in case (c), we have three possibilities:

(c1) P_∞, P_0, P_1 defined over k,

(c2) P_∞ defined over k and P_0, P_1 conjugated over a quadratic extension,

(c3) P_∞, P_0, P_1 conjugated over a cubic extension.

Clearly, the type of divisor and the structure of the support of D as a galois set are invariant by k-isomorphism; thus, the set $\mathcal{H}$ of k-isomorphy classes of smooth projective curves of genus 2 is the disjoint union of 5 subsets:

$$\mathcal{H} = \mathcal{H}_a \cup \mathcal{H}_b \cup \mathcal{H}_{c1} \cup \mathcal{H}_{c2} \cup \mathcal{H}_{c3}.$$

If $\mathrm{char}(k) \neq 2$, there are 11 possibilities for the structure of the support of D as a galois set, one for each partition of 6. We have a similar decomposition of $\mathcal{H}$ as the disjoint union of 11 subsets:

$$\begin{aligned}\mathcal{H} = \mathcal{H}_6 \cup \mathcal{H}_{5,1} \cup \mathcal{H}_{4,2} \cup \mathcal{H}_{4,1,1} \cup \mathcal{H}_{3,3} \cup \mathcal{H}_{3,2,1} \cup \mathcal{H}_{3,1,1,1} \\ \cup \mathcal{H}_{2,2,2} \cup \mathcal{H}_{2,2,1,1} \cup \mathcal{H}_{2,1,1,1,1} \cup \mathcal{H}_{1,1,1,1,1,1},\end{aligned}$$

where, for instance, $\mathcal{H}_{4,1,1}$ denotes the set of classes of curves in $\mathcal{H}$ having two Weierstrass points defined over k and four Weierstrass points defined over a quartic extension of k and forming a complete orbit under the action of $\mathrm{Gal}(\bar{k}/k)$.

We choose a point $\infty \in \mathbb{P}^1(k)$, and we call it infinity. This choice determines an embedding $\mathbb{A}^1 \subseteq \mathbb{P}^1$ and identifications $k(\mathbb{P}^1) = k(x)$, $\mathrm{Aut}(\mathbb{P}^1) = \mathrm{PGL}_2$. The function field $k(C)$, as a quadratic extension of $k(x)$, admits a generator $y \in k(C)$ satisfying:

$$\begin{aligned} y^2 &= f(x) \quad (\text{if } \mathrm{char}(k) \neq 2), \\ y^2 + y &= f(x) \quad (\text{if } \mathrm{char}(k) = 2), \end{aligned} \tag{3–1}$$

for some rational function $f(x) \in k(x)$. This equation for the function field of C is unique up to two actions: x can be replaced by any automorphism $\gamma(x) \in \mathrm{PGL}_2(k)$ and $f(x)$ can be replaced, respectively, by

$$\begin{aligned} &f(x)g(x)^2 && (\text{if } \mathrm{char}(k) \neq 2), \\ &f(x) + g(x) + g(x)^2 && (\text{if } \mathrm{char}(k) = 2), \end{aligned}$$

where $g(x) \in k(x)$ is an arbitrary rational function, $g(x) \neq 0$ in the odd characteristic case. Accordingly, one is able to exhibit a family of plane affine models containing all k-birational classes of curves of genus 2.

If $\mathrm{char}(k) \neq 2$, any projective smooth curve of genus 2 is k-isomorphic to the normalization of the projective closure of the plane affine curve C_0 defined by the equation $y^2 = f(x)$, where $f(x) = a_n x^n + \cdots + a_0 \in k[x]$ is a separable polynomial of degree 5 or 6. The curve C_0 is smooth and its closure $\tilde{C}$ in $\mathbb{P}^2$ has only one point at infinity, P_∞, which is a singular point. If $n = 5$, the point P_∞ has only one preimage in the normalization $C \to \tilde{C}$, which we still denote by P_∞; this point is a Weierstrass point and it is always defined over k. If $n = 6$, the point P_∞ has two preimages in C, which we denote by P_{∞_1}, P_{∞_2}; these points are permuted by ι and they are defined over k if and only if a_n is a square in k^*. Since the rest of the points of C are in bijection with the points in C_0, it is common to attach to these points of C the affine coordinates (x, y) of the corresponding points in C_0. In affine coordinates, the hyperelliptic involution is expressed by $\iota(x, y) = (x, -y)$.

If $\mathrm{char}(k) = 2$, any projective smooth curve of genus 2 is k-isomorphic to the normalization of the projective closure of the plane affine curve C_0 defined (after removal of denominators) by an equation:

(a) $y^2 + y = ax^5 + bx^3 + cx^2 + d, \qquad a \neq 0,$

(b) $y^2 + y = ax^3 + bx + \frac{c}{x} + d, \qquad ac \neq 0,$

(c1) $y^2 + y = ax + \frac{b}{x} + \frac{c}{x+1} + d, \qquad abc \neq 0,$

(c2) $y^2 + y = ax + \frac{bx+c}{Q(x)} + d, \qquad a \neq 0,\ (b, c) \neq (0, 0),$

(c3) $y^2 + y = \frac{ax^2+bx+c}{P(x)} + d, \qquad (a, b, c) \neq (0, 0, 0),$

where $Q(x)$, $P(x)$ are irreducible polynomials of respective degree 2,3. As before, one attaches to the points of C the affine coordinates of the corresponding affine model.

The set of "points at infinity" of C coincides with the set W of Weierstrass points, except for the model (c3), in which case

$$C(\bar{k}) \setminus C_0(\bar{k}) = W \cup \{P_{\infty_1}, P_{\infty_2}\},$$

with P_{∞_1}, P_{∞_2} permuted by ι; these two points are defined over k if and only if d belongs to the Artin-Schreier group: $AS(k) := \{\lambda + \lambda^2 \,|\, \lambda \in k\}$. In affine coordinates, the hyperelliptic involution is expressed by: $\iota(x, y) = (x, y + 1)$.

3.2 Quadratic Twist

The quadratic extensions of k are parameterized by $k^*/(k^*)^2$ if $\mathrm{char}(k) \neq 2$ (Kummer theory) and by $k/AS(k)$ if $\mathrm{char}(k) = 2$ (Artin-Schreier theory). If a smooth projective curve C of genus 2 is given by Equation (3–1), we define the twisted curve by an element $\lambda \in k^*/(k^*)^2$, respectively, $\lambda \in k/AS(k)$, as the curve C^λ determined by the equation

$$y^2 = \lambda f(x), \qquad \text{respectively,} \quad y^2 + y = f(x) + \lambda.$$

The curves C and C^λ are isomorphic over the quadratic extension of k determined by λ, but they are not necessarily k-isomorphic. This induces a well-defined action of $k^*/(k^*)^2$, respectively, $k/AS(k)$, on $\mathcal{H}$ and we denote by $\mathcal{H}^t$ the quotient set of classes of curves of genus 2 up to k-isomorphism and quadratic twist. The galois structure of the set of Weierstrass points is preserved by quadratic twist and we obtain an analogous decomposition for the set $\mathcal{H}^t$ as the disjoint union of 11, respectively, 5 subsets.

If $k = \mathbb{F}_q$ is a finite field, we have $k^*/(k^*)^2 \simeq \mathbb{Z}/2\mathbb{Z}$, respectively, $k/AS(k) \simeq \mathbb{Z}/2\mathbb{Z}$, according to the parity of q. Actually, if q is even, we have an exact sequence of additive groups

$$0 \longrightarrow \mathbb{F}_2 \longrightarrow \mathbb{F}_q \xrightarrow{AS} \mathbb{F}_q \xrightarrow{Tr} \mathbb{F}_2 \longrightarrow 0,$$

where $AS(\lambda) = \lambda + \lambda^2$. Thus, the subgroup $AS(\mathbb{F}_q)$ coincides with the set of elements of absolute trace zero.

Let $N_m(C) = \#C(\mathbb{F}_{q^m})$ be the number of rational points of C over the unique extension of degree m of k. If we denote by C' the nontrivial quadratic twist of C, we have

$$N_1(C) + N_1(C') = 2q + 2, \quad N_2(C) = N_2(C'),$$

or equivalently,

$$a_1 + a_1' = 0, \quad a_2 = a_2', \tag{3–2}$$

where a_1, a_2 and a_1', a_2' are the coefficients of the numerator of the zeta function, respectively, of C and C' (see, for example (1–2)).

3.3 Generating Curves of Genus 2 up to *k*-isomorphism and Quadratic Twist

If $\mathrm{char}(k) \neq 2$, the moduli functor of curves of genus 2 is the variety

$$\mathcal{M} = \begin{pmatrix} \mathbb{P}^1 \\ 6 \end{pmatrix} \backslash \mathrm{PGL}_2.$$

The set of k-points of this functor parameterizes smooth projective curves of genus 2 up to k-isomorphism and quadratic twist. More precisely, if we denote by

$$\mathbb{X} := \begin{pmatrix} \mathbb{P}^1(\bar{k}) \\ 6 \end{pmatrix}^{\mathrm{Gal}(\bar{k}/k)},$$

the set of families of six different points of $\mathbb{P}^1(\bar{k})$ which are invariant (as a family) under the galois action, we can consider the map

$$w : \mathcal{H}^t \longrightarrow \mathcal{M}(k) = \mathbb{X} \backslash \mathrm{PGL}_2(k), \tag{3–3}$$

which assigns to any curve C the set $\{x(P_1), \cdots, x(P_6)\}$ of images of the Weierstrass points $P_1, \cdots, P_6$ of C under any k-morphism, $x : C \longrightarrow \mathbb{P}^1$, of degree 2. This map w is well-defined and bijective. The inverse map sends $\{x_1, \cdots, x_6\}$ to the curve C defined by the equation

$$y^2 = \prod_{x_i \neq \infty} (x - x_i).$$

In exactly the same way as $\mathcal{H}$ and $\mathcal{H}^t$, the sets $\mathbb{X}$ and $\mathbb{X} \backslash \mathrm{PGL}_2(k)$ split as the union of 11 different subsets according to the galois structure of the sextuples of points. Clearly, the map w of (3–3) respects this decomposition. In fact, in an affine model, the Weierstrass points have coordinates $(x, 0)$ and the possible Weierstrass point at infinity P_∞ with image $x(P_\infty) = \infty$ is always defined over k.

In order to describe $\mathcal{H}^t$ when $\mathrm{char}(k) = 2$, we don't use the moduli space of curves of genus 2 (described in [Igusa 60]). Instead, we find for each of the cases (a), (b), (c1), (c2), and (c3) explicit conditions on the coefficients a, b, c, d of the equations, determining when two curves of the same type are k-isomorphic. Curves of type (a) are precisely those whose Jacobian is supersingular; this case has been thoroughly studied in [van der Geer and van der Vlugt 92]. For details concerning the other cases, see [Cardona et al. 02].

We have developed two independent MATHEMATICA subroutines that find unique representatives of the set $\mathcal{H}^t$ when $k = \mathbb{F}_q$ is the finite field with q elements. For q odd, the subroutine Gen2 finds representatives of the set $\mathbb{X}$ under the action of $\mathrm{PGL}_2(k)$. For the sake

Equation	N_1	N_2	a_1	a_2
type (a)				
$y^2+y=x^5$	3	5	0	0
$y^2+y=x^5+x^2$	5	9	2	4
$y^2+y=x^5+x^3$	5	5	2	2
$y^2+y=x^5+x^3+x^2$	3	9	0	2
type (b)				
$y^2+y=x^3+\frac{1}{x}$	4	4	1	0
$y^2+y=x^3+x+\frac{1}{x}$	2	8	−1	2
type (c1)				
$y^2+y=x+\frac{1}{x}+\frac{1}{x+1}$	3	3	0	−1
type (c2)				
$y^2+y=x+1/(x^2+x+1)$	3	7	0	1
$y^2+y=x+x/(x^2+x+1)$	5	7	2	3
type (c3)				
$y^2+y=1/(x^3+x+1)$	2	6	−1	1
$y^2+y=x/(x^3+x+1)$	4	10	1	3
$y^2+y=(x^2+x)/(x^3+x+1)$	6	6	3	5

TABLE 2. $q=2$.

of efficiency, we split the search of these representatives into 11 different cases, since for each different structure of the galois set, we use different procedures to lower the complexity of the search. For q even, the subroutine Gen2Ch2 works directly with the five types of generating equations, restricted always to the case $d=0$. We remark that since two triples of points of $\mathbb{P}^1$ with the same galois structure are in the same orbit under the action of $\mathrm{PGL}_2(\mathbb{F}_q)$, the quadratic and cubic irreducible polynomials $Q(x)$, $P(x)$ of cases (c2) and (c3) can be fixed a priori. These subroutines, as well as the package FF, can be downloaded at www.mat.uab.es/danielm.

Our programs also compute for each curve the numbers N_1, N_2 of points of the curve over the fields $\mathbb{F}_q, \mathbb{F}_{q^2}$ and the relevant coefficients a_1, a_2 of the numerator of the zeta function. For instance, we list in Table 2 the output for $q=2$ and in Table 3 the output for $q=3$.

Remark 3.1. For q odd, explicit formulas for $\sharp\mathcal{H}^t$ as a polynomial in q can be found in [López et al. 02]. By similar methods, we found formulas for the cardinality of each of the 11 subsets $\mathcal{H}^t_P$, where P is a partition of 6.

For q even, in [van der Geer and van der Vlugt 92] the authors find explicit formulas for $\sharp\mathcal{H}_a$ and, even more, for the number of curves in $\mathcal{H}_a$ with prescribed number N_1 of k-points. Formulas for $\sharp\mathcal{H}_b$ and $\sharp\mathcal{H}_{ci}$, $i=1,2,3$

Equation	N_1	N_2	a_1	a_2
21111				
$y^2=(1+x^2)x(1+x)(-1+x)$	4	6	0	−2
2211				
$y^2=(1+x^2)x(-1-x+x^2)$	6	10	2	2
$y^2=(1+x^2)(1+x)(-1+x+x^2)$	4	14	0	2
222				
$y^2=(1+x^2)(-1+x+x^2)(-1-x+x^2)$	8	14	4	10
42				
$y^2=(1+x^2)(-1-x^2+x^4)$	6	18	2	6
$y^2=(1+x^2)(1-x+x^2+x^4)$	6	14	2	4
$y^2=(1+x^2)(-1-x+x^4)$	4	14	0	2
$y^2=(1+x^2)(1-x+x^3+x^4)$	8	10	4	8
411				
$y^2=x(-1+x-x^2-x^3+x^4)$	4	10	0	0
$y^2=x(1+x+x^2+x^4)$	6	10	2	2
$y^2=x(1+x-x^3+x^4)$	4	6	0	−2
$y^2=x(-1-x^2+x^4)$	4	18	0	4
$y^2=x(-1+x+x^4)$	6	14	2	4
$y^2=x(1-x+x^2-x^3+x^4)$	6	18	2	6
33				
$y^2=(-1-x+x^3)(1-x+x^3)$	2	20	−2	7
$y^2=(-1-x+x^3)(1-x^2+x^3)$	4	12	0	1
$y^2=(-1-x+x^3)(-1+x^2+x^3)$	6	12	2	3
3111				
$y^2=x(x-1)(1+x-x^2+x^3)$	3	5	−1	−2
$y^2=x(x-1)(-1+x^2+x^3)$	5	13	1	2
321				
$y^2=(1+x^2)(1+x-x^2+x^3)$	5	13	1	2
$y^2=(1+x^2)(-1+x^2+x^3)$	3	9	−1	0
$y^2=(1+x^2)(-1-x-x^2+x^3)$	1	13	−3	6
$y^2=(1+x^2)(1-x+x^3)$	3	17	−1	4
6				
$y^2=1+x^2-x^4+x^6$	4	20	0	5
$y^2=1-x^2+x^6$	8	12	4	9
$y^2=-1+x+x^3+x^4+x^5+x^6$	6	16	2	5
$y^2=-1+x^5+x^6$	4	16	0	3
$y^2=-1-x^3-x^4+x^5+x^6$	2	12	−2	3
$y^2=-1+x+x^5+x^6$	4	8	0	−1
$y^2=1-x+x^2-x^3+x^5+x^6$	6	8	2	1
51				
$y^2=-1+x-x^2-x^4+x^5$	3	15	−1	3
$y^2=1-x+x^5$	7	15	3	7
$y^2=-1+x+x^3+x^5$	1	11	−3	5
$y^2=-1-x^3-x^4+x^5$	5	15	1	3
$y^2=1+x-x^2-x^3+x^5$	5	19	1	5
$y^2=-1-x-x^2+x^3-x^4+x^5$	3	7	−1	−1
$y^2=-1-x+x^2+x^3+x^5$	3	11	−1	1
$y^2=-1-x-x^4+x^5$	5	11	1	1

TABLE 3. $q=3$.

$q=2$	$y^2+y=1+(x^2+x)/(x^3+x+1)$
$q=3$	$y^2=-(x^2+1)(x^2+x-1)(x^2-x-1)$ $y^2=-(x^2+1)(x^4+x^3-x+1)$ $y^2=-x^6+x^2-1$
$q=4$	$y^2+y=s+x/(x^3+x+1),\ \ s^2=s+1$
$q=5$	$y^2=(2x^3+4x-2)(x^3-2x^2-1)$ $y^2=2x^6-2x^5+2x^4+x^3-x^2-2x+2$ $y^2=(2x^2+1)(x^4-2x^3+x^2-2x-2)$
$q=7$	$y^2=(-x^2+3)(x^2+1)(x^2+2)$ $y^2=-x^6+2x^4-3x^2-2$
$q=8$	$y^2+y=u+((u+u^2)+ux+ux^2)/(x^3+ux+u),\ \ u^3=u^2+1$
$q=9$	$y^2=s(x^3-x+1)(x^3-x-1),\ s^2=-1$
$q=11$	$y^2=(-x^2+2)(x^4-5x^3+x^2+x+4)$

TABLE 4.

can be found in [Cardona et al. 02]. Our numerical computations agree with all these results.

As a by-product of our search, we obtain the complete list of curves of genus 2 without rational points. By Weil's bound, any curve of genus 2 over $\mathbb{F}_q$ has rational points if $q>13$. By searching all curves for $q\leq 13$, we obtain

Theorem 3.2. *Any smooth projective curve C of genus 2 defined over a finite field $\mathbb{F}_q$, such that $C(\mathbb{F}_q)=\emptyset$, is $\mathbb{F}_q$-isomorphic to one of the curves listed in Table 4.*

The fact that $C(\mathbb{F}_{13})\neq\emptyset$ for all curves defined over $\mathbb{F}_{13}$ has already been observed by Stark [Stark 72].

4. ABELIAN SURFACES AS JACOBIANS

We are far from having a complete answer to the question of which isogeny classes of abelian surfaces contain a Jacobian. There is abundant literature about existence and nonexistence results for decomposable surfaces, with significant contributions by Serre, Hayashida-Nishi, Rück, Frey-Kani, Kani, Ibukiyama-Katsura, and Oort among others, although in some cases, the adaptation of the arguments to the finite field case is still to be done.

For simple surfaces, the situation is much clearer, principally because of a well-known result of Weil. Actually, we need a generalization of the classical result of [Weil 57], which can be easily deduced from the arguments of Section 5.10 of [Adleman and Huang 92]:

Theorem 4.1. (Weil, Adleman-Huang.) *Let A be a principally polarized abelian surface defined over a finite field k. If A is simple over the quadratic extension of k, then A is k-isomorphic to the Jacobian of a projective smooth curve of genus 2.*

Using this result, if A is simple over the quadratic extension of k, then the isogeny class of A contains a Jacobian if and only if it contains a principally polarized surface. This latter question has been completely solved in the ordinary case in [Howe 95]. Moreover, in Theorem 4.3 below, we use a criterion of Howe to prove that any simple surface of the family (M) of Theorem 2.9 is isogenous to a principally polarized surface. Thus, it remains to solve the question only for a scattered family of supersingular simple surfaces and for the simple sur-

faces with $a_1 = 0$, which, by Proposition 2.14, are the only nonsupersingular surfaces that decompose over the quadratic extension of k.

In any case, it is quite simple to carry out a computational exploration of the problem. We have written two programs Jac2,Jac2Ch2, which for a given (odd, respectively, even) prime power q display all isogeny classes of abelian surfaces A over $\mathbb{F}_q$, determine the decomposition type of A and count the number of projective smooth curves of genus 2 for which the Jacobian is isogenous to A.

As a first step, the program considers the pairs of integers (a_1, a_2) parameterizing all Weil polynomials (Lemma 2.1) and for each of them, it checks the conditions of Theorems 2.9 and 2.15 labeling each polynomial with one of the following symbols:

- **x** there exists no abelian surface corresponding to this pair (a_1, a_2)
- **a** absolutely simple
- **o** ordinary, simple, not absolutely simple
- **s** simple, supersingular
- **d** decomposes as $E_1 \times E_2$, with E_1, E_2 not $\mathbb{F}_q$-isogenous
- **e** decomposes as $E \times E$

This information is kept in the form of a matrix indexed by the values of (a_1, a_2), with the above symbols as entries. Once this matrix is obtained (with an insignificant expenditure of time), it is written as a first output of the program. Afterwards, the programs Gen2,Gen2Ch2 search for all curves of genus 2 over $\mathbb{F}_q$ and for each curve, they compute the pair (a_1, a_2) of relevant coefficients of the characteristic polynomial of its Jacobian and then add one to the entry $(|a_1|, a_2)$ of the matrix. Then, the programs produce as a second output the same matrix with the changes produced by counting the Jacobians. These subroutines can be downloaded at www.mat.uab.es/danielm.

For instance, for $q = 2, 3$, the two outputs of Jac2Ch2,Jac2 are given in Tables 5 and 6.

In the display of the matrix, the rows are indexed by increasing values of a_1, starting with $a_1 = 0$, whereas the columns are indexed by the values of a_2 within the bounds

$$2|a_1|\sqrt{q} - 2q \le a_2 \le \frac{a_1^2}{4} + 2q,$$

given by Lemma 2.1. To accommodate the reader we write the minimum value of a_2 corresponding to the first entry in the row at the beginning of each row.

a_1	min.a_2		
0	-4	sosodosde	**s o s** 1 1 1 **d e**
1	-1	oaadad	**o** 1 1 1 **d**
2	2	sade	1 1 **e**
3	5	od	1 **d**
4	8	e	**e**

TABLE 5. $q = 2$.

a_1	min.a_2		
0	-6	soodoosodsode	**s o o d** 2 1 1 1 2 1 1 1 **e**
1	-2	oadaaadad	1 1 1 2 2 2 1 1 **d**
2	1	oodaade	1 2 2 2 1 2 1
3	5	adad	1 1 1 **d**
4	8	ode	1 1 1
5	12	d	**d**
6	15	e	**e**

TABLE 6. $q = 3$.

Finally, the matrix has entries only with $a_1 \ge 0$ and the program takes into account only one curve for each pair C, C' of twisted curves. This is harmless after the following observation, which is an immediate consequence of Theorem 2.9, Theorem 2.15 and (3–2) of Section 3.2:

Lemma 4.2.

(i) For any $a_1, a_2 \in \mathbb{Z}$, the couples (a_1, a_2) and $(-a_1, a_2)$ have the same symbol **x,a,o,s,d,e** *attached as above.*

(ii) If C is a curve of genus 2 whose Jacobian corresponds to the couple (a_1, a_2), then the nontrivially twisted curve C' has Jacobian corresponding to the couple $(-a_1, a_2)$.

In particular, the figures occurring in the rows with $a_1 > 0$ give the exact number of k-isomorphy classes of curves whose Jacobian belongs to this isogeny class. Only in the row $a_1 = 0$ do the figures give the number of curves up to k-isomorphism and quadratic twist.

One can observe some regular behavior in the numerical results obtained by running the programs Jac2,Jac2Ch2 for all $q \le 49$.

4.1 Observations

For $q \le 49$, one can check that in the second output matrix:

1. There is no **a**.
2. In the last position of the odd rows, we find either an **x** or a **d**.

3. In the second position of the top row, we always find **o**. If $\mathrm{char}(k) \neq 2$, in the third position of the top row we find **o**, too.

4. Assume $q \geq 5$ and $p \neq 3$. Then, all other surfaces in the top row, apart from the two (one if $\mathrm{char}(k) = 2$) mentioned above, are Jacobians, with the only exception of $A = (0, -q)$ when q is not a square and $p \equiv 1 \pmod 3$ or $p = 2$, or when q is a square and $p \equiv 7 \pmod{12}$.

The first two observations can be generalized as follows:

Theorem 4.3. *Every absolutely simple abelian surface A defined over a finite field $\mathbb{F}_q$ is $\mathbb{F}_q$-isogenous to the Jacobian of a projective smooth curve of genus 2.*

Theorem 4.4. *Let a_1 be an odd integer, $|a_1| < 2[2\sqrt{q}]$. Let $a_2 = 2q + (a_1^2 - 1)/4$ be the largest integer such that (a_1, a_2) determine a Weil polynomial and assume that $(a_1 \pm 1)/2$ are q-Waterhouse numbers. Then, the abelian surface $A = (a_1, a_2)$ decomposes over $\mathbb{F}_q$ and it is not $\mathbb{F}_q$-isogenous to the Jacobian of a smooth projective curve of genus 2.*

Theorem 4.4 is an immediate consequence of a result of Serre ([Lauter 00], Lemma 1). For such a surface, we have $\Delta = 1$, so that A decomposes. Moreover, $\beta_1, \beta_2 = (a_1 \pm 1)/2$ are integers such that $\beta_1 - \beta_2 = \pm 1$; hence, the polynomial $(t - \beta_1)(t - \beta_2)$ factorizes in $\mathbb{Z}[t]$ as the product of two polynomials whose resultant is ± 1. By the result of Serre, $\pi_1, \bar{\pi}_1, \pi_2, \bar{\pi}_2$ cannot be the eigenvalues of Frobenius of a smooth projective curve of genus 2 defined over $\mathbb{F}_q$.

Theorem 4.4 can be reinterpreted in terms of number of points as follows: If we restrict our attention to curves C with a fixed value of $N_1 = \sharp C(\mathbb{F}_q)$, then the number N_2 of points of C over the quadratic extension is bounded by $a_2 \leq 2q + (a_1^2/4)$, which by (1–2) translates into

$$N_2 \leq 3q + (q+1)N_1 + \frac{q^2 + 1 - N_1^2}{2}.$$

Since $N_1 \equiv q \pmod 2$, the maximum possible value of N_2 would be

$$N_2 = 3q + (q+1)N_1 + \frac{q^2 - N_1^2}{2},$$

and Theorem 4.4 asserts that this value is never attained.

Theorem 4.3 is a consequence of Theorem 4.1, the work of [Howe 95], [Howe 96] and our characterization of the absolutely simple surfaces (Theorem 2.15).

Proof of Theorem 4.3: By Theorem 4.1, it is sufficient to show that any absolutely simple abelian surface A defined over $\mathbb{F}_q$ is $\mathbb{F}_q$-isogenous to a principally polarized one. If A is ordinary, this has been proved by Howe [Howe 95]. In fact, he proves that the parameters (a_1, a_2) of an ordinary abelian surface over $\mathbb{F}_q$ which is not isogenous to any principally polarized surface satisfy $q = a_1^2 - a_2$ and this implies that A decomposes over $\mathbb{F}_{q^3}$ by Proposition 2.13. For A nonordinary, Howe has found sufficient conditions for an abelian surface to be principally polarized, which are applicable in our case ([Howe 96], Prop. 7.2).

Let A be a nonordinary absolutely simple abelian surface. By Theorem 2.15, A is of type (M). The quartic field K generated by any root π of $f_A(t)$ is a CM field with $K^+ = \mathbb{Q}(\sqrt{\Delta})$ as the real quadratic subfield. The criterion of Howe asserts in this case that if there is a prime ideal that ramifies in K/K^+, or there is an inert prime ideal in K/K^+ dividing $\pi - \bar{\pi}$, then A is $\mathbb{F}_q$-isogenous to a principally polarized abelian surface. Let us check that this condition is always satisfied.

We denote by $\mathcal{O}, \mathcal{O}^+$ the respective rings of integers of K, K^+. Since $p \nmid \Delta$ and Δ is a quadratic residue modulo p (with $\Delta \equiv 1 \pmod 8$ if $p = 2$), the prime p decomposes in K^+:

$$p\mathcal{O}^+ = \wp\wp'. \tag{4–1}$$

On the other hand, $f_A(t)$ decomposes in $\mathbb{Q}_p[t]$ as

$$f_A(t) = (t^2 + \beta t + q)(t - \alpha_1)(t - \alpha_2),$$

with $t^2 + \beta t + q$ irreducible ([Rück 90], Lemma 3.2). Hence, p decomposes in $\mathcal{O}$ as

$$p\mathcal{O} = \mathcal{P}_1\mathcal{P}_2\mathcal{P}^2, \text{ or } p\mathcal{O} = \mathcal{P}_1\mathcal{P}_2\mathcal{P}_{(2)}. \tag{4–2}$$

From $f_A(t) \equiv t^3(t + a_1) \pmod p$, we get by an old result of Kummer that p and $\pi + a_1$ are generators of one of the prime ideals $\mathcal{P}_1, \mathcal{P}_2$; let's say: $\mathcal{P}_1 = (p, \pi + a_1)$. The two decompositions (4–1), (4–2) imply that one of the prime ideals of $\mathcal{O}^+$ above p decomposes in $\mathcal{O}$ as the product $\mathcal{P}_1\mathcal{P}_2$ and the other is either ramified or inert (it is easy to determine when it is inert or ramified in terms of δ). Since $\mathrm{Gal}(K/K^+) = \{1, \sigma\}$, where σ is complex conjugation, we know explicit generators for $\mathcal{P}_2 = \mathcal{P}_1^\sigma$ too: $\mathcal{P}_2 = (p, \bar{\pi} + a_1)$. In particular, neither $\mathcal{P}_1$ nor $\mathcal{P}_2$ can divide $\pi - \bar{\pi}$; for instance,

$$\mathcal{P}_1 \mid \pi - \bar{\pi} = (\pi + a_1) - (\bar{\pi} + a_1) \Longrightarrow \mathcal{P}_1 \mid (\bar{\pi} + a_1)$$
$$\Longrightarrow \mathcal{P}_1 \supseteq \mathcal{P}_2,$$

a_1	min. a_2		
0	-8	doxosoxdsoxodoxde	**d o x** 2 1 2 **x** 2 2 4 **x** 4 2 2 **x** 1 1
1	-4	doaaaadaaadad	**d** 1 2 2 2 2 2 4 1 3 2 2 **d**
2	0	daxasdxade	**d** 3 **x** 4 2 4 **x** 4 **d** 1
3	4	doaodad	**d** 2 2 4 1 2 **d**
4	8	daxde	1 2 **x** 2 1
5	12	dad	**d** 1 **d**
6	16	de	**d e**
7	20	d	**d**
8	24	e	**e**

TABLE 7. $q = 4$.

a_1	min. a_2		
0	-10	sooodsoooosdooosdoode	1 **o o** 1 2 1 1 1 4 2 2 3 5 1 3 3 3 2 2 1 2
1	-5	aoadaaaaadaaadad	1 1 1 2 3 4 3 2 5 3 2 2 3 3 2 **d**
2	-1	oaadaaaadaade	**o** 2 2 5 3 4 2 6 6 2 1 4 1
3	4	oadaoadad	3 2 1 4 4 1 2 2 **d**
4	8	oadoade	1 2 3 2 3 3 1
5	13	adsd	1 1 2 **d**
6	17	ode	1 1 1
7	22	d	**d**
8	26	e	**e**

TABLE 8. $q = 5$.

which is impossible. But, $p \mid \delta$ and $\mathrm{N}_{K/\mathbb{Q}}(\pi - \bar{\pi}) = \delta$ (see Lemma 4.5 below); hence the other prime in $\mathcal{O}$ above p must divide $\pi - \bar{\pi}$ and the criterion of Howe is satisfied. □

Lemma 4.5. *Let A be an abelian surface defined over $\mathbb{F}_q$ such that $f_A(t) \in \mathbb{Z}[t]$ is irreducible. Let $K = \mathbb{Q}(\pi)$ be the quartic field generated by a root π of $f_A(t)$ in $\bar{\mathbb{Q}}$. Then,*

$$N_{K/\mathbb{Q}}(\pi - \bar{\pi}) = \delta := (a_2 + 2q)^2 - 4qa_1^2.$$

Proof: If $\pi_1, \bar{\pi}_1, \pi_2, \bar{\pi}_2$ are the four roots of $f_A(t)$ in $\bar{\mathbb{Q}}$, we have:

$$f_A(t) = (t^2 + \beta_1 t + q)(t^2 + \beta_2 t + q),$$

where $\beta_i = \pi_i + \bar{\pi}_i$ are real numbers. The invariant δ is the product, $\delta = d_1 d_2$, of the two discriminants of these quadratic factors. Hence, $\pi_i - \bar{\pi}_i = \pm\sqrt{d_i}$ and

$$\begin{aligned} \mathrm{N}_{K/\mathbb{Q}}(\pi - \bar{\pi}) &= (\pi_1 - \bar{\pi}_1)(\bar{\pi}_1 - \pi_1)(\pi_2 - \bar{\pi}_2)(\bar{\pi}_2 - \pi_2) \\ &= (-d_1)(-d_2) = \delta. \end{aligned}$$ □

We have not been able to check if the third and fourth observations above are true in general or not. By Theorem 2.9, the abelian surfaces $A_1 = (0, -2q + 1)$ and (for q odd) $A_2 = (0, -2q + 2)$ are simple and ordinary. By ([Howe 95], §13) they are $\mathbb{F}_q$-isogenous to the generalized Jacobian of a *good curve* in the sense of Oort-Ueno, but as our tables show, they seem to be not $\mathbb{F}_q$-isogenous to the Jacobian of a smooth curve.

Actually, we have run a modified version of our programs centering the attention only in curves with $N_1 = q + 1$ (that is, $a_1 = 0$) and we have checked that observations 3 and 4 remain true for $q \leq 64$. The assertion of Obervation 3 has been proved recently by Howe. His proof that A_1 is not isogenous to a Jacobian is included in Section 6. For the surface A_2, see [Howe 02].

5. COMPUTATIONAL RESULTS

In Tables 7–14, we collect the output of the programs Jac2, Jac2Ch2 for $4 \leq q \leq 16$. For each q, the output consists of two matrices, indexed by pairs of integers (a_1, a_2), corresponding to Weil polynomials. The content of the matrices is explained at the beginning of Section 4. For $q \geq 11$ only the second matrix is displayed.

a_1	min.	a_2	
0	-14	soodooosoooodosoooodosoodoode	1 **o o** 1 3 1 1 **s** 4 4 2 2 7 2 5 4 7 2 4 5 7 3 2 3 7 2 4 2 2
1	-8	aadaaaaaaadaaaaadaaadad	1 2 2 2 3 4 8 2 4 5 4 8 2 5 8 6 4 4 4 2 3 3 **d**
2	-3	oadaaoaaadaaaadaade	1 4 4 4 2 8 6 6 2 1 4 6 4 4 6 8 4 3 4 2
3	2	oadaaaaadaaadad	1 3 2 3 6 6 2 5 6 4 4 4 4 2 **d**
4	8	odaaaadaade	3 6 6 2 4 6 6 2 5 4 3
5	13	adaaadad	1 2 4 2 3 3 1 **d**
6	18	odaade	2 4 2 2 4 1
7	24	dad	1 1 **d**
8	29	de	1 1
9	34	d	**d**
10	39	e	**e**

TABLE 9. $q = 7$.

a_1	min.	a_2	
0	-16	soxoxoxdsoxoxoxodoxoxoxdsoxoxoxde	1 **o x** 3 **x** 4 **x** 3 **s** 6 **x** 12 **x** 3 **x** 12 7 12 **x** 6 **x** 12 **x** 12 3 6 **x** 7 **x** 9 **x** 3 3
1	-10	xaaoxadaxaaaxadaxaaaxaaaxad	**x** 3 3 **o x** 6 4 6 **x** 6 6 9 **x** 6 6 9 **x** 12 6 9 **x** 6 6 9 **x** 3 **d**
2	-4	xaxaxdxaxaxaxaxaxdxaxe	**x** 10 **x** 9 **x** 12 **x** 12 **x** 18 **x** 18 **x** 18 **x** 6 **x** 18 **x** 7 **x** 3
3	1	oxaaaxaaaxodaxadax	3 **x** 3 6 6 **x** 15 6 3 **x** 9 3 12 **x** 3 3 6 **x**
4	7	asaxdxaxadaxdx	4 3 6 **x** 15 **x** 12 **x** 12 4 6 **x** 9 **x**
5	13	axadoxadax	3 **x** 6 3 6 **x** 6 3 3 **x**
6	18	xaxdxaxe	**x** 6 **x** 6 **x** 6 **x** 3
7	24	aaxad	1 3 **x** 3 **d**
8	30	xde	**x** 3 1
9	35	od	1 **d**
10	41	e	**e**

TABLE 10. $q = 8$.

a_1	min.	a_2	
0	-18	dooxooxoosodxooxoosodxooxoodooxodxode	**d o o x** 5 2 **x** 2 8 **s** 2 5 **x** 4 6 **x** 14 4 4 4 16 **x** 3 12 1 6 10 **d** 10 6 **x** 6 10 **x** 6 5 1
1	-12	daaaoaaaaadaaaaaaadaaaaadaaxdad	**d** 3 2 2 6 6 2 4 6 2 8 8 6 10 4 2 14 10 2 10 14 4 8 4 3 14 4 **x** 6 2 **d**
2	-6	doaxaaaaodaaaaaadaaaadaade	**d** 3 4 **x** 12 8 4 10 16 4 8 4 6 12 12 4 22 6 4 8 12 4 8 4 4 5
3	0	daaxaaxadsaaxadxaadad	1 2 6 **x** 6 10 **x** 6 12 4 4 12 **x** 7 10 **x** 8 10 **d** 4 **d**
4	6	dooaaaxdaaaadaade	1 6 6 4 12 4 **x** 14 14 2 10 8 6 4 4 2 8
5	12	daaaoadaaadad	**d** 5 2 2 11 4 2 4 6 2 4 4 **d**
6	18	daaxadxade	**d** 3 8 **x** 6 12 **x** 2 8 1
7	24	daaadad	**d** 2 2 2 3 2 **d**
8	30	daode	1 2 3 1 3
9	36	dad	**d** 1 **d**
10	44	de	**d** 1
11	48	d	**d**
12	54	e	**e**

TABLE 11. $q = 9$.

a_1	min.	a_2	
0	-22	1 **o o** 1 3 2 1 2 6 4 1 2 13 1 6 4 9 8 3 5 13 4 5 6 14 6 14 7 11 6 6 3 16 5 5 13 12 2 7 8 7 4 5 1 5
1	-15	1 4 2 4 4 3 7 5 5 10 9 4 9 12 8 4 6 16 8 10 6 8 13 4 12 14 12 6 13 8 5 10 4 10 5 4 5 **d**
2	-8	4 **o** 6 6 8 6 14 6 8 8 16 12 12 6 20 24 8 6 18 6 14 6 20 9 8 6 12 12 6 4 12 2
3	-2	2 2 2 8 12 2 8 8 4 14 4 4 20 8 6 15 10 4 12 12 6 10 4 3 8 4 **d**
4	5	4 9 4 9 4 14 4 14 14 12 4 9 11 13 8 6 12 15 4 6 6 5
5	12	2 5 7 7 4 9 10 8 4 4 8 4 4 5 4 5 **d**
6	18	2 6 4 2 16 9 4 11 12 6 10 2 4 5
7	25	3 4 4 3 2 4 5 4 2 **d**
8	32	3 4 6 2 4 6 3
9	38	2 1 2 3 **d**
10	45	1 2 1
11	51	1 **d**
12	58	1

TABLE 12. $q = 11$.

a_1	min. a_2	
0	-26	2 **o o** 1 5 1 2 3 4 4 2 6 9 **s** 3 4 17 4 9 4 16 8 4 6 14 8 12 13 9 6 6 12 20 4 12 8 30 2 7 10 16 12 7 5 11 9 7 5 13 4 5 5 4
1	-18	2 3 3 4 4 5 10 7 4 8 10 8 8 6 12 16 8 8 12 13 8 12 20 12 14 10 16 16 8 8 12 12 10 16 12 6 8 11 10 8 6 9 10 3 **d**
2	-11	2 6 4 8 6 16 7 4 8 24 6 18 6 23 12 12 12 12 20 8 8 28 16 28 6 12 18 6 10 32 18 12 4 18 12 8 3 8 6
3	-4	1 6 5 5 12 4 8 9 18 6 6 18 12 16 8 6 24 16 6 11 14 15 10 8 6 9 10 8 12 10 4 5 **d**
4	3	4 4 5 12 6 11 16 12 6 17 8 23 4 18 18 24 6 10 20 14 6 14 8 15 6 7 6 12
5	11	2 10 6 4 12 8 5 8 12 12 8 4 4 12 15 8 4 12 4 6 3 **d**
6	18	8 8 4 6 18 12 10 10 18 9 8 4 12 16 6 5 10 2
7	25	1 2 8 6 9 4 3 6 8 4 4 7 4 **d**
8	32	3 6 7 2 5 10 10 4 6 4 5
9	39	1 4 2 4 4 3 3 **d**
10	47	2 4 2 5 2
11	54	1 1 **d**
12	61	2 1
13	68	**d**
14	75	**e**

TABLE 13. $q = 13$.

a_1	min. a_2	
0	-32	1 **o x** 4 **x** 4 **x** 8 **x** 16 **x** 8 **x** 8 **x** 17 5 16 **x** 24 **x** 16 **x** 24 **x** 24 **x** 44 **x** 24 **x** 24 8 16 **x** 32 **x** 16 **x** 38 **x** 32 **x** 16 **x** 36 **x** 40 10 32 **x** 32 **x** 12 **x** 38 **x** 16 **x** 24 **x** 12 **x** 10 6
1	-24	**d** 4 **x** 8 4 8 **x** 8 4 10 **x** 16 8 12 **x** 16 4 16 **x** 16 12 16 **x** 24 8 25 **x** 36 8 20 **x** 32 4 28 **x** 16 16 32 **x** 16 8 16 **x** 20 8 32 **x** 24 2 18 **x** 24 8 8 **x** 8 **d**
2	-16	**x** 14 **x** 16 **x** 24 **x** 16 **x** 36 **x** 24 **x** 64 **x** 16 **x** 63 **x** 32 **x** 48 **x** 48 **x** 60 **x** 32 **x** 40 **x** 64 **x** 64 **x** 32 **x** 48 **x** 16 **x** 24 **x** 40 **x** 40 **x** 16 **x** 12
3	-8	**d o x** 8 6 20 **x** 8 8 20 **x** 24 8 20 **x** 40 **x** 16 **x** 16 16 16 **x** 24 12 28 **x** 40 8 24 **x** 24 4 34 **x** 24 10 8 **x** 16 4 8 **x**
4	0	**d** 8 **x** 32 **x** 16 **x** 32 **x** 36 **x** 48 **x** 32 **x** 32 10 36 **x** 72 **x** 32 **x** 32 **x** 28 **x** 68 **x** 16 **x** 32 d 16 **x** 20 **x**
5	8	**d** 16 **x** 4 8 8 **x** 24 4 24 **x** 16 8 16 **x** 16 4 16 **x** 32 10 16 **x** 8 4 20 **x** 16 4 12 **x**
6	16	**x** 12 **x** 24 **x** 40 **x** 24 **x** 40 **x** 32 **x** 24 **x** 32 **x** 48 **x** 16 **x** 40 **x** 16 **x** 8
7	24	1 4 **x** 16 4 12 **x** 16 4 29 **x** 8 8 24 **x** 8 1 8 **x** 4 **d**
8	32	4 8 **x** 16 **x** 16 **x** 34 **x** 24 **x** 16 **x** 16 **x** 16 5
9	40	**d** 6 **x** 12 4 12 **x** 8 4 15 **x** 8 **d**
10	48	**x** 12 **x** 8 **x** 16 **x** 8 **x** 6
11	56	**d** 4 **x** 4 2 4 **x**
12	64	**d** 4 **x** 8 **x**
13	72	**d** 2 **x**
14	80	**x** 1
15	88	**d**
16	96	1

TABLE 14. $q = 16$.

6. APPENDIX BY EVERETT W. HOWE

For every prime power q, let f_q denote the polynomial $x^4 + (1-2q)x^2 + q^2$. In Section 4 of this article, Maisner and Nart observe that for all prime powers $q \leq 64$, no genus-2 curve over $\mathbf{F}_q$ has characteristic polynomial f_q. (By the *characteristic polynomial* of a curve, we mean the characteristic polynomial of the Frobenius endomorphism of the Jacobian of the curve.) The purpose of this appendix is to prove that Maisner and Nart's observation holds for all prime powers q.

Theorem. *There is no curve of genus* 2 *over any finite field* $\mathbf{F}_q$ *whose characteristic polynomial is equal to* f_q.

Proof: Suppose, to obtain a contradiction, that C is a genus-2 curve over a finite field $\mathbf{F}_q$ whose characteristic polynomial is equal to f_q. Note that then $\#C(\mathbf{F}_q) = q+1$ and $\#C(\mathbf{F}_{q^2}) = (q-1)(q-3)$.

Let J be the Jacobian of C, let λ be the canonical principal polarization of J, let F be the Frobenius endomorphism of J, and let $V = q/F$ be the Verschiebung endomorphism of J. Since f_q is irreducible and its middle coefficient is coprime to q, we see that J is a simple ordinary abelian surface, and it follows that the ring $(\operatorname{End} J) \otimes \mathbf{Q}$ is equal to the field $\mathbf{Q}(F)$. In fact, this field is a totally imaginary quadratic extension of a totally real quadratic field, and general theory (see [Mumford 74, p. 201]) shows that the Rosati involution $x \mapsto x^\dagger$ on $\mathbf{Q}(F)$ is complex conjugation.

Let i be the endomorphism $F - V$ of J. It is easy to check that $i^2 = -1$, and it follows that $i^\dagger i = 1$. Thus i is an automorphism of J that respects the polarization λ, so i can be viewed as an automorphism of the

polarized abelian variety (J, λ). Since C is hyperelliptic, Torelli's theorem (see [Milne 86, p. 202]) shows that the natural map from the automorphism group of C to the automorphism group of (J, λ) is an isomorphism that takes the hyperelliptic involution to -1. Thus, the automorphism i of (J, λ) gives us an automorphism α of C, defined over $\mathbf{F}_q$, whose square is the hyperelliptic involution.

Let W denote the set of Weierstrass points of C, viewed as a set with an action of the absolute Galois group of $\mathbf{F}_q$. If P is a geometric point of C whose orbit under the action of α contains fewer than four points, then P must be fixed by $\alpha^2 = -1$, so P must lie in W. Thus, for every finite extension field k of $\mathbf{F}_q$ we have $\#C(k) \equiv \#W(k) \bmod 4$.

Suppose that q is odd. Then W consists of six points, and we will show that exactly two of these points are fixed by α.

Consider the map $C \to \mathbf{P}^1$ obtained from the hyperelliptic involution, and let W' denote the set of six points of $\mathbf{P}^1$ lying under the Weierstrass points of C. The automorphism α induces an involution β of $\mathbf{P}^1$ that takes the set W' to itself. Geometrically, this involution is conjugate to the involution $x \mapsto -x$, so if none of the points in W' were fixed by β the curve C would be isomorphic (over the algebraic closure of $\mathbf{F}_q$) to a curve of the form $y^2 = f(x^2)$, where f is a cubic polynomial. But then α would have to be of the form $(x, y) \mapsto (-x, \pm y)$, and such an automorphism has order two. Thus, β must fix at least one of the six points of W'. But the points not fixed by β come in plus/minus pairs, so there must be at least two points of W' fixed by β. Since $x \mapsto -x$ has exactly two fixed points in $\mathbf{P}^1$, there must be exactly two points of W' fixed by β. It follows that exactly two points of W are fixed by α, as claimed.

Since α is defined over $\mathbf{F}_q$, the two points of W fixed by α must be defined over $\mathbf{F}_{q^2}$. Thus, $\#W(\mathbf{F}_{q^2}) \geq 2$. But we also have

$$\#W(\mathbf{F}_{q^2}) \equiv \#C(\mathbf{F}_{q^2}) = (q-1)(q-3) \equiv 0 \bmod 4,$$

so we must have $\#W(\mathbf{F}_{q^2}) = 4$. But this is impossible, as one can see by asking where the other two points of W are defined. Thus, if q is odd, no curve can have characteristic polynomial f_q.

Suppose that q is a power of 2. Then q must be a multiple of 4, because if q were 2 the curve C would have -1 points over $\mathbf{F}_4$. We see that $\#W(\mathbf{F}_q) \equiv 1 \bmod 4$ and $\#W(\mathbf{F}_{q^2}) \equiv 3 \bmod 4$. But a genus-2 curve in characteristic 2 has at most three Weierstrass points, so C must have exactly three Weierstrass points, and exactly one of them is defined over $\mathbf{F}_q$.

Once again we let W' denote the points of $\mathbf{P}^1$ lying under the Weierstrass points of C and we let β be the involution of $\mathbf{P}^1$ obtained from α. Clearly β must fix the unique point of $W'(\mathbf{F}_q)$. But β cannot fix the other two points of W', because in that case β would be the identity on $\mathbf{P}^1$, and α could not have order four. Thus, β must swap the other two points of W'. It follows that over the algebraic closure of $\mathbf{F}_q$ we can write C as $y^2 + y = ax + b/x + b/(x+1)$, where we have chosen the coordinates so that β is given by $x \mapsto x + 1$. But then α must send (x, y) to $(x + 1, y + c)$ where $c^2 + c = a$, and this automorphism has order two. Once again we obtain a contradiction, and the theorem is proved. □

Maisner and Nart also note that for every odd prime power $q < 64$, no genus-2 curve over $\mathbf{F}_q$ has characteristic polynomial $g_q = x^4 + (2 - 2q)x^2 + q^2$. The obvious conjecture is that the same statement is true for all odd prime powers q. Unfortunately, the argument we used above cannot be easily modified to prove this conjecture; the critical fact we used was that the ring $\mathbf{Z}[F, V]$ contains a root of unity other than ± 1, and this is no longer true when we replace f_q with g_q in our argument. In a forthcoming paper [Howe 02], we will prove this conjecture using an argument that depends on the Brauer relations in a biquadratic number field.

ACKNOWLEDGMENTS

We thank the referee for several suggestions which have led us to give a more complete form to the paper. We are also grateful to Everett W. Howe for his extreme kindness in accepting that his proof of one of the questions raised by our computations appears as an appendix to the paper. The first author was supported by CONACYT; the second author was supported by DGI, BHA2000-0180.

REFERENCES

[Adleman and Huang 92] L. M. Adleman, M.-D. A. Huang. *Primality testing and abelian varieties over finite fields*, Lecture Notes in Mathematics 1512, Springer-Verlag, Berlin-Heidelberg, 1992.

[Cardona et al. 02] G. Cardona, E. Nart, and J. Pujolàs. *Curves of genus two over fields of even characteristic.* to appear, http://www.arxiv.org/math.NT/0210105.

[Guàrdia 98] J. Guàrdia. *Geometria Aritmètica en una família de corbes de gènere tres.* Tesi, Universitat de Barcelona, 1998.

[Howe 95] E. W. Howe. "Principally polarized abelian varieties over finite fields." *Transactions of the American Mathematical Society* **347** (1995), 2361–2401.

[Howe 96] E. W. Howe. "Kernels of polarizations of abelian varieties over finite fields." *Journal of Algebraic Geometry* **5** (1996), 583–608.

[Howe 02] E. W. Howe. "On the nonexistence of certain curves of genus two." *Compositio Mathematica*, to appear. arxiv:math.NT/0201311.

[Howe and Zhu 02] E. W. Howe and H. J. Zhu. "On the existence of absolutely simple abelian varieties of a given dimension over an arbitrary field." *Journal of Number Theory* **92** (2002), 139–163.

[Igusa 60] J.-I. Igusa. "Arithmetic variety of moduli for genus two." *Annals of Mathematics* **72** (1960), 612–649.

[Lachaud 91] G. Lachaud. "Artin-Schreier curves, exponential sums and the Carlitz-Uchiyama bound for geometric codes." *Journal of Number Theory* **39** (1991), 18–40.

[Lauter 00] K. Lauter. "Non-existence of a curve over $\mathbb{F}_3$ of genus 5 with 14 rational points." *Proceedings of the American Mathematical Society* **128**:2 (2000), 369–374.

[López et al. 02] A. López, D. Maisner, E. Nart, and X. Xarles. "Orbits of galois invariant n-sets of $\mathbb{P}^1$ under the action of PGL_2." *Finite Fields and Their Applications* **8** (2002), 193–206.

[Milne 86] J. S. Milne. "Jacobian varieties." in *Arithmetic Geometry*, (G. Cornell and J. H. Silverman, eds.), pp. 167–212, Springer-Verlag, New York, 1986.

[Mumford 74] David Mumford. *Abelian Varieties*, 2nd ed., Oxford University Press, Oxford, 1974.

[Rück 90] H.-G. Rück. "Abelian surfaces and Jacobian varieties over finite fields." *Compositio Mathematica* **76** (1990), 351–366.

[Stark 72] H. Stark. "On the Riemann hypothesis in hyperelliptic function fields." in *Analytic Number Theory*, Proceedings of Symposia in Pure Mathematics, Vol. XXIV, pp. 285–302, American Mathematical Society, Providence, RI, 1973.

[Tate 69] J. Tate. "Classes d'isogénie des variétés abéliennes sur un corps fini (d'après Honda)." in *Séminaire Bourbaki 1968/69, Exposé 352*, pp. 95–110, Lecture Notes in Mathematics 179, Springer-Verlag, Berlin 1971.

[van der Geer and van der Vlugt 92] G. van der Geer and M. van der Vlugt. "Supersingular curves of genus 2 over finite fields of characteristic 2." *Mathematische Nachrichten* **159** (1992), 73–81.

[Waterhouse 69] W. C. Waterhouse. "Abelian varieties over finite fields." *Annales Scientifiques de l'École Normale Supérieure (4)* **2** (1969), 521–560.

[Waterhouse and Milne 69] W. Waterhouse and J. Milne. "Abelian varieties over finite fields." in *1969 Number Theory Institute*, Proceedings of Symposia in Pure Mathematics, Vol. XX, pp. 53–64, American Mathematical Society, Providence, RI, 1971.

[Weil 57] A. Weil. "Zum Beweis des Torellischen Satzes." *Nachrichten der Akademie der Wissenschaften in Göttingen, Mathematisch-Physikalische Klasse IIa* (1957), 33–53.

[Xing 94] C. P. Xing. "The structure of the rational point groups of simple abelian varieties of dimension two over finite fields." *Archiv der Mathematik* **63** (1994), 427–430.

[Xing 96] C. P. Xing. "On supersingular abelian varieties of dimension two over finite fields." *Finite Fields and Their Applications* **2** (1996), 407–421.

Daniel Maisner, Departament de Matemàtiques Universitat Autònoma de Barcelona, Edifici C, 08193 Bellaterra, Barcelona, Spain (danielm@mat.uab.es)

Enric Nart, Departament de Matemàtiques Universitat Autònoma de Barcelona, Edifici C, 08193 Bellaterra, Barcelona, Spain (nart@mat.uab.es)

Everett W. Howe, Center for Communications Research, 4320 Westerra Court, San Diego, CA 92121-1967 (however@alumni.caltech.edu)

Received January 18, 2001; accepted in revised form November 21, 2001.

Reversible Complex Hénon Maps

C. R. Jordan, D. A. Jordan, and J. H. Jordan

CONTENTS

We identify and investigate a class of complex Hénon maps $H : \mathbb{C}^2 \to \mathbb{C}^2$ that are reversible, that is, each H can be factorized as RU where $R^2 = U^2 = \mathrm{Id}_{\mathbb{C}^2}$. Fixed points and periodic points of order two or three are classified in terms of symmetry, with respect to R or U, and as either elliptic or saddle points. We report on experimental investigation, using a Java applet, of the bounded orbits of H.

1. INTRODUCTION

For $\alpha, \beta \in \mathbb{C}$, the *Hénon map* $H_{\alpha,\beta} : \mathbb{C}^2 \to \mathbb{C}^2$ is defined by the rule

$$H_{\alpha,\beta}((z, w)) = (\alpha - \beta w - z^2, z). \tag{1–1}$$

If $\alpha, \beta \in \mathbb{R}$, then $H_{\alpha,\beta}$ restricts to the *real Hénon map* $H_{\alpha,\beta} : \mathbb{R}^2 \to \mathbb{R}^2$. Real and complex Hénon maps, and their history, are well-documented, see, for example, [Devaney 89, Hale and Koçak 91] for the real case, and [Friedland et al. 89, Bedford et al. 91, Bedford et al. 93, Hubbard and Oberste-Vorth 94, Oberste-Vorth 97, Smillie and Buzzard 97] for the complex case.

If R is an involution of $\mathbb{R}^n$ and $F : \mathbb{R}^n \to \mathbb{R}^n$ or if R is an involution of $\mathbb{C}^n$ and $F : \mathbb{C}^n \to \mathbb{C}^n$ then, following [Devaney 76, Devaney 84], F is *R-reversible* if $F^{-1} = RFR$. This is equivalent to requiring that RF is an involution or that $F = RU$ for some involution U.

Let R and S denote the involutions of $\mathbb{C}^2$ such that $R((z, w)) = (w, z)$ and $S(z, w) = (-w, -z)$, or, where appropriate, their restrictions to $\mathbb{R}^2$. If $\alpha \in \mathbb{R}$, the real Hénon maps $H_{\alpha,1}$ and $H_{\alpha,-1}$ are R-reversible and S-reversible, respectively, and are discussed in [Devaney 84] and [Devaney 89, Section 2.9, Exercises 21–34]. The only comment on reversible complex Hénon maps that we have found in the literature is a comment in [Friedland et al. 89], where a conjugate form of $H(\alpha, \beta)$ is used, that $H_{\alpha,\beta}$ is R-reversible if and only if $\beta = 1$ and S-reversible if and only if $\beta = -1$. For $\beta \in \mathbb{C}$, with $|\beta| = 1$, let R_β be the involution of $\mathbb{C}^2$ such that

$$R_\beta((z, w)) = (\beta\overline{w}, \beta\overline{z}). \tag{1–2}$$

2000 AMS Subject Classification: Primary 32H50, 37F10; Secondary 37C25, 37E15, 37F45

Keywords: Hénon map, reversibility, fixed points, periodic points, bounded orbits, ellipticity

1058-6458/2001 $0.50 per page
Experimental Mathematics **11**:3, page 339

FIGURE 1. Projections of orbits.

Then R_1 and R_{-1} have the same restrictions to $\mathbb{R}^2$ as R and S.

In Section 2, we shall see that $H_{\alpha,\beta}$ is R_β-reversible if and only if $\alpha \in \mathbb{R}\beta$. In this case, the involution $R_\beta H_{\alpha,\beta}$ is given by $(z,w) \mapsto (\beta\overline{z}, \beta\overline{\alpha} - \overline{w} - \beta\overline{z}^2)$. In [Devaney 76, Devaney 84] the involutions R are assumed to be diffeomorphisms. Although R_β is not a $\mathbb{C}$-diffeomorphism of $\mathbb{C}^2$, it is a $\mathbb{R}$-diffeomorphism when $\mathbb{C}^2$ is identified with $\mathbb{R}^4$, so the results of [Devaney 84] apply.

The role of the reflection $z \mapsto \beta\overline{z}$ in the involutions R_β and $R_\beta H_{\alpha,\beta}$ gives rise to orbits, with reflective symmetry, that can be quite striking in appearance. Figure 1 shows projections onto the z-plane of two examples.

The reversibility not only influences the geometry of orbits but facilitates calculation and analysis. In Section 3, we analyse the fixed points and determine when they are symmetric, for either of the involutions R_β or $R_\beta H_{\alpha,\beta}$, and when they are elliptic. We shall do the same for periodic points of order 2 or 3 in Section 4. Section 5 is concerned with local dynamics and establishes a sufficient condition for an orbit to be unbounded. This has been applied to plot bounded orbits. (A Java applet is available at http://www.shef.ac.uk/~daj/henon/H.html.) The final section reports on experimental observations of such orbits. For example, if β is a primitive mth root of unity, then $\mathrm{orb}((0,0))$, if bounded, appears to be dense in the union of m closed curves which are deformations of ellipses, becoming more deformed as $|\alpha|$ increases. We also comment on the influence on orbits of nearby periodic elliptic points and on bifurcation.

Our interest in Hénon maps arose from a problem in [Jordan 93, 3.3], a special case of which would ask whether, for a nonperiodic orbit $\{(z_n, w_n)\}_{n\in\mathbb{Z}}$ of the Hénon map, z_n could take the same value infinitely often.

2. REVERSIBILITY

Lemma 2.1. *Let $\alpha, \beta, \rho \in \mathbb{C}$, with $|\rho| = 1$, let $H = H_{\alpha,\beta}$, and let R_ρ be the involution of $\mathbb{C}^2$ such that $R_\rho((z,w)) = (\rho\overline{w}, \rho\overline{z})$. Then H is R_ρ-reversible if and only if $\beta = \rho$ and $\alpha \in \mathbb{R}\beta$.*

Proof: It is easily checked that $(R_\rho H_{\alpha,\beta})^2 = \mathrm{Id}_{\mathbb{C}^2}$ if and only if $\beta = \rho$ and $\alpha \in \mathbb{R}\beta$. □

2.1 Notation

The Euclidean norm on $\mathbb{C}^2 = \mathbb{R}^4$ will be denoted $||\ ||$. We denote by H the map $H_{\alpha,\beta}$, where $\beta = e^{i\theta}$ for some $\theta \in \mathbb{R}$ with $-\pi < \theta \leq \pi$, and $\alpha = r\beta$ for some $r \in \mathbb{R}$. The involutions R_β and $R_\beta H$ will be denoted by R and U, respectively. Thus H is R-reversible, $H = RU$ and

$$U((z,w)) = (\beta\overline{z}, \beta\overline{\alpha} - \overline{w} - \beta\overline{z}^2). \tag{2–1}$$

Here, $\beta\overline{z}$ is obtained from z by reflection in the line inclined at $\frac{\theta}{2}$ to the real axis. We call this line the *U-line*. For $n \in \mathbb{Z}$, $H^{-n}((0,0)) = UH^{n-1}((0,0))$, so the projection onto the z-plane of $\mathrm{orb}((0,0))$ is symmetrical about the U-line. This symmetry can be observed in Figure 1.

The space of parameters for which H is R-reversible is $\mathcal{P} := \{(\alpha,\beta) : |\beta| = 1, \alpha \in \mathbb{R}\beta\}$. If $r > 0$, $-\pi < \theta \leq \pi$ and $\alpha = re^{i\theta}$, then α determines two points $\mathrm{pos}(\alpha) := (re^{i\theta}, e^{i\theta})$ and $\mathrm{neg}(\alpha) := (re^{i\theta}, -e^{i\theta}) = (-re^{i(\theta\pm\pi)}, e^{i(\theta\pm\pi)})$ in $\mathcal{P}$.

For $P \in \mathbb{C}^2$, we say that P is *periodic of order n* if n is the least positive integer such that $H^n(P) = P$. The set of periodic points of a given order n is invariant under both R and U. A periodic point P is *U-symmetric*, resp. *R-symmetric*, if $U(P) = P$, respectively $R(P) = P$.

3. FIXED POINTS

3.1 Symmetry

Let

$$f(z) = z^2 + (\beta + 1)z - \alpha. \tag{3–1}$$

For $P = (z,w) \in \mathbb{C}^2$,

$$H(P) = P \Leftrightarrow w = z \text{ and } f(z) = 0. \tag{3–2}$$

Counting multiplicity, there are two fixed points determined by the zeros of f. Let P be a fixed point for H. Then $R(P) = U(P)$ is a fixed point. If $P = R(P) = U(P)$, we shall say that P is *symmetric*.

Theorem 3.1. *The fixed points of H are symmetric if and only if $r \geq -c^2$, where $c = \cos\frac{\theta}{2}$.*

Proof: The fixed points have the form (z,z), where $f(z) = 0$, and, as $U((z,z))$ is also fixed, $U((z,z)) =$

$(\beta\overline{z}, \beta\overline{z})$. For $v \in \mathbb{C}$,

$$f(ve^{\frac{i\theta}{2}}) = \beta(v^2 + 2cv - r). \tag{3–3}$$

Thus the zeros of f are $z = ve^{\frac{i\theta}{2}}$, where $v = -c \pm \sqrt{c^2 + r}$. The fixed points are symmetric if and only if these are on the U-line if and only if $r \geq -c^2$. □

3.2 Ellipticity

We now analyse the fixed points identified in Section 3.1 in terms of local dynamics. Let $P = (z, w) \in \mathbb{C}^2$, let n be a positive integer, and let $J_n(P)$ denote the Jacobian matrix of H^n at P. Then

$$J_1(P) = \begin{pmatrix} -2z & -\beta \\ 1 & 0 \end{pmatrix} \text{ and } \det J_1(P) = \beta. \tag{3–4}$$

Let P be periodic of order n. By (3–4), $|\det J_1(P)| = 1$, so $|\det J_n(P)| = 1$. Either both eigenvalues of $J_n(P)$ have modulus 1, in which case P is *elliptic*, or one has modulus > 1 and the other has modulus < 1, in which case P is a *saddle point*. The dynamics at saddle points is well understood in terms of the stable and unstable manifolds, e.g., [Bedford et al. 91, Fornæss 96, Smillie and Buzzard 97]. If $n \in \mathbb{N}$, then

$$RH^nR = H^{-n} = UH^nU \text{ and } UH^{-n} = H^{n-1}R. \tag{3–5}$$

Hence the stable and unstable manifolds are mapped to each other by R and by U. In a sense made precise in [Bedford et al. 93] or [Smillie and Buzzard 97, Corollary 13.4], most periodic points are saddle points.

Theorem 3.2. *Let $c = \cos\frac{\theta}{2}$.*

(i) If $r < -c^2$, that is, if the fixed points of H are not symmetric, then they are saddle points.

(ii) If $-c^2 \leq r \leq 1-2c$, then both fixed points are elliptic.

(iii) If $1-2c < r \leq 1+2c$, then one fixed point is elliptic and one is a saddle point.

(iv) If $r > 1+2c$, then both fixed points are saddle points.

Proof: Using (3–4), one shows that, for all $w, z \in \mathbb{C}$, the eigenvalues of $J_1((z, w))$ are $\lambda_1(z) = -z + \sqrt{z^2 - \beta}$ and $\lambda_2(z) = -z - \sqrt{z^2 - \beta}$. Let $z = ve^{\frac{i\theta}{2}}$, $v \in \mathbb{C}$. Then $\lambda_1(z), \lambda_2(z) = e^{\frac{i\theta}{2}}(-v \pm \sqrt{v^2 - 1})$. Hence $|\lambda_1(z)| = 1 = |\lambda_2(z)|$ if and only if $v \in \mathbb{R}$ and $v^2 \leq 1$.

By Theorem 3.1 and its proof, the fixed points have the form (z, z), where $z = (-c \pm \sqrt{c^2 + r})e^{\frac{i\theta}{2}}$. (i)-(iv) follow easily. □

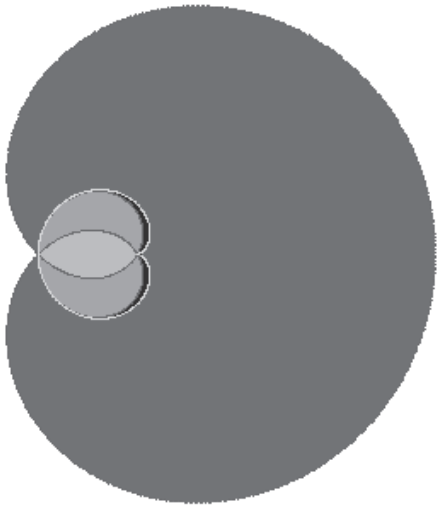

FIGURE 2. The regions where fixed points are elliptic.

Remark 3.3. Corresponding to $0 \neq \alpha \in \mathbb{C}$, there are four fixed points, P_1^+ and P_2^+ for $\text{pos}(\alpha)$, and P_1^- and P_2^- for $\text{neg}(\alpha)$. Number these so that $||P_1^+|| \geq ||P_2^+||$ and $||P_1^-|| \geq ||P_2^-||$. Figure 2 shows the values of α, determined by Theorem 3.2, for which there are elliptic fixed points. These are P_2^+ everywhere that is shaded, P_2^- everywhere except in the large outermost region, P_1^+ in the eye where the shading is lightest and P_1^- in the darkest regions.

3.3 Linearization

We now discuss orbits for the linearization L_H of H in the case where the fixed points are elliptic. This will provide a basis for discussions of orbits of H later in the paper. Thus

$$L_H((z, w)) = (-2\zeta z - \beta w, z) = (M(z, w)^T)^T,$$

where $P = (\zeta, \zeta)$ is an elliptic fixed point and $M = J_1(P)$. By Theorem 3.2 and its proof, the eigenvalues of M can be written in the form $\lambda_1 = e^{i(\frac{\theta}{2}+\phi)}$ and $\lambda_2 = e^{i(\frac{\theta}{2}-\phi)}$ for some $\phi \in \mathbb{R}$. If both $e^{i\theta}$ and $e^{i\phi}$ are roots of unity, then L_H has finite order and hence H cannot be locally conjugate to L_H. A condition on the eigenvalues under which L_H is locally conjugate to H is given by [Zehnder 77].

Suppose that β is a primitive mth root of unity, but that $e^{i\phi}$ is not a root of unity. Then $\lambda_2^m = e^{-im(\frac{\theta}{2}+\phi)} = \lambda_1^{-m}$. Let

$$D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$

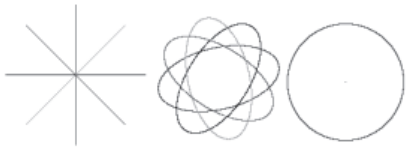

FIGURE 3. Projections of orbits for L_H with $\theta = \pi/4$.

If $z_1 z_2 \neq 0$, the orbit of (z_1, z_2) for the action of the group $\langle D^m \rangle$ lies on, and is dense in, the closed curve $\{(e^{it}z_1, e^{-it}z_2)\}$ and its orbit for the action of $\langle D \rangle$ is dense in the union of m such closed curves. Projections of such curves onto a complex line $\rho z_1 + \sigma z_2 = 0$ are (possibly degenerate) ellipses. The orbit of $(z_1, 0)$ or $(0, z_2)$ for the action of $\langle D \rangle$ is dense in a circle $\{(e^{it}z_1, 0)\}$ or $\{(0, e^{-it}z_2)\}$, in the z_1- or z_2-plane. Consequently, for $(z, w) \in \mathbb{C}^2$, the orbit for the action of $\langle L_H \rangle$ is dense either in the union of m (possibly degenerate) closed curves, whose projections onto the z-plane are ellipses, or in a single such curve. Examples of some orbits for L_H are shown in Figure 3.

If the eigenvectors generate a free abelian subgroup of $\mathbb{C}^*$ of rank 2, then orbits for the action of $\langle D \rangle$ are dense in two-dimensional tori $\{(e^{it}z_1, e^{iu}z_2)\}$, so orbits for the action of $\langle L_H \rangle$ are also dense in two-dimensional tori.

4. PERIODIC POINTS OF ORDER 2

4.1 Symmetry

For $P = (z, w) \in \mathbb{C}^2$, if $\beta \neq -1$,

$$H^2(P) = P \Leftrightarrow (1+\beta)w = \alpha - z^2 \text{ and } f(z)g(z) = 0, \tag{4–1}$$

where $f(z)$ is as in (3–1) and

$$g(z) := z^2 - (\beta+1)z + (1+\beta)^2 - \alpha. \tag{4–2}$$

Counting multiplicity, this determines the four points P such that $H^2(P) = P$, including the fixed points. Thus there are at most two periodic points of order 2. If P is a periodic point of order 2, then so are $H(P)$, $R(P)$, and $U(P)$. As z determines w, P is U-symmetric if and only if z is on the U-line. Also $R(P) \neq U(P)$, otherwise $H(P) = P$, so either P is R-symmetric and $U(P) = H(P)$, in which case $RH(P) = U(P) = H(P)$ and $H(P)$ is R-symmetric, or P is U-symmetric and $R(P) = H(P)$, which is U-symmetric.

Theorem 4.1. *Let $c = \cos\frac{\theta}{2}$.*

(i) If $r = 3c^2 \neq 0$, then H has no periodic points of order 2.

(ii) If $\beta \neq -1$ and $r \neq 3c^2$, then H has precisely two distinct periodic points of order 2. These are U-symmetric if $r > 3c^2$ and are R-symmetric if $r < 3c^2$.

(iii) If $\beta = -1$ and $r \neq 0$, then the periodic points of order 2 are the two points of the form $(z, -z)$ where $z^2 = \alpha$.

Proof: For $v \in \mathbb{C}$, $g(ve^{\frac{i\theta}{2}}) = 0 \Leftrightarrow v^2 - 2cv + 4c^2 - r = 0$ so the zeros of g are $z = ve^{\frac{i\theta}{2}}$ where $v = c \pm \sqrt{r - 3c^2}$.

(i) Suppose that $r = 3c^2 \neq 0$. Then $\beta \neq -1$ and the double zero $ce^{\frac{i\theta}{2}}$ of g is, by (3–3), a zero of f. For periodic points of order 1 or 2, w is determined by z so the solutions of $H^2((z, w)) = (z, w)$ are already solutions of $H((z, w)) = (z, w)$.

(ii) If $\beta \neq -1$ and $r \neq 3c^2$, then g and f have no common zero so H has two periodic points, (z_1, w_1) and (z_2, w_2), say, of order 2, with $z_1, z_2 = ve^{\frac{i\theta}{2}}$, where $v = c \pm \sqrt{r - 3c^2}$. The result follows.

(iii) This is routine. □

4.2 Ellipticity

Theorem 4.2. *With c as in Theorem 4.1, if $r \neq 3c^2$, then the periodic points of H of order 2 are elliptic if and only if $4c^2 - 1 \leq r \leq 4c^2$.*

Proof: Let $\{(z_i, w_i) : i = 1, 2\}$ be an orbit of period 2 under H. The Jacobian matrix of H^2 at (z_i, w_i) has trace $t = 4z_1 z_2 - 2\beta$ and determinant $d = \beta^2$. As z_1 and z_2 are the roots of g, $t = 4(1+\beta)^2 - 4\alpha - 2\beta = 2\beta b$, where $b = 8c^2 - 2r - 1$. Hence $t^2 - 4d = t^2 - 4\beta^2 = 4\beta^2(b^2 - 1)$. The eigenvalues are $\beta(-b \pm i\sqrt{1 - b^2})$ and these have modulus 1 if and only if $b^2 - 1 \geq 0$, that is if and only if $4c^2 - 1 \leq r \leq 4c^2$. □

4.3 Notation

For $n \geq 1$, let Q_n denote the set of all $(\alpha, \beta) \in \mathcal{P}$ for which H has an elliptic periodic point of order n. In the notation of Section 2.1, Figure 4 shows, in darker shading, respectively lighter shading, the values of α for which $\mathrm{pos}(\alpha) \in Q_2$, respectively $\mathrm{neg}(\alpha) \in Q_2$.

4.4 Points of Period 3

If P is periodic of order 3, then $HU(P) = R(P)$ so $R(P)$ and $U(P)$ must be in the same orbit. If there is

FIGURE 4. The set Q_2.

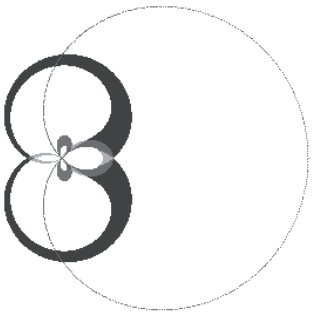

FIGURE 5. The set Q_3.

a symmetric periodic point of order 3, for either R or U, then there is an orbit of the form $\{P, H(P), H^2(P)\}$ where P is R-symmetric, $H(P)$ is U-symmetric, and $U(P) = H^2(P) = R(H(P))$. Call such an orbit *symmetric*. If there is no symmetric orbit then, for any periodic point P of order 3, $\mathrm{orb}(U(P)) = \mathrm{orb}(R(P)) \neq \mathrm{orb}(P)$.

For calculation of the periodic points (z, w) of order 3, there is a polynomial h, of degree 6, such that, with f as in (3–1), the zeros of hf determine the z-coordinates of the eight points, up to multiplicity, where $H^3(z, w) = (z, w)$. Except when $2\beta z^2+\beta^4-2\alpha\beta+1 = 0$, z determines w.

Following a suggestion of the referee, we have used the method described in [Giarrusso and Fisher 95] to factorize h as the product of the two cubics

$$z^3 - \Omega z^2 - (\alpha + \beta^2 - (\beta + 1)\Omega - \beta + 1)z \\ - \alpha(\beta + 1 - \Omega) + \beta^3 - \beta\Omega + 1, \qquad (4\text{–}3)$$

where Ω represents the sum of the z-coordinates of the three points, $(z_i, w_i), 1 \leq i \leq 3$, in an orbit of period 3 and is a root of the quadratic

$$\Omega^2 - (\beta + 1)\Omega + 2\beta^2 + 2 - 2\beta - \alpha. \qquad (4\text{–}4)$$

At each of these points, the Jacobian matrix J_3 has trace $-8z_1z_2z_3 + 2\beta(z_1 + z_2 + z_3) = -8(\alpha(\beta + 1 - \Omega) - \beta^3 + \beta\Omega - 1) + 2\beta\Omega$, and determinant β^3. Writing $z = ve^{\frac{i\theta}{2}}$, $\Omega = \Gamma e^{\frac{i\theta}{2}}$ and $c = \cos\frac{\theta}{2}$, the roots of (4–3) have the form $ve^{\frac{i\theta}{2}}$ where

$$v^3-\Gamma v^2-(r+4c^2-3-2\Gamma c)v+\Gamma r-\Gamma-2rc+8c^3-6c = 0 \qquad (4\text{–}5)$$

and

$$\Gamma = c \pm \sqrt{6 + r - 7c^2}. \qquad (4\text{–}6)$$

The eigenvalues of $J_3(z_i, w_i)$ are $e^{\frac{3i\theta}{2}}(u\pm\sqrt{u^2-1})$, where $u = (4r - 3)\Gamma - 8c(r + 3) + 32c^3$.

If $6+r \geq 7c^2$, each of the cubics in (4–5) has a real root so each orbit of period 3 contains a point (z, w) with z on the U-line. In the situation where z determines w, such a point must be U-symmetric. In the exceptional case, a lengthy calculation shows that the only examples of periodic points (z, w) of order 3 that are not U-symmetric, but for which z is on the U-line, occur with $\beta = 1$ and $\alpha \geq 1$, in which case $\{(z, z), (1 - z, z), (z, 1 - z)\}$ is a symmetric orbit, with $U((1 - z, z)) = (1 - z, z)$, when $z^2 = \alpha - 1$. Therefore, if $6 + r \geq 7c^2$, there are two symmetric orbits of order 3.

The points in the orbit of period 3 determined by either value of Γ are elliptic if $u \in \mathbb{R}$ and $u^2 \leq 1$. Note that $u \in \mathbb{R}$ if either $6 + r \geq 7c^2$ or $r = \frac{3}{4}$. In the latter case, there is a nonsymmetric orbit of elliptic points of order 3 when $28c^2 > 27$ and $-1 \leq 32c^3 - 30c \leq 1$. When $6+r = 7c^2$, there is a symmetric orbit $\mathcal{O}$ of period 3 and multiplicity 2. Here $u = \cos\frac{3\theta}{2}$ so the points in $\mathcal{O}$ are elliptic.

Points α for which there are elliptic points of period three, for $\mathrm{pos}(\alpha)$ or $\mathrm{neg}(\alpha)$, are shown, with various combinations indicated by different levels of shading, in Figure 5.

4.5 Elliptic Periodic Points and the Keep Set

The *forward and backward keep sets* K^+ and K^- of H are defined as follows:

$$K^+ = \{P \in \mathbb{C}^2 : \{H^n(P)\}_{n>0} \text{ is bounded}\};$$
$$K^- = \{P \in \mathbb{C}^2 : \{H^n(P)\}_{n<0} \text{ is bounded}\}.$$

For example, see [Hubbard and Oberste-Vorth 94]. The *keep set* K is $K^+ \cap K^-$. By (3–5), for $P \in \mathbb{C}^2$,

$$P \in K^+ \Leftrightarrow R(P) \in K^- \Leftrightarrow U(P) \in K.$$

Hence K is invariant under both U and R and if P is fixed by either R or U, then $P \in K^+ \Leftrightarrow P \in K$.

It is known, e.g., [Smillie and Buzzard 97, Theorem 13.2] that periodic saddle points must be on the boundary of K, so any periodic point $P \in \operatorname{Int} K$ must be elliptic. We thank the referee for pointing out that, at any such point, H is locally conjugate to its linearization. On a dense subset of $\mathcal{P}$, containing those points (α, β) where the eigenvalues of the Jacobian matrix J_1 are roots of unity at the fixed points, H cannot be locally conjugate to L_H at the fixed points, which are therefore not in $\operatorname{Int} K$.

Experiments investigating whether selected points close to periodic points of orders 1, 2, or 3 are in the keep set produce pictures remarkably similar to those in Figures 2, 4 and 5. It would be interesting to know for which elliptic periodic points P there exists a neighbourhood U of P such that $U \backslash K$ has measure zero. We would also be interested to know more about the sets Q_n and their union Q. In particular, how does the Lebesgue measure of Q_n behave as n increases and is the Lebesgue measure of Q finite? Or could Q be the whole of $\mathcal{P}$?

5. DYNAMICAL RELATIONS

The dynamics of the Hénon map are known to be similar to those of the horseshoe map, see [Smillie and Buzzard 97, Section 5] or [Oberste-Vorth 97, Section 4]. For the reversible Hénon maps considered here, it is possible to be precise about the bounds which occur.

Definition 5.1. For $0 \neq \alpha \in \mathbb{C}$, let

$$b_\alpha = 1 + \sqrt{1 + |\alpha|} \in \mathbb{R}, \tag{5–1}$$

and let

$$\begin{aligned} V &= \{(z,w) : |z| \leq b_\alpha \text{ and } |w| \leq b_\alpha\}; && \text{(5–2)} \\ V^+ &= \{(z,w) : |w| > b_\alpha \text{ and } |w| \geq |z|\}; && \text{(5–3)} \\ V^- &= \{(z,w) : |z| > b_\alpha \text{ and } |z| \geq |w|\}. && \text{(5–4)} \end{aligned}$$

We note that, for the involution R defined in (1–2), $R(V^+) = V^-, R(V^-) = R(V^+)$ and $R(V) = V$.

Proposition 5.2.

(i) $H(V^-) \subseteq V^-$.

(ii) If $P \in V^-$ *then* $||H^n(P)|| \to \infty$ *as* $n \to \infty$.

(iii) $H^{-1}(V^+) \subseteq V^+$.

(iv) If $P \in V^+$ *then* $||H^{-n}(P)|| \to \infty$ *as* $n \to \infty$.

(v) $H(V) \subseteq V \cup V^-$ *and* $H^{-1}(V) \subseteq V \cup V^+$.

(vi) $K \subseteq V$.

Proof: Note that

$$|\alpha| b_\alpha^2 - b_\alpha - 1 = b_\alpha. \tag{5–5}$$

(i) Let $(z,w) \in V^-$ and let $(u,v) = H((z,w))$. Then $|z| \geq |w|$ and $|z| \geq b_\alpha(1+\epsilon)$ for some $\epsilon > 0$. Now

$$\frac{|u|}{|z|} \geq \frac{1}{|z|}(|z|^2 - |w| - |\alpha|) \geq |z| - 1 - \frac{|\alpha|}{|z|}.$$

Using (5–5),

$$\begin{aligned} |z| - 1 - \frac{|\alpha|}{|z|} &\geq b_\alpha(1+\epsilon) - 1 - \frac{|\alpha|}{b_\alpha(1+\epsilon)} \\ &= \frac{(\epsilon^2 + 2\epsilon) b_\alpha - \epsilon + 1}{(1+\epsilon)} \geq 1 + 2\epsilon. \end{aligned}$$

Thus $|u| > |z| > b_\alpha$ and $|v| = |z| \leq |u|$ and so $(u,v) \in V^-$.

(ii) Let $P = (z,w)$ and $(u_n, v_n) = H^n((z,w))$. If $(z,w) \in V^-$, the above argument shows that $|v_{n+1}| = |u_n| \geq (1+2\epsilon)^n |z|$ for some $\epsilon > 0$.

(iii) Using (3–5), $H^{-1}(V^+) = RHR(V^+) = RH(V^-) \subseteq R(V^-) = V^+$.

(iv) follows from (ii) and (3–5), while (v) and (vi) are immediate from (i)-(iv). □

Remark 5.3. The bound b_α is, in a sense, best possible, for if $\alpha = \beta = -1$, then $(z,w) = (-1 - \sqrt{2}, 1 + \sqrt{2})$ is a fixed point of H and $b_\alpha = 1 + \sqrt{2}$.

6. ORBITS

Most of this section is concerned with experimental observations of orbits of $(0,0)$. Orbits in $\operatorname{Int} K$, for a class of volume-preserving maps including H, are discussed in [Bedford et al. 91, Appendix] where it is shown that the closure of the orbit of a generic point is a union of q k-dimensional tori for $k = 1$ or $k = 2$. From Section 3.3 and our experimentation, it appears that if β is a primitive mth root of unity and r is small, then $k = 1$ and $q = m$. However, it appears that, for larger r, q can be nm for an integer $n > 1$. In Section 6.3, we shall describe an example where $m = 25$, but q appears to be 1075. If β is not a root of unity, then k may be 2 and, although $q = 1$ for the linearization and for small r, experimentation suggests that q need not always be 1. Examples with $q > 1$ will be observed in Section 6.2 and Section 6.3.

We restrict our study to the case where $\alpha = re^{i\theta}$ with $r > 0$ so that α determines H. The observations below are based on a Java applet, available

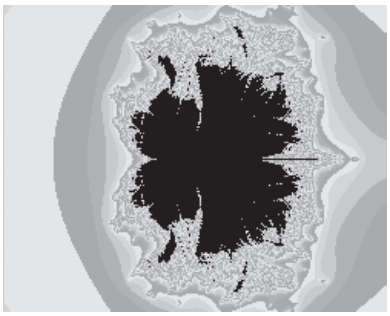

FIGURE 6. The set B.

at http://www.shef.ac.uk/~daj/henon/H.html, which, given r and θ, plots the z-projection $\Pi_z(\text{orb}((0,0))$. The set $B := \{\alpha : (0,0) \in K\}$ is shown in Figure 6. This was plotted using a Java applet based on the bound b_α from Proposition 5.2(ii). Note that if $|\alpha| > 3$, then $b_\alpha < |\alpha|$ and hence, since $H((0,0)) = (\alpha, 0) \in V^-$, $(0,0) \notin K$.

6.1 Coset Orbits

Let i and k be integers with $0 \le i < k$. The subset $\{H^{jk+i}((0,0)) : j \in \mathbb{Z}\}$ of orb$((0,0))$ will be denoted $\mathcal{O}_i^j$ and will be called the *i-th coset orbit* for the subgroup $\langle H^k \rangle$.

Suppose that β is a primitive mth root of unity. Recall from Section 3.3 that the orbit of a generic point under $\langle L_H \rangle$ is dense in a union of m closed curves whose z-projections are ellipses. For small r, $\Pi_z(\text{orb}((0,0)))$ appears to be dense in the union of m ovals, each corresponding to one of the coset orbits $\mathcal{O}_i^m$. In Figure 7, where $\theta = \frac{\pi}{3}, \frac{\pi}{2}, \frac{\pi}{3}$ and $m = 6, 4, 3$, respectively, and in Figure 8, where $\theta = \frac{\pi}{2}$ and $m = 4$, each coset orbit is shaded differently.

As r increases towards the boundary of B, the m ovals lose their convexity and smoothness, but remain closed curves. For example, see the top two coset orbits in Figure 14.

Where the line $\{xe^{i\theta} : x \ge 0\}$ crosses the boundary of B, the closed curves are distorted ovals for α close to the boundary, but closer to their original oval shape away from the boundary. Figure 9 shows the coset orbits $\mathcal{O}_3^4$ for

FIGURE 7. Orbits for $m = 6, 4, 3$ respectively.

FIGURE 8. Orbits for $r = 0.1, 0.24, 0.246, 0.249; \theta = \frac{\pi}{2}$.

FIGURE 9. Coset orbits for $r = 0.1, 0.23, 0.24, 0.3; \theta = \frac{\pi}{2}$.

$r = 0.1, 0.23, 0.24, 0.3$, and $\theta = \pi/2$. When $r = 0.2462$, the orbit is unbounded. However, the orbit appears to be bounded for $r = 0.24853$ (see Figure 10) and has a reasonably simple shape for $r = 0.3$.

For fixed small r, the eccentricity of the ovals decreases with θ. This can be seen in Figure 7. If β is not a root of unity, the pictures generated by the applet are consistent with orb$((0,0))$ being dense in a finite union of two-dimensional tori. For example, see the orbits in Figure 1.

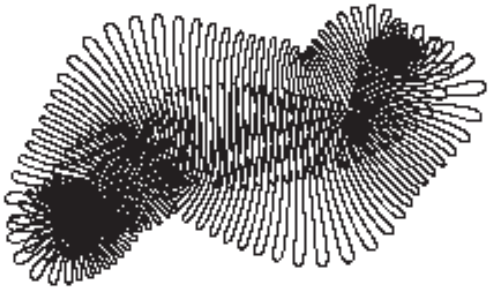

FIGURE 10. A coset orbit for $r = 0.24853$ and $\theta = \frac{\pi}{2}$.

6.2 Islands

It seems likely that the boundedness of the orbits discussed above is influenced by elliptic fixed points close to the centre of the ovals. Orbits of points close to this fixed point are similar in shape to those of $(0,0)$. There are points on the set B where orb$((0,0))$ appears to be influenced by elliptic periodic points of order greater than one. For example, for the point $\alpha = e^{\pi i/3}$, which appears to be on the boundary of an "island" of B, $(0,0)$ is an elliptic periodic point of order 4. For other val-

FIGURE 11. Four coset orbits with a close-up of one.

ues of α on this island, the orbits of $(0,0)$ appear to be influenced by such periodic points. Figure 11 shows orb$((0,0))$ when $r = 1$ and $\beta = e^{0.32\pi i}$, a primitive 25th root of unity. On the left are the four coset orbits for H^4 with a close up of one of these, decomposed as the union of 25 coset orbits for H^{100}, on the right. This suggests that there may be an orbit $\{P_1, P_2, P_3, P_4\}$ of elliptic periodic points of order 4, such that there exist neighbourhoods $N(P_1), N(P_2), N(P_3), N(P_4)$ with orb$((0,0)) \subset \bigcup_{1\le i\le 4} N(P_i)$ and $H(N(P_i)) \subset N(P_{i+1 \bmod 4})$.

Values of the parameters at which we have observed similar behaviour are shown on the left of Table 1. Figure 12 shows orbits for the first three rows of the left hand table.

FIGURE 12. Orbits for $\alpha = 0.55e^{\frac{4\pi i}{9}}, \alpha = 0.939e^{\frac{\pi i}{5}}$, $\alpha = 0.9385e^{0.187\pi i}$.

6.3 Bifurcation

Within the period 4 island, there is some bifurcation. Figure 13 shows the 12 coset orbits for H^{12}, where $\alpha = 0.98e^{0.305\pi i}$, on the left, and $\alpha = 0.95e^{0.3016\pi i}$, indicating bifurcation from 4 to 12. The coset orbits for $\alpha = 0.98e^{0.305\pi i}$ are smoother than those for $\alpha = 0.95e^{0.3016\pi i}$. Other values of α for which we have observed bifurcation are shown on the right in Table 1.

If $\theta = 0.16\pi$, so that $m = 25$, and r is about 0.82 then the coset orbits for H^{25} appear as 25 closed curves. However for $r = 0.83$, these each bifurcate into 43 closed curves, suggesting that the closure of orb$((0,0))$ is the union of 1075 1-dimensional tori. The $11th$ coset orbits for $r = 0.82$, 0.826 and 0.83 are shown in Figure 14.

r	θ	period
0.55	$4\pi/9$	5
0.939	0.2π	11
0.9385	0.187π	116
0.790666	0.295704π	21
0.788	0.36363636π	9
0.696059	0.520135π	10

r	θ	bifurcation
0.987	0.30631π	$12 \mapsto 60$
0.983	0.306π	$12 \mapsto 444$
0.943	0.19825π	$11 \mapsto 99$
0.962	0.198π	$11 \mapsto 253$
0.661	0.495π	$5 \mapsto 90$
0.661	0.49498971π	$90 \mapsto 360$

TABLE 1. Parameters for periodic behaviour and bifurcation.

FIGURE 13. Bifurcation into 12.

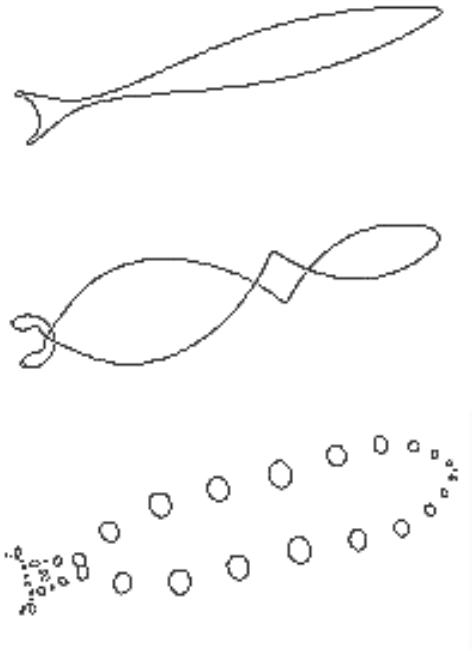

FIGURE 14. Bifurcation of one coset orbit.

REFERENCES

[Bedford et al. 91] E. Bedford and J. Smillie. "Polynomial diffeomorphisms of $\mathbb{C}^2$. II: stable manifolds and recurrence." *J. Amer. Math. Soc.* **4** (1991), 657–679.

[Bedford et al. 93] E. Bedford, M. Lyubich and J. Smillie. "Distribution of periodic points of polynomial diffeomorphisms of $\mathbb{C}^2$." *Invent. Math.* **114** (1993), 277–288.

[Devaney 76] R. Devaney. "Reversible diffeomorphisms and flows." *Trans. Amer. Math. Soc.* **218** (1976), 89–113.

[Devaney 84] R. Devaney. "Homoclinic bifurcations and the area-conserving Hénon map." *J. Differential Equations* **51** (1984), 254–266.

[Devaney 89] R. Devaney. *An Introduction to Chaotic Dynamical Systems*, 2nd edition, Addison Wesley, Redwood City, 1989.

[Fornæss 96] J. E. Fornæss. *Dynamics in Several Complex Variables.* CBMS Regional Conference Series in Mathematics, **87**, Amer. Math. Soc., Providence, 1996.

[Friedland et al. 89] S. Friedland and J. Milnor. "Dynamical properties of plane polynomial automorphisms." *Ergod. Th.& Dynam. Sys.* **9** (1989), 67– 99.

[Giarrusso and Fisher 95] D. Giarrusso and Y. Fisher. "A parameterization of the period 3 hyperbolic components of the Mandelbrot set." *Proc. Amer. Math. Soc.* **123** (1995), 3731–3737.

[Hale and Koçak 91] J. Hale and H. Koçak. *Dynamics and Bifurcations.* Springer-Verlag, New York, 1991.

[Jordan 93] D. A. Jordan. "Iterated skew polynomial rings and quantum groups." *J. Algebra* **156** (1993), 194–218.

[Hubbard and Oberste-Vorth 94] J. H. Hubbard and R. W. Oberste-Vorth. "Hénon mappings in the complex domain I: the global topology of dynamical space." *Publ. Math. IHES* **79** (1994), 5–46.

[Oberste-Vorth 97] R. W. Oberste-Vorth. "An introduction to multi-dimensional complex dynamics: Hénon mappings in $\mathbb{C}^2$." *Nonlinear analysis, Methods & Applications* **30** (1997), 2143–2154.

[Smillie and Buzzard 97] J. Smillie and G. T. Buzzard. "Complex Dynamics in Several Variables." in *Flavors of Geometry*, S. Levy, ed., pp. 117–150, Cambridge University Press, 1997.

[Zehnder 77] E. Zehnder. "A simple proof of a theorem by C. L. Siegel." in *Geometry and Topology*, J. Palis and M.́do Carmo, eds., Lecture Notes in Math., vol. 597, pp. 855–866, Springer-Verlag, New York, 1977.

C. R. Jordan, The Open University in Yorkshire, 2 Trevelyan Square, Boar Lane, Leeds LS1 6ED, UK
(c.r.jordan@open.ac.uk)

D. A. Jordan, Department of Pure Mathematics, University of Sheffield, Hicks Building, Sheffield S3 7RH, UK
(d.a.jordan@sheffield.ac.uk)

J. H. Jordan, Department of Probability and Statistics, University of Sheffield, Hicks Building, Sheffield S3 7RH, UK
(jonathan.jordan@sheffield.ac.uk)

Received November 27, 2000; accepted in revised form November 28, 2001.

Le critère de Beurling et Nyman pour l'hypothèse de Riemann: aspects numériques

Bernard Landreau et Florent Richard

SOMMAIRE

2000 AMS Subject Classification: Primary 11M26; Secondary 46E99

Keywords: Riemann Hypothesis, Nyman-Beurling Criterion

Soit $\mathcal{B}$ le sous-espace de $\mathcal{H} = L^2(0,+\infty)$ composé des fonctions f telles que $f(t) = \sum_{k=1}^{n} c_k \rho\left(\frac{\theta_k}{t}\right)$, $n \in \mathbb{N}$, $c_k \in \mathbb{C}$, $0 < \theta_k \leq 1$, pour $1 \leq k \leq n$, où $\rho(t)$ désigne la partie fractionnaire de t. Notons aussi χ la fonction caractéristique de l'intervalle]0,1]. Un résultat bien connu de Nyman et Beurling [Nyman 50, Beurling 55] implique que l'hypothèse de Riemann est vraie si et seulement si $d(\chi, \mathcal{B}) = 0$. Nous présentons ici divers résultats numériques concernant l'approximation de χ par des éléments de $\mathcal{B}$.

Let $\mathcal{B}$ be the subspace of $\mathcal{H} = L^2(0,+\infty)$ consisting of the functions f such that $f(t) = \sum_{k=1}^{n} c_k \rho\left(\frac{\theta_k}{t}\right)$, $n \in \mathbb{N}$, $c_k \in \mathbb{C}$, $0 < \theta_k \leq 1$, for $1 \leq k \leq n$, where $\rho(t)$ denotes the fractional part of t. We also denote by χ the characteristic function of $(0,1]$. A well known result of Nyman and Beurling [Nyman 50, Beurling 55] implies that the Riemann hypothesis holds if and only if $d(\chi, \mathcal{B}) = 0$. We present several numerical results about the approximation of χ by elements of $\mathcal{B}$.

1. INTRODUCTION

On considère l'espace de Hilbert $\mathcal{H} = L^2(0,+\infty)$ et le sous-espace $\mathcal{B}$ de $\mathcal{H}$ des fonctions de la forme

$$f(t) = \sum_{k=1}^{n} c_k \rho\left(\frac{\theta_k}{t}\right),$$

$n \in \mathbb{N}$, $c_k \in \mathbb{C}$, $0 < \theta_k \leq 1$, pour $1 \leq k \leq n$, où $\rho(t)$ désigne la partie fractionnaire de t. On note χ la fonction caractéristique de l'intervalle $]0,1]$.

Il résulte des travaux de Nyman et Beurling [Nyman 50, Beurling 55] que l'hypothèse de Riemann équivaut au fait que χ est limite dans $\mathcal{H}$ d'une suite d'éléments de $\mathcal{B}$, autrement dit au fait que $d(\chi, \mathcal{B}) = 0$ où d désigne la distance naturelle sur $\mathcal{H}$ induite par le produit scalaire $< f, g >= \int_0^{+\infty} f(t)\bar{g}(t)dt$.

Si l'on note, pour $0 < \lambda \leq 1$, $\mathcal{B}_\lambda$ le sous-espace de $\mathcal{B}$ des fonctions f telles que $\min_{1 \leq k \leq n} \theta_k \geq \lambda$ et

$D(\lambda) = d(\chi, \mathcal{B}_\lambda)$, le théorème de Beurling et Nyman affirme donc l'équivalence entre l'hypothèse de Riemann et la convergence de $D(\lambda)$ vers 0 quand λ tend vers 0.

On dispose sur $D(\lambda)$ de la minoration suivante établie par L. Báez-Duarte, M. Balazard, E. Saias et le premier auteur [Báez-Duarte et al. 00].

Théorème 1.1. (Báez-Duarte et al.)

$$\liminf_{\lambda\to 0} D(\lambda)\sqrt{\log(1/\lambda)} \geq C,$$

où la constante C est définie par

$$C := \left(\sum_\beta \frac{1}{|\beta|^2}\right)^{\frac{1}{2}}$$

la sommation portant sur les zéros β de la fonction ζ de partie réelle $\frac{1}{2}$, chaque zéro étant compté une seule fois, quel que soit son ordre de multiplicité.

Signalons que ce résultat a été récemment amélioré par J.-F. Burnol [Burnol 01] qui a établi le théorème suivant.

Théorème 1.2. (Burnol.)

$$\liminf_{\lambda\to 0} D(\lambda)\sqrt{\log(1/\lambda)} \geq \left(\sum_\beta \frac{m(\beta)^2}{|\beta|^2}\right)^{\frac{1}{2}}$$

où la sommation porte toujours sur les zéros β de la fonction ζ de partie réelle $\frac{1}{2}$ et $m(\beta)$ désigne la multiplicité de β.

On remarquera que d'une part si l'hypothèse de Riemann est fausse les deux théorèmes précédents sont triviaux puisque le membre de gauche vaut $+\infty$. D'autre part, on sait [Rosser 39, p. 29] que sous l'hypothèse de Riemann on a

$$\sum_\beta \frac{m(\beta)}{|\beta|^2} = 2 + \gamma - \log 4\pi,$$

où γ désigne la constante d'Euler. Cela permet donc d'affirmer finalement à partir du résultat de Burnol que

Théorème 1.3.

$$\liminf_{\lambda\to 0} D(\lambda)\sqrt{\log(1/\lambda)} \geq \sqrt{2 + \gamma - \log 4\pi}.$$

Les résultats numériques mentionnés en [Báez-Duarte et al. 00] et développés dans ce qui suit ont conduit leurs auteurs à formuler la conjecture suivante.

Conjecture 1.4. *On a*

$$\lim_{\lambda\to 0} D(\lambda)\sqrt{\log(1/\lambda)} = \sqrt{2 + \gamma - \log 4\pi}.$$

Nous présentons ici un certain nombre de résultats numériques concernant l'approximation de la fonction χ par des éléments de $\mathcal{B}$.

2. PRODUITS SCALAIRES

La plupart des calculs présentés ici nécessitent l'évaluation des produits scalaires entre les fonctions g_θ définies par $g_\theta(t) := \rho(\theta/t)$ pour $\theta > 0$.

On utilise pour cela les formules[1] explicites. de Vassiounine [Vassiounine 96] suivantes s'appliquant aux fonctions $e_n(t) = g_{\frac{1}{n}}(t) = \rho\left(\frac{1}{nt}\right)$, $n \geq 1$.

Théorème 2.1. (Vassiounine.) *On a pour $n, m \geq 1$*

$$< e_n, e_m > = \frac{\log(2\pi) - \gamma}{2}\left(\frac{1}{n} + \frac{1}{m}\right) + \frac{m-n}{2mn}\log\left(\frac{n}{m}\right) - \frac{\pi\omega}{2nm}\sum_{k=1}^{n_0-1}\rho\left(\frac{km_0}{n_0}\right)\cot\frac{\pi k}{n_0} - \frac{\pi\omega}{2mn}\sum_{k=1}^{m_0-1}\rho\left(\frac{kn_0}{m_0}\right)\cot\frac{\pi k}{m_0},$$

où $\omega = (n, m)$, $n = \omega n_0$, $m = \omega m_0$ et γ désigne la constante d'Euler.

A partir de ces formules, il est aisé d'établir les résultats plus généraux suivants.

Proposition 2.2. *On a*

(i) $< g_\theta, g_\theta >= \theta < e_1, e_1 >= \theta(\log(2\pi) - \gamma)$ *pour tout* $\theta > 0$;

(ii) $< g_{\frac{p}{q}}, g_{\frac{p'}{q'}} >= pp' < e_{p'q}, e_{pq'} >$, *pour* $p, p', q, q' \geq 1$, $p \leq q$, *et* $p' \leq q'$;

(iii) $< \chi, g_\theta >= \theta(-\log\theta + 1 - \gamma)$, *pour tout* θ, $0 < \theta \leq 1$.

3. CALCULS DE LA DISTANCE d_n

D'après la formule d'inversion de Mœbius et le théorème des nombres premiers, on a

$$\sum_{k=1}^{\infty}\mu(k)\rho\left(\frac{1}{kt}\right) = -1 \quad \text{pour tout } t > 0.$$

[1] En fait, Vassiounine a donné des formules pour $< e_n - e_1/n, e_m - e_1/m >$.

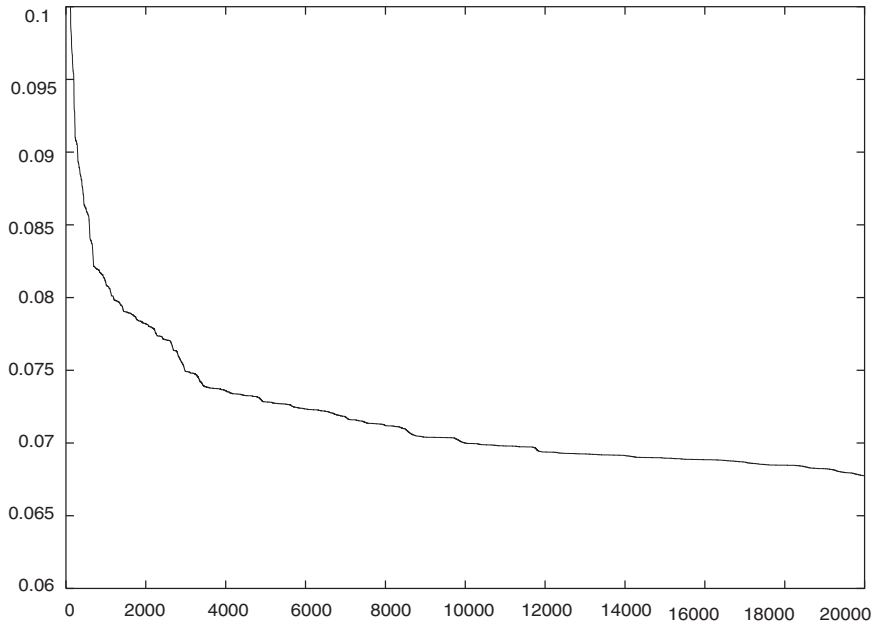

FIGURE 1. La distance d_n de 1 à 20 000.

Une première approche naturelle consiste donc à se restreindre aux fonctions de $\mathcal{B}$ ayant des paramètres θ rationnels et plus précisément de la forme $\frac{1}{k}$, $k \in \mathbb{N}^*$. On considère pour cela les fonctions e_k définies par $e_k(t) = \rho(\frac{1}{kt})$ pour $k \geq 1$ et on note V_n le sous-espace vectoriel engendré par la famille $(e_1, \ldots, e_n)$. On s'intéresse alors à la distance $d_n := d(\chi, V_n)$ dans $\mathcal{H}$.

Les résultats numériques qui suivent plaident en faveur de la conjecture suivante déjà énoncée en [Báez-Duarte et al. 00] et similaire à la conjecture précédente.

Conjecture 3.1. *On a*

$$d_n^2 \sim \frac{2 + \gamma - \log 4\pi}{\log n} \qquad \textit{quand } n \to +\infty.$$

Rappelons que la convergence de d_n vers zéro entraîne automatiquement l'hypothèse de Riemann puisque $D(\frac{1}{n}) \leq d_n$ mais en revanche la réciproque n'est pas claire[2].

Des calculs sur d_n ont déjà été présentés en [Báez-Duarte et al. 00], nous les avons prolongés jusqu'à $n = 20\,000$. La méthode est fondée sur une orthonormalisation de Gram-Schmidt de la base des $(e_k)_{k\geq 1}$. Nous avons également utilisé d'autres méthodes de calculs comme la formule

$$d_n^2 = \frac{\det \mathrm{Gram}\,(e_1, \ldots, e_n, \chi)}{\det \mathrm{Gram}\,(e_1, \ldots, e_n)},$$

ou encore la méthode du gradient conjugué ou bien encore la méthode d'orthogonalisation QR; ces méthodes sont nettement plus lentes mais permettent de confirmer (au moins jusqu'à $n = 10\,000$) les résultats précédemment trouvés.

On peut observer, Figure 1, que la décroissance de d_n est relativement lente et irrégulière.

En grossissant le graphe, par exemple entre $0,07$ et $0,08$, on observe, Figure 2, une succession de ruptures de pente difficiles à interpréter. En effet, s'il est clair, que pour les petites valeurs de n, ce sont les nombres premiers qui provoquent une pente importante pour le graphe de d_n, cela devient beaucoup plus complexe par la suite et mériterait certainement une étude approfondie.

Expérimentalement, la suite $d_n\sqrt{\log n}$ converge et l'on peut calculer aisément par la méthode des moindres carrés le nombre réel $a = a_N$ qui minimise $\sum_{n=1}^{N} |d_n - a/\sqrt{\log n}|^2$. On trouve pour $N = 20\,000$, $a_N \approx 0,21377$. La proximité de ce nombre avec $(2 + \gamma - \log 4\pi)^{\frac{1}{2}} \approx 0,21492$ apporte du crédit aux conjectures précédentes. La comparaison du graphe de d_n et de celui de sa valeur asymptotique conjecturée, Figure 3, explique le léger décalage de la constante a_N, sans que celui-ci ne soit vraiment significatif.

4. CALCULS DE LA PROJECTION ORTHOGONALE DE χ SUR V_n

Dans cette section, nous nous proposons d'étudier la projection orthogonale de χ notée p_n sur le sous-espace V_n.

[2]En fait, la question est maintenant réglée, cf Addendum.

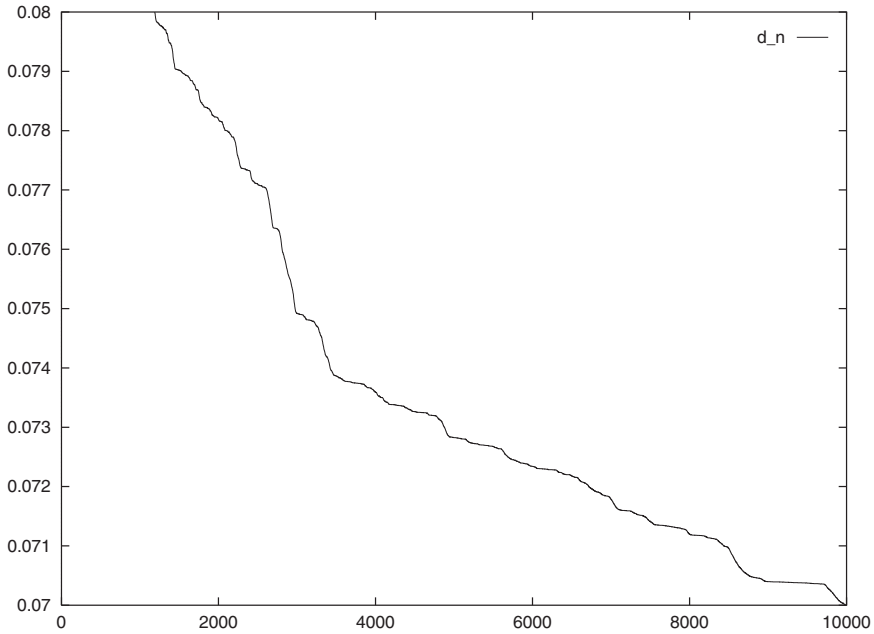

FIGURE 2. La distance d_n de 1 à 10 000, fenêtre $[0,07; 0,08]$.

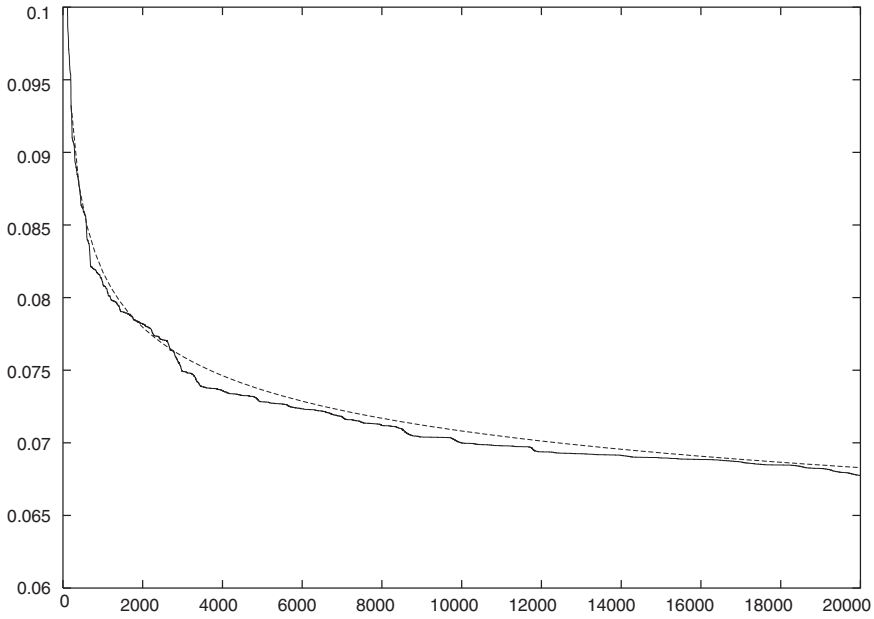

FIGURE 3. La distance d_n de 1 à 20 000 et $\sqrt{2+\gamma-\log 4\pi}/\sqrt{\log n}$.

Nous nous sommes intéressés principalement aux coordonnées de p_n dans la base des $(e_k)_{1\leq k\leq n}$. On écrit pour cela

$$p_n = a_{n,1}e_1 + a_{n,2}e_2 + \cdots + a_{n,n}e_n.$$

On s'intéresse ici au comportement des $a_{n,k}$. On s'aperçoit rapidement sur les premières valeurs, Table 1, que les $a_{n,k}$, pour k fixé, convergent expérimentalement vers $-\mu(k)$ lorsque n tend vers l'infini. Cela ne surprend pas: d'une part, comme nous l'avons déjà mentionné, la suite de fonctions $u_n := -\sum_{k=1}^{n} \mu(k)e_k$ converge simplement vers χ sur $]0,+\infty[$, d'autre part il résulte des travaux exposés en [Báez-Duarte 01] que l'hypothèse de Riemann implique cette propriété de convergence de $a_{n,k}$ vers $-\mu(k)$. Malheureusement la suite de fonctions (u_n) ne converge pas vers χ dans $\mathcal{H}$ [Báez-Duarte 01].

$n\backslash k$	1	2	3	4	5	6	7	8	9	10
1	0,335									
2	-0,829	1,900								
3	-0,977	1,138	1,349							
4	-0,925	0,856	1,049	0,700						
5	-0,927	0,859	0,863	0,296	0,751					
6	-0,931	0,863	0,880	0,302	0,777	-0,060				
7	-0,939	0,891	0,894	0,195	0,673	-0,399	0,614			
8	-0,938	0,894	0,897	0,172	0,671	-0,410	0,575	0,081		
9	-0,937	0,896	0,888	0,179	0,660	-0,416	0,567	0,030	0,086	
10	-0,939	0,901	0,881	0,170	0,695	-0,416	0,573	0,047	0,152	-0,126
11	-0,945	0,918	0,890	0,145	0,734	-0,487	0,545	0,029	0,106	-0,416
12	-0,945	0,918	0,890	0,144	0,735	-0,490	0,546	0,029	0,106	-0,418
13	-0,944	0,916	0,891	0,155	0,724	-0,466	0,510	0,026	0,090	-0,422
14	-0,950	0,926	0,890	0,138	0,735	-0,517	0,615	0,019	0,102	-0,421
15	-0,950	0,924	0,895	0,130	0,749	-0,519	0,599	0,042	0,100	-0,416
16	-0,951	0,925	0,900	0,128	0,759	-0,528	0,629	-0,025	0,106	-0,414
17	-0,952	0,927	0,908	0,125	0,768	-0,546	0,625	0,015	0,055	-0,422
18	-0,952	0,927	0,909	0,125	0,770	-0,551	0,626	0,021	0,043	-0,422
19	-0,953	0,929	0,907	0,122	0,773	-0,543	0,619	0,019	0,068	-0,454
20	-0,953	0,925	0,909	0,130	0,768	-0,554	0,630	0,019	0,026	-0,381

TABLE 1. Les coefficients $a_{n,k}$ pour $1 \leq n \leq 20$ et $1 \leq k \leq 10$

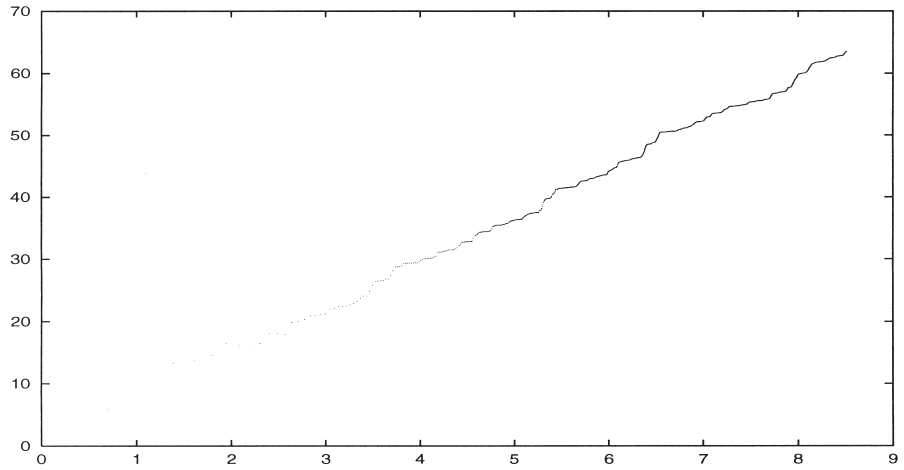

FIGURE 4. Les coefficients $b_{n,1}^{-1}$ en fonction de $\log n$ de 1 à 5 000.

On peut penser pour $a_{n,k}$ à une expression de la forme $-\mu(k)(1-\frac{\log k}{\log n})$. En effet A. Selberg utilise dans [Selberg 46] une approximation de $1/\zeta(s)$ sur la droite critique par des polynômes de Dirichlet de la forme $\sum_{k=1}^{n}\mu(k)(1-\frac{\log k}{\log n})\frac{1}{k^s}$, ce qui incite à tenter d'approximer χ par la suite de fonctions

$$-\sum_{k=1}^{n}\mu(k)(1-\frac{\log k}{\log n})e_k.$$

Ceci nous amène à étudier, à k fixé, plus précisément les quantités $b_{n,k} := |a_{n,k}+\mu(k)|$.

Prenons pour commencer $k=1$, le tracé de $b_{n,1}^{-1}$ en fonction de $\log n$, Figure 4, met en évidence, hormis quelques valeurs initiales très particulières, la quasi-linéarité en $\log n$. Cependant, il est clair que la formule $\frac{1}{\log k}$ pour la pente ne convient pas pour cette valeur particulière de k.

Un calcul de la meilleure constante c, au sens des moindres carrés, telle que $b_{n,1} = c/\log n$ donne la valeur $c_1 \approx 0,133$. On observe ce même comportement comme le montre la Figure 5 pour $k = 2, 3, 4, 5$ et $k \geq 6$ à condition que k soit sans facteur carré.

En revanche, pour les k plus grands que 4 et ayant un facteur carré, par exemple pour $k=8$, Figure 6, on obtient pour $a_{n,8}$ un graphe qui tend vers zéro en oscillant de façon irrégulière.

On pense naturellement à calculer pour chaque k sans facteur carré la meilleure constante c_k telle que $b_{n,k} = \frac{c_k}{\log n}$. Pour cela on minimise par exemple la quantité

$$\sum_{n=1}^{N}(b_{n,k}^{-1} - c\log n)^2.$$

Les résultats sont donnés Table 2 pour $1 \leq k \leq 19$, $\mu(k) \neq 0$ et $N = 5000$.

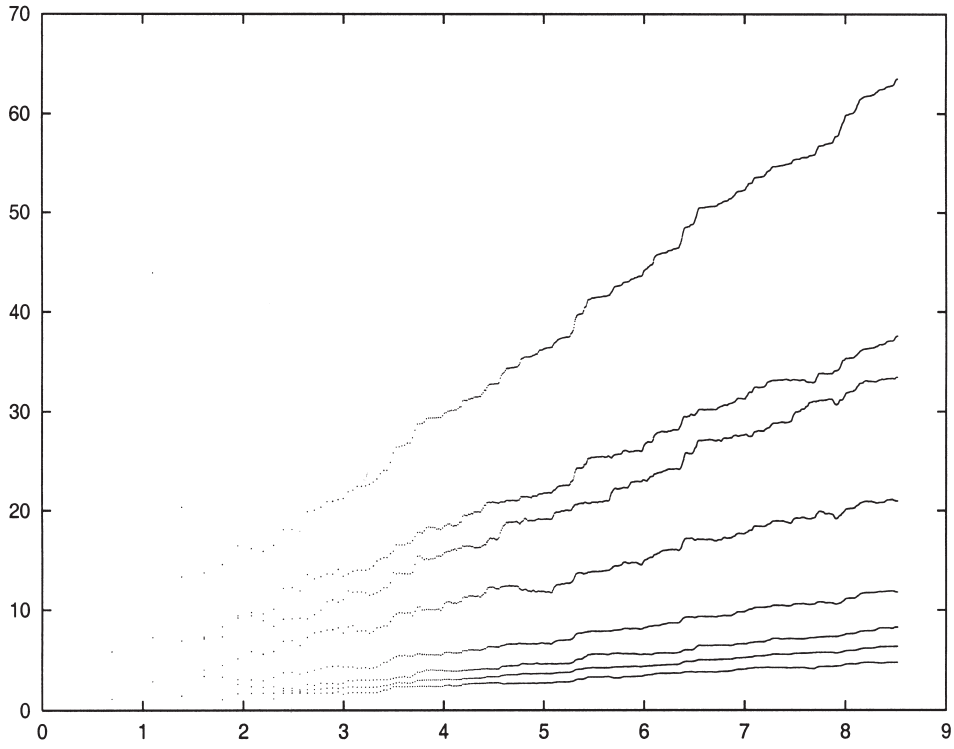

FIGURE 5. Les coefficients $b_{n,k}^{-1}$ en fonction de $\log n$, $k = 1, 2, 3, 4, 5, 6, 7, 10$ et $1 \leq n \leq 5000$.

L'étude de la dépendance de ces constantes en fonction de $\log k$, Figure 7, met en évidence une proportionnalité avec $\log k$ mais avec une sérieuse dispersion pour les grandes valeurs de k.

On peut aussi d'un autre côté, pour apprécier le rôle des différentes fonctions e_k, considérer à n fixé, la suite des entiers k ordonnés suivant l'ordre décroissant des $|a_{n,k}|$. On obtient par exemple pour $n = 100$ le classement suivant.

1 2 3 5 7 6 11 13 10 17 15 14 19 23 21 33 22 29 31 37 26 47 39 65 41 35 74 43 53 59 46 55 34 51 77 38 57 30 67 97 61 42 70 87 66 58 71 73 79 78 62 82 89 69 91 85 94 93 95 83 86 92 90 68 84 44 36 12 81 4 28 88 54 18 40 99 56 80 100 52 63 72 60 9 48 98 49 45 25 50 16 27 32 75 24 20 96 64 76 8.

La structure multiplicative des entiers y apparaît clairement sans pour autant qu'il se dégage une règle simple d'ordonnancement.

k	1	2	3	5	6	7	10
c_k	0,133	0,233	0,253	0,711	1,337	1,048	1,748

k	11	13	14	15	17	19
c_k	1,487	1,641	2,092	1,881	1,960	2,021

TABLE 2. Les constantes c_k pour $1 \leq k \leq 19$, $\mu(k) \neq 0$, calculées pour $N = 5000$.

Pour conclure cette section, disons que l'expression $-\mu(k)(1 - \frac{\log k}{\log n})$ pour le coefficient $a_{n,k}$ est effectivement une formule qui colle grossièrement au comportement asymptotique de $a_{n,k}$ quand n et k tendent vers l'infini, elle n'en reste pas moins approximative et nécessiterait une nette amélioration. On verra un peu plus loin ce que donne l'étude directe de la suite de vecteurs définis par cette expression.

5. CALCULS DE LA DISTANCE DE χ À CERTAINES SUITES

Une autre approche consiste à étudier des suites candidates pour converger dans $\mathcal{H}$ vers χ. On considère tout d'abord la suite naturelle de vecteurs

$$u_n = -\sum_{k=1}^{n} \mu(k) e_k.$$

On observe sur la Figure 8 que, ainsi qu'il a été mentionné plus haut, la suite u_n ne converge pas vers χ dans $\mathcal{H}$, cependant la distance $d(\chi, u_n)$ semble rester bornée. On notera de plus une corrélation certaine et finalement assez naturelle entre les variations de $d(\chi, u_n)$ et les variations de la fonction sommatoire M de la fonction de Mœbius, cela apparaît clairement en comparant le graphe

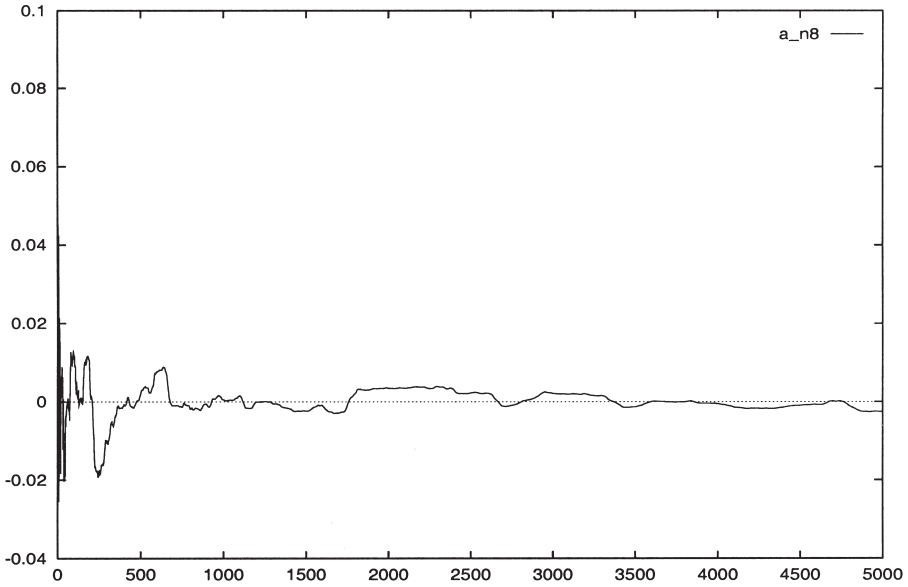

FIGURE 6. Les coefficients $a_{n,8}$ en fonction de n, $8 \leq n \leq 5000$.

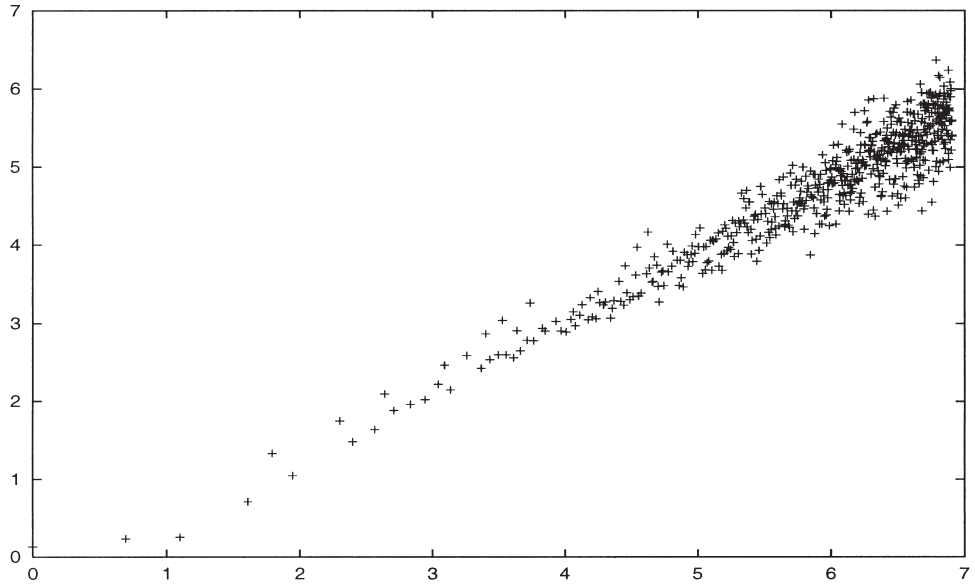

FIGURE 7. Les constantes c_k en fonction de $\log k$ pour $1 \leq k \leq 1000$, $\mu(k) \neq 0$, calculées pour $N = 5000$.

de $d(\chi, u_n)$ avec celui de $x \mapsto |M(x)|/\sqrt{x}$. On considère alors la suite de vecteurs

$$v_n = -\sum_{k=1}^{n} \mu(k) \left(1 - \frac{\log k}{\log n}\right) e_k,$$

ce qui semble être un bon candidat compte-tenu de l'étude de la projection orthogonale de χ sur V_n réalisée à la section précédente.

Le graphe de la distance $d(\chi, v_n)$ comparé à celui de d_n est le suivant, Figure 9.

On constate maintenant une décroissance générale vers zéro mais disons à "bonne" distance de d_n. On peut encore approcher un peu plus près le graphe de d_n. Après diverses expérimentations, il apparaît que la suite définie par

$$w_n = -e_1 - (1 + \frac{1}{\log n}) \sum_{k=2}^{n} \mu(k) \left(1 - \frac{\log k}{\log n}\right) e_k.$$

réalise une bonne performance. Le graphe de la distance $d(\chi, w_n)$ comparé à celui de d_n et $d(\chi, v_n)$ est donné Figure 10.

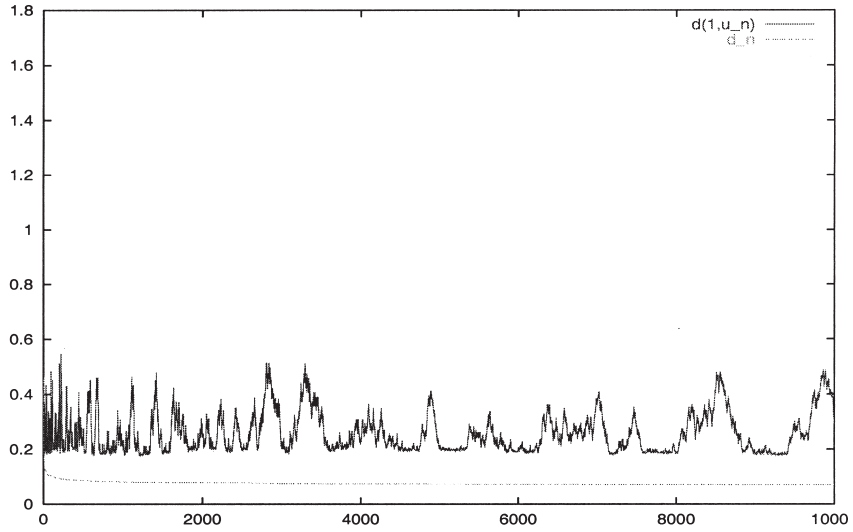

FIGURE 8. Distances d_n et $d(\chi, u_n)$, $1 \leq n \leq 10\,000$.

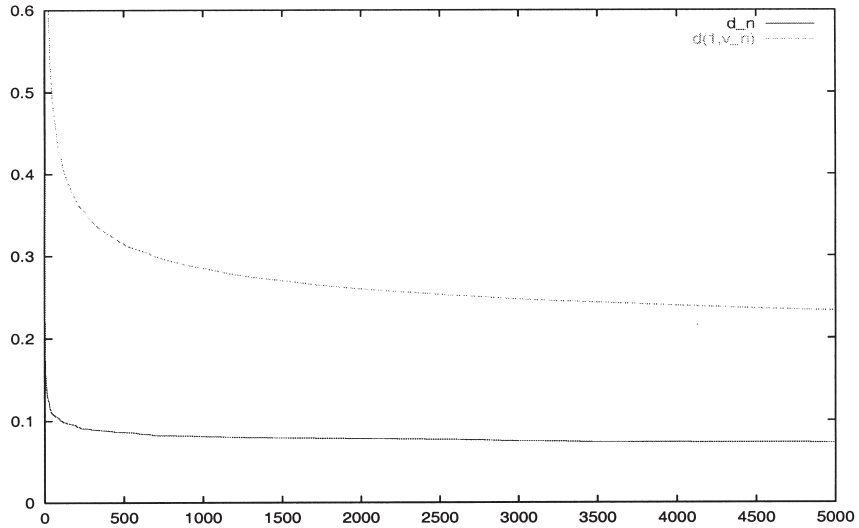

FIGURE 9. Comparaison de d_n et $d(\chi, v_n)$ pour $n \leq 5000$.

Le graphe est maintenant nettement plus proche de celui de d_n. Malgré diverses tentatives nous n'avons pu trouver de suites véritablement meilleures pour approcher la fonction χ. Compte-tenu de ces résultats, on peut raisonnablement conjecturer que les suites v_n et w_n convergent vers χ dans $\mathcal{H}$; il reste donc à étudier les expressions $\|\chi - v_n\|^2$ ou $\|\chi - w_n\|^2$ ce qui n'est pas aisé malgré leurs formes tout à fait explicites obtenues grâce aux formules de Vassiounine déjà mentionnées.

6. RECHERCHE DES MEILLEURS θ

6.1 Introduction

Dans ce qui précède, nous nous sommes toujours jusque là restreints à l'étude de l'approximation de χ par des combinaisons linéaires de fonctions $e_k(t) = \rho(\frac{1}{kt})$. Dans cette section, nous nous proposons maintenant d'élargir cette étude au cas plus général où les paramètres θ sont quelconques dans $]0, 1]$. On considère pour cela pour $N \geq 1$ le sous-ensemble $\mathcal{B}_N$ de $\mathcal{B}$ des fonctions de la forme

$$f(t) = \sum_{i=1}^{N} c_i g_{\theta_i}(t),\ c_i \in \mathbb{C}, \theta_i \in]0, 1], \text{ pour } 1 \leq i \leq N,$$

où $g_\theta(t) := \rho\left(\frac{\theta}{t}\right)$.

Notre démarche va consister à trouver expérimentalement à N fixé les $(\theta_i)_{1 \leq i \leq N}$ qui engendrent par combinaisons linéaires les meilleures approximations de χ. Autrement dit, en notant

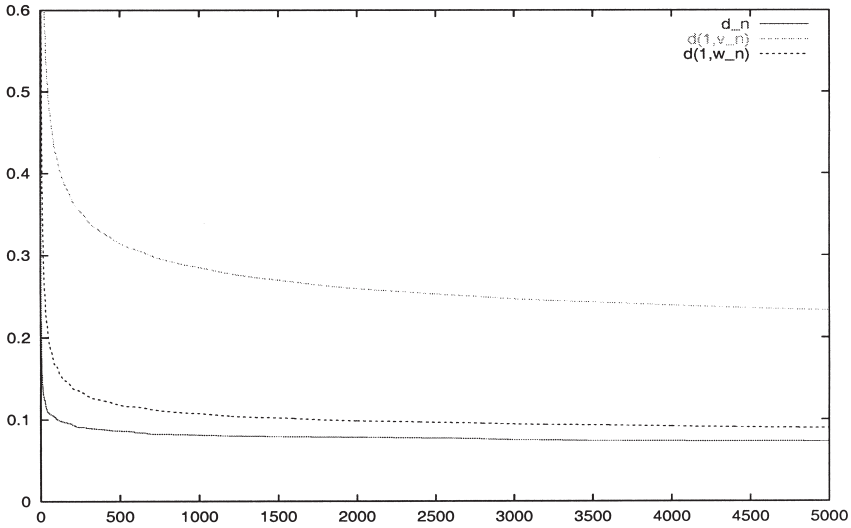

FIGURE 10. Comparaison de d_n, $d(\chi, v_n)$ et $d(\chi, w_n)$ pour $n \leq 5000$.

$V(\theta_1, \ldots, \theta_N) = \text{Vect}\,(g_{\theta_1}, \ldots, g_{\theta_N})$, nous recherchons les $(\theta_i)_{1 \leq i \leq N}$ rendant minimale la distance $d(\chi, V(\theta_1, \ldots, \theta_N))$. Signalons de suite que $\mathcal{B}_N$ n'étant pas un sous-espace vectoriel de $\mathcal{H}$, l'existence d'un N-uplet $(\theta_1, \ldots, \theta_N)$ tel que $d(\chi, \mathcal{B}_N) = d(\chi, V(\theta_1, \ldots, \theta_N))$ n'est pas assurée.

Notons que par un résultat classique de géométrie euclidienne on a

$$d(\chi, V(\theta_1, \ldots, \theta_N)) = \frac{\det \text{Gram}\,(g_{\theta_1}, \ldots, g_{\theta_N}, \chi)}{\det \text{Gram}\,(g_{\theta_1}, \ldots, g_{\theta_N})},$$

où $\text{Gram}\,(v_1, \ldots, v_n)$ désigne la matrice de Gram des vecteurs v_i.

6.2 Cas d'un seul θ

Dans le cas particulier où $N = 1$, on obtient aisément le résultat suivant.

Théorème 6.1. *La distance de χ à la droite engendrée par g_θ est minimale pour $\theta = \theta_0 := e^{-1-\gamma} \approx 0,207$ et on a alors*

$$d(\chi, \mathbb{C}g_{\theta_0}) = d(\chi, \mathcal{B}_1) = \sqrt{1 - \frac{4e^{-1-\gamma}}{\log 2\pi - \gamma}} \approx 0,587.$$

La démonstration se fait simplement en étudiant les variations de la fonction

$$\theta \mapsto \frac{\det \text{Gram}\,(g_\theta, \chi)}{\det \text{Gram}\,(g_\theta)} = \frac{< g_\theta, g_\theta > - < \chi, g_\theta >^2}{< g_\theta, g_\theta >}$$

dont on obtient une expression explicite à l'aide de la proposition de la section 2.

6.3 Cas $N = 2$

Dans ce qui suit, on s'intéresse à l'existence puis à la détermination expérimentale d'un couple (θ_1, θ_2) minimisant la fonctionnelle $F_2(\theta_1, \theta_2) := d(\chi, \text{Vect}(g_{\theta_1}, g_{\theta_2}))$.

Proposition 6.2. *Il existe une fonction f de $\mathcal{B}_2$ telle que*

$$\|\chi - f\|_2 = \inf_{\varphi \in \mathcal{B}_2} \|\chi - \varphi\|_2.$$

Une preuve de ce résultat a été donnée indépendamment par E. Saias [Balazard et Saias 98] et V. Vassiounine dans un cadre légèrement différent, une démonstration complète tirée de celle de Saias est exposée par le second auteur en [Richard 00].

Pour trouver expérimentalement un couple $(\theta_1, \theta_2) \in]0, 1]^2$ qui minimise la fonctionnelle $F_2 : (\theta_1, \theta_2) \longmapsto d\,(\chi, \text{Vect}\,(g_{\theta_1}, g_{\theta_2}))$, nous avons tout d'abord réalisé une représentation graphique de la surface définie par F_2, Figure 11. Nous avons utilisé pour cela le logiciel Matlab; la fonction F_2 est évaluée en calculant les déterminants des matrices de Gram correspondantes. Le maillage choisi est décimal avec un pas de $0,01$. L'étoile sur la figure signale le point minimum observé sur le maillage choisi, en l'occurence au point $(1, \frac{1}{3})$.

On observe clairement une suite de minima locaux qui se répartissent sur un faisceau de droites d'équations $\theta_2 = \frac{1}{n}\theta_1$, $n \geq 1$.

On observe également sur les bords de la surface par un effet de continuité lorsque θ_1 ou θ_2 tend vers zéro le graphe de la fonction $\theta \mapsto d(\chi, \mathbb{C}g_\theta)$ étudiée à la section précédente.

Nous avons ensuite entrepris un calcul de minimisation plus poussé. On cherche précisément le minimum de la

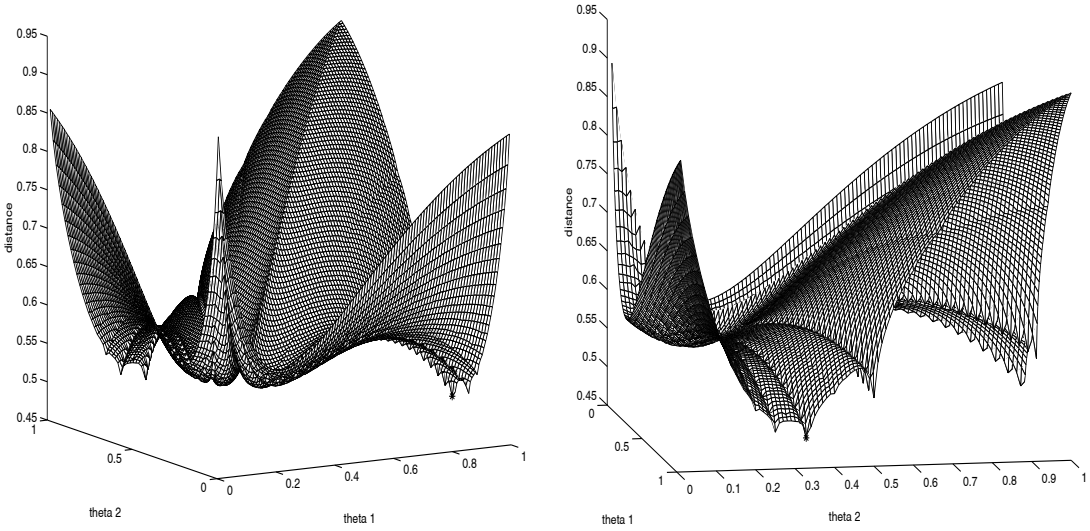

FIGURE 11. La distance de χ à $\mathrm{Vect}(g_{\theta_1}, g_{\theta_2})$ (vue de l'origine, de côté).

fonction $F_2(\theta_1, \theta_2)$ sur l'ensemble $D_Q = \{(\theta_1, \theta_2) \in \mathbb{Q} \times \mathbb{Q} \mid \theta_1 = \frac{p_1}{q_1}, \theta_2 = \frac{p_2}{q_2}, 1 \leq p_i \leq q_i \leq Q, i = 1, 2\}$ où $Q \geq 1$ est fixé. Les distances sont calculées cette fois par la méthode d'orthonormalisation de Gram-Schmidt (le calcul des déterminants de Gram est en effet bien trop coûteux) de façon similaire aux calculs conduits pour la distance d_n dans la troisième section.

On trouve alors, en effectuant les calculs jusqu'à une borne $Q = 200$, comme valeurs réalisant le minimum de F_2 en permanence le couple $(1, \frac{1}{3})$ ce qui corrobore le résultat lu sur le graphe un peu plus haut et donne une distance minimale $F_2(1, \frac{1}{3}) \approx 0,483$.

On peut noter que si le minimum de F_2 était atteint avec au moins l'un des θ_i irrationnel, on devrait alors observer en augmentant suffisamment la borne Q, par densité de $\mathbb{Q}$ et par continuité de F_2, des θ rationnels proches de la valeur idéale donnant de meilleures valeurs de la distance que le couple $\left(1, \frac{1}{3}\right)$. Or il n'en est rien, du moins jusqu'à $Q = 200$.

Tout laisse donc à penser au vu de ces résultats que les meilleurs θ sont les rationnels 1 et $\frac{1}{3}$.

6.4 Cas $N > 2$

On prolonge maintenant l'étude précédente au cas de plus de deux paramètres θ. On s'emploie à déterminer le minimum de la fonction

$$F_N : (\theta_1, \ldots, \theta_N) \longmapsto d\left(\chi, \mathrm{Vect}\left(g_{\theta_1}, \ldots, g_{\theta_N}\right)\right).$$

L'existence d'un N-uplet réalisant le minimum n'est plus assuré et ne semble pas facile à établir dans le cas général.

Remarquons que l'on a

$$\left\| \chi - \sum_{i=1}^{N} c_i g_{\theta_i} \right\|^2 \geq \int_1^\infty \left| \sum_{i=1}^{N} c_i \rho\left(\frac{\theta_i}{t}\right) \right|^2 dt = \int_1^\infty \left| \sum_{i=1}^{N} c_i \theta_i \times \frac{1}{t} \right|^2 dt = \left| \sum_{i=1}^{N} c_i \theta_i \right|^2.$$

Ainsi, il sera intéressant de regarder la valeur de $\sum_{i=1}^{N} c_i \theta_i$ qui doit, si l'hypothèse de Riemann est vraie, tendre vers 0 lorsque N tend vers l'infini pour le N-uplet des meilleurs θ_i.

Dans tout ce qui suit, nous comparerons les résultats avec ceux obtenus pour $d_N = d(\chi, \mathrm{Vect}(e_1, \ldots, e_N))$.

Dans un premier temps, nous avons effectué, compte-tenu des approximations de χ étudiées précédemment, une recherche des meilleurs θ parmi les inverses d'entiers. Nous avons ensuite élargi notre recherche à tous les θ rationnels. Dans les deux cas, la recherche est menée parmi les rationnels de dénominateur borné.

6.4.1 Minimisation sur les inverses d'entiers. Nous recherchons ici à N fixé les $(\theta_i)_{1 \leq i \leq N}$ qui minimisent F_N avec $\theta_i = \frac{1}{q_i}$, $1 \leq q_i \leq Q$, $1 \leq i \leq N$.

Voici, Table 3, les résultats obtenus. Il est évident que lorsque N augmente, pour des raisons de temps de calcul, nous sommes malheureusement obligés de réduire sérieusement la borne Q, ce qui limite la portée des résultats.

A priori, l'étude de la section 4 nous laisse penser que l'on va retrouver en premier les inverses de nombres premiers entre lesquels vont s'intercaler les inverses de nom-

N	Q	$\theta_i,\ i = 1 \dots N$	distance	d_N	$\sum_{i=1}^{N} c_i\theta_i$
2	500	$1, \frac{1}{3}$	0,4825	0,5385	0,09106
3	500	$1, \frac{1}{2}, \frac{1}{3}$	0,3063	0,3063	0,04161
4	100	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{5}$	0,1999	0,2552	0,01624
5	50	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}$	0,1725	0,1900	0,01057
6	50	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}$	0,1609	0,1897	0,00977
7	50	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{11}$	0,1537	0,1559	0,00828
8	25	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{10}, \frac{1}{11}$	0,1475	0,1554	0,00788
9	25	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{10}, \frac{1}{11}$	0,1438	0,1550	0,00730
10	25	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{10}, \frac{1}{11}, \frac{1}{13}, \frac{1}{14}$	0,1399	0,1542	0,00709
11	25	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{10}, \frac{1}{11}, \frac{1}{13}, \frac{1}{14}, \frac{1}{17}$	0,1362	0,1430	0,00654
12	25	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{10}, \frac{1}{11}, \frac{1}{13}, \frac{1}{14}, \frac{1}{17}$	0,1334	0,1430	0,00612
13	25	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{10}, \frac{1}{11}, \frac{1}{13}, \frac{1}{14}, \frac{1}{15}, \frac{1}{17}$	0,1310	0,1408	0,00594
14	25	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{10}, \frac{1}{11}, \frac{1}{13}, \frac{1}{14}, \frac{1}{15}, \frac{1}{17}, \frac{1}{21}$	0,1310	0,1360	0,00589
15	25	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{10}, \frac{1}{11}, \frac{1}{13}, \frac{1}{14}, \frac{1}{15}, \frac{1}{17}, \frac{1}{19}, \frac{1}{21}$	0,1270	0,1354	0,00559

TABLE 3. Meilleurs θ_i inverses d'entiers, $2 \le N \le 15$.

bres composés sans facteur carré. C'est effectivement le cas mais l'on note tout de même quelques particularités. Il y a la curieuse inversion d'ordre pour $\frac{1}{2}$ et $\frac{1}{3}$ au départ. Le rationnel $\frac{1}{4}$ apparaît pour $N = 9$. Un fait marquant est que la suite des ensembles $(\{\theta_1, \dots, \theta_k\})_{k\in\mathbb{N}}$ n'est pas une suite croissante pour l'inclusion. par exemple, $\frac{1}{4}$ apparaît pour $N = 9$, disparaît pour $N = 10$ puis réapparaît pour $N = 12$. On notera que l'expression $\sum_{i=1}^{N} c_i\theta_i$ semble tendre vers zéro en décroissant.

6.4.2 Minimisation sur les rationnels. Nous cherchons maintenant à déterminer à N fixé les $(\theta_i)_{1\le i\le N}$ qui minimisent $F_N(\theta_1, \dots, \theta_N)$, où les θ_i sont rationnels de la forme $\theta_i = \frac{p_i}{q_i}$, $1 \le p_i \le q_i \le Q$, $1 \le i \le N$.

Encore une fois, et plus que dans la recherche précédente, pour des raisons évidentes de temps, les bornes des dénominateurs sont relativement petites ce qui limite considérablement la zone d'investigation.

Voici cependant, Table 4, les résultats obtenus.

La question importante, objet de cette seconde recherche, est de savoir si des rationnels autres que les inverses d'entiers peuvent jouer un rôle important dans l'approximation de χ. La réponse est positive puisque l'on note avec attention l'apparition des rationnels $\frac{3}{11}, \frac{5}{13}, \frac{5}{11} \dots$. On note également que certaines valeurs de θ apparaissent puis disparaissent (par exemple $\frac{5}{13}$) de sorte que la suite des ensembles $(\{\theta_1, \dots, \theta_k\})_{k\in\mathbb{N}}$ n'est toujours pas une suite croissante pour l'inclusion. Les inverses d'entiers ayant un facteur carré n'apparaissent pas, du moins pour $N \le 10$.

Il apparaît alors intéressant et complémentaire d'étudier numériquement la distance δ_n définie par $\delta_n = d(\chi, W_n)$ où $W_n = \mathrm{Vect}(g_{p/q}, 1 \le p \le q \le n)$ et de la comparer à d_n. Notons que l'on a immédiatement compte-tenu des différentes définitions

$$D(1/n) \le \delta_n \le d_n.$$

Le calcul de cette distance, mené comme celui de d_n, ne peut malheureusement être fait que pour des entiers n relativement petits; en effet les rationnels θ considérés constituent la suite de Farey de rang n dont le cardinal croît de façon quadratique en n.

Signalons simplement que l'on observe jusqu'à $n \le 200$ (soit plus de 12 000 rationnels) un comportement similaire à celui de d_n mais avec toutefois des valeurs de δ_n bien inférieures à celles de d_n. Cela vient confirmer les calculs précédents qui ont mis en lumière le rôle non négligeable des θ de la forme p/q avec $p > 1$.

Disons pour conclure cette section que le comportement des meilleurs θ reste, malgré une stabilité relative évidente des suites obtenues, pour l'instant encore assez mystérieux. Il est difficile de dégager une conjecture claire pour la suite de ces meilleurs θ et, en particulier, en dépit de l'apparition des rationnels comme par exemple $\frac{3}{11}$ ou $\frac{5}{13}$, il est tout à fait possible qu'asymptotiquement les θ de la forme $\frac{1}{k}$ soient les meilleurs possibles.

ADDENDUM

Depuis la date de soumission de ce travail, un nouveau résultat important a été établi par L. Báez-Duarte [Báez-Duarte 02]. Si l'on note $\mathcal{B}^{nat}$ le sous-espace de $\mathcal{B}$ engendré par les fonctions $t \mapsto \rho(\frac{1}{nt})$, $n \ge 1$, alors l'hypothèse de

N	Q	$\theta_i,\ i = 1 \dots N$	distance	d_N	$\sum_{i=1}^{N} c_i\theta_i$
3	40	$1, \frac{1}{2}, \frac{1}{3}$	0,3063	0,3063	0,04161
4	40	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{5}$	0,1999	0,2552	0,01624
5	25	$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}$	0,1725	0,1900	0,01057
6	18	$1, \frac{1}{2}, \frac{1}{3}, \frac{3}{11}, \frac{1}{5}, \frac{1}{7}$	0,1607	0,1897	0,01238
7	15	$1, \frac{1}{2}, \frac{5}{13}, \frac{1}{3}, \frac{3}{11}, \frac{1}{5}, \frac{1}{7}$	0,1506	0,1559	0,01687
8	15	$1, \frac{1}{2}, \frac{5}{11}, \frac{1}{3}, \frac{3}{11}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}$	0,1436	0,1554	0,01673
9	13	$1, \frac{1}{2}, \frac{5}{13}, \frac{1}{3}, \frac{3}{13}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{11}$	0,1374	0,1550	0,01543
10	13	$1, \frac{1}{2}, \frac{5}{13}, \frac{1}{3}, \frac{3}{13}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{10}, \frac{1}{11}$	0,1330	0,1542	0,01494

TABLE 4. Meilleurs θ_i rationnels, $3 \leq N \leq 10$.

Riemann équivaut au fait que $d(\chi, \mathcal{B}^{nat}) = 0$ i.e $\chi \in \overline{\mathcal{B}^{nat}}$ ou encore que la suite d_n étudiée dans la section 3 tende vers 0 à l'infini. Cela vient prouver le fait qu'il suffit, pour approcher la fonction χ, de considérer les combinaisons linéaires de fonctions g_θ où les θ sont rationnels et de la forme $\frac{1}{n}$, $n \geq 1$.

REMERCIEMENTS

Les auteurs tiennent à remercier spécialement Luis Báez-Duarte, Michel Balazard et Eric Saias pour toutes les discussions fructueuses et les encouragements qui ont conduit à ce travail.

Les auteurs remercient également les referees qui, par leurs remarques et leurs suggestions, ont permis d'améliorer la qualité de l'article.

RÉFÉRENCES

[Báez-Duarte 01] L. Báez-Duarte. "Arithmetical Aspects of Beurling's Real Variable Reformulation of the Riemann Hypothesis." Document de travail, communication privée.

[Báez-Duarte 02] L. Báez-Duarte. "A strengthening of the Nyman-Beurling criterion for the Riemann Hypothesis." *arXiv:math.NT/0205003*, à paraître dans *Atti Accad. Naz. Lincei Rend. Mat. Appl.*

[Báez-Duarte et al. 00] L. Báez-Duarte, M. Balazard, B. Landreau et E. Saias. "Notes sur la fonction ζ de Riemann, 3." *Advances in Math.* **149** (2000), 130–144.

[Balazard et Saias 98] M. Balazard et E. Saias. Notes manuscrites, (1998).

[Beurling 55] A. Beurling. "A closure problem related to the Riemann Zeta-function." *Proc. Nat. Acad. Sci.* **41** (1955), 312–314.

[Burnol 01] J.-F. Burnol. "A lower bound in an approximation problem involving the zeros of the Riemann zeta function." *Advances in Math.* **170** (2002), 56–70.

[Nyman 50] B. Nyman. *On the One-Dimensional Translation Group and Semi-Group in Certain Function Spaces*, Thèse, Uppsala, 1950.

[Richard 00] F. Richard. *L'hypothèse de Riemann selon Beurling et Nyman*, mémoire de DEA, Université Bordeaux I, 2000.

[Rosser 39, p. 29] B. Rosser. "The nth prime is greater than $n \log(n)$." *Proc. London Math. Soc. (2)* **45** (1939), 21–44.

[Selberg 46] A. Selberg. "The zeta-function ant the Riemann hypothesis." in C. R. Dixième congrès Math. Scandinaves, Copenhague, (1946), 187–200.

[Titchmarsh 86] E. C. Titchmarsh. *The theory of the Riemann zeta-function*, (revised by D.R.Heath-Brown), Clarendon Press, Oxford, 1986.

[Vassiounine 96] Vasyunin V. I., "On a biorthogonal system related with the Riemann hypothesis." *St. Petersburg Math. J.* **7**: 3 (1996), 405–419.

Bernard Landreau, Laboratoire d'algorithmique arithmétique et de théorie des nombres de Bordeaux, UMR CNRS 5465, Université Bordeaux I, 351, crs de la libération, 33 405 Talence Cedex, France (landreau@math.u-bordeaux.fr)

Florent Richard, Laboratoire d'algorithmique arithmétique et de théorie des nombres de Bordeaux, UMR CNRS 5465, Université Bordeaux I, 351, crs de la libération, 33 405 Talence Cedex, France (richard.flo@free.fr)

Received July 18, 2001; accepted in revised form May 1, 2002.

Computing a Glimpse of Randomness

Cristian S. Calude, Michael J. Dinneen, and Chi-Kou Shu

CONTENTS

A Chaitin Omega number is the halting probability of a universal Chaitin (self-delimiting Turing) machine. Every Omega number is both *computably enumerable* (the limit of a computable, increasing, converging sequence of rationals) and *random* (its binary expansion is an algorithmic random sequence). In particular, every Omega number is strongly noncomputable. The aim of this paper is to describe a procedure, that combines Java programming and mathematical proofs, to compute the *exact values of the first 64 bits of a Chaitin Omega:*

0000001000000100000110001000011010001111110010111011101000010000.

Full description of programs and proofs will be given elsewhere.

1. INTRODUCTION

Any attempt to compute the uncomputable or to decide the undecidable is without doubt challenging, but hardly a new endeavor (see, for example, [Marxen and Buntrock 90], [Stewart 91], [Casti 97]). This paper describes a hybrid procedure (which combines Java programming and mathematical proofs) for computing the *exact values of the first 64 bits of a concrete Chaitin Omega number*, Ω_U, the halting probability of the universal Chaitin (self-delimiting Turing) machine U, see [Chaitin 90a]. Note that any Omega number is not only noncomputable, but random, making the computing task even more demanding.

Computing lower bounds for Ω_U is not difficult: we just generate more and more halting programs. Are the bits produced by such a procedure exact? *Hardly.* If the first bit of the approximation happens to be 1, then yes, it is exact. However, if the provisional bit given by an approximation is 0, then, due to possible overflows, nothing prevents the first bit of Ω_U from being either 0 or 1. This situation extends to other bits as well. Only an initial run of 1s may give exact values for some bits of Ω_U.

The paper is structured as follows. Section 2 introduces the basic notation. Computably enumerable (c.e.) reals, random reals, and c.e. random reals are presented

2000 AMS Subject Classification: Primary 68Q30; Secondary 68Q17

Keywords: Chaitin Omega number, halting problem, algorithmic randomness

1058-6458/2001 $0.50 per page

in Section 3. Various theoretical difficulties preventing the exact computation of any bits of an Omega number are discussed in Section 4. The register machine model of Chaitin [Chaitin 90a] is discussed in Section 5. In Section 6 we summarize our computational results concerning the halting programs of up to 84 bits long for U. They give a lower bound for Ω_U which is proved to provide the exact values of the first 64 digits of Ω_U in Section 7.

Chaitin [Chaitin 00b] has pointed out that the self-delimiting Turing machine constructed in the preliminary version of this paper [Calude et al. 00] is universal in the sense of Turing (i.e., it is capable to simulate any self-delimiting Turing machine), but it is not universal in the sense of algorithmic information theory because the "price" of simulation is not bounded by an additive constant; hence, the halting probability is not an Omega number (but a computably enumerable real with some properties close to randomness). The construction presented in this paper is a self-delimiting Turing machine. Full details will appear in [Shu 03].

2. NOTATION

We will use notation that is standard in algorithmic information theory and assume familiarity with Turing machine computations, computable and computably enumerable (c.e.) sets (see, for example, [Bridges 94], [Odifreddi 99], [Soare 87], [Weihrauch 87]), and elementary algorithmic information theory (see, for example, [Calude 94]).

By $\mathbf{N}, \mathbf{Q}$, we denote the set of nonnegative integers (natural numbers) and rationals, respectively. If S is a finite set, then $\#S$ denotes the number of elements of S. Let $\Sigma = \{0,1\}$ denote the binary alphabet. Let Σ^* be the set of (finite) binary strings, and Σ^ω the set of infinite binary sequences. The length of a string x is denoted by $|x|$. A subset A of Σ^* is *prefix-free* if whenever s and t are in A and s is a prefix of t, then $s = t$.

For a sequence $\mathbf{x} = x_0x_1 \cdots x_n \cdots \in \Sigma^\omega$ and an nonnegative integer $n \geq 1$, $\mathbf{x}(n)$ denotes the initial segment of length n of $\mathbf{x}$ and x_i denotes the ith digit of $\mathbf{x}$, i.e. $\mathbf{x}(n) = x_0x_1 \cdots x_{n-1} \in \Sigma^*$. Due to Kraft's inequality, for every prefix-free set $A \subset \Sigma^*$, $\Omega_A = \sum_{s \in A} 2^{-|s|}$ lies in the interval $[0,1]$. In fact Ω_A is a probability: Pick, at random using the Lebesgue measure on $[0,1]$, a real α in the unit interval and note that the probability that some initial prefix of the binary expansion of α lies in the prefix-free set A is exactly Ω_A.

Following Solovay [Solovay 75, Solovay 00], we say that C is a *(Chaitin)* (self-delimiting Turing) *machine*, shortly, a *machine*, if C is a Turing machine processing binary strings such that its program set (domain) $PROG_C = \{x \in \Sigma^* \mid C(x) \text{ halts}\}$ is a prefix-free set of strings. Clearly, $PROG_C$ is c.e.; conversely, every prefix-free c.e. set of strings is the domain of some machine. The *program-size complexity* of the string $x \in \Sigma^*$ (relatively to C) is $H_C(x) = \min\{|y| \mid y \in \Sigma^*,\ C(y) = x\}$, where $\min \emptyset = \infty$. A major result of algorithmic information theory is the following invariance relation: We can effectively construct a machine U (called *universal*) such that for every machine C, there is a constant $c > 0$ (depending upon U and C) such that for every $x, y \in \Sigma^*$ with $C(x) = y$, there exists a string $x' \in \Sigma^*$ with $U(x') = y$ (U simulates C) and $|x'| \leq |x| + c$ (the overhead for simulation is no larger than an additive constant). In complexity-theoretic terms, $H_U(x) \leq H_C(x) + c$. Note that $PROG_U$ is c.e., but not computable.

If C is a machine, then $\Omega_C = \Omega_{PROG_C}$ represents its halting probability. When $C = U$ is a universal machine, then its halting probability Ω_U is called a *Chaitin Ω number*, shortly, *Ω number*.

3. COMPUTABLY ENUMERABLE AND RANDOM REALS

Reals will be written in binary, so we start by looking at random binary sequences. Two complexity-theoretic definitions can be used to define random sequences (see [Chaitin 75, Chaitin 00a]): an infinite sequence $\mathbf{x}$ is *Chaitin random* if there is a constant c such that $H(\mathbf{x}(n)) > n - c$, for every integer $n > 0$, or, equivalently, $\lim_{n\to\infty} H(\mathbf{x}(n)) - n = \infty$. Other *equivalent* definitions include the Martin-Löf [Martin-Löf 66, Martin-Löf 66] definition using statistical tests (*Martin-Löf random sequences*), the Solovay [Solovay 75] measure-theoretic definition (*Solovay random sequences*), and the Hertling and Weihrauch [Hertling and Weihrauch 98] topological approach to define randomness (*Hertling–Weihrauch random sequences*). Independent proofs of the equivalence between the Martin-Löf and Chaitin definitions have been obtained by Schnorr and Solovay, see [Chaitin 90a, Chaitin 01]. In what follows, we will simply call "random" a sequence satisfying one of the above equivalent conditions. Their equivalence motivates the following "randomness hypothesis" ([Calude 00]): *A sequence is "algorithmically random" if it satisfies one of the above equivalent conditions.* Of course, randomness implies strong noncomputability (see, for example, [Calude 94]), but the converse is false.

A real α is random if its binary expansion $\mathbf{x}$ *(i.e.* $\alpha = 0.\mathbf{x}$*) is random.* The choice of the binary base does not play any role, see [Calude and Jürgensen 94], [Hertling and Weihrauch 98], [Staiger 91]: *randomness is a property of reals not of names of reals.*

Following Soare [Soare 69], a real α is called c.e. if there is a computable, increasing sequence of rationals which converges (*not necessarily computably*) to α. We will start with several characterizations of c.e. reals (see [Calude et al. 01]). If $0.\mathbf{y}$ is the binary expansion of a real α with infinitely many ones, then $\alpha = \sum_{n \in X_\alpha} 2^{-n-1}$, where $X_\alpha = \{i \mid y_i = 1\}$.

Theorem 3.1. *Let α be a real in $(0, 1]$. The following conditions are equivalent:*

(i) There is a computable, nondecreasing sequence of rationals which converges to α.

(ii) The set $\{p \in \mathbf{Q} \mid p < \alpha\}$ of rationals less than α is c.e..

(iii) There is an infinite prefix-free c.e. set $A \subseteq \Sigma^*$ with $\alpha = \Omega_A$.

(iv) There is an infinite prefix-free computable set $A \subseteq \Sigma^*$ with $\alpha = \Omega_A$.

(v) There is a total computable function $f : \mathbf{N}^2 \to \{0, 1\}$ such that

(a) If for some k, n we have $f(k, n) = 1$ and $f(k, n + 1) = 0$, then there is an $l < k$ with $f(l, n) = 0$ and $f(l, n + 1) = 1$.

(b) We have: $k \in X_\alpha \iff \lim_{n\to\infty} f(k, n) = 1$.

We note that following Theorem 3.1, (v), given a computable approximation of a c.e. real α via a total computable function f, $k \in X_\alpha \iff \lim_{n\to\infty} f(k, n) = 1$; the values of $f(k, n)$ may oscillate from 0 to 1 and back; we will not be sure that they stabilized until 2^k changes have occurred (of course, there need not be so many changes, but in this case, there is no guarantee of the exactness of the value of the kth bit).

Chaitin [Chaitin 75] proved the following important result:

Theorem 3.2. *If U is a universal machine, then Ω_U is c.e. and random.*

The converse of Theorem 3.2 is also true: It has been proved by Kučera and Slaman [Kučera and Slaman 01] based on work reported in [Calude et al. 01] (see also [Calude and Chaitin 99], [Calude 02a], [Downey 02]):

Theorem 3.3. *Let $\alpha \in (0, 1)$. The following conditions are equivalent:*

(i) The real α is c.e. and random.

(ii) For some universal machine U, $\alpha = \Omega_U$.

4. THE FIRST BITS OF AN OMEGA NUMBER

We start by noting that

Theorem 4.1. *Given the first n bits of Ω_U, one can decide whether $U(x)$ halts or not on an arbitrary string x of length at most n.*

The first 10,000 bits of Ω_U include a tremendous amount of mathematical knowledge. In Bennett's words [Bennett and Gardner 79]:

> [Ω] embodies an enormous amount of wisdom in a very small space . . . inasmuch as its first few thousands digits, which could be written on a small piece of paper, contain the answers to more mathematical questions than could be written down in the entire universe.
>
> Throughout history mystics and philosophers have sought a compact key to universal wisdom, a finite formula or text which, when known and understood, would provide the answer to every question. The use of the Bible, the Koran and the I Ching for divination and the tradition of the secret books of Hermes Trismegistus, and the medieval Jewish Cabala exemplify this belief or hope. Such sources of universal wisdom are traditionally protected from casual use by being hard to find, hard to understand when found, and dangerous to use, tending to answer more questions and deeper ones than the searcher wishes to ask. The esoteric book is, like God, simple yet undescribable. It is omniscient, and transforms all who know it . . . Omega is in many senses a cabalistic number. It can be known of, but not known, through human reason. To know it in detail, one would have to accept its uncomputable digit sequence on faith, like words of a sacred text.

It is worth noting that even if we get, by some kind of miracle, the first 10,000 digits of Ω_U, the task of solving the problems whose answers are embodied in these bits is computable, but unrealistically difficult: The time it takes to find all halting programs of length less than n from $0.\Omega_0\Omega_2 \ldots \Omega_{n-1}$ grows faster than any computable function of n.

Computing some initial bits of an Omega number is even more difficult. According to Theorem 3.3, c.e. random reals can be coded by universal machines through their halting probabilities. How "good" or "bad" are these names? In [Chaitin 75] (see also [Chaitin 97, Chaitin 99]), Chaitin proved the following:

Theorem 4.2. *Assume that ZFC[1] is arithmetically sound.[2] Then, for every universal machine U, ZFC can determine the value of only finitely many bits of Ω_U.*

In fact, one can give a bound on the number of bits of Ω_U which ZFC can determine; this bound can be explicitly formulated, but it *is not computable.* For example, in [Chaitin 97] Chaitin described, in a dialect of Lisp, a universal machine U and a theory T, and proved that U can determine the value of at most $H(T)+15,328$ bits of Ω_U; $H(T)$ is the program-size complexity of the theory T, an *uncomputable* number.

Fix a universal machine U and consider all statements of the form

$$\text{"The } n^{th} \text{ binary digit of the expansion of } \Omega_U \text{ is } k\text{"}, \tag{4–1}$$

for all $n \geq 0, k = 0, 1$. How many theorems of the form (4–1) can ZFC prove? More precisely, is there a bound on the set of nonnegative integers n such that ZFC proves a theorem of the form (4–1)? From Theorem 4.2, we deduce that ZFC can prove only finitely many (true) statements of the form (4–1). This is Chaitin information-theoretic version of Gödel's incompleteness (see [Chaitin 97, Chaitin 99]):

Theorem 4.3. *If ZFC is arithmetically sound and U is a universal machine, then almost all true statements of the form (4–1) are unprovable in ZFC.*

Again, a bound can be explicitly found, but not effectively computed. Of course, for every c.e. random real α, we can construct a universal machine U such that $\alpha = \Omega_U$ and ZFC is able to determine finitely (but as many as we want) bits of Ω_U.

A machine U for which Peano Arithmetic can prove its universality and ZFC cannot determine more than the initial block of 1 bits of the binary expansion of its halting probability, Ω_U, will be called *Solovay machine.*[3]

[1] Zermelo set theory with choice.
[2] That is, any theorem of arithmetic proved by ZFC is *true.*
[3] Clearly, U depends on ZFC.

To make things worse Calude [Calude 02b] proved the following result:

Theorem 4.4. *Assume that ZFC is arithmetically sound. Then, every c.e. random real is the halting probability of a Solovay machine.*

For example, if $\alpha \in (3/4, 7/8)$ is c.e. and random, then in the worst case, ZFC can determine its first two bits (11), but no more. For $\alpha \in (0, 1/2)$, we obtained Solovay's Theorem [Solovay 00]:

Theorem 4.5. *Assume that ZFC is arithmetically sound. Then, every c.e. random real $\alpha \in (0, 1/2)$ is the halting probability of a Solovay machine which cannot determine any single bit of α. No c.e. random real $\alpha \in (1/2, 1)$ has the above property.*

The conclusion is that the worst fears discussed in the first section proved to materialize: In general, only the initial run of 1s (if any) can be exactly computed.

5. REGISTER MACHINE PROGRAMS

We start with the register machine model used by Chaitin [Chaitin 90a]. Recall that any register machine has a finite number of registers, each of which may contain an arbitrarily large nonnegative integer. The list of instructions is given below in two forms: our compact form and its corresponding Chaitin [Chaitin 90a] version. The main difference between Chaitin's implementation and ours is in the encoding: we use 7 bit codes instead of 8 bit codes.

L: ? L1 **(L: GOTO L1)**

This is an unconditional branch to L1. L1 is a label of some instruction in the program of the register machine.

L: ∧ R L1 **(L: JUMP R L1)**

Set the register R to be the label of the next instruction and go to the instruction with label L1.

L: @ R **(L: GOBACK R)**

Go to the instruction with a label which is in R. This instruction will be used in conjunction with the jump instruction to return from a subroutine. The instruction is illegal (i.e., run-time error occurs) if R has not been explicitly set to a valid label of an instruction in the program.

L: = R1 R2 L1 **(L: EQ R1 R2 L1)**

This is a conditional branch. The last 7 bits of register R1 are compared with the last 7 bits of register R2. If they are equal, then the execution continues at the instruction with label L1. If they are not equal, then execution continues with the next instruction in sequential order. R2 may be replaced by a constant which can be represented by a 7-bit ASCII code, i.e., a constant from 0 to 127.

L: # R1 R2 L1 **(L: NEQ R1 R2 L1)**

This is a conditional branch. The last 7 bits of register R1 are compared with the last 7 bits of register R2. If they are not equal, then the execution continues at the instruction with label L1. If they are equal, then execution continues with the next instruction in sequential order. R2 may be replaced by a constant which can be represented by a 7-bit ASCII code, i.e., a constant from 0 to 127.

L:) R **(L: RIGHT R)**

Shift register R right 7 bits, i.e., the last character in R is deleted.

L: (R1 R2 **(L: LEFT R1 R2)**

Shift register R1 left 7 bits, add to it the rightmost 7 bits of register R2, and then shift register R2 right 7 bits. The register R2 may be replaced by a constant from 0 to 127.

L: & R1 R2 **(L: SET R1 R2)**

The contents of register R1 are replaced by the contents of register R2. R2 may be replaced by a constant from 0 to 127.

L: ! R **(L: READ R)**

One bit is read into the register R, so the numerical value of R becomes either 0 or 1. Any attempt to read past the last data-bit results in a run-time error.

L: / **(L: DUMP)**

All register names and their contents, as bit strings, are written out. This instruction is also used for debugging.

L: % **(L: HALT)**

Halts the execution. This is the last instruction for each register machine program.

A *register machine program* consists of a finite list of labeled instructions from the above list, with the restriction that the `HALT` instruction appears only once, as the last instruction of the list. The data (a binary string) follows immediately the `HALT` instruction. The use of undefined variables is a run-time error. A program not reading the whole data, or attempting to read past the last data-bit, results in a run-time error. Because of the position of the `HALT` instruction and the specific way data is read, register machine programs are Chaitin machines.

To be more precise, we present a context-free grammar $G = (N, \Sigma, P, S)$ in Backus-Naur form which generates the register machine programs.

(1) N is the finite set of nonterminal variables:

$$N = \{S\} \cup INST \cup TOKEN$$

$$INST = \{\langle \mathrm{RMS}_{\mathrm{Ins}}\rangle, \langle ?_{\mathrm{Ins}}\rangle, \langle \wedge_{\mathrm{Ins}}\rangle, \langle @_{\mathrm{Ins}}\rangle, \langle =_{\mathrm{Ins}}\rangle, \langle \#_{\mathrm{Ins}}\rangle, \langle)_{\mathrm{Ins}}\rangle, \langle (_{\mathrm{Ins}}\rangle, \langle \&_{\mathrm{Ins}}\rangle, \langle !_{\mathrm{Ins}}\rangle, \langle /_{\mathrm{Ins}}\rangle, \langle \%_{\mathrm{Ins}}\rangle\}$$

$$TOKEN = \{\langle \textsc{Data}\rangle, \langle \textsc{Label}\rangle, \langle \textsc{Register}\rangle, \langle \textsc{Constant}\rangle, \langle \textsc{Special}\rangle, \langle \textsc{Space}\rangle, \langle \textsc{Alpha}\rangle, \langle \mathrm{LS}\rangle\}$$

(2) Σ, the alphabet of the register machine programs, is a finite set of terminals, disjoint from N:

$$\Sigma = \langle \textsc{Alpha}\rangle \cup \langle \textsc{Special}\rangle \cup \langle \textsc{Space}\rangle \cup \langle \textsc{Digit}\rangle$$

$$\langle \textsc{Alpha}\rangle = \{a, b, c, \ldots, z\}$$

$$\langle \textsc{Special}\rangle = \{:, /, ?, \wedge, @, =, \#,), (, \&, !, \%\}$$

$$\langle \textsc{Space}\rangle = \{\text{‘space’,‘tab’}\}$$

$$\langle \textsc{Digit}\rangle = \{0, 1, \ldots, 9\}$$

$$\langle \textsc{Constant}\rangle = \{d \mid 0 \le d \le 127\}$$

(3) P (a subset of $N \times (N \cup \Sigma)^*$) is the finite set of rules (productions):

$$S \to \langle \mathrm{RMS}_{\mathrm{Ins}}\rangle^* \langle \%_{\mathrm{Ins}}\rangle \langle \textsc{Data}\rangle$$

$$\langle \textsc{Data}\rangle \to (0|1)^*$$

$$\langle \textsc{Label}\rangle \to 0 \mid (1|2|\ldots|9)(0|1|2|\ldots|9)^*$$

$$\langle \mathrm{LS}\rangle \to : \langle \textsc{Space}\rangle^*$$

$$\langle \textsc{Register}\rangle \to \langle \textsc{Alpha}\rangle(\langle \textsc{Alpha}\rangle \cup (0|1|2|\ldots|9))^*$$

$$\langle \mathrm{RMS}_{\mathrm{Ins}}\rangle \to \langle ?_{\mathrm{Ins}}\rangle \mid \langle \wedge_{\mathrm{Ins}}\rangle \mid \langle @_{\mathrm{Ins}}\rangle \mid \langle =_{\mathrm{Ins}}\rangle \mid \langle \#_{\mathrm{Ins}}\rangle \mid \langle)_{\mathrm{Ins}}\rangle \mid \langle (_{\mathrm{Ins}}\rangle \mid \langle \&_{\mathrm{Ins}}\rangle \mid \langle !_{\mathrm{Ins}}\rangle \mid \langle /_{\mathrm{Ins}}\rangle$$

`(L: HALT)`

$$\langle \%_{\mathrm{Ins}}\rangle \to \langle \textsc{Label}\rangle \langle \mathrm{LS}\rangle \%$$

`(L: GOTO L1)`

$$\langle ?_{\mathrm{Ins}}\rangle \to \langle \textsc{Label}\rangle \langle \mathrm{LS}\rangle ? \langle \textsc{Space}\rangle^* \langle \textsc{Label}\rangle$$

`(L: JUMP R L1)`

$$\langle \wedge_{\mathrm{Ins}}\rangle \to \langle \textsc{Label}\rangle \langle \mathrm{LS}\rangle \wedge \langle \textsc{Space}\rangle^* \langle \textsc{Register}\rangle \langle \textsc{Space}\rangle^+ \langle \textsc{Label}\rangle$$

`(L: GOBACK R)`

$$\langle @_{\mathrm{Ins}}\rangle \to \langle \textsc{Label}\rangle \langle \mathrm{LS}\rangle @ \langle \textsc{Space}\rangle^* \langle \textsc{Register}\rangle$$

`(L: EQ R 0/127 L1 or L: EQ R R2 L1)`

$$\langle =_{\mathrm{Ins}}\rangle \to \langle \textsc{Label}\rangle \langle \mathrm{LS}\rangle = \langle \textsc{Space}\rangle^* \langle \textsc{Register}\rangle \langle \textsc{Space}\rangle^+ \langle \textsc{Constant}\rangle \langle \textsc{Space}\rangle^+ \langle \textsc{Label}\rangle \mid \langle \textsc{Label}\rangle \langle \mathrm{LS}\rangle = \langle \textsc{Space}\rangle^* \langle \textsc{Register}\rangle \langle \textsc{Space}\rangle^+ \langle \textsc{Register}\rangle \langle \textsc{Space}\rangle^+ \langle \textsc{Label}\rangle$$

```
                (L: NEQ R 0/127 L1 or L: NEQ R R2 L1)
⟨#Ins⟩ → ⟨Label⟩⟨LS⟩#⟨Space⟩*⟨Register⟩⟨Space⟩+
         ⟨Constant⟩⟨Space⟩+⟨Label⟩ | ⟨Label⟩⟨LS⟩#
         ⟨Space⟩*⟨Register⟩⟨Space⟩+⟨Register⟩⟨Space⟩+
         ⟨Label⟩

                (L: RIGHT R)
⟨)Ins⟩ → ⟨Label⟩⟨LS⟩)⟨Space⟩*⟨Register⟩

                (L: LEFT R L1)
⟨(Ins⟩ → ⟨Label⟩⟨LS⟩(⟨Space⟩*⟨Register⟩⟨Space⟩+
         ⟨Constant⟩ | ⟨Label⟩⟨LS⟩(⟨Space⟩*⟨Register⟩
         ⟨Space⟩+⟨Register⟩

                (L: SET R 0/127 or L: SET R R2)
⟨&Ins⟩ → ⟨Label⟩⟨LS⟩&⟨Space⟩*⟨Register⟩⟨Space⟩+
         ⟨Constant⟩ | ⟨Label⟩⟨LS⟩&⟨Space⟩*⟨Register⟩
         ⟨Space⟩+⟨Register⟩

                (L: READ R)
⟨!Ins⟩ → ⟨Label⟩⟨LS⟩!⟨Space⟩*⟨Register⟩

                (L: DUMP)
⟨/Ins⟩ → ⟨Label⟩⟨LS⟩/
```

(4) $S \in N$ is the start symbol for the set of register machine programs.

It is important to observe that the above construction is *universal* in the sense of algorithmic information theory (see the discussion at the end of Section 1). Register machine programs are self-delimiting because the `HALT` instruction is at the end of any valid program. Note that the data, which immediately follows the `HALT` instruction, is read bit by bit with no endmarker: This type of construction was first programmed in Lisp by Chaitin [Chaitin 90a, Chaitin 00a].

To minimize the number of programs of a given length that need to be simulated, we have used "canonical programs" instead of general register machines programs. A *canonical program* is a register machine program in which (1) labels appear in increasing numerical order starting with 0; (2) new register names appear in increasing lexicographical order starting from 'a'; (3) there are no leading or trailing spaces; (4) operands are separated by a single space; (5) there is no space after labels or operators; and (6) instructions are separated by a single space. Note that *for every register machine program, there is a unique canonical program which is equivalent to it*, that is, both programs have the same domain and produce the same output on a given input. If x is a program and y is its canonical program, then $|y| \leq |x|$.

Here is an example of a canonical program:

```
0:!a 1:^b 4 2:!c 3:?11 4:=a 0 8 5:&c 110
6:(c 101 7:@b 8:&c 101 9:(c 113 10:@b 11:%10
```

To facilitate the understanding of the code, we rewrite the instructions with additional comments and spaces:

```
0:! a          // read the first data bit into register a
1:^ b 4        // jump to a subroutine at line 4
2:! c          // on return from the subroutine call c is
               //written out
3:? 11         // go to the halting instruction
4:= a 0 8      // the right most 7 bits are compared with
               // 127; if they are equal, then go to label 8
5:& c 'n'      // else, continue here and
6:( c 'e'      // store the character string 'ne' in register
7:@ b          // c; go back to the instruction with label 2
               // stored in register b
8:& c 'e'      // store the character string 'eq' in
               // register c
9:( c 'q'
10:@ b
11:%           // the halting instruction
10             // the input data
```

For optimization reasons, our particular implementation designates the first maximal sequence of `SET/LET` instructions as (static) register preloading instructions. We "compress" these canonical programs by (1) deleting all labels, spaces and the colon symbol with the first nonstatic instruction having an implicit label 0, (2) separating multiple operands by a single comma symbol, and (3) replacing constants with their ASCII numerical values. The compressed format of the above program is

```
!a^b,4!c?11=a,0,8&c,110(,c,101@b&,c,101(,c,113@b%10
```

Note that compressed programs are canonical programs because during the process of "compression," everything remains the same except for the elimination of space. Compressed programs use an alphabet with 49 symbols (including the halting character). The length is calculated as the sum of the program length and the data length (7 times the number of characters). For example, the length of the above program is $7 \times (49 + 2) = 357$.

For the remainder of this paper, we will be focusing on compressed programs.

6. SOLVING THE HALTING PROBLEM FOR PROGRAMS UP TO 84 BITS

A Java version interpreter for register machine compressed programs has been implemented; it imitates Chaitin's universal machine in [Chaitin 90a]. This interpreter has been used to test the Halting Problem for all register machine programs of at most 84 bits long. The results have been obtained according to the following procedure:

Program plus data length	Number of halting programs	Program plus data length	Number of halting programs
7	1	49	1012
14	1	56	4382
21	3	63	19164
28	8	70	99785
35	50	77	515279
42	311	84	2559837

TABLE 1. Distribution of halting programs.

1. Start by generating all programs of 7 bits and test which of them stops. All strings of length 7 which can be extended to programs are considered prefixes for possible halting programs of length 14 or longer; they will simply be called *prefixes*. In general, all strings of length n which can be extended to programs are *prefixes* for possible halting programs of length $n+7$ or longer. *Compressed prefixes* are prefixes of compressed (canonical) programs.

2. Testing the Halting Problem for programs of length $n \in \{7, 14, 21, \ldots, 84\}$ was done by running all candidates (that is, programs of length n which are extensions of prefixes of length $n-7$) for up to 100 instructions, and proving that any generated program which does not halt after running 100 instructions never halts. For example, (uncompressed) programs that match the regular expression `"0:\^ a 5.* 5:\? 0"` never halt on any input.

For example, the programs `"!a!b!a!b/%10101010"` and `"!a?0%10101010"` produce run-time errors; the first program "under reads" the data and the second one "over reads" the data. The program `"!a?1!b%1010"` loops.

One would naturally want to know the shortest program that halts with more than 100 steps. If this program is larger than 84 bits, then all of our looping programs never halt. The trivial program with a sequence of 100 dump instructions runs for 101 steps, but can we do better? The answer is yes. The following family of programs $\{P_1, P_2, \ldots\}$ recursively counts to 2^i, but has linear growth in size. The programs P_1 through P_4 are given below:[4]

```
/&a,0=a,1,5&a,1?2%
/&a,0&b,0=b,1,6&b,1?3=a,1,9&a,1?2%
/&a,0&b,0&c,0=c,1,7&c,1?4=b,1,10&b,1?3=a,1,13&a,1?2%
/&a,0&b,0&c,0&d,0=d,1,8&d,1?5=c,1,11&c,1?4=b,1,14&b,1
    ?3=a,1,17&a,1?2%
```

[4]In all cases the data length is zero.

In order to create the program P_{i+1} from P_i only 4 instructions are added, while updating "goto" labels.

The running time $t(i)$ (excluding the halt instruction) of program P_i is found by the following recurrence: $t(1) = 6$, $t(i) = 2 \cdot t(i-1) + 4$. Thus, since $t(4) = 86$ and $t(5) = 156$, P_5 is the smallest program in this family to exceed 100 steps. The size of P_5 is 86 bytes (602 bits), which is smaller than the trivial dump program of 707 bits. What is the smallest program that halts after 100 steps is an open question. A hybrid program, given below, created by combining P_2 and the trivial dump programs is the smallest known.

```
&a,0/&b,0//////////////////////=b,1,26&b,1?2=a,1,29
    &a,1?0%
```

This program of 57 bytes (399 bits) runs for 102 steps. Note that the problem of finding the smallest program with the above property is undecidable (see [Chaitin 99]).

The distribution of halting compressed programs of up to 84 bits for U, the universal machine processing compressed programs, is presented in Table 1. All binary strings representing programs have the length divisible by 7.

7. THE FIRST 64 BITS OF Ω_U

Computing all halting programs of up to 84 bits for U seems to give the exact values of the first 84 bits of Ω_U. However, this is false! To understand the point, let's first ask ourselves whether the converse implication in Theorem 4.1 is true? The answer is *negative*. Globally, if we can compute all bits of Ω_U, then we can decide the Halting Problem for every program for U and conversely. However, if we can solve for U the Halting Problem for all programs up to N bits long, we might not still get any exact value for any bit of Ω_U (less all values for the first N bits). Reason: A large set of very long halting programs can contribute to the values of more significant bits of the expansion of Ω_U.

$$\begin{aligned}
\Omega_U^{7} &= 0.0000001\\
\Omega_U^{14} &= 0.00000010000001\\
\Omega_U^{21} &= 0.000000100000010000011\\
\Omega_U^{28} &= 0.0000001000000100000110001000\\
\Omega_U^{35} &= 0.00000010000001000001100010000110010\\
\Omega_U^{42} &= 0.000000100000010000011000100001101000110111\\
\Omega_U^{49} &= 0.0000001000000100000110001000011010001111101110100\\
\Omega_U^{56} &= 0.00000010000001000001100010000110100011111100101100011110\\
\Omega_U^{63} &= 0.000000100000010000011000100001101000111111001011101100111011100\\
\Omega_U^{70} &= 0.000000100000010000011000100001101000111111001011101110011001111001001\\
\Omega_U^{77} &= 0.000000100000010000011000100001101000111111001011101110100000111000001010011111\\
\Omega_U^{84} &= 0.000000100000010000011000100001101000111111001011101110100001000001111011011011011101
\end{aligned}$$

TABLE 2. Successive approximations for Ω_U.

So, to be able to compute the exact values of the first N bits of Ω_U, we need to be able to *prove* that longer programs do not affect the first N bits of Ω_U. And, fortunately, this is the case for our computation. Due to our specific procedure for solving the Halting Problem discussed in Section 6, any compressed halting program of length n has a compressed prefix of length $n-7$. This gives an upper bound for the number of possible compressed halting programs of length n.

Let Ω_U^n be the approximation of Ω_U given by the summation of all halting programs of up to n bits in length. Compressed prefixes are partitioned into two cases—ones with a `HALT` (%) instruction and ones without. Hence, halting programs may have one of the following two forms: either "$x\,y$ `HALT` u," where x is a prefix of length k not containing `HALT`, y is a sequence of instructions of length $n-k$ not containing `HALT`, and u is the data of length $m \geq 0$, or "$x\,u$," where x is a prefix of length k containing one occurrence of `HALT` followed by data (possibly empty) and u is the data of length $m \geq 1$. In both cases, the prefix x has been extended by at least one character. Accordingly, the "tail" contribution to the value of

$$\Omega_U = \sum_{n=0}^{\infty} \sum_{\{|w|=n,\, U(w)\,\text{halts}\}} 2^{-|w|}$$

is bounded from above by the sum of the following two convergent series (which reduce to two independent sums of geometric progressions):

$$\sum_{m=0}^{\infty}\sum_{n=k}^{\infty} \underbrace{\#\{x \mid \text{prefix } x \text{ not containing HALT}, |x|=k\}}_{x} \cdot \underbrace{48^{n-k}}_{y} \cdot \underbrace{1}_{\text{HALT}} \cdot \underbrace{2^m}_{u} \cdot 128^{-(n+m+1)},$$

and

$$\sum_{m=0}^{\infty} \underbrace{\#\{x \mid \text{prefix } x \text{ containing HALT}, |x|=k\}}_{x} \cdot \underbrace{2^m}_{u} \cdot 128^{-(m+k)}.$$

The number 48 comes from the fact that the alphabet has 49 characters and the last instruction before the data is `HALT` (%).

There are 402906842 prefixes not containing `HALT` and 1748380 prefixes containing `HALT`. Hence, the "tail" contribution of all programs of length 91 or greater is bounded by:

$$\begin{aligned}
&\sum_{m=0}^{\infty}\sum_{n=13}^{\infty} 402906842 \cdot 48^{n-13} \cdot 2^m \cdot 128^{-(n+m+1)}\\
&\quad + \sum_{m=0}^{\infty} 1748380 \cdot 2^m \cdot 128^{-(m+13)}\\
&= 402906842 \cdot \frac{64}{128 \cdot 48^{13}} \cdot \sum_{n=13}^{\infty} \left(\frac{48}{128}\right)^n\\
&\quad + 1748380 \cdot \frac{1}{63 \cdot 128^{13}} < 2^{-68}, \qquad (7\text{–}1)
\end{aligned}$$

that is, the first 68 bits of Ω_U^{84} "may be" correct by our method. Actually, we do not have 68 correct bits, but *only* 64 because adding a 1 to the 68th bit may cause an overflow up to the 65th bit. From (7–1) it follows that no other overflows may occur.

The list in Table 2 presents the main results of the computation:
The exact bits are underlined in the 84 approximation:

$$\Omega_U^{84} = 0.\underline{000000100000010000011000100001101000111111001011101110100001000001111011011011011101}$$

In summary, the first 64 exact bits of Ω_U are:

0000001000000100000110001000011010001111110010111011101000010000

8. CONCLUSIONS

The computation described in this paper is the first attempt to compute some initial exact bits of a random real. The method, which combines programming with mathematical proofs, can be improved in many respects. However, due to the impossibility of testing that long looping programs never actually halt (the undecidability of the Halting Problem), the method is essentially non-scalable.

As we have already mentioned, solving the Halting Problem for programs of up to n bits might not be enough to compute exactly the first n bits of the halting probability. In our case, we have solved the Halting Problem for programs of at most 84 bits, but we have obtained only 64 exact initial bits of the halting probability.

Finally, there is no contradiction between Theorem 4.5 and the main result of this paper. Ω's are halting probabilities of Chaitin universal machines, and each Ω is the halting probability of an infinite number of such machines. Among them, there are those (called Solovay machines in [Calude 02b]) which are in a sense "bad," as ZFC cannot determine more than the initial run of 1s of their halting probabilities. But the same Ω can be defined as the halting probability of a Chaitin universal machine which is not a Solovay machine, so ZFC, if supplied with that different machine, may be able to compute more (but always, as Chaitin proved, only finitely many) digits of the same Ω. Such a machine has been used for the Ω discussed in this paper.

All programs used for the computation as well as all intermediate and final data files (3 giga-bytes in gzip format) can be found at ftp://ftp.cs.auckland.ac.nz/pub/CDMTCS/Omega/

ACKNOWLEDGMENTS

We thank Greg Chaitin for pointing out an error in our previous attempt to compute the first bits of an Omega number [Calude and Dineen 00], and for continuous advice and encouragement.

REFERENCES

[Bennett and Gardner 79] C. H. Bennett, M. Gardner. "The random number omega bids fair to hold the mysteries of the universe." *Scientific American* **241** (1979), 20–34.

[Bridges 94] D. S. Bridges. *Computability—A Mathematical Sketchbook*, Springer Verlag, Berlin, 1994.

[Calude 94] C. S. Calude. *Information and Randomness. An Algorithmic Perspective*. Springer-Verlag, Berlin, 1994.

[Calude 00] C. S. Calude. "A glimpse into algorithmic information theory." in *Logic, Language and Computation*, P. Blackburn, N. Braisby, L. Cavedon, A. Shimojima (eds.), Volume 3, CSLI Series, pp. 67–83, Cambridge University Press, Cambridge, 2000.

[Calude 02a] C. S. Calude. "A characterization of c.e. random reals." *Theoret. Comput. Sci.*, **217** (2002), 3–14.

[Calude 02b] C. S. Calude. "Chaitin Ω numbers, Solovay machines and incompleteness." *Theoret. Comput. Sci.* **284** (2002), 269–277.

[Calude and Chaitin 99] C. S. Calude and G. J. Chaitin. "Randomness everywhere." *Nature*, **400**:22 (July 1999), 319–320.

[Calude et al. 00] C. S. Calude, M. J. Dinneen, and C. Shu. "Computing 80 Initial Bits of A Chaitin Omega Number: Preliminary Version." *CDMTCS Research Report* **146** (2000)

[Calude et al. 01] C. S. Calude, P. Hertling, B. Khoussainov, and Y. Wang. "Recursively enumerable reals and Chaitin Ω numbers." in *Proceedings of the 15th Symposium on Theoretical Aspects of Computer Science (Paris)*, M. Morvan, C. Meinel, D. Krob (eds.), pp. 596–606, Springer–Verlag, Berlin, 1998. Full paper in *Theoret. Comput. Sci.* **255** (2001), 125–149.

[Calude and Jürgensen 94] C. Calude and H. Jürgensen. "Randomness as an invariant for number representations." in *Results and Trends in Theoretical Computer Science*, H. Maurer, J. Karhumäki, G. Rozenberg (eds.), pp. 44–66, Springer-Verlag, Berlin, 1994.

[Casti 97] J. L. Casti. "Computing the uncomputable." *The New Scientist*, **154/2082** (17 May 1997), 34.

[Chaitin 75] G. J. Chaitin. "A theory of program size formally identical to information theory." *J. Assoc. Comput. Mach.* **22** (1975), 329–340. (Reprinted in: [Chaitin 90b], 113–128)

[Chaitin 90a] G. J. Chaitin. *Algorithmic Information Theory*, Cambridge University Press, Cambridge, 1987. (Third printing 1990)

[Chaitin 90b] G. J. Chaitin. *Information, Randomness and Incompleteness, Papers on Algorithmic Information Theory*, World Scientific, Singapore, 1987. (2nd ed., 1990)

[Chaitin 97] G. J. Chaitin. *The Limits of Mathematics*. Springer-Verlag, Singapore, 1997.

[Chaitin 99] G. J. Chaitin. *The Unknowable*, Springer-Verlag, Singapore, 1999.

[Chaitin 00a] G. J. Chaitin. *Exploring Randomness*, Springer-Verlag, London, 2000.

[Chaitin 00b] G. J. Chaitin. Personal communication to C. S. Calude, November 2000.

[Chaitin 01] G. J. Chaitin. Personal communication to C. S. Calude, December 2001.

[Downey 02] R. G. "Downey. Some Computability-Theoretical Aspects of Reals and Randomness." *CDMTCS Research Report* **173** (2002).

[Hertling and Weihrauch 98] P. Hertling and K. Weihrauch. "Randomness spaces." in *Automata, Languages and Programming, Proceedings of the 25th International Colloquium, ICALP'98* (Aalborg, Denmark), K. G. Larsen, S. Skyum, and G. Winskel (eds.), pp. 796–807, Springer-Verlag, Berlin, 1998.

[Kučera and Slaman 01] A. Kučera and T. A. Slaman. "Randomness and recursive enumerability." *SIAM J. Comput.*, **31**:1 (2001), 199–211.

[Martin-Löf 66] P. Martin-Löf. *Algorithms and Random Sequences*, Erlangen University, Nürnberg, Erlangen, 1966.

[Martin-Löf 66] P. Martin-Löf. "The definition of random sequences." *Inform. and Control* **9** (1966), 602–619.

[Marxen and Buntrock 90] H. Marxen and J. Buntrock. "Attacking the busy beaver 5." *Bull EATCS* **40** (1990), 247–251.

[Odifreddi 99] P. Odifreddi. *Classical Recursion Theory*, North-Holland, Amsterdam, Vol.1, 1989, Vol. 2, 1999.

[Shu 03] C. Shu. *Computing Exact Approximations of a Chaitin Omega Number*, Ph.D. Thesis, University of Auckland, New Zealand, 2003.

[Soare 69] R. I. Soare. "Recursion theory and Dedekind cuts." *Trans. Amer. Math. Soc.* **140** (1969), 271–294.

[Soare 87] R. I. Soare. *Recursively Enumerable Sets and Degrees*, Springer-Verlag, Berlin, 1987.

[Solovay 75] R. M. Solovay. *Draft of a paper (or series of papers) on Chaitin's work . . . done for the most part during the period of Sept.–Dec. 1974*, unpublished manuscript, IBM Thomas J. Watson Research Center, Yorktown Heights, New York, May 1975, 215 pp.

[Solovay 00] R. M. Solovay. "A version of Ω for which ZFC can not predict a single bit." in *Finite Versus Infinite. Contributions to an Eternal Dilemma*, C.S. Calude, G. Păun (eds.), pp. 323–334, Springer-Verlag, London, 2000.

[Staiger 91] L. Staiger. "The Kolmogorov complexity of real numbers." in *Proc. Fundamentals of Computation Theory*, Lecture Notes in Comput. Sci. No. 1684, G. Ciobanu and Gh. Păun (eds.), pp. 536–546, Springer–Verlag, Berlin, 1999.

[Stewart 91] I. Stewart. "Deciding the undecidable." *Nature* **352** (1991), 664–665.

[Weihrauch 87] K. Weihrauch. *Computability*, Springer-Verlag, Berlin, 1987.

Cristian S. Calude, Department of Computer Science, University of Auckland, Private Bag 92019, Auckland, New Zealand (cristian@cs.auckland.ac.nz)

Michael J. Dinneen, Department of Computer Science, University of Auckland, Private Bag 92019, Auckland, New Zealand (mjd@cs.auckland.ac.nz)

Chi-Kou Shu, Department of Computer Science, University of Auckland, Private Bag 92019, Auckland, New Zealand (cshu004@cs.auckland.ac.nz)

Received January 31, 2002; accepted February 11, 2002.

Computing Kazhdan-Lusztig Polynomials for Arbitrary Coxeter Groups

Fokko du Cloux

CONTENTS

2000 AMS Subject Classification: Primary 20C08; Secondary 20C40, 20F55, 68R15

Keywords: Kazhdan-Lusztig polynomials, computational group theory

Let (W, S) be an arbitrary Coxeter system, $y \in S^*$. We describe an algorithm which will compute, directly from y and the Coxeter matrix of W, the interval from the identity to y in the Bruhat ordering, together with the (partially defined) left and right actions of the generators. This provides us with exactly the data that are needed to compute the Kazhdan-Lusztig polynomials $P_{x,z}$, $x \leq z \leq y$. The correctness proof of the algorithm is based on a remarkable theorem due to Matthew Dyer.

1. INTRODUCTION

Let (W, S) be a Coxeter system, i.e., a group W together with a presentation of the form $\langle S \mid (st)^{m_{s,t}} = e \ \ (s,t \in S)\rangle$, where S is a finite set which we shall usually just consider to be $\{1, \ldots, n\}$; e is the identity element in W; and $(m_{s,t})$ is a symmetric matrix with values in $\{1, 2, \ldots\} \cup \{\infty\}$, such that $m_{s,s} = 1$ for all $s \in S$, $m_{s,t} \geq 2$ for $s \neq t$; $m_{s,t} = \infty$ simply means that the corresponding relation is to be omitted. The cardinality of S is called the *rank* of the group; sometimes we omit S from the notation and simply say that W is a Coxeter group. We refer to [Humphreys 90] for general information about Coxeter groups. Examples of Coxeter groups are Weyl groups of finite-dimensional or Kač-Moody semisimple Lie algebras, and finite groups generated by reflections in Euclidian space; other examples may be realized as discrete groups generated by reflections in hyperbolic space.

In their seminal paper [Kazhdan and Lusztig 79], Kazhdan and Lusztig have defined for each pair of elements (x, y) in W such that $x \leq y$ in the Bruhat ordering (to be defined below), a polynomial $P_{x,y} \in \mathbf{Z}[q]$. We will use in Section 4 the recursion formula which, in principle, leads to the computation of $P_{x,y}$. If W is the Weyl group of a finite-dimensional or Kač-Moody Lie algebra $\mathfrak{g}$, the Kazhdan-Lusztig polynomials of W hold the key to the representation theory of $\mathfrak{g}$, and also to the geometry of

1058-6458/2001 $0.50 per page
Experimental Mathematics **11**:3, page 371

the corresponding Schubert varieties. In this case, it is known that the coefficients of $P_{x,y}$ are nonnegative integers. For other W, very little is known about the $P_{x,y}$ (in particular, the positivity of their coefficients remains conjectural); they probably point to a yet-to-be-discovered geometry and/or representation theory.

Due to the fundamental importance of Kazhdan-Lusztig polynomials, and to the difficulty in computing them by hand except in very special cases, great efforts have been made to implement their calculation on a computer. As far as I know, previous attempts to achieve this (including my own) have always proceeded in three stages: (a) implement the group W, either by way of a linear representation, or combinatorially; (b) deal with the Bruhat ordering; and (c) implement the actual recursion. In practice, the main burden of the computation falls on the Bruhat order routines, but in turn, the efficiency of these routines will depend on how well we have been able to solve stage (a). In view of these difficulties, most computations I am aware of have been restricted to finite Coxeter groups (for the best programs, up to a group order of about half a million, say). The exceptions are the results posted by Mark Goresky [Goresky 96] on his homepage, concerning the first few hundred elements of the affine Weyl groups associated to root systems of rank ≤ 3, and $\tilde{A}_3$, and Bill Casselman's Kazhdan-Lusztig programs, which he has used to compute one- and two-sided cells in various finite, affine, and hyperbolic Coxeter groups (see [Casselman 00] for an application, and Casselman's homepage [Casselman 01]). Goresky's files contain the necessary data for the description of the singularities of the Schubert variety associated to $y \in W$, for small y; basically, this involves the computation of the element c_y in the Kazhdan-Lusztig basis (see Section 4.2), and pruning the result a little to extract the irreducible components of the stratification defined by equality of polynomials (a rough version of equisingularity.)

In this paper, we shall present an algorithm which, for any Coxeter group W, constructs directly from the Coxeter matrix $(m_{s,t})$ and a word $a = (s_1, \ldots, s_p)$ in the generators, the interval $[e, y]$ in the Bruhat ordering, where $y = s_1 \ldots s_p$ is the element of W represented by a, together with the (partially defined) left and right actions of all the generators on $[e, y]$, without any prior implementation of the group operations. This provides us with exactly the data we need to compute $P_{x,z}$ for all $x \leq z \leq y$ using the recursion formula of Kazhdan and Lusztig. The algorithm is based on the analysis of the structure of Bruhat intervals and some other Bruhat-like posets in [du Cloux 00]; the essential ingredient in its correctness proof is a remarkable theorem due to Matthew Dyer in his thesis [Dyer 87], which we shall discuss in Section 3.

We have made a preliminary implementation of this algorithm in a demonstration program available from the `Coxeter` homepage [du Cloux 01]. This program will compute the basis element c_y, and the data which would appear in Goresky's files describing the singularity of the corresponding Schubert cell whenever this makes sense, for an element y of moderate length in an arbitrary Coxeter group W, of rank less than 16, say (more details on the scope of the program are given in Section 6.) The performance limitations of the demonstration program are mainly due to the simple-minded routines used to access the Bruhat order within the Kazhdan-Lusztig computation. By using ideas from [du Cloux 99], we believe that its performance (regarding both speed and memory usage) could be improved considerably—we hope to include a full-fledged implementation in some future version of `Coxeter`.

2. DEFINITION AND ELEMENTARY PROPERTIES OF THE BRUHAT ORDERING

2.1 Posets

For any poset P, and $x \leq y$ in P, we denote by $[x, y]$ the set of $z \in P$ such that $x \leq z \leq y$. Let us say that P is *locally finite*, if all intervals $[x, y]$ in P are finite; assume from now on that P is locally finite. A *chain* in P is a totally ordered subset; a chain is *maximal* if it is not properly contained in any other chain. Define the length of a chain to be its cardinality minus one. We say that P is *graded*, if for all $x \leq y \in P$, all maximal chains in $[x, y]$ have the same length; this is also sometimes called the Jordan-Dedekind condition. Let us assume that furthermore P has a smallest element $\mathbf{0}$; then we define the *length function* l on P by setting $l(x)$ to be the length of the maximal chains in $[\mathbf{0}, x]$. For $x \leq y$ in P, we also define the length of the interval $[x, y]$ to be the length of its maximal chains; it is easy to see that the length of $[x, y]$ is equal to $l(y) - l(x)$. The *atoms* of an interval $[x, y] \subset P$, $x < y$, are the $z \in [x, y]$ such that $l(z) = l(x) + 1$; similarly, the coatoms of $[x, y]$ are the $z \in [x, y]$ such that $l(z) = l(y) - 1$. For $x \in P$, we will speak about the coatoms of x instead of the coatoms of $[\mathbf{0}, x]$ and denote their set $\mathrm{coat}(x)$. A *decreasing subset* of P is a subset Q such that if $y \in Q$ and $x \leq y$, then $x \in Q$; since P has a smallest element $\mathbf{0}$, this means simply that Q is a union of intervals $[\mathbf{0}, y]$.

2.2 Coxeter Groups

We denote S^* the free monoid generated by S, i.e., the set of all words in the alphabet S. To avoid confusion between elements of S^* and elements of W, we will write a word $a \in S^*$ as $a = (s_1, \ldots, s_p)$; if $x = s_1 \ldots s_p$ is the corresponding element in W, we say that a is an *expression* for x. The smallest possible number of letters in an expression for x is called the *length* of x, and denoted $l(x)$; an expression for x is called *reduced*, if it has exactly $l(x)$ letters. It is an elementary fact about Coxeter groups that for $x \in W$, $s \in S$, we have either $l(xs) = l(x) + 1$, in which case we write $xs > x$, or $l(xs) = l(x) - 1$, in which case we write $xs < x$, and of course, we have an analogous statement for left multiplications.

2.3 Bruhat Ordering

The following proposition can be used to define the Bruhat ordering :

Proposition 2.1. *There exists a unique ordering on W, which we call the Bruhat ordering, and denote $\leq$, such that* (a) *$(W, \leq)$ has smallest element e* (b) *for each $x \in W$, $s \in S$ such that $l(xs) < l(x)$, $[e, x]$ is the (non-necessarily disjoint) union of $[e, xs]$ and $[e, xs]s$.*

Proof: Uniqueness is clear, since all intervals $[e, x]$ are uniquely defined by induction on the length of x, and the knowledge of these intervals is enough to determine the poset structure. For the existence, the main point is to show that $[e, xs] \cup [e, x]s$ is independent of the choice of s, so that it can be used as a definition of $[e, xs]$. This can be proved directly by an elementary argument using dihedral groups, but we will omit the proof, since it follows from the usual definition of the Bruhat ordering (see the Proposition in [Humphreys 90], Section 5.9). □

2.4 First Properties of the Bruhat Ordering

The Bruhat ordering satisfies the following properties:

(a) The previous usage of $<$ in Section 2.2 is compatible with Proposition 2.1: if $x \in W$ and $s \in S$ are such that $l(xs) < l(x)$, then $xs \in [e, x]$, hence $xs < x$ for the Bruhat ordering.

(b) If $a = (s_1, \ldots, s_p)$ is any reduced expression for x, the interval $[e, x]$ is the set of all elements $z \in W$ of the form $z = s_{j_1} \ldots s_{j_q}$, where $1 \leq j_1 < \ldots < j_q \leq p$ (the elements corresponding to the so-called *subexpressions* of a.) This follows easily from (b) in proposition 2.1 by induction on the length of x; in particular, our definition of the Bruhat ordering is equivalent to the usual one (see for instance [Humphreys 90], Section 5.9).

(c) $x \to x^{-1}$ is an automorphism of the Bruhat ordering (this follows from (b) above).

(d) The Bruhat ordering also satisfies the property analogous to (b) in Proposition 2.1 for left multiplications (this follows from (c)).

(e) If $x \in W$, $s \in S$ are such that $xs < x$, the interval $[e, x]$ is stable under right multiplication by s; this is clear from the equality $[e, x] = [e, xs] \cup [e, xs]s$. The analogous property holds for left multiplications.

2.5 Further Properties of the Bruhat Ordering

Two other essential properties of the Bruhat ordering lie slightly deeper than the previous observations. The first one is that the Bruhat ordering is *graded* (see [Humphreys 90], Section 5.11); moreover, the length function on W as a poset coincides with the length function previously defined for elements of W. The second one is that the poset W is *Eulerian*. This means that we have $\chi_{[x,y]} = 0$ for each $x < y$ in W, where $\chi_{[x,y]}$ is the "Euler characteristic" defined by

$$\chi_{[x,y]} = \sum_{x \leq z \leq y} (-1)^{l(z)-l(x)}.$$

Equivalently, the Möbius function of W is given by $\mu(x, y) = (-1)^{(l(y)-l(x))}$ for all $x < y$ in W (see [Stanley 97], Sections 3.14 and 3.7, for more details on Eulerian posets and Möbius functions.) In this form, this was proved by Verma [Verma 71].

An elementary consequence is that all intervals $[x, y]$ for which $l(y) - l(x) = 2$ have exactly four elements: x, y, and exactly two intermediary elements z and z'. Also, it follows that for any $x < y$ in W, the cardinality of $[x, y]$ is *even*.

3. DYER'S THEOREM

3.1 Dihedral Coxeter Groups

A *dihedral* Coxeter group is simply a Coxeter group of rank 2, i.e., for which the generating set S has two elements s, t. If $m = m_{s,t}$ is finite, then W is a finite group of order $2m$; if m is infinite, then W is the infinite dihedral group, isomorphic to the (nontrivial) semidirect product of $\mathbf{Z}$ with $\mathbf{Z}/2\mathbf{Z}$. The Bruhat ordering of a dihedral group is particularly easy to describe: there are exactly two elements in each length j such that $0 < j < m$, one in length 0, and one in length m if $m < \infty$; and all elements in length $j > 0$ are comparable to all elements

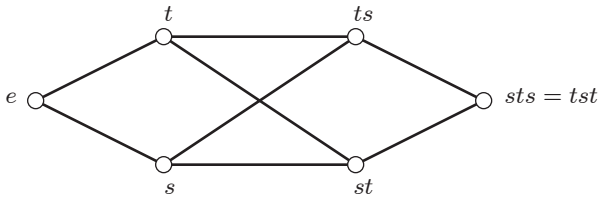

FIGURE 1. Hasse diagram of the Bruhat ordering in type A_2.

in length $j-1$. For example, if $m=3$, we get the familiar picture of the Bruhat ordering on the Weyl group of type A_2 (also the symmetric group on three letters), shown in Figure 1.

We say that an interval $[x,y]$ of length >1 in an arbitrary Coxeter group W is dihedral if it is isomorphic as a poset to the Bruhat ordering on a (finite) dihedral group. Let us also make the convention that intervals of length 0 (one-element sets) and of length 1 (two-element sets) are dihedral. Then it is easy to see that any subinterval in a dihedral interval is again dihedral. We shall say that $y \in W$ is dihedral, if the interval $[e,y]$ is dihedral; equivalently, there exist s,t in S such that y belongs to the subgroup of W generated by s and t.

3.2 Dihedral Subgroups

We now explain a striking result due to Matthew Dyer (one of the many striking results contained in his thesis [Dyer 87]) that imposes very important restrictions on the Bruhat ordering of a Coxeter group. For its proof (which we reproduce here since this result of Dyer's apparently has not been included in his publications), we first need to review some of his other results.

The *reflections* in W are defined to be the conjugates of the elements of S; so the set T of reflections is a finite union of conjugacy classes of elements of order 2 in W. The *canonical geometrical realisation* of W is defined as follows. Consider the real vector space $V=\mathbf{R}^S$, of dimension n, and its standard basis $(e_s)_{s\in S}$. Endow V with the unique symmetric bilinear form $\langle\ ,\ \rangle$ for which $\langle e_s, e_t\rangle = -\cos(\frac{\pi}{m_{s,t}})$. Then there is a unique injective homomorphism from W into the orthogonal group of V taking s to the orthogonal reflection with respect to the hyperplane $e_s^\perp$ ([Humphreys 90], Section 5.3); this is the canonical geometrical realization. We define the *roots* of W to be the conjugates under W of the basis elements e_s, and denote the set of roots by Φ. Since $s.e_s = -e_s$, we have $\Phi = -\Phi$. Reflections in W will correspond to orthogonal reflections with respect to elements of Φ; more precisely, if for each $t\in T$ we define its reflection line L_t to be the (-1)-eigenspace of the action of t in V, then $t \to L_t \cap \Phi$ is a bijection from T to pairs of opposite elements in Φ.

We will define a dihedral subgroup of W to be any subgroup generated by two distinct reflections. We will say that two dihedral subgroups are *commensurate*, if they are contained in a common dihedral subgroup. For any dihedral group D, it is easy to see that for any choice of reflections t_1, t_2 generating D, $D\cap T$ is exactly the set of elements of odd length with respect to the chosen generators, and also the set of conjugates of t_1 and t_2 in D. Hence, all the reflection lines L_t, $t\in D\cap T$, lie in the two-dimensional subspace V_D of V spanned by L_{t_1} and L_{t_2}; this shows immediately that V_D does not depend on the choice of generators. If a dihedral subgroup is contained in another, it is clear that the corresponding two-dimensional subspaces are the same; hence, the same is true for commensurate dihedral subgroups. The converse also holds, and follows from the proof of the following lemma [Dyer 87, Corollary 3.18]:

Lemma 3.1. *Each dihedral subgroup of W is contained in a unique maximal one.*

Proof: Let α, β be two nonproportional elements of Φ, and let V_1 be the two-dimensional subspace of V spanned by α and β. Let $\Phi_1 = \Phi\cap V_1$, and let D be the subgroup of W generated by the reflections r_γ, $\gamma\in\Phi_1$, where for each unit vector $u\in V$, we denote by r_u the reflection $x\to x-2\langle u,x\rangle u$. It will suffice to show that D is dihedral. If $\langle\ ,\ \rangle_1$ is the restriction of $\langle\ ,\ \rangle$ to V_1, there are three cases to consider: (a) $\langle\ ,\ \rangle_1$ is positive definite; (b) $\langle\ ,\ \rangle_1$ is nonzero positive degenerate; (c) $\langle\ ,\ \rangle_1$ has signature $(1,-1)$. In cases (a) and (c), take $V_2 = V_1^\perp$; in case (b), choose a subspace $V_2\subset V_1^\perp$ such that $V = V_1\oplus V_2$. Then in all cases, D acts trivially on V_2. So D is contained in $\mathbf{O}(V_1)\times\{\mathrm{Id}_{V_2}\}\subset\mathbf{O}(V)$.

But it is well known ([Bourbaki 68], Chapter V, n^o 4.4, Corollary 3) that W is a discrete subgroup of $\mathbf{O}(V)$; hence, D may be identified with a discrete subgroup of $\mathbf{O}(V_1)$. Moreover, as a Lie group $\mathbf{O}(V_1)$ always has the form $\mathbf{Z}_2\ltimes A$, where A is isomorphic to $\mathbf{R}/\mathbf{Z}$ in case (a), to $\mathbf{R}$ in case (b), and to $\mathbf{R}^\times$ in case (c), with $\mathbf{Z}_2$ acting by inversion; in case (c), D is contained in $\mathbf{Z}_2\ltimes\mathbf{R}_+^*$. Then from elementary topological considerations, one sees that in all three cases D is a dihedral subgroup of W, finite in case (a), infinite in the two other cases. □

3.3 The Reflection Subgroup of a Bruhat Interval

In fact, Dyer [Dyer 90] Theorem 3.3 shows that if R is any (finite, say) subset of T, the subgroup W' of W generated

by R is always a Coxeter group in its own right (this was proved independently by Deodhar in [Deodhar 89]). More precisely, he defines a canonical subset $S' \subset T \cap W'$ such that S' is a set of Coxeter generators for W'; he shows that $|S'| \leq |R|$, although it is certainly possible that $|S'| > |S|$. In particular, when W' is dihedral, S' is a two-element set. We shall call such a subgroup W' a *reflection subgroup* of W.

It is a well known and useful fact in the theory of Coxeter groups ([Humphreys 90] section 5.12) that if I is any subset of S, and if W_I is the subgroup of W generated by I, then any (left, say) coset $W_I x$ of W_I in W contains a unique element of minimal length. This was extended in [Dyer 90] Corollary 3.4 to the case of an arbitrary reflection subgroup W'.

Let $x < y$ in W be such that $l(x) = l(y) - 1$. Then it is an easy consequence of 2.4 (b) that if $y = s_1 \dots s_p$ is a reduced decomposition, x may be obtained by erasing a single s_j from the decomposition (and in fact, j is unique.) This may also be expressed by saying that there exists a reflection $t \in T$ such that $x = yt$ (take $t = s_p \dots s_{j+1} s_j s_{j+1} \dots s_p$); so we have $y^{-1}x = x^{-1}y \in T$.

The following theorem regroups the main results of Dyer's study of the reflection subgroups arising from maximal chains in subintervals [Dyer 91, Proposition 2.1]:

Theorem 3.2. *Let $x < y$ in W, and let $x = x_0 < x_1 < \dots < x_m = y$ be a maximal chain in $[x, y]$, so that $m = l(y) - l(x)$. For $1 \leq j \leq m$, set $t_j = x_{j-1}^{-1} x_j$, and let $W' \subset W$ be the subgroup generated by the t_j. Then*

(a) W' is independent of the choice of the maximal chain.

(b) Let z be the unique element of minimal length in $W'x$; then $[x, y]$ is contained in the left coset $W'z$, and the map $u \to uz^{-1}$ defines a poset isomorphism from $[x, y]$ to $[x', y'] \subset W'$, where $x' = xz^{-1}$, $y' = yz^{-1}$, and $[x', y']$ is the interval in the Bruhat ordering defined by the canonical generating set S' of W'.

3.4 Dyer's Characterization of Dihedral Intervals

The above theorem is the essential ingredient in the proof of the theorem on which our algorithm is based, and which may be stated as follows :

Theorem 3.3. ([Dyer 87] Proposition 7.25.) *Let (W, S) be an arbitrary Coxeter system, and let $[x, y]$ be a Bruhat interval in W of length at least two. Then the following are equivalent :*

(i) $[x, y]$ has two atoms;

(ii) $[x, y]$ has two coatoms;

(iii) $[x, y]$ is dihedral.

Proof: Of course (iii)⇒(i) and (iii)⇒(ii) are trivial. Let us prove, for instance that (ii)⇒(iii). We argue by induction on $l(y) - l(x)$. If $l(y) - l(x)$ is two, there is nothing to prove. So we may assume that $l(y) - l(x)$ is at least three.

Let t, t' be the two reflections of W taking y to the two coatoms of $[x, y]$ (see Section 3.3). From Lemma 3.1, the dihedral subgroup $\langle t, t' \rangle$ is contained in a unique maximal dihedral subgroup D. From Theorem 3.2, it suffices to prove that the subgroup generated by all the $u^{-1}v$, $u < v$ in $[x, y]$, $l(u) = l(v) - 1$, is dihedral; and since any reflection subgroup (containing at least two reflections) of a dihedral group is again dihedral, it will suffice to show that $u^{-1}v \in D$ for all such u, v.

We argue by induction on $l(y) - l(u)$. If $l(y) - l(u) = 1$, we have $u^{-1}v \in \{t, t'\}$, so there is nothing to prove. Assume $l(y) - l(u) > 1$, and let z be an atom of $[v, y]$, so that $l(z) - l(u) = 2$. Write $[u, z] = \{u, v, v', z\}$, and let $t_1 = u^{-1}v$, $t_2 = v^{-1}z$, $t_2' = v'^{-1}z$. Then we know from Theorem 3.2 that the two dihedral subgroups $\langle t_1, t_2 \rangle$ and $\langle t_2, t_2' \rangle$ are commensurate (in fact the latter is contained in the former); but from the inductive hypothesis, $\langle t_2, t_2' \rangle \subset D$; so $\langle t_1, t_2 \rangle \subset D$ as well, and in particular $t_1 \in D$. The proof of (i)⇒(iii) is entirely similar. □

Corollary 3.4. *Let $y \in W$ be non-dihedral (see Section 3.1), and let $s \in S$. Then if $zs < s$ for all $z \in \mathrm{coat}(y)$ except at most one, we have $ys < y$; the analogous property holds for left multiplications.*

Proof: We consider the case of right multiplications. If $y = e$, there is nothing to prove. So let $y > e$, and assume that $ys > y$. If we would have $zs < z$ for *all* $z \in \mathrm{coat}(y)$, then from Section 2.4 (e), $[e, y[:= [e, y] \setminus \{y\}$ would be stable under right multiplication by s; but on $[e, y[$ this defines an involution without fixed points, contradicting the fact that $|[e, y[|$ is odd, since $|[e, y]|$ is even from Section 2.5.

Hence, there is a unique $z \in \mathrm{coat}(y)$ such that $zs > z$. Now any $x < y$ other than z satisfies $x \leq z'$ for some $z' \in \mathrm{coat}(y)$, $z' \neq z$: this is clear if $x \in \mathrm{coat}(y)$, and otherwise follows from the fact that $[x, y]$ has at least two coatoms if $l(y) - l(x) > 1$. Hence $[e, y] \setminus \{z, y\}$ is stable under right multiplication by s, and $[e, ys]$ contains exactly two elements not already in $[e, y]$, *viz.* zs and ys; in

particular, $\mathrm{coat}(ys) = \{y, zs\}$, which from Theorem 3.3 implies that $[e, ys]$ is dihedral; but then $[e, y]$ is dihedral as well, contradicting our assumption on y. □

4. KAZHDAN-LUSZTIG POLYNOMIALS

4.1 Recursion Formulae

We refer to the original paper [Kazhdan and Lusztig 79] or to [Humphreys 90], Chapter 7, for the proper definition and proofs of the basic properties of the Kazhdan-Lusztig polynomials. Our goal here is to recall the recursion formulæ from [Kazhdan and Lusztig 79] which we use to compute these polynomials, in order to assess the data which will be needed in the process.

Let q be an indeterminate. Then it is clear that there is at most one family $P_{x,y}$ of elements of $\mathbf{Z}[q]$, defined for $x \leq y$ in W, satisfying the following requirements :

(a) $P_{x,x} = 1$ for all $x \in W$;

(b) if $x < y$, and $s \in S$, are such that $ys < y$ and $xs > x$, $P_{x,y} = P_{xs,y}$;

(b') if $x < y$, and $s \in S$, are such that $sy < y$ and $sx > x$, $P_{x,y} = P_{sx,y}$;

(c) if $x < y$, and $s \in S$, are such that $ys < y$ and $xs < x$, and x is not comparable to ys, $P_{x,y} = P_{xs,ys}$;

(d) if $x < y$, and $s \in S$, are such that $ys < y$, $xs < x$, and $x \leq ys$, we have

$$P_{x,y} = P_{xs,ys} + qP_{x,ys} - \sum_{\substack{x \leq z < ys \\ zs < z}} q^{\frac{1}{2}(l(y)-l(z))} \mu(z, ys) P_{x,z},$$

where $\mu(z, ys)$ is the coefficient of degree $\frac{1}{2}(l(ys) - l(z) - 1)$ in $P_{z,ys}$ (defined to be 0 if $l(ys) - l(z)$ is even); it is not hard to show by induction that, in fact, $P_{x,y}$ is at most of degree $\frac{1}{2}(l(y) - l(x) - 1)$ when $x < y$.

4.2 Kazhdan-Lusztig Basis

In fact, the $P_{x,y}$ are (up to a degree shift) the coordinates of the elements of a remarkable basis of the Hecke algebra of W, the so-called Kazhdan-Lusztig basis. There is one such basis element c_y for each y in W, and in many applications it is the knowledge of such a c_y which is required. In other words, a common requirement will be the computation of the $P_{x,y}$ for a fixed y, and all $x \leq y$. The whole recursion then takes place in the interval $[e, y]$, which is finite even if the group is infinite. We see that in order to carry out the recursion, we will need the following:

(a) an enumeration of the interval $[e, y]$, and a description of the Bruhat ordering on it;

(b) in this enumeration, the action of the generators $s \in S$ on the left and on the right.

Note that the action of each $s \in S$ on $[e, y]$ is only partially defined. In fact, it follows from Proposition 2.1 that it is everywhere defined (on the right, say) if and only if $ys < y$; in other words, if and only if s belongs to the so-called right descent set of y (see Section 4.3 below). Otherwise, xs remains within $[e, y]$ if and only if there exists $z \geq x$ in $[e, y]$ such that $zs < z$. We will say that this is the *domain* of the right action of s; it is a (possibly empty) decreasing subset of $[e, y]$.

4.3 Descent Sets

For each $y \in W$, we define $L(y)$ $(R(y))$ to be the set of $s \in S$ such that $sy < y$ $(ys < y)$; we set $LR(y) = L(y) \coprod R(y) \subset S \coprod S$. We say that $L(y)$ $(R(y), LR(y))$ is the left (right, two-sided) descent set of y. These descent sets play an important role in the theory.

We notice that the knowledge of $LR(x)$ for each $x \leq y$ suffices to determine the left and right actions of all generators on $[e, y]$ (this remark could be used if a compact data encoding is required, but it is likely to be unpractical for systematic computations). Indeed, encoding $LR(x)$ as a sequence of $2|S|$ bits in the obvious way, and looking at the bit position for, say, the right action of a generator s, we see that if the bit for x is set, meaning $xs < x$, there will be exactly one among the coatoms z of x for which the corresponding bit is *not* set; indeed, $[e, x]$ is stable under right multiplication by s, so $zs > z$ happens if and only if $zs = x$, hence $z = xs$ (these are precisely the remarks underlying the axiomatization in [du Cloux 00]).

4.4 Outline of the Computation

Assuming we have solved (a) and (b) above, the only further (and obvious) idea used in our program is to remember the values of all the polynomials already computed. More precisely, assume that a polynomial $P_{x,z}$ is required for $x \leq z \leq y$. The program first reduces to the case where $LR(x) \supset LR(z)$ (this amounts to putting ourselves in cases (c) or (d) of Section 4.1). Then it looks up a list of such x; the first time this occurs for a given z, it actually has to make the list, which involves extracting the subinterval $[e, z]$—we will come back to this problem, which is probably the most time-consuming part of the computation, in the next section. If the polynomial has already been computed, it will find it there; otherwise, it

uses the recursion formula, potentially triggering many other computations, finds the requested $P_{x,z}$, and writes it down.

5. DESCRIPTION OF THE MAIN ALGORITHM

5.1 Data Structures

We shall now describe an algorithm which takes as input an arbitrary word over the alphabet S, and produces as output the poset $[e, y]$, and for each $x \in [e, y]$, the "shift-table" of x recording the result of the left and right actions of the generators on x. The only other datum that this algorithm needs is the Coxeter matrix of W, which is assumed to have been somehow read into memory. In other words, for s, t in $\{1, \ldots, n\}$, we may call a function $\texttt{Cox}(s, t)$, which will return $m_{s,t}$ (with some convention for representing ∞; in our program, it is represented by 0).

An enumeration of $[e, y]$ simply means an identification of $[e, y]$ with the numbers 0 to $N - 1$, where N is the cardinality of $[e, y]$. We view this as a function ν from $[e, y]$ to $[0, N[\subset \mathbf{N}$. We shall always require that the enumeration be *increasing*: i.e., if $x < z$ in $[e, y]$, we want to have $\nu(x) < \nu(z)$. Increasing enumerations exist for any finite poset. On the other hand, we do not insist that the enumeration be length-first; in other words, we do not require that $l(x) < l(z)$ implies $\nu(x) < \nu(z)$. It is always possible to do a length-sort if and when this becomes necessary (for instance, in order to have pleasant output.) In particular, the fact that ν is increasing implies that $\nu(e) = 0$.

In order to represent a graded poset, it is enough to give for each x in $[e, y]$ the list of coatoms of x; this will be empty if and only if $x = e$.

To summarize, we wish to find the cardinality N of $[e, y]$, and construct the following data:

(a) for each $x \in [0, N[$, the list $\texttt{coat}[x]$ of the elements in $[0, x[$ which correspond to coatoms of x;

(b) for each $x \in [0, N[$, the length $\texttt{length}[x]$;

(c) for each $x \in [0, N[$, the shift table $\texttt{shift}[x]$; this is the datum of $2n$ elements of $[0, N[\cup \infty$, corresponding to the left and right action of the generators, where ∞ is a special value, indicating that the corresponding shift is undefined (in practice, it is convenient to take for ∞ a value which is larger than any legal value for x).

(d) for each $x \in [0, N[$, the descent set $\texttt{LR}[x]$; of course these are trivially deduced from the shift tables, and we have seen that the converse is also true; however, it is convenient to have them both available.

Notice that these data imply the enumeration of the poset: clearly, if all left shifts are known, it is a trivial matter to write down for each $x \in [0, N[$ a reduced expression for the corresponding group element, and indeed the `ShortLex` normal form (this is by definition the lexicographically smallest reduced expression, corresponding to the ordering of S implied by its identification with $\{1, \ldots, n\}$.) Conversely, if we are given a reduced expression of an element of $[e, y]$, following the right shifts from the identity, we find the corresponding $x \in [0, N[$ (in fact, this works for any expression, reduced or not, provided the path does not take us outside of $[e, y]$.) So we will leave ν implicit in the sequel.

At the beginning of the algorithm, the data are initialized to their values for $y = e$: we have $N = 1$, the coatom list of $x = 0$ is empty, the length of $x = 0$ is 0, the shifts are all set to ∞, and the descent set $\texttt{LR}[0]$ is empty. We shall describe in the next sections the main loop of the algorithm, passing from the data for $[e, y]$ to the data for $[e, ys]$, where s is a new letter in our input word. A very nice feature of the construction is that, apart from some undefined shifts becoming defined, it fully preserves the data already constructed for $[e, y]$, at least if $ys > y$ (which should be the "hard" case, and always happens if the word is reduced.) So once we have determined the size of the bigger interval, we simply resize our lists and fill in the new part.

5.2 Poset Structure

It is easy to find out how many new elements have to be added to our interval. Indeed, from Section 2.4 (e), we see that for each $x \in [e, ys]$ we have $x \leq y$ or $xs \leq y$ or both; hence the new elements are the $x = x's$, where x' runs through the elements in $[e, y]$ for which $x's$ is not already in $[e, y]$, i.e., for which the right shift of x' by s is undefined. From this, we get the new value of N, and we can fill in the length function and the right shifts by s (note that right shift by s is everywhere defined on $[e, ys]$.)

Now we may construct the coatom lists of the new elements as follows ([du Cloux 00], Section 2.9 and Proposition 2.14.) Traverse the list of new elements in increasing order. For each x, let $x' = xs$; then the coatoms of x are x', and the $z = z's$, where z' runs through the coatoms of x' in $[e, y]$ for which $z's > z'$. Since the coatoms of x are all represented by strictly smaller integers, it is clear that the enumeration is indeed increasing.

5.3 Shifts

It remains to explain how to find the shifts other than right multiplication by s. Let σ be such a shift, i.e., σ is either right multiplication by some $t \neq s$, or left multiplication by any t. Before we start, all the $\sigma(x)$ for the newly created x are set to undefined. Let $\Delta \subset [e, y]$ be the previous domain of σ, and let $\Delta' \subset [e, ys]$ be the new one. Then we need to define σ on $\Delta' \setminus \Delta$; this will involve some of the newly created elements, and some already existing elements for which σ was previously undefined. But we notice that, in fact, $\Delta' \setminus \Delta = \coprod\{\sigma(x), x\}$, where x runs through the set of newly created elements such that $\sigma(x) < x$. So if for each newly created element x we are able to decide whether $\sigma(x) < x$ or $\sigma(x) > x$, we are done: if $\sigma(x) > x$, the corresponding shift is left undefined for the time being; if $\sigma(x) < x$, there is a unique coatom z of x such that $\sigma(z) > z$ (in fact, $\sigma(z)$ will have the value ∞ at this point), and we set $\sigma(x) = z$, $\sigma(z) = x$.

So again, we traverse the list of newly created elements in increasing order. If $l(x) = 1$ (which can happen only if x is the first new element and s is a generator which had not occurred before), then $x = s$, and the only σ other than right shift by s that can take it down is left shift by s; so we conclude directly in this case. Assume from now on that $l(x) > 1$. We have already remarked that if $\sigma(x) < x$, there is a single coatom z of x such that $\sigma(z) > z$; so if there are at least two such coatoms, we may conclude without further ado that $\sigma(x) > x$. Otherwise, from the Theorem in Section 3.4, we can decide if x is dihedral or not by looking at the cardinality of `coat`$[x]$. If x is nondihedral, we conclude from Corollary 3.4 that $\sigma(x) < x$. If x is dihedral, it is easy to conclude directly: the other generator involved in x is the unique element t in `R`$[xs]$. Since we have assumed that $\sigma(z) < z$ for one of the two coatoms of x, σ is either left multiplication by s, or right or left multiplication by t. If $l(x) = m_{s,t}$, $\sigma(x) < x$; otherwise, $xt > x$, and $sx < x$, $tx > x$ if $l(x)$ is odd, $sx > x$, $tx < x$ if $l(x)$ is even (note that this is the *only* place in the whole algorithm where the Coxeter matrix comes in.) This concludes the main loop of the algorithm.

5.4 Nonreduced Expressions

Each time the algorithm reads a new generator from its input word, we are in the situation where the data for the interval $[e, y]$ have been constructed, and we want the data for $[e, ys]$. We have assumed so far that $ys > y$ (note that from the shift tables for y, which are available when we read s, we can immediately determine whether or not this is the case.) If, on the contrary, $ys < y$, we are in the situation where we have to restrict to a subinterval $[e, ys]$ of the already constructed interval. There is a straightforward way of extracting the subinterval "backwards" from the knowledge of the coatom lists. In our program, however, we prefer to imitate the above procedure, using for instance the `ShortLex` normal form for ys, to extract $[e, ys]$ in the form of a list of elements in $[e, y]$. Then as the construction proceeds, we have to deal with the situation where we only enlarge a *subinterval* of our current poset, which can no longer be assumed to be an interval, but rather an arbitrary finite decreasing subset of W; in fact, this causes no trouble at all, and can be handled exactly as above.

Note that there are efficient methods available to construct *a priori* the normal form of an arbitrary element in an arbitrary Coxeter group (see [Casselman 02] for a nice exposition of the practical implementation of the ideas from Brink and Howlett in [Brink and Howlett 93]). So the trouble caused by nonreduced expressions could in principle be avoided entirely. However, as we have seen in Section 4.4, it will in any case be necessary to extract many subintervals of the form $[e, x]$ in the course of the Kazhdan-Lusztig computations, so this problem has to be addressed one way or another.

6. SCOPE AND FURTHER DEVELOPMENTS

6.1 Poset Memory Requirements

We would like to conclude with a few informal remarks on the resources required by this algorithm. In our experience, given enough memory, time has never been a factor in Kazhdan-Lusztig computations. It only becomes a problem if we are unable to keep in memory the tables described in this article, or the polynomials already computed. Assuming for simplicity that everything is represented by 32-bit unsigned integers, the memory requirements for the table constructions are not hard to evaluate. The limiting factor is the size N of the actual poset $[e, y]$ we are constructing.

The sizes of the length, descent, and shift tables are exactly N, N, and $2nN$, respectively (if we assume that the rank does not exceed 16, so that the descent sets can be represented by a single word.). The size of the coatom lists is harder to predict, but a rule-of-thumb for the most common cases (where there is enough commutativity around) is that $2nN$ is a reasonable estimate for the total number of coatoms (for groups like the free Coxeter group, where all the coefficients $m_{s,t}$ are infinite, the

$y_1 = 21321324321323432132 \in F_4$; (length 21)
$y_2 = 2121321213212432121321234321213212343212132123432121321234321213212 \in H_4$; (length 56)
$y_3 = (321212)^8.3.(212123)^8 \in \tilde{G}_2$ (length 97);
$y_4 = (12345)^8 \in \tilde{A}_4$ (length 40).

FIGURE 2. Some Coxeter group elements.

$$\begin{pmatrix} 1 & 3 & 2 & 2 \\ 3 & 1 & 4 & 2 \\ 2 & 4 & 1 & 3 \\ 2 & 2 & 3 & 4 \end{pmatrix} \quad \begin{pmatrix} 1 & 5 & 2 & 2 \\ 5 & 1 & 3 & 2 \\ 2 & 3 & 1 & 3 \\ 2 & 2 & 3 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 6 & 2 \\ 6 & 1 & 3 \\ 2 & 3 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 3 & 2 & 2 & 3 \\ 3 & 1 & 3 & 2 & 2 \\ 2 & 3 & 1 & 3 & 2 \\ 2 & 2 & 3 & 1 & 3 \\ 3 & 2 & 2 & 3 & 1 \end{pmatrix}$$

FIGURE 3. Coxeter matrices of F_4, H_4, $\tilde{\mathrm{G}}_2$, $\tilde{\mathrm{A}}_4$.

size would be much larger). In addition, there is an unavoidable overhead of $2N$ elements, because we have to indicate somehow the number of elements in each coatom list, and we need a pointer to the beginning of each list. So we end up with an estimate of $4(n+1)N$ long integers, which seems to be pretty well-validated by experience.

6.2 Kazhdan-Lusztig Memory Requirements

Of course, we will still need a sizeable amount of additional memory for the Kazhdan-Lusztig computations. In order to give an idea of how big this requirement might be, we print a table obtained for a few typical cases.

We consider the elements $y_1, \ldots, y_4$ defined in Figure 2 (Here we use the usual Coxeter matrices for F_4, H_4, $\tilde{\mathrm{G}}_2$ and $\tilde{\mathrm{A}}_4$, see Figure 3.). The elements in F_4 and H_4 are (at least in our experience) among the worst possible ones: in fact, they are elements of longest length with the property of having one-element left and right descent sets, which are moreover equal. The element chosen in $\tilde{G}_2$ also has this property. The element in $\tilde{A}_4$ has one-element left and right descent sets, but they are unequal.

The breakup of the corresponding memory costs has been collected in Table 1. A few explanatory comments may be in order regarding this table. The memory cost for the poset construction is computed as explained above (on a machine where pointers also have a size of 4 bytes.) The number of Hasse edges is, in fact, the total number of coatoms for the various lists `coat`$[x]$; notice that our heuristic bound of $2nN$ holds in these examples.

As we have explained in Section 4.4, in order to record the polynomials already computed, we maintain an array of N rows, where row z holds the $P_{x,z}$ for the elements $x \leq z$, which are in "extremal position" with respect to z; a row is allocated if and when a $P_{x,z}$ is actually required. The allocation requires eight bytes for each polynomial: four to record the value of x, and four for a pointer to the actual polynomial, which is written down in full only once. The header of each row also requires eight bytes. In fact, we maintain another such table for recording μ-coefficients (when a μ-coefficient is required, we try as much as possible to compute it without unnecessarily computing full polynomials; also, μ-coefficients are essential information in many applications.) We allocate a μ-coefficient only if in addition to x being extremal with respect to z, the length difference is odd and ≥ 3. Again, each allocation takes up eight bytes. Finally, we do not allocate the row for z if $z^{-1} < z$ in our enumeration; it is reasonably easy to deduce row z from row z^{-1} if one maintains a (partially defined) table of inverses, at an additional cost of N words. The cost of the memory tables is the sum of the $4N$ words for the headers, the space for the allocation of the nonempty rows, and the space for the table of inverses.

When many distinct polynomials appear, the space used for recording them becomes an important part of the memory requirement (sometimes the dominant part.) The cost of storing them is $\deg(P) + 2$ words for each P, plus the cost of a searching structure to access the store efficiently (in our case, a hash table). The total cost of the computation is the sum of the costs for the poset, memory tables, and polynomial storage.

6.3 Conclusion

The upshot of this analysis is that on a computer with 512 Mb of memory available, the computation of the basis vector c_y should go through for a size of $[e, y]$ of about 2^{18}

	$y_1 \in F_4$	$y_2 \in H_4$	$y_3 \in \tilde{G}_2$	$y_4 \in \tilde{A}_4$
N	988	14 042	9 276	56 410
Hasse edges	5 244	98 357	53 925	496 734
memory cost for poset construction (bytes)	68 400	1 067 444	586 740	5 145 896
Kazhdan-Lusztig polynomials allocated	1383	379 991	572 003	2 221 661
Kazhdan-Lusztig polynomials computed	887	246 895	416 994	1 569 005
mu coefficients allocated	410	181 387	269 985	1 042 705
mu coefficients computed	146	107 148	191 284	619 255
memory cost for memory tables (bytes)	34 104	4 771 864	6 921 424	27 243 128
distinct polynomials found	135	67 864	190 860	23 946
$\sum \deg(P) + 2$	728	653 508	2 992 386	240 106
memory cost for polynomial storage (bytes)	3992	3 156 944	13 496 424	1 151 992
total cost for computation (bytes)	106 496	8 996 252	21 004 588	33 541 016

TABLE 1. Memory costs for some typical computations.

to 2^{20} if the rank is not bigger than 8, say (this will not be possible with the demonstration program, however; a more careful implementation of the Kazhdan-Lusztig computation will be needed.) In practice, this seems to correspond to elements of length around 40 or 50, unless the rank of the group is small, where lengths might get larger (but not much larger than 100, except in very easy cases.)

REFERENCES

[Bourbaki 68] N. Bourbaki. *Groupes et algèbres de Lie, Chap. 4-6.* Hermann, Paris, 1968.

[Brink and Howlett 93] B. Brink and D. Howlett. "A finiteness property and an automatic structure for Coxeter groups." *Math. Ann.* **296** (1993), 179–190.

[Casselman 00] W. Casselman. "Verifying Kottwitz' conjecture by computer." *Representation Theory* **4** (2000), 32–45.

[Casselman 01] W. Casselman. www.math.ubc.ca/people/faculty/cass/ cass.html.

[Casselman 02] W. Casselman. "Computation in Coxeter Groups I. Multiplication." Electron. J. Combin. **9**:1 (2002), Research Paper 25, 22 pages.

[Deodhar 89] V. Deodhar. "A note on subgroups generated by reflections in Coxeter groups." *Arch. Math. (Basel)* **53**:6 (1989), 543–546.

[du Cloux 01] F. du Cloux. `Coxeter`, demonstration version. available from http://www.desargues.univ-lyon1.fr/home/ducloux/coxeter.html.

[du Cloux 99] F. du Cloux. "A transducer approach to Coxeter groups." *J. Symbolic Computation* **27** (1999), 311–324.

[du Cloux 00] F. du Cloux. "An abstract model for Bruhat intervals." *Europ. J. Combinatorics* **21** (2000), 197–222.

[Dyer 87] M. Dyer. *Hecke algebras and reflections in Coxeter groups.* PhD thesis, University of Sydney, 1987.

[Dyer 90] M. Dyer. "Reflection subgroups of Coxeter systems." *J. Algebra* **135** (1990), 57–73.

[Dyer 91] M. Dyer. "On the "Bruhat graph" of a Coxeter system." *Compositio Math.* **78** (1991), 185–191.

[Goresky 96] M. Goresky. "Tables of Kazhdan-Lusztig polyomials." available from http://www.math.ias.edu/∼goresky.

[Humphreys 90] J.E. Humphreys. *Reflection Groups and Coxeter Groups.* Cambridge University Press, Cambridge, UK, 1990.

[Kazhdan and Lusztig 79] D. Kazhdan and G. Lusztig. "Representations of Coxeter groups and Hecke algebras." *Invent. Math.* **53** (1979), 165–184.

[Stanley 97] R.P. Stanley. *Enumerative Combinatorics*. Cambridge University Press, Cambridge, UK, 1997.

[Verma 71] D.-N. Verma. "Möbius inversion for the Bruhat ordering on a Weyl group." *Ann. Sci. Ec. Norm. Sup.* **4** (1971), 393–398.

Fokko du Cloux, Institut Girard Desargues, UMR 5028 CNRS, Université Lyon-I, 69622 Villeurbanne Cedex France (ducloux@desargues.univ-lyon1.fr)

Received July 17, 2001; accepted in revised form November 15, 2001.

Sur un système fibré lié à la suite des nombres premiers

Alain Costé

SOMMAIRE

Nous étudions le système dynamique défini par la transformation $\Phi :]0,1] \longrightarrow]0,1]$ où $\Phi(x) = px - 1$ si $x \in]1/p, 1/q]$, q et p étant deux nombres premiers consécutifs. La question de l'existence d'une mesure absolument continue invariante par Φ est reliée par un argument de chaîne de Markov à une conjecture concernant un ensemble de suites de nombres premiers. Cette hypothèse est corroborée par des simulations de type Monte-Carlo. Nous montrons que cela entraîne la stabilité statistique de Φ sur l'intervalle $]0, 2/3]$. En utilisant des arguments heuristiques nous définissons des versions simplifiées de l'opérateur de Perron-Frobenius associé à Φ. Cela nous permet de construire à l'aide de Maple une densité de probabilité présentant une bonne adéquation expérimentale avec les histogrammes des orbites issues de constantes fondamentales.

We study the dynamical system defined by the transformation $\Phi :]0,1] \longrightarrow]0,1]$ where $\Phi(x) = px - 1$ if $x \in]1/p, 1/q]$, q and p being two consecutive prime numbers. The problem of the existence of an invariant absolutely continuous measure by Φ is related via a Markov chain argument to a conjecture concerning a set of prime number sequences. This hypothesis is corroborated by Monte Carlo simulations. We prove that this implies the statistical stability of the transformation Φ on the interval $]0, 2/3]$. By using heuristical arguments, we define simplified versions of the Perron-Frobenius operator associated to Φ. Using Maple, we construct a probability density presenting a good experimental fit with the histograms of orbits stemming from fundamental constants.

1. INTRODUCTION

Un système fibré est la donnée d'un ensemble I, le plus souvent un intervalle de $\mathbb{R}$ et de deux applications $\Phi\colon I \longrightarrow I$, (que l'on nomme plutôt transformation) et $P\colon I \longrightarrow \mathbb{N}$ vérifiant la propriété que la restriction de Φ à tout $P^{-1}(n)$ est injective. L'exemple classique est le système lié au développement décimal où: $I = [0,1[$, $P\colon x \longrightarrow [10x]$ et $\Phi\colon x \longrightarrow 10x - [10x]$. Le système fibré lié au développement en fraction continue est défini

2000 AMS Subject Classification: Primary 37E05; Secondary 37A45
Keywords: Prime numbers, Markov chain

1058-6458/2001 $0.50 per page

sur $I =]0,1]$ par Φ: $x \longrightarrow 1/x - [1/x]$ et P: $x \longrightarrow [1/x]$. Nous renvoyons le lecteur à [Schweiger 95] pour une documentation complète sur le sujet.

Un système fibré est un système dynamique discret. Etant donnée une condition initiale x dans I la suite des itérés (x,$\Phi(x)$,$\Phi^2(x)$,,...) de x sous l'action de Φ forme une orbite dont le comportement asymptotique est le principal objet d'étude. Cette orbite engendre la suite $n \longrightarrow P_n(x) = P(\Phi^n(x))$ que l'on appelle développement de x. Une suite de $\mathbb{N}$ qui est le développement d'un élément $x \in I$ est appelée suite admissible. La transmuée de la transformation Φ par l'application P est le shift à droite sur l'ensemble des suites admissibles. On comprend donc l'intérêt de décrire complétement ce dernier ensemble. Dans les deux exemples précédents l'ensemble des suites admissibles est de la forme $A^{\mathbb{N}}$ où A est un sous ensemble de $\mathbb{N}$. Il n'en n'est pas toujours ainsi par exemple pour le système d'Engel noté $\mathcal{E}$ [Schweiger 95]. Celui-ci est défini sur $]0,1]$ par

$$\Phi(x) = (k+1)x - 1 \ \text{ sur } \ ]\frac{1}{k+1}, \frac{1}{k}].$$

Il se trouve que les suites admissibles de $\mathcal{E}$ sont exactement les suites d'entiers croissantes.

Dans ce travail nous étudions le système inspiré de $\mathcal{E}$ où l'on remplace la suite des entiers par celle des nombres premiers. Ainsi nous considérons l'application Φ définie par $x \longrightarrow px - 1$ sur l'intervalle $[1/p, 1/q[$ si $q < p$ sont deux nombres premiers consécutifs. L'intervalle $]0,1]$ est stable par Φ car d'après un théorème de Tschebychef on a toujours $p/q - 1 \leq 1$ ou encore $p \leq 2q$.

A l'instar de ce qui se passe pour $\mathcal{E}$ on s'attend à ce qu'une suite admissible soit croissante. Il n'en est rien. En fait une telle suite présente de brusques sauts suivis de descentes progressives avec une tendance marquée à revenir vers les nombres 3, 5, 7 et 11. Nous nous sommes alors intéressés aux problèmes suivants: (1) Caractérisation des suites admissibles, (2) Existence d'une densité stationnaire, (3) Ergodicité du système.

Voici maintenant exposées les grandes lignes de notre travail. Du fait que l'on ait l'inégalité plus fine $p \leq 5/3q$ (si $q < p$ sont deux nombres premiers consécutifs avec $2 \leq q$), l'intervalle $I =]0, 2/3]$ est stable par Φ. De plus on voit facilement que toutes les orbites restent dans I à partir d'un certain rang. Aussi du point de vue de la théorie ergodique seul le système restreint à cet intervalle est digne d'intérêt. Ce faisant on exclut simplement les suites admissibles commençant par une série de nombres 2. Par ailleurs on s'aperçoit que par rapport aux problèmes posés un certain ensemble joue un rôle central. Il s'agit de l'ensemble $\mathcal{O}$ défini comme la réunion des orbites issues des nombres $p/q - 1$ (où $p < q$ sont deux premiers consécutifs avec $2 \leq q$). Cet ensemble est en effet très riche car le caractère premier des dénominateurs précédents fait que ces orbites sont deux à deux disjointes.

Au paragraphe 2 nous montrons qu'une suite admissible tronquée (ou D-suite) est une concaténation de débuts de développements de nombres $1/p$ (p premier) avec une condition aux indices de discontinuité (appelés indices de saut). Une telle suite est donc accompagnée d'une "ombre" définie comme la suite de $\mathcal{O}$ composée de la juxtaposition des morceaux d'orbite associés aux nombres $1/p$ mentionnés précédemment. Le dernier terme de cette "ombre" que l'on appelle résultant de la D-suite est une notion qui se révèle très utile au paragraphe 3.

L'idée centrale du paragraphe 3 réside dans la constatation que l'opérateur de Perron-Frobenius T associé à Φ laisse invariant le sous-espace de $L^1(I)$ constitué des fonctions de la forme $f = \sum_{r\in\mathcal{O}} \alpha_r \ \mathbb{1}_{]0,r]}$ où $\sum \mid \alpha_r \mid r < \infty$. La matrice $\mathcal{Z}$ de T réduit à cet espace dans la base de Schauder $\left(\frac{1}{r}\mathbb{1}_{]0,r]}\right)_{r\in\mathcal{O}}$ est colonne-stochastique. Nous introduisons la chaîne de Markov $\mathcal{P}$ dont l'ensemble des états est $\mathcal{O}$ et dont les probabilités de passage sont données par les coefficients de $\mathcal{Z}$. Cette chaîne est d'un grand intérêt pour l'étude des propriétés du système. A partir d'une distribution invariante $\alpha = [\alpha_r]_{r\in\mathcal{O}}$ de $\mathcal{P}$ on obtient une densité stationnaire de T en posant $f = \sum_{r\in\mathcal{O}} \alpha_r/r \ \mathbb{1}_{]0,r]}$. Comme $\mathcal{P}$ est irréductible et apériodique l'existence d'une telle distribution invariante est équivalente à l'ergodicité de la chaîne, cette propriété étant encore équivalente à la non nullité de la limite quand $n \longrightarrow \infty$ d'un quelconque des coefficients diagonaux de la matrice puissance nième de $\mathcal{Z}$. Dans le but de tester cette hypothèse nous établissons d'abord une formule exprimant le coefficient générique de la matrice $\mathcal{Z}^n$ par une somme portant sur les D-suites liées aux indices du coefficient par le biais de leurs résultants. Ainsi par exemple si β_n est le coefficient diagonal de $\mathcal{Z}^n$ correspondant à $2/3$ on a

$$\beta_n = \sum \left\{ \prod_{k=1}^{n} \frac{1}{p_k}, \ (p_1, p_2, \ldots, p_n) \right.$$
$$\left. \text{est une } D\text{-suite de résultant } 2/3 \right\}.$$

En exploitant le fait que le résultant d'une D-suite est lié à la longueur du cylindre engendré par cette D-suite nous obtenons une expression du coefficient diagonal de $\mathcal{Z}^n$ comme probabilité pour qu'un nombre tiré au hasard dans un certain intervalle ait un développement

de longueur n de résultant donné. Cela nous donne la possibilité de déterminer ce coefficient par la méthode de Monte Carlo. Au vu de simulations sur Maple nous nous permettons d'avancer que la limite de β_n est non nulle et qu'elle est de l'ordre de 0,145, ce qui est confirmé par le calcul approché de la densité stationnaire aux paragraphes 4 et 5. Ainsi nous conjecturons que $\mathcal{P}$ est ergodique. Nous montrons que sous cette hypothèse la transformation Φ est statistiquement stable.

Le but du paragraphe 4 est la détermination d'une approximation la plus fine possible de la densité stationnaire $g = \sum \alpha_r / r \ \mathbb{1}_{]0,r]}$ de T. Nous proposons deux méthodes qui sont complémentaires et dont l'élaboration est fondée sur une démarche heuristique conduisant à des approximations sans estimation effective des termes d'erreur. Nous utilisons le procédé dû à Ulam [Ulam 60] d'approximation matricielle de l'opérateur T en apportant l'innovation consistant à conserver le terme reste sous forme d'un opérateur intégral.

L'opérateur T est la somme d'une série d'opérateurs dont le terme général fait apparaître la suite

$$n \longrightarrow u_n = \frac{p_n}{p_{n-1}} - 1$$

où p_n est le nombre premier de rang n. Un entier N étant fixé nous définissons l'opérateur tronqué T_N par la somme partielle d'indice N de la série ci-dessus et nous notons $\widetilde{T}_N$ l'opérateur reste $T - T_N$. Dans la première méthode nous nous appuyons sur le théorème des nombres premiers pour justifier le remplacement de u_n par sa "moyenne" $1/n$. Dans la deuxième méthode nous utilisons le modèle probabiliste de H. Cramér [Cramér 36] pour déterminer l'espérance de $\widetilde{T}_N f(t)$, expression que nous adoptons comme nouvelle approximation de cet opérateur. Nous poursuivons la simplification en transformant les deux versions précédentes de $\widetilde{T}_N$ en un opérateur intégral $\widetilde{T}_N^{(i)}$, ($i = 1$ ou $i = 2$ suivant la méthode utilisée). Pour cela nous faisons intervenir une fonction G fournissant un "lissage" correct de la suite $n \longrightarrow p_n$. L'opérateur $T_N + \widetilde{T}_N^{(i)}$ laisse pratiquement stable un sous-espace $\mathcal{L}_N^{(i)}$ de $L^1(I)$ somme directe $\mathcal{E}_N + F^{(i)}$, où $\mathcal{E}_N$ est l'espace de dimension finie engendré par les $\mathbb{1}_{]0,r]}$, r parcourant la réunion des orbites issues des nombres $p_n/p_{n-1} - 1$ (pour $n \leq N$) et $F^{(i)}$ est un espace de fonctions continues liées à la fonction primitive de $x \longrightarrow 1/G(x)$. De plus l'expression de cet opérateur est suffisamment simple pour se prêter à une programmation sous Maple. Le vecteur propre de valeur propre dominante de $T_N + \widetilde{T}_N^{(i)}$ est déterminé par approximations successives. Nous définissons $g_N^{(i)}$ comme le normalisé dans $L^1(I)$ de ce vecteur propre. Nous observons une bonne convergence expérimentale de chaque suite $(g_N^{(i)})$ vers la densité stationnaire g, la convergence étant plus rapide avec la seconde méthode. De plus au vu de l'expression de $g_N^{(i)}$ nous sommes amenés à conjecturer que $k = lim_{t \longrightarrow 0} g(t)/ln(t)$ existe avec k de l'ordre de -1,3. Au dernier paragraphe nous nous intéressons au développement de quelques constantes fondamentales et nous présentons le résultat des calculs des coefficients liés à la densité stationnaire.

2. LE SYSTÈME FIBRÉ ET LA DYNAMIQUE SYMBOLIQUE ASSOCIÉE

La lettre p désigne toujours un nombre premier différent de 1 et lorsque $p > 2$, $p^{\sim}$ désigne le nombre premier immédiatement inférieur à p. Nous définissons les fonctions P et Φ sur $]0,1]$ par

$$P(x) = \min(\{p; 1/x < p\}) \quad \text{et} \quad \Phi(x) = P(x)x - 1.$$

La fonction Φ prend ses valeurs dans $]0,1]$ car d'après un théorème de Tschebychef on a $p < 2p^{\sim}$, (pour $p > 2$). On peut donc itérer Φ. Si l'on part de $x > 2/3$, la suite $\Phi^n(x)$ commence par décroître et elle finit par prendre une valeur dans l'intervalle $I =]0, 2/3]$. De plus ce dernier intervalle est stable par Φ car on a l'inégalité plus fine $p \leq 5/3p^{\sim}$ pour $p > 2$ (cette inégalité peut se démontrer à partir de l'encadrement effectif du nombre premier de rang n rappelé au début du §4). C'est pourquoi seul le système dynamique défini par la restriction de Φ à I est pris en considération. Il s'agît d'un système fibré σ-affine ([Schweiger 95]) dont la partition associée est constituée des intervalles $]1/p, 1/p^{\sim}]$ (pour $p > 2$) et de l'intervalle $]1/2, 2/3]$. Aussi afin d'harmoniser les notations, nous posons $p^{\sim} = 3/2$ (au lieu de $p^{\sim} = 1$) lorsque $p = 2$.

Nous notons pour tout x de I et tout entier $n \geq 0$, $P_n(x) = P(\Phi^n(x))$.

Définition 2.1. Etant donné $x \in I$, la suite $[P_n(x)]_{n \geq 0}$ est appelée développement du nombre x et pour tout $N \geq 1$ la suite $[P_n(x)]_{0 \leq n \leq N-1}$ est appelée développement d'ordre N de x.

Proposition 2.2. *Soit $x \in I$ et soit $(p_k)_{k \geq 0}$ son développement. On a pour tout $n \geq 0$,*

$$x = \sum_{\ell=0}^{n} \prod_{k=0}^{\ell} \frac{1}{p_k} + \Phi^{n+1}(x) \prod_{k=0}^{n} \frac{1}{p_k}. \qquad (2\text{–}1)$$

De plus

$$x = \sum_{\ell=0}^{\infty} \prod_{k=0}^{\ell} \frac{1}{p_k}. \qquad (2\text{–}2)$$

Preuve: La relation (2–1) est vraie pour $n = 0$. En rapprochant l'égalité (2–1) supposée établie pour n de $\Phi^{n+1}(x) = 1/p_{n+1}(1 + \Phi^{n+2}(x))$, on voit qu'elle est encore vraie pour $n + 1$. Ce qui démontre (2–1) par récurrence. La formule (2–2) en découle immédiatement. □

Corollaire 2.3. *Une condition nécessaire et suffisante pour que le développement de $x \in I$ soit périodique à partir d'un certain rang est que x soit rationnel.*

Preuve: La condition est suffisante car si x est de la forme n/m où $n, m \in \mathbb{N}^*$, alors pour tout k, $\Phi^k(x)$ est de la forme n'/m où $n' \in \mathbb{N}^*$ et $n' < m$. Par suite il existe k et $h > 0$ tels que $\Phi^k(x) = \Phi^{k+h}(x)$. Le développement de x est donc $h-$périodique à partir du rang k. La condition est nécessaire. En effet si le développement de x est $h-$périodique, il résulte de la proposition 2.2 que

$$x = \sum_{\ell=0}^{h-1} \prod_{k=0}^{\ell} \frac{1}{p_k} \Big(1 - \prod_{k=0}^{h-1} \frac{1}{p_k}\Big)^{-1}.$$

Donc x est rationnel. □

Rappelons que la lettre p désigne toujours un nombre premier et $p^\sim$ le nombre premier immédiatement inférieur à p si $p > 2$ et $3/2$ si $p = 2$.

Etant donné $k \in \mathbb{N}$, nous posons

$$r(p,k) = \Phi^k(1/p^\sim) \qquad \text{et} \qquad s(p,k) = P(r(p,k)).$$

On a en particulier $r(p,1) = p/p^\sim - 1$ et $s(p,0) = p$.

Par commodité d'écriture une suite finie ou infinie est notée $(p_k)_{0\le k<n}$ où $n \in \mathbb{N} \cup \{\infty\}$; elle est aussi notée $(p_0, \ldots, p_{n-1})$ lorsque n est fini.

Nous posons aussi

$$H(p_0, \ldots, p_{n-1}) = \sum_{\ell=0}^{n-1} \prod_{k=0}^{\ell} \frac{1}{p_k}.$$

2.1 Caractérisation des suites développements de nombres

Définition 2.4. (Indice de bifurcation.) Soient $x, y \in I$, $x < y$. On appelle indice de bifurcation du couple (x,y) le plus petit entier $n = n(x,y)$ tel que $P_n(x) \neq P_n(y)$.

Lemme 2.5. *Pour tout $x, y \in I$, $x < y$, si $n = n(x,y)$ est l'indice de bifurcation de (x,y), on a: $P_n(y) < P_n(x)$.*

Preuve: On montre par récurrence la propriété $(\mathcal{P}_m)$: pour tout couple $(x,y) \in I^2$, $x < y$, tel que $n(x,y) = m$, on a $P_m(y) < P_m(x)$.

$(\mathcal{P}_0)$ est clairement vraie. Supposons $(\mathcal{P}_m)$ vraie et soient x, $y \in I$, $x < y$, tels que $n(x,y) = m + 1$, (définition 2.4). On a par hypothèse: $P_0(x) = P_0(y) = p_0$; donc $\Phi(x) = p_0 x - 1 < p_0 y - 1 = \Phi(y)$. Comme l'indice de bifurcation du couple $(\Phi(x), \Phi(y))$ est égal à m, on peut lui appliquer la propriété $(\mathcal{P}_m)$, ce qui donne $P_{m+1}(y) < P_{m+1}(x)$. Donc $(\mathcal{P}_{m+1})$ est vraie. □

Définition 2.6. (Indice de saut.) Soit $S = (p_k)_{h\le k<n}$, où $h \in \mathbb{N}$ et $n \in \mathbb{N} \cup \{\infty\}$, une suite de nombres premiers. On appelle indice de saut de S toute valeur prise par la suite finie ou infinie $(k_i)_{0\le i<m}$ définie par récurrence par $k_0 = h$, et pour tout $i \ge 1$:

$$k_i = \min(\{k; k_{i-1} < k < n \quad \text{et} \quad s(p_{k_{i-1}}, k - k_{i-1}) \neq p_k\}),$$

si l'ensemble précédent est non vide.

Définition 2.7. (*D*-suite.) On dit qu'une suite $(p_k)_{0\le k<n}$ finie ou infinie de nombres premiers est une D-suite si pour tout couple d'indices de saut consécutifs (k_1, k_2) de cette suite on a $p_{k_2} > s(p_{k_1}, k_2 - k_1)$.

Exemple 2.8. (11,2), (2,5) et (11,2,11) sont des D-suites, mais (11,2,5) et (11,2,7) n'en sont pas. Une suite dont tous les éléments sont égaux à $p \neq 2$ est une D-suite. Par contre (2,2) n'en est pas une.

Proposition 2.9. *La suite développement d'un réel $x \in I$ est une D-suite.*

Preuve: Soit $x \in I$ et soit $(p_k)_{k\ge 0}$ son développement. Soient $k_1 < k_2$ deux indices de saut consécutifs de cette suite. En considérant $\Phi^{k_1}(x)$ au lieu de x on peut supposer que $k_1 = 0$. Posons $y = 1/p_0^\sim$. Si $x = y$ la conclusion est immédiate car il n'y a pas de deuxième saut. Sinon $x < y$ et par définition de k_2, pour tout $k < k_2$, $p_k = P_k(x) = P_k(y) = s(p_0, k)$ et $P_{k_2}(x) \neq P_{k_2}(y)$. Donc k_2 est l'indice de bifurcation de (x,y). Par suite d'après le lemme 2.5, $p_{k_2} > s(p_0, k_2)$. □

Proposition 2.10. *Soit $S = (p_k)_{0\le k<n}$ une D-suite. Alors pour tout $h < n$ la suite tronquée $S' = (p_k)_{h\le k<n}$ est une D-suite. De plus les indices de saut de S supérieurs ou égaux à h sont aussi des indices de saut de S'.*

Preuve: Soit k_0 le plus grand des indices de saut de S strictement inférieurs à h. Si les indices de saut de S sont tous strictement inférieurs à h, alors S' est le développement d'ordre $n-h$ de $r(p_{k_0}, h-k_0)$. C'est donc une D-suite d'après la proposition 2.9. Dans le cas contraire, soit k_1 le plus petit indice de saut de S minoré par h. Si $k_1 = h$ la conclusion est immédiate. Autrement considérons la suite $S_1 = (p_h, \dots, p_{k_1-1})$. Par définition de k_0 et de k_1, S_1 est le développement d'ordre $k_1 - h$ de $r\big(p_{k_0}, h-k_0\big)$. C'est donc une D-suite et par conséquent l'inégalité de la définition 2.7 est vérifiée pour tout couple d'indices de saut de S' strictement inférieurs à k_1. Soit k_2 le plus grand des indices de saut de S_1. Par définition de k_2 on a:

(1) $s\big(p_{k_2}, k-k_2\big) = p_k$, pour tout k tel que $k_2 \le k < k_1$.

De plus, par définition de k_0 et de k_1 :

(2) $s\big(p_{k_0}, k-k_0\big) = p_k$, pour tout k tel que $k_0 \le k < k_1$.

Montrons que

(3) $s\big(p_{k_2}, k_1-k_2\big) \le s\big(p_{k_0}, k_1-k_0\big)$.

L'égalité (2) avec $k = k_2$ implique $r\big(p_{k_0}, k_2-k_0\big) \le 1/p^{\sim}_{k_2}$. Si la relation précédente est une égalité, alors (3) est évidemment aussi une égalité. Autrement, compte tenu de (1) et (2), l'inégalité (3) est une conséquence du lemme 2.5 appliqué au couple $\Big(r\big(p_{k_0}, k_2-k_0\big), 1/p^{\sim}_{k_2}\Big)$. Comme $s\big(p_{k_0}, k_1-k_0\big) < p_{k_1}$, on a donc: $s\big(p_{k_2}, k_1-k_2\big) < p_{k_1}$. On en déduit que k_1 est l'indice de saut de la suite S' qui suit immédiatement k_2 et cette inégalité montre que S' est une D-suite. En effet, à partir de k_1 les indices de saut de S' sont les mêmes que ceux de S. Ce qui justifie le dernière assertion de l'énoncé. □

Lemme 2.11. *Soit $(p_k)_{0\le k\le n-1}$ une D-suite finie. Alors $H(p_0, \dots, p_{n-1}) < 1/p^{\sim}_0$.*

Preuve: Si $n = 1$ le lemme est clairement vrai. Supposons-le établi pour tout entier $m \le n$ et soit $(p_k)_{0\le k\le n}$ une D-suite de longueur $n+1$. Si 0 est le seul indice de saut de cette suite, celle-ci est le développement d'ordre $n+1$ de $1/p^{\sim}_0$, donc l'inégalité est vraie. Autrement soit k_1 le deuxième indice de saut de la suite. D'après la proposition 2.10, la suite $(p_k)_{k_1\le k\le n-1}$ est une D-suite (de longueur inférieure à n). Donc par l'hypothèse de récurrence: $H\big(p_{k_1}, \dots, p_{n-1}\big) < 1/p^{\sim}_{k_1}$. On voit en utilisant l'égalité

$$H(p_0, \dots, p_{n-1}) = H\big(p_0, \dots, p_{k_1-1}\big) + H\big(p_{k_1}, \dots, p_{n-1}\big) \prod_{k=0}^{k_1-1} \frac{1}{p_k},$$

que

$$\begin{aligned} H(p_0, \dots, p_{n-1}) &< H(p_0, \dots, p_{k_1-1}) + 1/p^{\sim}_{k_1} \prod_{k=0}^{k_1-1} \frac{1}{p_k} \\ &\le H\big(p_0, \dots, p_{k_1-1}, s(p_0, k_1)\big); \end{aligned}$$

la dernière inégalité découlant du fait que $s(p_0, k_1) \le p^{\sim}_{k_1}$, puisque par définition de k_1, $s(p_0, k_1) < p_{k_1}$. Or, toujours par définition de k_1, $\big(p_0, \dots, p_{k_1-1}, s\big(p_0, k_1\big)\big)$ est le développement d'ordre $k_1 + 1$ de $1/p^{\sim}_0$. Par suite $H\big(p_0, \dots, p_{k_1-1}, s(p_0, k_1)\big) < 1/p^{\sim}_0$. D'où finalement $H(p_0, \dots, p_{n-1}) < 1/p^{\sim}_0$. Le lemme est ainsi démontré par récurrence. □

Théorème 2.12. *Une condition nécessaire et suffisante pour qu'une suite de nombres premiers $(p_k)_{k\ge 0}$ soit le développement d'un nombre réel appartenant à I est qu'elle soit une D-suite.*

Preuve: La condition est nécessaire d'après la proposition 2.9. Soit $(p_k)_{k\ge 0}$ une D-suite. Montrons que cette suite est le développement du nombre $x = \lim_n H(p_0, \dots, p_n)$. Il résulte du lemme 2.11 que $1/p_0 < x \le 1/p^{\sim}_0$. Donc $p_0 = P_0(x)$. On en déduit que $\Phi(x) = p_0 x - 1$. Mais $p_0 x - 1 = \lim_n H(p_1, \dots, p_n)$. On voit par une récurrence sur k que $p_k = P_k(x)$ pour tout k. □

2.2 Notion de résultant d'une *D*-suite.

Les résultats établis dans ce paragraphe sont utiles pour la suite.

Notation. Etant donné une D-suite finie $S = (p_k)_{0\le k<n,}$ on note

$$J(S) = \{x \in I \ ; P_k(x) = p_k \text{ pour tout k tel que } 0 \le k < n\}$$

et on appelle **résultant de S** le nombre $\mathrm{res}(S) = r\big(p_{k_{m-1}},\ n - k_{m-1}\big)$ où k_{m-1} est le plus grand indice de saut de S. (Dans [Schweiger 95] les ensembles $J(S)$ sont appelés des cylindres).

Théorème 2.13. *Soit $S = (p_k)_{0\le k<n}$ une D-suite finie. Alors on a*

$$J(S) =]H(p_0, \dots, p_{n-1}),\ H(p_0, \dots, p_{k_{m-1}},\ p^{\sim}_{k_{m-1}})],$$

où k_{m-1} est le plus grand indice de saut de la suite $(p_k)_{0\le k<n}$.

Preuve: Soit $x \in J(S)$. On a d'après la proposition 2.2:

$$x = H(p_0,\dots,p_{k_{m-1}-1}) + \Phi^{k_{m-1}}(x)\prod_{k=0}^{k_{m-1}-1}\frac{1}{p_k}.$$

Par ailleurs le développement de $\Phi^{k_{m-1}}(x)$ commence par $(p_{k_{m-1}},\dots,p_{n-1})$. Donc $H(p_{k_{m-1}},\dots,p_{n-1}) < \Phi^{k_{m-1}}(x) \le 1/p^{\sim}_{k_{m-1}}$. En rapprochant cette double inégalité de la relation précédente on en déduit l'encadrement souhaité de x.

Dans le but d'établir l'inclusion inverse montrons d'abord que

$$H\big(p_0,\dots,p_{k_{m-1}-1},p^{\sim}_{k_{m-1}}\big) \le 1/p^{\sim}_0. \tag{2–3}$$

Du fait de l'existence du saut en k_{m-1}, il résulte du théorème 2.12 que si l'on complète $(p_0,\dots,p_{k_{m-1}-1})$ par le développement de $1/p^{\sim}_{k_{m-1}}$ on obtient une D-suite qui est le développement de $H\big(p_0,\dots,p_{k_{m-1}-1},p^{\sim}_{k_{m-1}}\big)$, ce qui prouve (2–3).

Supposons d'abord que $k_{m-1}=0$. La suite $(p_k)_{0\le k<n}$ est alors le développement d'ordre n de $1/p^{\sim}_0$. On en déduit que $]H(p_0,\dots,p_{n-1}),1/p^{\sim}_0] \subset J(S)$.

Supposons maintenant que $k_{m-1}>0$. Soit $x \in$

$$\big]H(p_0,\dots,p_{n-1}),H(p_0,\dots,p_{k_{m-1}-1},p^{\sim}_{k_{m-1}})\big].$$

D'après (2–3) on a: $p_0 = P_0(x)$. De plus $p_0x-1 \in]H(p_1,\dots,p_{n-1}),H(p_1,\dots,p_{k_{m-1}-1},p^{\sim}_{k_{m-1}})]$. Or d'après la proposition 2.2, k_{m-1} est aussi le dernier indice de saut de $(p_1,\dots,p_{n-1})$. Par conséquent en supposant le théorème établi pour les suites de longueur $n-1$ on voit que le développement de p_0x-1 commence par $(p_1,\dots,p_{n-1})$, donc celui de x commence par $(p_0,\dots,p_{n-1})$. Ce qui implique l'inclusion souhaitée pour les suites de longueur n. Le théorème est ainsi démontré par récurrence. □

Proposition 2.14. *Soit $S=(p_k)_{0\le k<n}$ une D-suite finie. Alors $J(S)$ est un intervalle de longueur* $\operatorname{res}(S)\prod_{k=0}^{n-1}\frac{1}{p_k}$.

Preuve: Par le théorème 2.13 on sait que $J(S)$ est un intervalle de longueur

$$\big(1/p^{\sim}_{k_{m-1}} - H(p_{k_{m-1}},\dots,p_{n-1})\big)\prod_{k=0}^{k_{m-1}-1}\frac{1}{p_k},$$

où k_{m-1} est le plus grand indice de saut de S. Mais comme $(p_{k_{m-1}},\dots,p_{n-1})$ est le développement d'ordre $n-k_{m-1}$ de $1/p^{\sim}_{k_{m-1}}$, on a d'après la proposition 2.10:

$$1/p^{\sim}_{k_{m-1}} = H(p_{k_{m-1}},\dots,p_{n-1}) + r(p_{k_{m-1}},n-k_{m-1})\prod_{k=k_{m-1}}^{n-1}\frac{1}{p_k}.$$

□

3. LA CHAÎNE DE MARKOV $\mathcal{P}$

Rappelons que la lettre p désigne toujours un nombre premier différent de 1 et $p^{\sim}$ désigne le nombre premier immédiatement inférieur à p si $p>2$ et $3/2$ si $p=2$.

Le fait que les extrémités des intervalles du système fibré soient toutes (sauf une) des inverses de nombres premiers apporte ceci de particulier que les orbites finies issues de ces nombres sont deux à deux disjointes, mise à part l'exception due à l'égalité $r(2,2)=r(5,1)$. Un sous-ensemble de la réunion de ces orbites joue un rôle essentiel dans la suite. Il s'agit de l'ensemble $\mathcal{O}$ qui est la réunion des orbites des nombres $p/p^{\sim}-1$. Nous commençons par établir une bijection de $\mathcal{O}$ sur un ensemble de couples (p,k) via l'application $(p,k)\longrightarrow r(p,k)$. Pour cela nous introduisons la fonctions h suivante:

pour $p\neq 2$ et $p\neq 5$,

$$h(p) = \min\{k\ge 2,\ \text{tel qu'il existe j, } 1\le j<k \text{ avec } r(p,j)=r(p,k)\}$$

et enfin $h(2)=h(5)=2$.

Nous pouvons donc identifier $\mathcal{O}$ à l'ensemble $\{(p,k)\ ;\ p \text{ premier}, 1\le k\le h(p)-1\}$.

Nous définissons aussi la fonction j suivante:

pour $p\neq 2$ et $p\neq 5$,

$$j(p) = \text{ l'unique entier k, } 1\le k<h(p) \text{ tel que } r(p,k)=r\big(p,h(p)\big)$$

et enfin $j(2)=j(5)=1$.

(En général on a pour $p>5$:

$$r\big(p,h(p)-1\big)=\frac{1}{p^{\sim}},$$

donc $j(p)=1$. Le plus petit p tel que $1<j(p)$ est $p=19$).

Nous désignons par λ la mesure de Lebesgue. Rappelons que $I=]0,2/3]$.

L'application Φ définie au §2 est non singulière, en ce sens que pour tout borélien A de I, $\lambda(A)=0 \Rightarrow \lambda\big(\Phi^{-1}(A)\big)=0$. On peut donc considérer l'opérateur de

$$A_{p,k}^{q,\ell} = \begin{cases} \dfrac{1}{p} & \text{si} \quad p > s(q,\ell) & & \text{et} \quad k=1, \\ \dfrac{1}{s(q,\ell)} & \text{si} \quad p=q>5, & \ell < h(q)-1 & \text{et} \quad k=\ell+1, \\ & \text{ou si} \quad p=q>5, & \ell = h(q)-1 & \text{et} \quad k=j(q), \\ & \text{ou si} \quad p=s(q,1), q\le 5, & \ell=1 & \text{et} \quad k=1, \\ 0 & \text{autrement.} & & \end{cases}$$

TABLE 1.

Perron-Frobenius T de $L^1(I,\lambda)$ dans $L^1(I,\lambda)$ associé à Φ. D'après [Schweiger 95] on a pour tout $f \in L^1(I,\lambda)$,

$$Tf(t) = \sum_p \frac{1}{p} f\Big(\frac{1+t}{p}\Big) \mathbb{1}_{\left]0,p/p^{\sim}-1\right]}(t), \tag{3–1}$$

cette série étant absolument convergente pour la norme L^1. En particulier on a pour tout $s \in I$:

$$T(\mathbb{1}_{]0,s]}) = \frac{1}{P(s)}\ \mathbb{1}_{]0,\Phi(s)]} + \sum_{p>P(s)} \frac{1}{p}\ \mathbb{1}_{]0,p/p^{\sim}-1]}, \tag{3–2}$$

(pour la définition de $P(s)$ voir le début du §2).

Introduisons le sous-espace $\mathcal{E}$ de $L^1(I,\lambda)$ constitué des fonctions f de la forme

$$f = \sum_{(p,k)\in\mathcal{O}} \alpha_{p,k}\ \mathbb{1}_{]0,r(p,k)]}$$

où $\sum \mid \alpha_{p,k} \mid r(p,k) < \infty$. Pour simplifier l'écriture nous écrivons par la suite $e_{p,k} = \mathbb{1}_{]0,r(p,k)]}$.

Le lemme suivant montre que la famille $\big(e_{p,k}\big)_{(p,k)\in\mathcal{O}}$ constitue une base de Schauder de $\mathcal{E}$.

Lemme 3.1. *Soit $(r_n)_{n\ge 0}$ une suite injective de $\mathbb{R}_+^*$ et soit $(\alpha_n)_{n\ge 0}$ une suite de réels telle que $\sum \mid \alpha_n \mid r_n < \infty$. On suppose que la fonction $f = \sum \alpha_n \mathbb{1}_{]0,r_n]}$ est presque partout nulle. Alors $\alpha_n = 0$ pour tout n.*

Preuve: Puisque f est continue à gauche, l'hypothèse implique que f est identiquement nulle. Or du fait de l'injectivité de $(r_n)_{n\ge 0}$, on a pour tout n : $f(r_n) - f(r_n + 0) = \alpha_n$. □

On voit d'après (3–2) que T laisse $\mathcal{E}$ invariant. On peut donc parler de la matrice infinie $\mathcal{A} = \big[A_{p,k}^{q,\ell}\big]_{[(p,k),(q,\ell)]\in\mathcal{O}^2}$ de l'opérateur $T_{|\mathcal{E}}$ dans la base $(e_{p,k})_{(p,k)\in\mathcal{O}}$. Ainsi, pour tout $(q,\ell) \in \mathcal{O}$:

$$Te_{q,\ell} = \sum_{(p,k)\in\mathcal{O}} A_{p,k}^{q,\ell}\ e_{p,k}.$$

On obtient les $A_{p,k}^{q,\ell}$ à partir de (3–2) (voir Table 1).

En particulier $A_{3,1}^{3,1} = \dfrac{1}{3}$, $A_{2,1}^{5,1} = \dfrac{1}{2}$, $A_{5,1}^{2,1} = \dfrac{1}{5}$.

Puisque T est une isométrie sur $L^1(I,\lambda)^+$ on a pour tout $(q,\ell) \in \mathcal{O}$:

$$r(q,\ell) = \sum_{(p,k)\in\mathcal{O}} r(p,k)\ A_{p,k}^{q,\ell}.$$

Soit la matrice infinie

$$\mathcal{Z} = \big[Z_{p,k}^{q,\ell}\big]_{[(p,k),(q,\ell)]\in\mathcal{O}^2},$$

où

$$Z_{p,k}^{q,\ell} = A_{(p,k)}^{(q,\ell)} \frac{r(p,k)}{r(q,\ell)}.$$

La relation précédente montre que $\mathcal{Z}$ est colonne-stochastique en ce sens que la somme des éléments de chacune de ses colonnes est égale à 1. Cela nous permet de définir la chaîne de Markov que nous notons $\mathcal{P}$ dont $\mathcal{O}$ est l'ensemble des états et dont les $Z_{p,k}^{q,\ell}$ sont les probabilités de passage de (q,ℓ) à (p,k).

Lemme 3.2. *Pour tout $p \ge 5$, on a $p^{\sim} > s(p,1)$.*

Preuve: On a

$$r(p,1) \ge \frac{2}{p^{\sim}} > \frac{1}{(p^{\sim})^{\sim}}.$$

Donc $s(p,1) \le (p^{\sim})^{\sim}$. D'où $s(p,1) < p^{\sim}$. □

Proposition 3.3. *La chaîne de Markov $\mathcal{P}$ est irréductible et apériodique.*

Preuve: Il s'agit de montrer que pour tout couple d'états $[(q,\ell),(p,k)]$ il existe une suite finie $[(p_i,k_i)]_{0\le i\le n}$ telle que $(p_0,k_0) = (q,\ell)$, $(p_n,k_n) = (p,k)$ et que pour tout i, $0 \le i \le n-1$, $A_{p_{i+1},k_{i+1}}^{p_i,k_i} \ne 0$.

Remarquons d'abord que tout état est accessible à partir de (5,1); en effet on a pour tout $p:\ A_{p,1}^{5,1} \neq 0$ et pour tout $k', 1 \leq k' < h(p) - 1\ :\ A_{p,k'+1}^{p,k'} \neq 0$.

Pour conclure il suffit d'établir que (5,1) est accessible à partir de tout état (q,ℓ). Dans ce but choisissons p_1 tel que $p_1 > s(q,\ell)$. Alors $A_{p_1,1}^{q,\ell} \neq 0$. Si $p_1 = 5$ le résultat est évident. Autrement définissons $(p_i)_{2\leq i\leq n}$ telle que $p_n = 5$ et que $p_i = p_{i-1}^{\sim}$ pour tout $i,\ 2 \leq i \leq n$, (on rappelle que $p^{\sim}$ désigne le nombre premier immédiatement inférieur à p). Par le lemme 3.2 on a pour tout $i,\ 2 \leq i \leq n\ :\ p_i > s(p_{i-1},1)$, d'où $A_{p_i,1}^{p_{i-1},1} \neq 0$. La suite $\big((q,\ell),(p_1,1),(p_2,1),\ldots,(p_n,1)\big)$ établit donc un lien entre (q,ℓ) et (5,1). □

3.1 La chaîne $\mathcal{P}$ est-elle ergodique?

Suivant le vocabulaire de [Feller 68] une distribution invariante de $\mathcal{P}$ est une famille $[\alpha_{p,k}]_{(p,k)\in\mathcal{O}}$ de $\mathbb{R}^+$ telle que $\sum \alpha_{p,k} = 1$ et $\mathcal{Z}\alpha = \alpha$.

La proposition suivante (qui est une conséquence directe du lemme 3.1) établit le lien entre distribution invariante de $\mathcal{P}$ et densité de probabilité stationnaire de T :

Proposition 3.4. *Soit $\alpha = [\alpha_{p,k}]_{(p,k)\in\mathcal{O}}$ où $\alpha_{p,k} \in \mathbb{R}^+$ et $\sum \alpha_{p,k} = 1$. Une condition nécessaire et suffisante pour que $\mathcal{Z}\alpha = \alpha$, c'est-à-dire que α soit une distribution invariante de $\mathcal{P}$, est que la fonction $f = \sum \frac{\alpha_{p,k}}{r(p,k)}\ \mathbb{1}_{]0,r(p,k)]}(t)$ soit une densité stationnaire de T.*

Puisque $\mathcal{P}$ est irréductible et apériodique, d'après [Feller 68, théorèmes XV.5 et XV.7] une condition nécessaire et suffisante pour qu'il existe une distribution invariante de $\mathcal{P}$ est qu'il existe au moins un état (p,i) tel que $\lim_n {}^nZ_{p,i}^{p,i} \neq 0$, (où l'on note ${}^nZ_{p,i}^{q,j}$ le coefficient générique de la matrice puissance n-ième de $\mathcal{Z}$). De plus si cette limite est non nulle pour un état, elle est aussi non nulle pour tout autre état. On dit alors que la chaîne est ergodique et dans ce cas la famille $[\alpha_{p,k}]_{(p,k)\in\mathcal{O}}$ où $\alpha_{p,k} = \lim_n {}^nZ_{p,k}^{p,k}$ constitue la seule distribution invariante de $\mathcal{P}$.

Notre but dans la suite de ce paragraphe est d'obtenir une expression des coefficients ${}^nZ_{p,k}^{p,k}$ en termes de D-suites, ce qui donne un moyen de tester l'hypothèse d'ergodicité de $\mathcal{P}$ à l'aide de simulations numériques.

Définition 3.5. Soit $c = [(q,\ell),(p,k)] \in \mathcal{O}^2$. On dit qu'une suite finie $(p_i)_{1\leq i\leq n}$ est c-compatible si la suite $\big(q, s(q,1),\ldots,s(q,\ell-1),p_1,p_2,\ldots,p_n\big)$ est une D-suite de résultant égal à $r(p,k)$.

Proposition 3.6. *Soit $c = [(q,\ell),(p,k)] \in \mathcal{O}^2$. Etant donné une suite finie $S = (p_1,\ldots,p_n)$, on définit $j = \min(\{1 \leq i \leq n\ ;\ s(q,\ell-1+i) \neq p_i\})$ si l'ensemble précédent est non vide et $j = \infty$ sinon. Alors une condition nécessaire et suffisante pour que S soit c-compatible est qu'elle soit une D-suite et qu'elle satisfasse l'une ou l'autre des conditions suivantes:*

(a) $j = \infty$ et $r(q,\ell+n) = r(p,k)$.

(b) $j \leq n, \qquad s(q,\ell-1+j) < p_j$ et $\mathrm{res}(S) = r(p,k)$.

Preuve: Supposons que S soit c-compatible. D'après la proposition 2.10, S est une D-suite. Si de plus $j = \infty$ cela entraîne que la suite $S' = (q, s(q,1),\ \ldots, s(q,\ell-1), p_1,\ldots,p_n)$ est le développement d'ordre $n+\ell$ de $\frac{1}{q^{\sim}}$, donc $\mathrm{res}(S') = r(q,\ell+n)$; ainsi S satisfait (a). Si $j < \infty$ alors $j \leq n$ et le deuxième indice de saut de la suite S' définie plus haut est $j+\ell-1$; donc $s(q,\ell-1+j) < p_j$. De plus d'après la proposition 2.10, j est un indice de saut de S. Donc S et S' ont les mêmes résultants; ainsi S satisfait (b). La réciproque est une conséquence immédiate de la proposition 2.10. □

Définition 3.7. On dit qu'une suite $[(q_i,\ell_i)]_{0\leq i\leq n}$ de $\mathcal{O}$ est une n-chaîne reliant (q,ℓ) et (p,k) si $(q_0,\ell_0) = (q,\ell)$, $(q_n,\ell_n) = (p,k)$ et $A_{q_i,\ell_i}^{q_{i-1},\ell_{i-1}} \neq 0$ pour tout i tel que $1 \leq i \leq n$.

Nous notons $Ch\big(n,(q,\ell),(p,k)\big)$ l'ensemble des n-chaînes reliant (q,ℓ) et (p,k).

Proposition 3.8. *Soit $c = [(q,\ell),(p,k)] \in \mathcal{O}^2$. Pour tout $n \geq 1$, l'application notée Δ_n qui a une n-chaîne $[(q_i,\ell_i)]_{0\leq i\leq n}$ reliant (q,ℓ) et (p,k) associe la suite $\Big(\big[A_{q_i,\ell_i}^{q_{i-1},\ell_{i-1}}\big]^{-1}\Big)_{1\leq i\leq n}$ est une bijection de $Ch\big(n,(q,\ell),(p,k)\big)$ sur l'ensemble des D-suites c-compatibles de longueur n.*

Preuve: L'application Δ_n est injective car étant donné $(q,\ell) \in \mathcal{O}$, si pour $(p,k) \in \mathcal{O}$ on a $A_{p,k}^{q,\ell} \neq 0$, alors (p,k) est déterminé par $A_{p,k}^{q,\ell}$. D'autre part lorsque (p,k) parcourt l'ensemble des éléments de $\mathcal{O}$ tels que $A_{p,k}^{q,\ell} \neq 0$, le nombre $\big[A_{p,k}^{q,\ell}\big]^{-1}$ parcourt l'ensemble des p_1 tels que $p_1 \geq s(q,\ell)$, c'est-à-dire tels que $\big(q, s(q,1),\ldots,s(q,\ell-1),p_1\big)$ soit une D-suite, dont le résultant est nécessairement $r(p,k)$. La proposition est donc vraie lorsque $n = 1$. Supposons-la établie pour $n \geq 1$. Soient $c = [(q,\ell),(p,k)] \in \mathcal{O}^2$ et $[(q_i,\ell_i)]_{0\leq i\leq n+1} \in Ch\big(n+1,(q,\ell),(p,k)\big)$.

Posons $p_i = \left[A_{q_i,\ell_i}^{q_{i-1},\ell_{i-1}}\right]^{-1}$, $1 \le i \le n+1$. Montrons que la suite $S = (p_i)_{1\le i\le n+1}$ est une D-suite c-compatible.

Discutons suivant les quatre cas où $A_{q_1,\ell_1}^{q,\ell} \neq 0$:

(1) Si $q > 5$, $q_1 = q$, $\ell < h(q) - 1$, $\ell_1 = \ell + 1$,

alors $p_1 = s(q,\ell) = s(q_1, \ell_1 - 1)$; or d'après l'hypothèse de récurrence la suite $(q_1, s(q_1,1), \ldots, s(q_1,\ell_1 - 1), p_2, \ldots, p_{n+1})$ est une D-suite de résultant $r(p,k)$. Donc S est une D-suite c-compatible (définition 3.7).

(2) Si $q > 5$, $q_1 = q$, $\ell = h(q) - 1$, $\ell_1 = j(q)$,

alors $p_1 = s(q,\ell)$ et par hypothèse $(p_2, \ldots, p_{n+1})$ est une D-suite $[(q, j(q)), (p,k)]$-compatible. Puisque $s(q, h(q) - 1 + i) = s(q, j(q) - 1 + i)$ pour tout $i \ge 1$, on voit par la proposition 3.6 que S est c-compatible.

(3) Si $q_1 > s(q,\ell)$, $\ell_1 = 1$,

alors $p_1 = q_1$ et par hypothèse $(p_1, p_2, \ldots, p_{n+1})$ est une D-suite de résultant $r(p,k)$. Dans ce cas $(q, s(q,1), \ldots, s(q,\ell-1), p_1, p_2, \ldots, p_{n+1})$ est encore une D-suite dont l'indice ℓ est un indice de saut (sachant que 0 est son premier indice), de sorte que son résultant est toujours $r(p,k)$.

(4) Si $q \le 5$, $q_1 = s(q,1)$, $\ell = 1$, $\ell_1 = 1$,

alors $p_1 = s(q,1)$ et par hypothèse $(p_1, p_2, \ldots, p_{n+1})$ est une D-suite de résultant $r(p,k)$. Puisque $s(p_1, i) = s(q, i+1)$ pour tout $i \ge 1$, on voit que $(q, p_1, p_2, \ldots, p_{n+1})$ est encore une D-suite de même résultant $r(p,k)$.

Montrons que inversement toute D-suite $S = (p_i)_{1\le i\le n+1}$ c-compatible est l'image par Δ_{n+1} d'une $(n+1)$-chaîne reliant (q,ℓ) et (p,k).
Par hypothèse la suite $S' = (q, s(q,1), \ldots, s(q,\ell-1), p_1, p_2, \ldots, p_{n+1})$ est une D-suite de résultant $r(p,k)$.

Discutons suivant les valeurs possibles de p_1:

1. Si $p_1 = s(q,\ell)$ et $\ell < h(q) - 1$,

   alors $(p_2, \ldots, p_{n+1})$ est $[(q,\ell+1), (p,k)]$-compatible et d'après l'hypothèse de récurrence il existe une n-chaîne $(q_i,\ell_i)_{1\le i\le n+1}$ telle que pour tout $2 \le i \le n+1$,

   $$\left[A_{q_i,\ell_i}^{q_{i-1},\ell_{i-1}}\right]^{-1} = p_i,$$

   que $(q_1,\ell_1) = (q,\ell+1)$ et que $(q_{n+1},\ell_{n+1}) = (p,k)$. La suite $(q_i,\ell_i)_{0\le i\le n+1}$ où $(q_0,\ell_0) = (q,\ell)$ est une $(n+1)$-chaîne reliant (q,ℓ) et (p,k) et l'image par Δ_{n+1} de cette $(n+1)$-chaîne est bien la suite $(p_i)_{1\le i\le n+1}$.

2. Si $p_1 = s(q,\ell)$ et $\ell = h(q) - 1$ (ce qui inclus les cas où $q \le 5$ et $\ell = 1$),

   alors $(p_2, \ldots, p_{n+1})$ est $[(q, j(q)), (p,k)]$-compatible et comme précédemment on voit que la suite $(p_i)_{1\le i\le n+1}$ est définie par une $(n+1)$-chaîne reliant (q,ℓ) et (p,k).

3. Si $p_1 > s(q,\ell)$,

   alors d'après la proposition 2.10, la suite $(p_1, p_2, \ldots, p_{n+1})$ est une D-suite de même résultant que la suite S'. Donc $(p_2, \ldots, p_{n+1})$ est $[(p_1,1), (p,k)]$-compatible. Par hypothèse de récurrence il existe une n-chaîne $(q_i,\ell_i)_{1\le i\le n+1}$ telle que pour tout $2 \le i \le n+1$,

   $$\left[A_{q_i,\ell_i}^{q_{i-1},\ell_{i-1}}\right]^{-1} = p_i,$$

   que $(q_1,\ell_1) = (p_1,1)$ et que $(q_{n+1},\ell_{n+1}) = (p,k)$. Comme $A_{p_1,1}^{q,\ell} = \frac{1}{p_1}$, on obtient une $(n+1)$-chaîne reliant (q,ℓ) et (p,k) en prolongeant la précédente par $(q_0,\ell_0) = (q,\ell)$; de plus l'image par Δ_{n+1} de cette $(n+1)$-chaîne est bien la suite $(p_i)_{1\le i\le n+1}$.

La proposition est donc démontrée par récurrence. □

Le résultat suivant est une conséquence directe de ce qui précède:

Proposition 3.9. *Soit $[(q,\ell),(p,k)] \in \mathcal{O}^2$. Alors pour tout $n \ge 1$, si ${}^nA_{p,k}^{q,\ell}$ désigne le coefficient générique de la puissance n-ième de la matrice $\mathcal{A}$, on a*

$${}^nA_{p,k}^{q,\ell} = \sum \left\{ \prod_{k=1}^{n} \frac{1}{p_k} \,;(p_1, p_2, \ldots, p_n) \right.$$
$$\left. \text{est une } D\text{-suite } [(q,\ell),(p,k)]\text{-compatible} \right\}.$$

Nous avons vu plus haut que la question de l'ergodicité de la chaîne $\mathcal{P}$ se ramène à la non nullité de $\lim_n {}^nZ_{5,1}^{5,1}$. Or on a ${}^nZ_{5,1}^{5,1} = {}^nA_{5,1}^{5,1}$. De plus une condition nécessaire et suffisante pour qu'une suite $(p_1, p_2, \ldots, p_n)$ soit $[(5,1),(5,1)]$-compatible est qu'elle soit une D-suite de résultant $\frac{2}{3}$. Ainsi d'après la proposition 3.9 nous pouvons écrire:

$${}^nZ_{5,1}^{5,1} = \sum \left\{ \prod_{k=1}^{n} \frac{1}{p_k},\ (p_1, p_2, \ldots, p_n) \right.$$
$$\left. \text{est une } D\text{-suite de résultant } 2/3 \right\}.$$

Les valeurs approchées de ${}^nZ_{5,1}^{5,1}$ pour $n = 2$, 3, 4, 5, obtenues par calcul direct sont respectivement $0,253\ldots$ $0,183\ldots$, $0,171\ldots$, $0,160\ldots$.

Le nombre de termes dans la somme ci-dessus est asymptotiquement d'ordre plus grand que γ^{n^2} pour un certain $\gamma > 1$. Le calcul des ${}^nZ_{5,1}^{5,1}$ est tout de même possible par le procédé de Monte Carlo. Pour cela on nous nous appuyons sur les propositions suivantes:

Proposition 3.10. *Soit $n \geq 1$. Si x est pris au hasard dans I suivant la loi uniforme, la probabilité pour que la D-suite formée des n premiers nombres de son développement soit de résultant 2/3 est égale à ${}^nZ_{5,1}^{5,1}$.*

Preuve: Suivant les notations du paragraphe 2.2, l'ensemble des $x \in I$ ayant la propriété de l'énoncé est égal à la réunion disjointe des $J(S)$ où S parcourt l'ensemble des D-suites de longueur n et de résultant 2/3. Le résultat découle alors de la proposition 3.9 et de la proposition 2.14. □

Proposition 3.11. *Soient $(p,k) \in \mathcal{O}$ et $n \geq 1$. Soit $I(p,k) = \Big]H\big(p, s(p,1), \ldots, s(p,k-1)\big)\ ,\ \dfrac{1}{p^{\sim}}\Big]$.*

Si x est pris au hasard dans $I(p,k)$ suivant la loi uniforme, la probabilité pour que la suite formée des $k+n$ premiers nombres de son développement soit de résultant $r(p,k)$ est égale à ${}^nZ_{p,k}^{p,k}$.

Preuve: D'après le théorème 2.13, l'ensemble des $x \in I(p,k)$ vérifiant la propriété de l'énoncé est la réunion disjointe des ensembles $J(S)$ où S parcourt l'ensemble des D-suite de longueur $n+k$, de résultants $r(p,k)$ et commençant par $(p, s(p,1), \ldots, s(p,k-1))$. Il résulte de la proposition 2.14 que la mesure de Lebesgue de cet ensemble est égale à

$$r(p,k) \prod_{i=0}^{k-1} \frac{1}{s(p,i)} \sum \left\{ \prod_{j=1}^{n} \frac{1}{p_j} \,;(p_1, p_2, \ldots, p_n) \right.$$
$$\left. \text{est } [(p,k),(p,k)]\text{-compatible} \right\}.$$

Mais d'après les propositions 2.14 et 3.9, ce dernier nombre est égal au produit de la longueur de $I(p,k)$ par ${}^nZ_{p,k}^{p,k}$. □

Les calculs par Maple suivant ce procédé de ${}^nZ_{5,1}^{5,1}$ pour $n = 100, 200$ et 300 font apparaître des valeurs comprises entre $0,145$ et $0,148$.

D'autre part le calcul approché de la densité stationnaire

$$g = \sum_{(p,i)\in\mathcal{O}} \frac{\alpha_{p,i}}{r(p,i)} e_{p,i}$$

par la deuxième méthode du paragraphe suivant donne $\alpha_{5,1} = 0,145\ \ldots$ Tous ces éléments nous amènent à énoncer la conjecture suivante:

Conjecture 3.12. *La chaîne $\mathcal{P}$ est ergodique.*

3.2 Conséquences de l'hypothèse de l'ergodicité de la chaîne $\mathcal{P}$

Dans ce paragraphe nous admettons l'hypothèse de l'ergodicité de la chaîne $\mathcal{P}$. Soit

$$g = \sum_{(p,k)\in\mathcal{O}} \frac{\alpha_{p,k}}{r(p,k)}\ e_{p,k}$$

la densité de probabilité stationnaire de T (proposition 3.4). La transformation Φ préserve donc la mesure $\mu = g\lambda$. Nous nous intéressons aux propriétés asymptotiques du couple (Φ, μ).

Lemme 3.13. *Soit $A \subset I$ tel que $\Phi^{-1}(A) = A$. Soit $x \in I$ et soit $S = (p_k)_{0\leq k\leq n-1}$ le développement d'ordre n de x. Alors*

$$A \cap \big]H(S), x\big] = H(S) + \Big(\prod_{k=0}^{n-1} \frac{1}{p_k}\Big)\ \big(A \cap \big]0, \Phi^n(x)\big]\Big).$$

Preuve: L'égalité précédente avec $n = 1$ s'écrit

$$A\ \cap\ \big]\frac{1}{p_0}, x\big] = \frac{1}{p_0} + \frac{1}{p_0}\big(A\ \cap\ ]0, \Phi(x)]\big).$$

Afin de la prouver, remarquons que par hypothèse la restriction de Φ à $\big]\dfrac{1}{p_0}, x\big]$ concide avec la fonction $t \longrightarrow p_0 t - 1$. Donc

$$\Phi^{-1}(A)\ \cap\ \big]\frac{1}{p_0}, x\big] = \big(\frac{1}{p_0} + \frac{1}{p_0}A\big)\ \cap\ \big]\frac{1}{p_0}, x\big].$$

Comme $\big]\dfrac{1}{p_0}, x\big] = \dfrac{1}{p_0} + \dfrac{1}{p_0}\,\big]0, \Phi(x)\big]$ et que $\Phi^{-1}(A) = A$, l'égalité est vraie pour $n = 1$. La démonstration se poursuit en effectuant une récurrence sur n. □

Lemme 3.14. *Soit $A \subset I$ tel que $\Phi^{-1}(A) = A$. Alors pour toute D-suite finie $S = (p_k)_{0\leq k<n}$, on a:*

$$A \cap J(S) = H(S) + \Big(\prod_{k=0}^{n-1} \frac{1}{p_k}\Big)\ \big(A\ \cap\ ]0, \mathrm{res}(S)]\big).) \quad (3\text{–}3)$$

Preuve: La relation est vraie si $n = 1$. On a en effet:

$$S = (p_0), \quad J(S) = \big]\frac{1}{p_0}, \frac{1}{p_0^{\sim}}\big]$$

et

$$\Phi^{-1}(A) \cap J(S) = \big]\frac{1}{p_0}, \frac{1}{p_0^{\sim}}\big] \cap \Big(\frac{1}{p_0} + \frac{1}{p_0}A\Big) = \frac{1}{p_0} + \frac{1}{p_0}\Big(A \cap \big]0, r(p_0, 1)\big]\Big).$$

Supposons le lemme établi pour $n \geq 1$. Soit $S = (p_k)_{0\leq k<n+1}$ une D-suite de longueur $n+1$. Puisque

$$J(S) \subset \big]\frac{1}{p_0}, \frac{1}{p_0^{\sim}}\big],$$

on a

$$\Phi^{-1}(A) \cap J(S) = \big(\frac{1}{p_0} + \frac{1}{p_0}A\big) \cap J(S).$$

Si S est le développement d'ordre $n+1$ de $\frac{1}{p_0^{\sim}}$ l'égalité (3–3) résulte du lemme 3.13. Autrement S possède au moins deux indices de saut. Alors d'après la proposition 2.10 le plus grand indice de saut k_{m-1} de la D-suite $S' = (p_k)_{1\leq k<n+1}$ est le même que celui de S ; par suite d'après le théorème 2.13,

$$J(S) = \frac{1}{p_0} + \frac{1}{p_0}J(S'),$$

d'où

$$\Phi^{-1}(A) \cap J(S) = \frac{1}{p_0} + \frac{1}{p_0}\big(A \cap J(S')\big).$$

Par l'hypothèse de récurrence,

$$A \cap J(S') = H(S') + \Big(\prod_{k=1}^{n}\frac{1}{p_k}\Big)\big(A \cap]0, \mathrm{res}(S')]\big).$$

Comme $\mathrm{res}(S) = \mathrm{res}(S')$ et que $A = \Phi^{-1}(A)$, on en déduit que (3-3) est vraie pour $n+1$. Le lemme est donc établi par récurrence. □

Proposition 3.15. *En supposant l'ergodicité de la chaîne $\mathcal{P}$, la transformation Φ est ergodique.*

Preuve: Soit A borélien de I tel que $\Phi^{-1}(A) = A$. D'après la proposition 2.14 et le lemme 3.14, on a pour toute D-suite finie S:

$$\frac{\lambda\big(A \cap J(S)\big)}{\lambda\big(J(S)\big)} = \frac{\lambda\big(A \cap]0, \mathrm{res}(S)]\big)}{\mathrm{res}(S)}. \qquad (3\text{–}4)$$

Par ailleurs d'après un théorème de Lebesgue, on a pour λ-presque tout $x \in I$:

$$\lim_{\substack{y\longrightarrow x,\ y<x\\ z\longrightarrow x,\ x<z}} \frac{\lambda(A \cap [y, z])}{z - y} = \mathbb{1}_A(x).$$

Il en résulte que si pour $n \geq 0$ et $x \in I$, $S_n(x)$ désigne le développement d'ordre n de x, on a pour λ-presque tout $x \in I$:

$$\lim_n \frac{\lambda\big(A \cap J(S_n(x))\big)}{\lambda\big(J(S_n(x))\big)} = \mathbb{1}_A(x). \qquad (3\text{–}5)$$

Montrons que pour λ-presque tout x, $\mathrm{res}\big(S_n(x)\big) = \frac{2}{3}$ pour une infinité de n. La mesure $\mu = g\lambda$ qui est invariante par Φ, est équivalente à λ. Il découle du théorème de l'éternel retour de Poincaré que pour λ-presque tout $x \in I$ la trajectoire $[\Phi^n(x)]_{n\geq 0}$ issue de x visite une infinité de fois l'intervalle

$$\big]\frac{1}{3} + \frac{1}{3.5}, \ \frac{1}{3} + \frac{1}{3.3}\big];$$

cela entraîne que dans le développement de x, la séquence (3,5) apparaît une infinité de fois ; comme $5 > s(3, 1)$ cela prouve notre assertion. Par suite d'après (3–4) et (3–5), $\mathbb{1}_A(x)$ est λ-presque partout égale à une constante, c'est-à-dire que $A = I$ ou $A = \phi$ λ-pp. Ce qui démontre l'ergodicité du couple (Φ, μ). □

Le reste de ce sous-paragraphe vient en complément d'une première version de notre travail. Nous améliorons le résultat précédent en montrant que sous l'hypothèse de l'ergodicité de la chaîne $\mathcal{P}$ le couple (Φ, μ) est exact. Pour cela nous commençons par établir une formule permettant d'obtenir une expression simple de la puissance nième de l'opérateur de Perron-Frobenius T; ce qui permet d'utiliser les résultats de [Lasota and Mackey 94] dans le but de démontrer le caractère statistiquement stable de la transformation Φ.

Rappelons que pour une suite de nombres $S = (p_k)_{0\leq k\leq n-1}$, nous notons

$$H(S) = \sum_{\ell=0}^{n-1}\ \prod_{k=0}^{\ell}\ \frac{1}{p_k}.$$

Nous écrivons aussi

$$\Pi(S) = \prod_{k=0}^{n-1}\ \frac{1}{p_k}.$$

Pour les définitions de $res(S)$ et de $J(S)$ lorsque S est une D-suite nous renvoyons au début du §2.2. Nous notons $\mathcal{D}_n$ l'ensemble des D-suites de longueur n.

Proposition 3.16. *Soient $f \in L^1(I,\lambda)$ et $n \geq 1$. On a pour presque tout $t \in I$:*

$$T^n f(t) = \sum_{S\in\mathcal{D}_n} \Pi(S) f\big(H(S)+t\Pi(S)\big)\mathbb{1}_{]0,res(S)]}(t). \tag{3–6}$$

Preuve: La relation (3–6) est vraie pour $n = 1$, car c'est la formule classique donnant l'expression de T (formule (3–1) au début de ce paragraphe). En effet le résultant d'une D-suite de longueur 1 à savoir (p) est égal à $r(p,1)$. Appliquons l'opérateur T aux deux membres de la relation (3–6) supposée vraie pour $n \geq 1$. On obtient:

$$T^{n+1} f(t) = \sum_{S\in\mathcal{D}_n}\sum_{p} \Pi([S,p]) f\big(H([S,p]) + t\Pi([S,p])\big)\mathbb{1}_{\left]0,max\left(0,min\left(res(S)p-1,r(p,1)\right)\right)\right]}(t), \tag{3–7}$$

où $[S,p]$ désigne la suite S augmentée de l'élément p. On a tenu compte du fait que

$$\mathbb{1}_{]0,res(S)]}\left(\frac{1+t}{p}\right)\mathbb{1}_{]0,r(p,1)]}(t) = \mathbb{1}_{\left]0,max\left(0,min\left(res(S)p-1,r(p,1)\right)\right)\right]}(t).$$

La fonction indicatrice ci-dessus que l'on note $g_{[S,p]}$ est non identiquement nulle si et seulement si $\frac{1}{p} < res(S)$, ce qui revient au même de dire que la suite $[S,p]$ est une D-suite (définition 2.7). Supposons qu'il en soit ainsi. Deux cas se présentent alors:

(1) $\dfrac{1}{p} < res(S) \leq \dfrac{1}{p^{\sim}}$: on a

$$res([S,p]) = res(S)p - 1 \leq r(p,1)$$

car le plus grand indice de saut de $[S,p]$ est le même que celui de S (définition 2.6); d'où $g_{[S,p]} = \mathbb{1}_{]0,res([S,p])]}$.

(2) $\dfrac{1}{p^{\sim}} < res(S)$: on a

$$res([S,p]) = r(p,1) < res(P)p - 1$$

car le plus grand indice de saut de $[S,p]$ est son dernier indice. Par conséquent on a encore $g_{[S,p]} = \mathbb{1}_{]0,res([S,p])]}$.

Ainsi la relation (3–7) n'est autre que la relation (3–6) où n est remplacé par $n+1$. □

Remarque 3.17. La proposition précédente permet une démonstration plus directe de la proposition 3.9.

Nous notons $D(I)$ l'ensemble des densités de probabilité sur I, la mesure de référence étant λ. Pour tout $f \in D(I)$ et tout $n \geq 1$ nous définissons $U_n f \in D(I)$ par:

$$U_n f = \sum_{S\in\mathcal{D}_n} \Big(\int_{J(S)} f(u)du\Big)\frac{1}{res(S)}\mathbb{1}_{]0,res(S)]}.$$

Proposition 3.18. *Pour tout $f \in D(I)$, $\lim_n \|T^n f - U_n f\|_1 = 0$.*

Preuve: Soit $f \in D(I)$ et $n \geq 1$. D'après la proposition 3.16 et par définition de $U_n f$ on a:

$$\|T^n f - U_n f\|_1 \leq \sum_{S\in\mathcal{D}_n} \int_0^{res(S)} \Big| \Pi(S) f\big(H(S)+t\Pi(S)\big) - \frac{1}{res(S)}\int_{J(S)} f(u)du \Big| dt.$$

En effectuant le changement de variable $v = H(S) + t\Pi(S)$ dans les intégrales de la somme du second membre ci-dessus, tenant compte du fait que $J(S) = \big]H(S), H(S)+res(S)\Pi(S)\big]$ (théorème 2.13), on obtient:

$$\|T^n f - U_n f\|_1 \leq \sum_{S\in\mathcal{D}_n} \int_{J(S)} \Big| f(v) - \frac{1}{res(S)}\int_{J(S)} f(u)du \Big| dv.$$

On reconnaît au second membre: $\|f - E(f/\mathcal{B}_n)\|_1$ où $\mathcal{B}_n$ désigne la tribu finie engendrée par les $J(S)$, S parcourant $\mathcal{D}_n$; $E(f/\mathcal{B}_n)$ étant l'espérance conditionnelle de f par rapport à $\mathcal{B}_n$. Du fait même de l'existence de la correspondande bijective entre I et l'ensemble des D-suites, la tribu engendrée par $\bigcup_n \mathcal{B}_n$ est la tribu de Borel de I. Or par un théorème classique on a $\lim_n \|f - E(f/\mathcal{B}_n)\|_1 = 0$ pour tout $f \in L^1(I,\lambda)$. □

Pour les définitions de l'ensemble $\mathcal{O}$, de l'espace $\mathcal{E}$ et de la matrice $\mathcal{Z}$ nous renvoyons au début de ce paragraphe. Soit F l'application de $l^1(\mathcal{O})$ dans $L^1(I,\lambda)$ définie par: $(\alpha_{p,k})_{(p,k)\in\mathcal{O}} \longrightarrow \sum \frac{\alpha_{p,k}}{r(p,k)} e_{p,k}$ (on rappelle que $e_{p,k}$ désigne $\mathbb{1}_{]0,r(p,k)]}$). L'image de F est l'espace $\mathcal{E}$. Nous avons vu que cet espace est stable par l'opérateur de Perron-Frobenius T associé à Φ. Soit $T_{|\mathcal{E}}$ la réduction de T à $\mathcal{E}$. La matrice de $T_{|\mathcal{E}}$ dans la base de Schauder normalisée $\big(e_{p,k}/r(p,k)\big)_{(p,k)\in\mathcal{O}}$ n'est autre que $\mathcal{Z}$. Soit $K = F^{-1}T_{|\mathcal{E}}F$ l'opérateur de $l^1(\mathcal{O})$ dans $l^1(\mathcal{O})$ transmué de $T_{|\mathcal{E}}$ par l'application F. La matrice de K dans la base canonique étant aussi $\mathcal{Z}$ on voit que K est Markovien.

Nous notons $D(\mathcal{O})$ l'ensemble des densités sur $\mathcal{O}$, la mesure de référence étant bien entendu la mesure ν telle que $\nu(\{c\}) = 1$ pour tout $c \in \mathcal{O}$. D'après l'hypothèse d'ergodicité de la chaîne de Markov $\mathcal{P}$ Il existe une densité stationnaire unique α pour K. De plus pour toute autre densité β de $D(\mathcal{O})$ on a $K^n\beta \longrightarrow \alpha$ simplement sur $\mathcal{O}$ (voir le début du §3.1).

On voit facilement que K est constrictif [Lasota and Mackey 94, definition 5.3.2]. En effet soit B un sous ensemble fini de $\mathcal{O}$ tel que $\sum_{(p,k)\in B} \alpha_{p,k} > 1/2$. Pour toute densité β de $D(\mathcal{O})$ il existe un entier $n_0(\beta)$ tel que pour tout $n \geq n_0(\beta)$, $\sum_{(p,k)\in B}(K^n\beta)_{p,k} > 1/2$. En choisissant δ quelconque avec $0 < \delta < 1$, on a pour tout $E \subset \mathcal{O}$ tel que $\nu(E) < \delta$ et tout $n \geq n_0(\beta)$, $\sum_{(p,k)\in(\mathcal{O}/B)\cup E}(K^n\beta)_{p,k} < 1/2$; ce qui montre bien la constrictivité de K.

On peut alors appliquer à K le théorème 5.6.1 de [Lasota and Mackey 94]; on voit ainsi que la suite $\{K^n\}$ est asymptotiquement stable [Lasota and Mackey 94, definition 5.6.1], à savoir que pour tout $\beta \in D(\mathcal{O})$, $K^n\beta \longrightarrow \alpha$ dans $l^1(\mathcal{O})$.

Soit $g = \sum_{(p,k)\in\mathcal{O}} \frac{\alpha_{p,k}}{r(p,k)} e_{p,k}$ la densité stationnaire de T (proposition 3.4). Il résulte de ce qui précède que pour tout $f \in D(I)\cap\mathcal{E}$, $T^n f \longrightarrow g$ dans $L^1(I,\lambda)$. Le théorème suivant affirme qu'il en est de même pour tout $f \in D(I)$; ce qui veut dire que Φ est statistiquement stable [Lasota and Mackey 94, definition 5.6.2].

Théorème 3.19. *Sous l'hypothèse de l'ergodicité de la chaîne de Markov $\mathcal{P}$, la transformation Φ est statistiquement stable.*

Preuve: Soient $f \in D(I)$ et $\epsilon > 0$. D'après la proposition 3.18 il existe un entier n et $f^* \in D(I)\cap\mathcal{E}$ tels que $\|T^n f - f^*\|_1 < \epsilon/2$. D'après la discussion précédente il existe k tel que $\|T^k f^* - g\|_1 < \epsilon/2$. D'où $\|T^{n+k} f - g\|_1 < \epsilon$; on en déduit que pour tout $f \in D(I)$, $T^n f \longrightarrow g$ dans $L^1(I,\lambda)$. □

D'après [Lasota and Mackey 94, proposition 5.6.2] on peut alors énoncer le résultat suivant:

Corollaire 3.20. *Sous l'hypothèse d'ergodicité de la chaîne de Markov $\mathcal{P}$, le couple (Φ,μ) où $\mu = g\lambda$ est exact et en particulier mélangeant.*

3.3 Relations vérifiées par une distribution invariante de $\mathcal{P}$

Dans ce paragraphe nous déterminons les rapports $\frac{\alpha_{p,k}}{\alpha_{p,1}}$ pour $(p,k) \in \mathcal{O}$ en supposant que α est une distribution invariante de $\mathcal{P}$. Ces relations sont utilisées au §5 dans les algorithmes de calcul de la densité stationnaire g.

Rappelons que d'après la relation (3–2) du début de ce paragraphe on a pour tout $(p,k) \in \mathcal{O}$:

$$T(e_{p,k}) = \begin{cases} \dfrac{1}{s(p,k)} e_{p,k+1} + \displaystyle\sum_{p'>s(p,k)} \frac{1}{p'} e_{p',1} & \text{si } k < h(p)-1, \\ \dfrac{1}{s(p,k)} e_{p,j(p)} + \displaystyle\sum_{p'>s(p,k)} \frac{1}{p'} e_{p',1} & \text{si } k = h(p)-1. \end{cases}$$

Reportons cette expression dans la somme $\sum_{(p,k)\in\mathcal{O}} \frac{\alpha_{p,k}}{r(p,k)} T(e_{p,k})$ et après regroupement des termes semblables, identifions le résultat à $g = \sum \frac{\alpha_{p,k}}{r(p,k)} e_{p,k}$ (voir lemme 3.1).

On voit alors que pour tout $p > 5$ et tout k, $2 \leq k \leq h(p)-1$:

$$\frac{\alpha_{p,k}}{r(p,k)} = \begin{cases} \dfrac{\alpha_{p,k-1}}{s(p,k-1)\ r(p,k-1)} & \text{si } k \neq j(p), \\ \dfrac{\alpha_{p,j(p)-1}}{r\big(p,j(p)-1\big)s\big(p,j(p)-1\big)} & \\ \quad + \dfrac{\alpha_{p,h(p)-1}}{r\big(p,h(p)-1\big)s\big(p,h(p)-1\big)} & \\ & \text{si } k = j(p) \text{ et } j(p) \geq 2. \end{cases}$$

$\big($D'ailleurs il apparaît aussi que $4\alpha_{2,1} = \alpha_{5,1}\big)$.

On en déduit les relations suivantes:

$$\forall \ell,\ \ 2 \leq \ell < j(p), \quad \frac{\alpha_{p,\ell}}{\alpha_{p,1}} = \frac{r(p,\ell)}{r(p,1)} \prod_{k=1}^{\ell-1} \frac{1}{s(p,k)},$$

$$\forall \ell,\ \ j(p) < \ell < h(p), \quad \frac{\alpha_{p,\ell}}{\alpha_{p,j(p)}} = \frac{r(p,\ell)}{r(p,j(p))} \prod_{k=j(p)}^{\ell-1} \frac{1}{s(p,k)},$$

et si $j(p) \geq 2$,

$$\frac{\alpha_{p,j(p)}}{\alpha_{p,1}} = \frac{r(p,j(p))}{r(p,1)} \prod_{k=1}^{j(p)-1} \frac{1}{s(p,k)} \Big(1 - \prod_{k=j(p)}^{h(p)-1} \frac{1}{s(p,k)}\Big)^{-1}.$$

4. CALCUL APPROCHÉ DE LA DENSITÉ STATIONNAIRE

L'hypothèse de l'ergodicité de $\mathcal{P}$ étant admise, le but de cette partie est la détermination d'une approximation la plus fine possible de la densité stationnaire $g = \sum \alpha_{p,k}/r(p,k)\ \mathbb{1}_{]0,r(p,k)]}$. Les deux méthodes exposées ci-dessous sont complémentaires et sont fondées sur une démarche heuristique qui utilise la méthode de [Ulam 60] d'approximation matricielle de l'opérateur T tout en

conservant le terme reste sous la forme d'un opérateur intégral.

Rappelons la formule (3–1) du paragraphe 3:

$$Tf(t) = \sum_{n=1}^{\infty} \frac{1}{p_n} f\Big(\frac{1+t}{p_n}\Big) \mathbb{1}_{\left]0, p_n/p_n^{\sim}-1\right]}(t),$$

où p_n désigne le nombre premier de rang n.

Un entier N étant fixé, nous définissons l'opérateur tronqué T_N par la somme partielle d'indice N de la série ci-dessus et nous notons $\widetilde{T}_N$ l'opérateur reste $T - T_N$.

Une première étape consiste à remplacer la série définissant $\widetilde{T}_N f(t)$ par une intégrale. Pour cela nous utilisons la fonction G sur $\mathbb{R}_+$ définie par (log désignant toujours par la suite le logarithme népérien):

$$G(x) = (x + 2,1362)\Big[\log(x + 2,1362) + \log\log(x + 2,1362) - \frac{21}{22}\Big],$$

fonction qui présente l'intérêt de fournir un lissage correct de la suite $n \longrightarrow p_n$. Elle s'obtient en partant de l'encadrement [Ellison et Mendès France 75, p. 25]: $\forall n \geq 21$,

$$n\Big(\log\, n + \log\log\, n - \frac{3}{2}\Big) < p_n < n\Big(\log\, n + \log\log\, n - \frac{1}{2}\Big)$$

et en procédant à l'aide de Maple à un ajustement empirique de son graphe avec celui de la suite $n \longrightarrow p_n$ pour $n \leq 30\ 000$.

D'autre part grâce à la formule ([Ellison et Mendès France 75], p. 32) (où $a = 0,261\ldots$):

$$\sum_{p<x} \frac{1}{p} = a + \log\log(x) + O\Big(\frac{1}{\log x}\Big),$$

on a pour x assez grand:

$$\sum_{1\leq n\leq x} \frac{1}{p_n} \simeq a + \log\log\big(G(x)\big).$$

En fait nous n'utilisons que les premiers termes du développement asymptotique du deuxième membre, à savoir la fonction

$$L(x) = a + \log\log(x) + \frac{\log\log(x)}{\log(x)} + \frac{\log\log(x) - \frac{3}{4} - \frac{1}{2}\log\log(x)^2}{\log(x)^2}.$$

L'autre intérêt de L est de fournir une bonne approximation d'une primitive de $\dfrac{1}{G(x)}$ pour x grand.

4.1 Description de la première méthode d'approximation. Conjecture concernant la partie principale de la densité stationnaire au voisinage de 0.

Cette méthode est fondée sur le raisonnement heuristique suivant: d'après le théorème des nombres premiers (voir [Ellison et Mendès France 75]) $p_n - p_{n-1}$ est "en moyenne" égal à $\log n$ et comme $p_n \sim n \log n$ on voit que $\dfrac{p_n}{p_n^{\sim}} - 1$ se comporte "en moyenne" comme $\dfrac{1}{n}$. Cela nous amène pour la première méthode à utiliser la forme simplifiée $\widetilde{T}_N^{(1)}$ de l'opérateur $\widetilde{T}_N$ définie par:

$$\widetilde{T}_N^{(1)} f(t) = \mathbb{1}_{\left]0, \frac{1}{N}\right]}(t) \sum_{N\leq n<\frac{1}{t}} \frac{1}{p_n} f\Big(\frac{1+t}{p_n}\Big),$$

ou encore sous forme intégrale:

$$\widetilde{T}_N^{(1)} f(t) = \mathbb{1}_{\left]0, \frac{1}{N}\right]}(t) \int_N^{1/t} f\Big(\frac{1+t}{G(x)}\Big) \frac{dx}{G(x)};$$

ce qui donne, en faisant $t = 0$ sous l'intégrale puis en effectuant le changement de variable $u = L(x)$ (sachant que $du = \frac{dx}{G(x)}$):

$$\widetilde{T}_N^{(1)} f(t) = \mathbb{1}_{]0, \frac{1}{N}]}(t) \int_{L(N)}^{L(1/t)} f\Big(\frac{1}{G\big(L^{-1}(u)\big)}\Big) du. \quad (4–1)$$

Remarque 4.1. Dans le cas particulier où $f = \mathbb{1}_{]0,s]}$ avec $P(s) < p_{N+1}$ on justifie directement l'approximation

$$\widetilde{T}_N^{(1)} f(t) \simeq \Big[L(\frac{1}{t}) - L(N)\Big]\ \mathbb{1}_{\left]0, \frac{1}{N}\right]}(t) \quad (4–2)$$

en s'appuyant sur le fait que dans ce cas on a d'après la formule (3–2) du début du §3:

$$\widetilde{T}_N f(t) = \sum_{n\geq N+1} \frac{1}{p_n}\ \mathbb{1}_{\left]0, p_n/p_n^{\sim}-1\right]}(t).$$

Cela s'applique en particulier au cas où $f = \mathbb{1}_{]0,r(p,k)]}$ avec $p \leq p_N$ et $1 \leq k \leq h(p) - 1$.

En effet du fait que $\dfrac{1}{p^{\sim}} \leq r(p,k)$, on a $P(r(p,k)) \leq p < p_{N+1}$.

Pour la définition de la fonction h, ainsi que celles de $\mathcal{E}$, de $\mathcal{O}$, de $e_{p,k}$ et de $\mathcal{A}$ nous renvoyons au début du §3. D'après (4–1) et la remarque précédente un vecteur propre de $T_N + \widetilde{T}_N^{(1)}$ est de la forme

$$f(t) = \sum_{(p,k)\in\mathcal{O},\ p\leq p_N} \alpha_{p,k}\, e_{p,k}(t) + \mathbb{1}_{]0,\frac{1}{N}]}(t) \int_{L(N)}^{L(1/t)} \varphi(x)dx,$$

où $\alpha_{p,k} \in \mathbb{R}$ et φ est une fonction continue sur $[L(N), \infty[$.

4.1.1 Notation. Nous écrivons $O_N = \{(p,k) \in \mathcal{O};\ p \leq p_N\}$ et nous notons E_N le vecteur ligne dont les composantes sont les fonctions $e_{p,k}$, $(p,k) \in O_N$. Ainsi pour tout vecteur colonne $V = [\alpha_{p,k}]_{(p,k)\in O_N}$, la somme $\sum_{(p,k)\in O_N} \alpha_{p,k}\ e_{p,k}(t)$ s'écrit plus simplement $E_N(t)V$. D'autre part $\mathcal{E}_N$ désigne le sous-espace de $\mathcal{E}$ engendré par les $e_{p,k}$, $(p,k) \in O_N$ et pr_N est la projection canonique de $\mathcal{E}$ sur $\mathcal{E}_N$. Enfin $\mathcal{A}_N$ est la matrice de l'endomorphisme $pr_N \circ T_N \mid \mathcal{E}_N$ dans la base $[e_{p,k}]_{(p,k)\in O_N}$; (cette matrice est donc un bloc de la matrice $\mathcal{A}$). Dans les propositions 4.1 et 4.4 plus bas $\delta_{p,1}$ désigne le vecteur colonne élémentaire de $\mathbb{R}^{O_N}$ associé à $e_{p,1}$, $(p \leq N)$. Nous adoptons la notation $x \vee y$ pour $\max(x,y)$.

Soit $\mathcal{L}_N^{(1)}$ le sous-espace de $L^1(I,\lambda)$ constitué des fonctions f pour lesquelles il existe un couple (V,φ) où $V \in \mathbb{R}^{O_N}$ et φ est une fonction continue, tel que:

$$f(t) = E_N(t)V + \mathbb{1}_{]0,\frac{1}{N}]}(t) \int_{L(N)}^{L(1/t)} \varphi(x)dx. \qquad (4\text{–}3)$$

En première approximation $\mathcal{L}_N^{(1)}$ est invariant par $T_N + \widetilde{T}_N^{(1)}$. Plus précisément nous énonçons la proposition suivante:

Proposition 4.2. *L'opérateur $T_N + \widetilde{T}_N^{(1)}$ est asymptotiquement proche de l'endomorphisme noté Γ_N^1 de $\mathcal{L}_N^{(1)}$ qui à $f \in \mathcal{L}_N^{(1)}$ représenté selon (4-3) par le couple (V,φ), associe l'élément représenté par (W,ψ) où:*

$$W = \mathcal{A}_N V + \sum_{1\leq n\leq N} \Big[\frac{1}{p_n} \int_{L(N)}^{L(N\vee p_n)} \varphi(x)dx\Big]\delta_{p_n,1}$$

et

$$\psi(x) = s(V) + \int_{L(N)}^{x} \varphi(u)du, \quad x \geq L(N),$$

($s(V)$ désignant la somme des composantes de V).

Preuve: Soit $f \in \mathcal{L}_N^{(1)}$; posons $f_1(t) = E_N(t)V$ et

$$f_2(t) = \mathbb{1}_{]0,\frac{1}{N}]}(t) \int_{L(N)}^{L(1/t)} \varphi(x)dx.$$

Il résulte de (4–2) (voir la remarque plus haut) que

$$Tf_1(t) \simeq E_N(t)\mathcal{A}_N V + s(V)\mathbb{1}_{]0,\frac{1}{N}]}(t) \int_{L(N)}^{L(1/t)} du.$$

D'autre part on a pour tout $u \geq L(N)$: $L^{-1}(u) \geq N$, d'où $G\big(L^{-1}(u)\big) \geq G(N) \geq N$ et par suite

$$f_2\big(\frac{1}{G(L^{-1}(u))}\big) = \int_{L(N)}^{L\left(G(L^{-1}(u))\right)} \varphi(x)dx.$$

Comme $L\big(G\big(L^{-1}(u)\big)\big) = u + ue^{-u}(1+o(1))$, on a pratiquement $L\big(G\big(L^{-1}(u)\big)\big) = u$. Ainsi:

$$\widetilde{T}_N^{(1)} f_2(t) \simeq \mathbb{1}_{]0,\frac{1}{N}]}(t) \int_{L(N)}^{L(1/t)} \Big(\int_{L(N)}^{u} \varphi(x)dx\Big)du.$$

Enfin en simplifiant

$$T_N f_2(t) \simeq \sum_{1\leq n\leq N} \frac{1}{p_n} f_2\Big(\frac{1}{p_n}\Big) e_{p_n,1}(t),$$

on voit que Tf_2 est proche de la fonction de $\mathcal{L}_N^{(1)}$ représentée selon (4–3) par le couple

$$\Big(\sum_{n=1}^{N} \frac{1}{p_n} f_2\Big(\frac{1}{p_n}\Big)\delta_{p_n,1}\ ,\quad \int_{L(N)}^{x} \varphi(u)du\Big).$$

En ajoutant les deux approximations de Tf_1 et Tf_2 ainsi définies on obtient la fonction $\Gamma_N^1 f$ de $\mathcal{L}_N^{(1)}$ représentée par le couple (W,ψ) de l'énoncé. □

On obtient une approximation $g_N^{(1)}$ de g sous forme du vecteur propre normalisé de Γ_N^1 de valeur propre dominante μ_N. Le couple (V,φ) représentant $g_N^{(1)}$ vérifie:

$$\begin{cases} s(V) + \displaystyle\int_{L(N)}^{x} \varphi(u)du = \mu_N \varphi(x) \\ \text{et } \mathcal{A}_N V + \displaystyle\sum_{1\leq n\leq N} \Big(\frac{1}{p_n} \int_{L(N)}^{L(N\vee p_n)} \varphi(x)dx\Big)\delta_{p_n,1} = \mu_N V. \end{cases} \qquad (4\text{–}4)$$

On en déduit d'une part que pour tout $x \geq L(N)$,

$$\varphi(x) = \frac{1}{\mu_N}\ s(V) \exp\Big(\frac{1}{\mu_N}\big(x - L(N)\big)\Big). \qquad (4\text{–}5)$$

Soit d'autre part $D_N(\mu_N)$ la matrice carrée construite sur l'ensemble d'indices O_N telle que pour tout $n \leq N$ sa ligne d'indice $(p_n,1)$ soit formée de

$$\frac{1}{p_n}\Big[\exp\Big(\frac{1}{\mu_N}\big(L(N\vee p_n) - L(N)\big)\Big) - 1\Big]$$

répété $card(O_N)$ fois, tandis que toutes ses autres lignes sont nulles.

Compte tenu de (4–5) la relation (4–4) conduit à

$$[\mathcal{A}_N + D_N(\mu_N)]V = \mu_N V. \qquad (4\text{–}6)$$

La valeur de μ_N maximale pour laquelle l'équation (4–6) admet une solution non nulle est calculée par approximations successives , ce qui détermine simultanément V. On obtient alors φ à partir de (4–5) et enfin $g_N^{(1)}$ par normalisation dans $L^1(I,\lambda)$ de la fonction

$t \longrightarrow E_N(t)V + \mathbb{1}_{]0,\frac{1}{N}]}(t)\ s(V)\Big[\exp\Big(\frac{1}{\mu_N}\big(L(1/t) - L(N)\big)\Big) - 1\Big]$.

Nous trouvons $\mu_{200} = 0,997\ldots$ et $\mu_{600} = 0,999\ldots$. Il y a donc tout lieu de penser que $\mu_N \longrightarrow 1$ et puisque $\exp(L(1/t)) \sim \log(1/t)$, nous énonçons la conjecture suivante:

Conjecture 4.3. *$k = lim_{t\longrightarrow 0} g(t)/log(1/t)$ existe et $k \neq 0$.*

Avec $N = 600$ nous obtenons $k \simeq 0,78$. Une meilleure estimation de cette constante est donnée au prochain paragraphe.

4.2 Seconde méthode d'approximation fondée sur l'heuristique de H.Cramér

Afin d'accélérer la convergence de g_N vers g nous définissons une approximation plus fine de l'opérateur T en introduisant un terme qui prend en compte les fluctuations erratiques de la suite

$$n \longrightarrow \Big(\frac{p_n}{p_n^{\sim}} - 1\Big).$$

Pour cela nous utilisons le modèle heuristique de [Cramér 36] dont voici la description. Soit $(U_n)_{n>2}$ un suite d'urnes contenant des boules noires et blanches, la probabilité de tirer une blanche de U_n étant $\frac{1}{log\ n}$. Considérons l'expérience aléatoire consistant à tirer de manière indépendante une boule dans chaque urne. Soit pour tout entier $n \geq 1$ la variable aléatoire $\mathbb{X}_n$ donnant le rang du nième tirage produisant une boule blanche. Selon cette heuristique, la suite $(p_n)_{n\geq 1}$ est vue comme une réalisation du processus $(\mathbb{X}_n)_{n\geq 1}$. A partir de ce modèle H.Cramér [19] a montré que pour tout $c \geq 0$ la probabilité de l'évènement $p_{n+1} - p_n > c\ logp_n^2$ est équivalente à Ap_n^{-c}, (on a $A \simeq 1 - 0,4c$ expérimentalement); de plus on a presque sûrement $\mathbb{X}_{n+1} - \mathbb{X}_n \leq log\mathbb{X}_n^2$ pour tout n.

Par conséquent l'inégalité suivante est pratiquement toujours vérifiée:

$$\frac{2}{p_n^{\sim}} \leq \frac{p_n}{p_n^{\sim}} - 1 \leq \frac{\log p_n^2}{p_n}.$$

Nous posons pour $x > 1$:

$$\mu(x) = \frac{2}{G(x)} \quad \text{et} \quad \nu(x) = \frac{\log G(x)^2}{G(x)},$$

et plus bas μ^{-1}, ν^{-1} désignent les fonctions inverses de μ et de ν sur I.

Nous notons pour $t \in I$ et $x > 1$:

$$\gamma(t,x) = \frac{tG(x)}{logG(x)^2} \quad \text{et}$$
$$F(t,x) = 1 - G(x)^{-\gamma(t,x)}(1 - 0,4\gamma(t,x)).$$

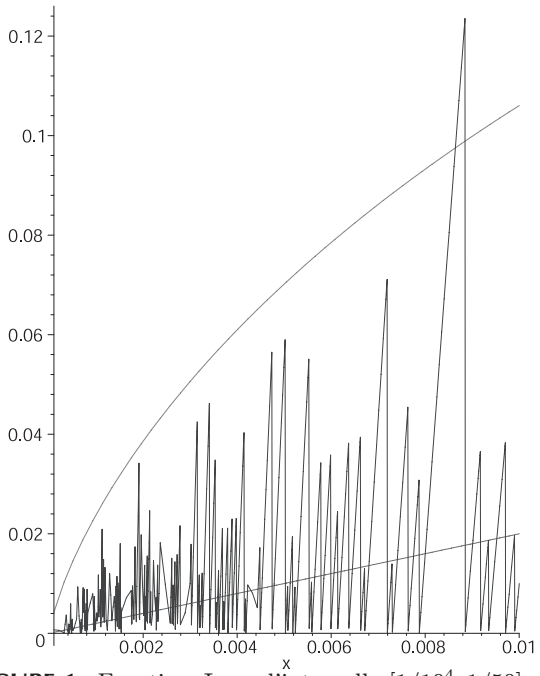

FIGURE 1. Fonction Φ sur l'intervalle $[1/10^4, 1/50]$ avec les fonctions $f_1 : x \to 2x$ et $f_2 : x \to 1/2\ xlog(x)^2$; ($f_1(x)$ et $f_2(x)$ sont les parties principales de $\mu(n)$ et $1/2\ \nu(n)$, si n est l'entier tel que $p_n = P(x)$).

Rappelons enfin que $\frac{p_n}{p_n^{\sim}} - 1$ s'écrit aussi $r(p_n, 1)$.

D'après la discussion précédente, pour $t \in I$ et pour n assez grand, la probabilité de l'événement $t \leq r(p_n, 1)$ est égale à $p_n^{-\gamma(t,n)}(1 - 0,4\gamma(t,n))$. Par conséquent, pour $N \in \mathbb{N}$ et $t \in I$ fixés, l'espérance de $\widetilde{T}_N f(t)$ pour $f \in L^1(I,\lambda)$ est donnée par

$$\sum_{N<n}^{n\leq N\vee\mu^{-1}(t)} \frac{1}{p_n} f\Big(\frac{1+t}{p_n}\Big) + \sum_{N\vee\mu^{-1}(t)<n}^{n\leq N\vee\nu^{-1}(t)} f\Big(\frac{1+t}{p_n}\Big) p_n^{-(1+\gamma(t,n))}(1 - 0,4\gamma(t,n)).$$

Nous définissons l'opérateur $\widetilde{T}_N^{(2)}$ en posant $\widetilde{T}_N^{(2)} f(t)$ égale à l'expression ci-dessus. Puis comme au paragraphe précédent nous remplaçons les sommes $\sum_{n\geq N}$ par des intégrales en y effectuant la même simplification ($t = 0$ sous l'intégrale); cela donne après changement de variable:

$$\widetilde{T}_N^{(2)} f(t) = \int_{L(N)}^{L\left(N\vee\nu^{-1}(t)\right)} f\Big(\frac{1}{G\big(L^{-1}(u)\big)}\Big) du - \int_{L\left(N\vee\mu^{-1}(t)\right)}^{L\left(N\vee\nu^{-1}(t)\right)} f\Big(\frac{1}{G\big(L^{-1}(u)\big)}\Big) F\Big(t, L^{-1}(u)\Big) du. \tag{4–7}$$

Posons

$$R_N(f,t)=\frac{\displaystyle\int_{L\left(N\vee\mu^{-1}(t)\right)}^{L\left(N\vee\nu^{-1}(t)\right)} f\Big(\frac{1}{G\big(L^{-1}(u)\big)}\Big)F\Big(t,L^{-1}(u)\Big)du}{\displaystyle\int_{L\left(N\vee\mu^{-1}(t)\right)}^{L\left(N\vee\nu^{-1}(t)\right)} f\Big(\frac{1}{G\big(L^{-1}(u)\big)}\Big)du.}$$

Du fait que selon la conjecture précédente g est asymptotiquement proportionnelle à $-\log$ et que d'autre part $R_N(.,.)$ est homogène de degré 0 par rapport à son premier argument, on a $R_N(g,t)\simeq R_N(-\log,t)$. De plus $\log\Big(G\big(L^{-1}(u)\big)\Big)$ est pratiquement égal à e^u (voir paragraphe précédent). Par suite on a:

$$R_N(-\log,t)\simeq\frac{\displaystyle\int_{L\left(N\vee\mu^{-1}(t)\right)}^{L\left(N\vee\nu^{-1}(t)\right)} e^u F\Big(t,L^{-1}(u)\Big)du}{\Big(\exp\big(L\big(N\vee\nu^{-1}(t)\big)\big)-\exp\big(L\big(N\vee\mu^{-1}(t)\big)\big)\Big).}$$

Pour simplifier l'écriture nous notons par la suite

$$\theta_N(t)=R_N(-\log,t).$$

Soit T_N^* l'opérateur positif sur $L^1(I,\lambda)$ défini par

$$\begin{aligned}T_N^*f(t)=T_Nf(t)&+\int_{L(N)}^{L\left(N\vee\nu^{-1}(t)\right)} f(\frac{1}{G\big(L^{-1}(u)\big)})du\\&-\theta_N(t)\int_{L\left(N\vee\mu^{-1}(t)\right)}^{L\left(N\vee\nu^{-1}(t)\right)} f(\frac{1}{G\big(L^{-1}(u)\big)})du.\end{aligned}\tag{4–8}$$

D'après ce qui précède on peut chercher une approximation $g_N^{(2)}$ de g comme vecteur propre de T_N^*. Dans ce but nous procédons à une ultime simplification de T_N^* afin de permettre le calcul de ses éléments propres par un programme sous Maple. Soit $\mathcal{L}_N^{(2)}$ le sous-espace de $L^1(I,\lambda)$ constitué des fonctions de la forme:

$$\begin{aligned}f(t)=E_N(t)V&+P\Big(L\big(N\vee\nu^{-1}(t)\big)\Big)-P\big(L(N)\big)\\&-\theta_N(t)\Big[P\Big(L\big(N\vee\nu^{-1}(t)\big)\Big)-P\Big(L\big(N\vee\mu^{-1}(t)\big)\Big)\Big],\end{aligned}\tag{4–9}$$

où V est un vecteur colonne de $\mathbb{R}^{O_N}$ et $u\longrightarrow P(u)$ est une fonction continue sur $[L(N),\infty[$.

En première approximation $\mathcal{L}_N^{(2)}$ est invariant par T_N^*. Plus précisément nous énonçons la proposition suivante:

Proposition 4.4. *L'opérateur T_N^* est asymptotiquement proche de l'endomorphisme Γ_N^* de $\mathcal{L}_N^{(2)}$ qui à $f\in\mathcal{L}_N^{(2)}$ représentée selon (4–9) par le couple $\big(V,P\big)$ associe l'élément représenté par le couple (W,Q) tel que:*

$$\begin{aligned}W=&\mathcal{A}_NV+\sum_{n=1}^{N}\frac{1}{p_n}\Big[P\big(L\big(N\vee\nu^{-1}(\tfrac{1}{p_n})\big)\big)-P\big(L(N)\big)\\&-\theta_N(\tfrac{1}{p_n})\Big(P\big(L\big(N\vee\nu^{-1}(\tfrac{1}{p_n})\big)\big)\\&-P\big(L\big(N\vee\mu^{-1}(\tfrac{1}{p_n})\big)\big)\Big)\Big]\,\delta_{p_n,1},\end{aligned}$$

et $Q(u)$ est une primitive de la fonction

$$\begin{aligned}u\longrightarrow\ &s(V)+P\big(M(u)\big)-P\big(L(N)\big)\\&-\widetilde{\theta}_N(u)\Big[P\big(M(u)\big)-P\big(m(u)\big)\Big];\end{aligned}$$

où $s(V)$ désigne la somme des composantes de V et M,m et $\widetilde{\theta}_N$ sont les fonctions définies par:

$$M(u)=L\left(N\vee\nu^{-1}\Big(\frac{1}{G\big(L^{-1}(u)\big)}\Big)\right),$$

$$m(u)=L\left(N\vee\mu^{-1}\Big(\frac{1}{G\big(L^{-1}(u)\big)}\Big)\right),$$

$$\widetilde{\theta}_N(u)=\theta_N\Big(\frac{1}{G\big(L^{-1}(u)\big)}\Big).$$

(Rappelons que $\delta_{p_n,1}$ désigne le vecteur colonne élémentaire de $\mathbb{R}^{O_N}$ associé à $e_{p_n,1}$).

Preuve: Soit $f\in\mathcal{L}_N^{(2)}$. Ecrivons $f=f_1+f_2$ où $f_1(t)=E_N(t)V$ et

$$\begin{aligned}f_2(t)=&P\Big(L\big(N\vee\nu^{-1}(t)\big)\Big)-P\big(L(N)\big)\\&-\theta_N(t)\Big[P\Big(L\big(N\vee\nu^{-1}(t)\big)\Big)-P\Big(L\big(N\vee\mu^{-1}(t)\big)\Big)\Big].\end{aligned}$$

On a pour tout $(p,k)\in O_N$:

$$\frac{1}{G(N)}\le\frac{1}{p^{\sim}}\le r(p,k)$$

et par conséquent pour tout $u\ge L(N)$,

$$\frac{1}{G(L^{-1}(u))}\le r(p,k).$$

Donc

$$\begin{aligned}T_N^*(e_{p,k})(t)=T_N(e_{p,k})(t)&+L\big(N\vee\nu^{-1}(t)\big)-L(N)\\&-\theta_N(t)\Big(L\big(N\vee\nu^{-1}(t)\big)-L\big(N\vee\mu^{-1}(t)\big)\Big).\end{aligned}$$

On en déduit que $T_N^*f_1\simeq\Gamma_N^*f_1$ où $\Gamma_N^*f_1$ est représentée selon (4–9) par $\Big(\mathcal{A}_NV,\ s(V)\int du\Big)$.

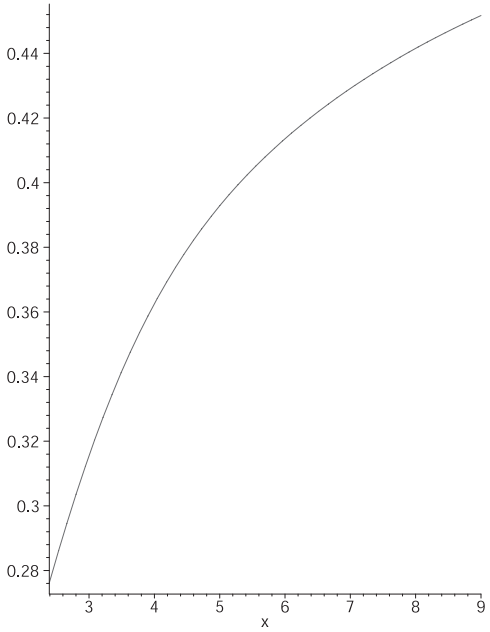

FIGURE 2. Fonction $\widetilde{\theta}_{600}$.

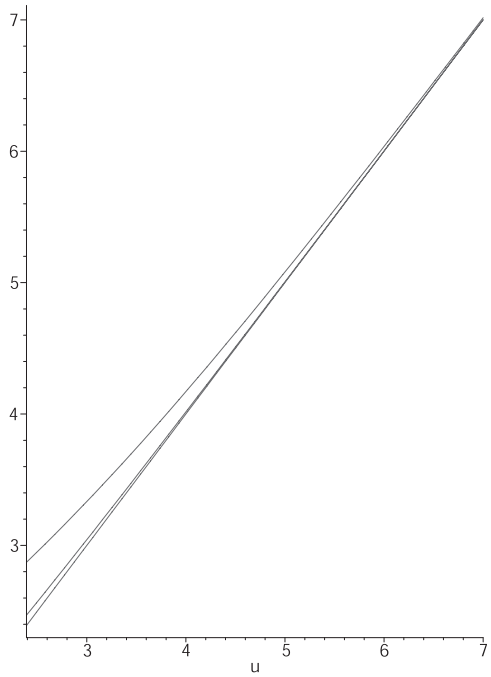

FIGURE 3. Fonctions $u \to u, u \to m(u), u \to M(u)$.

D'autre part après avoir effectué la simplification

$$T_N f_2 \simeq \sum_{n=1}^{N} \frac{1}{p_n} f_2\big(\frac{1}{p_n}\big) e_{p_n,1},$$

on a d'après (4–8): $T_N^* f_2 \simeq \Gamma_N^*(f_2)$, où $\Gamma_N^*(f_2)$ est représentée selon (4–9) par le couple formé du vecteur

$$\sum_{n=1}^{N} f_2\big(\frac{1}{p_n}\big)\delta_{p_n,1}$$

et d'une fonction primitive de

$$u \longrightarrow P\big(M(u)\big) - P\big(L(N)\big) - \widetilde{\theta}_N(u)\Big[P\big(M(u)\big) - P\big(m(u)\big)\Big].$$

Ainsi la fonction $\Gamma_N^* f = \Gamma_N^* f_1 + \Gamma_N^* f_2$ approximant $T_N^* f$ est bien représentée par le couple (W, ψ) de l'énoncé. □

Pour déterminer $g_N^{(2)}$ nous procédons par approximations successives en calculant

$$\lim_n \Gamma_N^{*n} f_0 / \|\Gamma_N^{*n} f_0\|_1,$$

où f_0 est la fonction $\mathbb{1}_{]0,2/3]}$, c'est-à-dire la fonction représentée selon (4–9) par $(\delta_{3,1}, 0)$. La convergence est relativement rapide car 20 itérations de Γ_N^* suffisent pour atteindre un résultat stationnaire.

Les calculs avec $N = 600$ conduisent à un couple (V, P) représentant $g_N^{(2)}$ selon (4–9) vérifiant $P(u) \simeq \lambda \exp u$ pour $u \geq 9$, avec $\lambda = 1,28\ldots$ Comme $L(\nu^{-1}(t)) = \log\log(1/t) + o(1)$ et que $\log(\nu^{-1}(t)) - \log(\mu^{-1}(t)) = O(\log\log(1/t))$ on voit que $g_{600}^{(2)}(t) \sim k \log(1/t)$ avec $k \simeq 1,28$.

En posant $u_0 = L(600)$, on a $P(u_0)\exp(-u_0) \simeq 0,78$. C'est la valeur de k obtenue par la méthode précédente. On voit que les deux méthodes d'approximation de g se complètent et ne sont pas contradictoires. Les suites $g_N^{(1)}$ et $g_N^{(2)}$ semblent bien converger vers la même limite g, la deuxième convergeant nettement plus vite que la première, (voir les comparaisons des graphes au §5).

5. RESULTAT DES CALCULS SOUS MAPLE

Dans cette partie nous apportons quelques précisions concernant l'algorithme de développement d'un nombre et nous donnons quelques exemples de développement de constantes fondamentales. Puis nous présentons les graphes des approximations de la densité stationnaire g calculées par les deux méthodes exposées au §4 en les comparant aux histogrammes des orbites issues des constantes précédentes, ceci afin de vérifier via le théorème de Birkhoff l'ergodicité du système.

5.1 Exemples de développements de nombres irrationnels

Etant donné un nombre réel x, on sait par la proposition 2.14 que son développement d'ordre $n+1$ (noté $S_n = (p_0 \ldots p_n)$) définit une approximation rationnelle

de x avec une précision inférieure à $1,5\prod_{i=0}^{n} p_i$. En admettant l'hypothèse de l'ergodicité de Φ , la valeur moyenne de $\prod_{i=0}^{n} p_i$ est environ égale à $\exp(nE(log(P))$; où $E(\log(P))$ est l'espérance par rapport à $\mu = g\lambda$ de la fonction $t \to \log(P(t))$. Or on a $\exp(E(log(P)) = 12$ environ. Par conséquent la précision apportée par le développement d'ordre n d'un nombre est en moyenne $1,5.\ 12^{-n}$. Inversement l'ordre moyen du développement permettant d'obtenir une approximation de x à la précision 10^{-n} est approximativement $n\frac{\log(10)}{\log(12)} = 0,92n$.

5.1.1 Description de l'algorithme. Afin d'économiser l'espace mémoire nous déterminons le développement d'un nombre irrationnel x par paquets successifs de longueur pratiquement constante. Pour cela nous nous appuyons sur le fait que si $S_n = (p_0 \dots p_n)$ est le développement d'ordre $n+1$ de x, le reste de son développement coïncide avec celui de la partie fractionnaire de $x \prod_{i=0}^{n} p_i$(voir théorème 2.13).

Voici les articulations de l'algorithme:

Une précision $\varepsilon = 10^{-n}$ étant fixée, si α_n est l'approximation décimale par défaut de x à ε près, on sait (voir proposition 2.14) que le développement S_k d'ordre k de α_n coïncide avec celui de x tant que $\alpha_n + \varepsilon$ appartient à $J(S_k)$. Si k_n est le maximum des k satisfaisant la propriété précédente, S_{k_n} constitue une première liste de longueur environ égale à $0,92n$. On reprend alors avec la partie fractionnaire de $x \prod_{i=0}^{k_n} p_i$ et ainsi de suite: si z est le produit des nombres premiers déja déterminés, le paquet suivant s'obtient par le développement de l'approximation à ε près de la partie fractionnaire de zx en appliquant la même clause d'arrêt. Ce procédé évite de manipuler de trop grands nombres; la part la plus importante du temps de calcul étant prise par la détermination des parties fractionnaires des produits zx.

5.1.2 Développement du nombre $\pi - 3$. Nous avons déterminé le développement d'ordre 142310 de π-3, ceci afin d'observer l'apparition d'un premier supérieur à 10^6. Cela se produit pour la première fois au rang 134396 avec le nombre 1508449. Cette attente est relativement longue car le rang moyen du premier superieur à 10^6 dans un développement générique est de l'ordre de 46000 environ (calculé à partir de $g_{600}^{(2)}$). Un autre fait surprenant est l'apparition au 66ième rang du nombre 58889; la probabilité de voir apparaître un si grand nombre dans les 66 premiers est inférieure à 2%. Voici le début du développement:

[11, 2, 11, 5, 5, 2, 5, 3, 17, 11, 3, 3, 11, 3, 3, 11, 5, 3, 23, 7, 5, 97, 29, 37, 107, 127, 29, 17, 409, 127, 11, 29, 5, 67, 19, 43, 31, 19, 103, 59, 29, 7, 3, 11, 11, 5, 47, 29, 11, 3, 5, 5, 3, 17, 5, 29, 11, 3, 3, 3, 3, 5, 5, 61, 151, 58889, 1877, 983, 757, 163, 103, 79, 17, 11, 2, 13]

Le nombre 58889 apparaît une seconde fois au 125706ième rang. Voici l'extrait du développement concerné:

[23, 11, 5, 5, 149, 53, 43, 27103, 2281, 58889, 23039, 3037, 347, 83, 23, 163, 37, 7, 79, 19, 17, 5, 5, 3, 17, 5, 149, 31, 37, 7, 5, 17, 37, 7, 17, 13, 83, 23, 19, 37, 7, 37, 127, 17, 5, 3, 3, 7, 5, 5, 3, 5, 2, 5, 5, 2]

Voici l'extrait du développement où apparaît 1508449:

[7, 7, 53, 29, 5, 7, 17, 7, 3, 7, 5, 37, 7, 23, 13, 23, 11, 11, 3, 3, 11, 7, 7, 7, 7, 7, 3, 29, 7, 5, 29, 1508449, 53407, 10037, 389, 67, 29, 11, 2, 43, 37, 13, 137, 61, 47, 89, 19, 17, 5, 7, 5, 17, 5, 5, 11, 2]

Voici rangée dans l'ordre croissant la liste (avec répétition) des plus grands nombres de ce développement:

[9767, 9803, 10037, 10037, 10061, 10061, 10061, 10169, 10177, 10211, 10331, 10391, 10391, 10427, 10589, 10601, 10739, 10831, 10903, 10937, 10957, 10973, 11069, 11239, 11251, 11423, 11587, 11617, 11777, 11801, 11863, 11863, 12211, 12227, 12473, 12497, 12637, 12653, 12757, 12809, 12821, 12823, 13033, 13249, 13553, 13613, 13649, 13669, 13721, 13751, 14057, 14281, 14369, 14771, 14879, 15013, 15031, 15101, 15187, 15307, 15391, 15493, 15641, 15887, 15901, 16127, 16127, 16267, 16301, 16349, 16519, 16691, 16811, 16921, 17377, 17597, 17597, 18119, 18251, 18301, 18583, 18593, 18911, 19009, 19069, 19211, 19469, 19489, 19661, 19961, 19961, 19973, 20849, 20849, 21169, 21191, 21247, 21401, 21599, 21751, 22247, 22277, 22291, 22303, 22637, 22669, 22807, 23039, 23909, 24197, 24953, 25031, 25219, 25793, 26107, 26141, 26399, 27059, 27103, 27611, 27647, 27743, 27917, 28027, 28387, 29017, 29363, 29641, 30071, 30241, 30403, 30427, 31601, 32297, 32687, 32779, 33317, 33487, 36493, 37619, 38011, 39133, 39503, 40087, 40387, 40529, 41351, 42101, 42443, 42487, 43189, 44867, 46171, 46261, 47857, 47911, 48731, 49843, 50207, 53407, 54139, 54941, 58477, 58889, 58889, 59207, 60493, 60821, 61297, 61487, 62927, 66553, 67103, 68729, 69481, 76667, 78607, 80447, 83389, 86923, 87337, 98867, 102139, 102407, 107071, 109919, 127781, 143827, 153871, 167861, 206233, 271357, 294179, 321467, 326219, 326737, 334379, 398569, 817123, 900539, 1508449]

5.1.3 Développement du nombre $e - 2$. Ce développement présente la particularité de produire

relativement tôt un nombre supérieur à 10^7, à savoir 30225161 au rang 11063.

Voici l'extrait du développement où ce nombre apparaît:

[5, 3, 17, 7, 5, 31, 19, 373, 2203, 131, 89, 17, 11, 2, 37, 11, 2, 157, 37, 11, 3, 3, 3, 7, 67, 163, 31, 331, 6547, 15373, 681259, 1540477, 30225161, 17403227, 636263, 1454347, 820399, 25679, 10831, 1381, 223, 37, 7, 5, 5, 53, 53, 787, 103, 149, 191, 29, 11, 17]

Voici rangée dans l'ordre croissant la liste des plus grand nombres du développement d'ordre 11972 de $e-2$:

[3613, 3907, 3989, 3989, 4027, 4211, 4507, 4861, 5009, 5113, 5209, 5227, 5417, 5897, 5903, 6547, 6947, 7297, 7369, 7649, 8147, 8609, 8693, 9151, 9257, 9463, 9533, 10831, 12889, 12889, 14243, 14537, 15217, 15373, 18127, 19949, 24631, 25537, 25679, 26879, 69467, 89041, 106781, 202049, 636263, 681259, 820399, 1454347, 1540477, 17403227, 30225161].

5.1.4 Développement de $\sqrt{2}-1$. Bien que l'on observe ici relativement tôt l'apparition d'un grand nombre (842321) au 18989ième rang, il faut attendre le 63626ième rang avec 3355487 pour dépasser 10^6. Et presque aussitôt après apparaît le nombre 24581083.

Voici l'extrait concerné:

[3, 5, 2, 29, 11, 3, 7, 37, 11, 2, 29, 17, 5, 3, 11, 149, 19, 17, 11, 2, 3355487, 240967, 31963, 39079, 41729, 4889, 2887, 1511, 211, 37, 127, 19, 37, 19, 23, 7, 7, 5, 2, 5, 5, 67, 239, 67, 83, 47, 17, 5, 29, 11, 5, 2, 13, 7, 7, 5, 5, 3, 3, 7, 5, 5, 113, 31, 23, 17, 5, 3, 3, 3, 3, 23, 7, 17, 23, 5, 37, 29, 5, 5, 7, 11, 3, 3, 11, 3, 7, 29, 31, 23, 17, 5, 3, 7, 5, 3, 11, 11, 5, 3, 59, 11, 11, 5, 2, 5, 5, 53, 37, 53, 11, 11, 11, 2, 11, 223, 53, 29, 37, 1187, 727, 2251, 331, 31, 29, 5, 5, 79, 23, 29, 11, 2, 11, 5, 2, 5, 2, 5, 7, 3229, 24581083, 1664987, 3730033, 432569, 67987, 15823, 64187, 14489, 5953, 2357, 439, 149, 17, 11, 19]

Voici les plus grands nombres:

[22783, 23399, 26203, 27031, 27581, 28283, 29683, 29717, 30241, 31963, 33247, 34313, 34403, 34457, 34667, 35831, 36373, 38933, 39079, 41729, 44159, 52067, 55619, 57587, 59651, 61331, 61751, 62141, 64187, 67073, 67987, 85009, 98711, 107941, 113497, 120811, 134581, 135277, 144289, 148639, 176041, 180797, 220973, 240967, 338803, 432569, 842321, 1664987, 3355487, 3730033, 24581083].

5.1.5 Développement de π^2-9. Le rang du premier supérieur à 10^6(2411821) est 12374.

Voici l'extrait concerné:

[5, 5, 3, 79, 17, 11, 31, 181, 233, 97, 311, 2411821, 1046447, 43987, 4943, 1171, 787, 79, 67, 11, 19, 47, 13, 23, 7, 5, 17, 5, 29, 5, 29, 11, 7, 3, 11, 5, 2, 5, 5, 5, 29, 5, 11, 7, 11, 5]

Voici les plus grands nombres:

[4567, 4643, 4751, 4861, 4943, 5381, 5387, 5897, 6079, 6101, 6143, 6373, 6521, 6761, 6803, 7103, 7207, 7219, 7321, 7789, 7817, 7853, 8243, 8951, 9767, 9941, 10289, 10477, 11299, 12197, 12889, 13421, 14591, 17939, 20287, 21157, 21893, 22769, 24989, 29059, 30089, 32507, 35933, 36479, 43987, 52769, 64373, 68729, 76913, 1046447, 2411821]

5.1.6 Développement du nombre $(\sqrt{5}-1)/2$. Les plus grands nombres du developpement d'ordre 103706:

[38011, 38197, 39839, 40627, 41047, 44371, 44579, 44617, 44983, 46021, 46091, 46261, 46663, 47591, 49801, 51001, 53087, 53993, 55229, 59729, 61001, 63719, 65921, 66683, 67939, 68863, 69767, 74071, 75431, 76819, 77417, 79133, 82531, 84481, 91453, 95737, 98443, 100447, 102101, 106163, 149993, 160309, 171271, 218599, 227453, 229979, 253381, 292021, 300277, 377021, 652903]

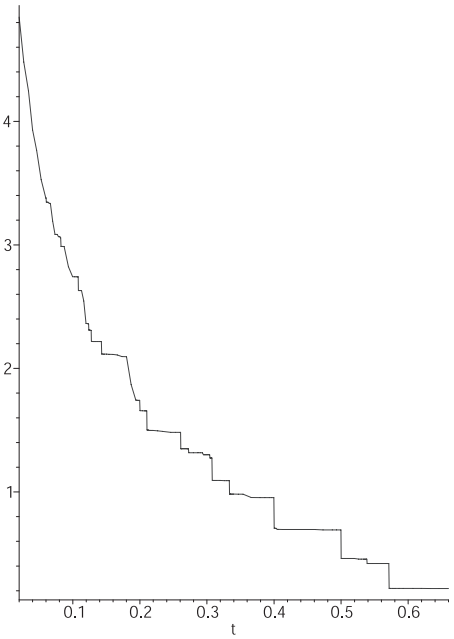

FIGURE 4. Fonction $g_{600}^{(2)}$ sur [0.02,1.5].

5.2 Calcul de la densité stationnaire et des coefficients qui en dépendent

Commençons par apporter quelques précisions concernant la programmation. Les fonctions intervenant dans

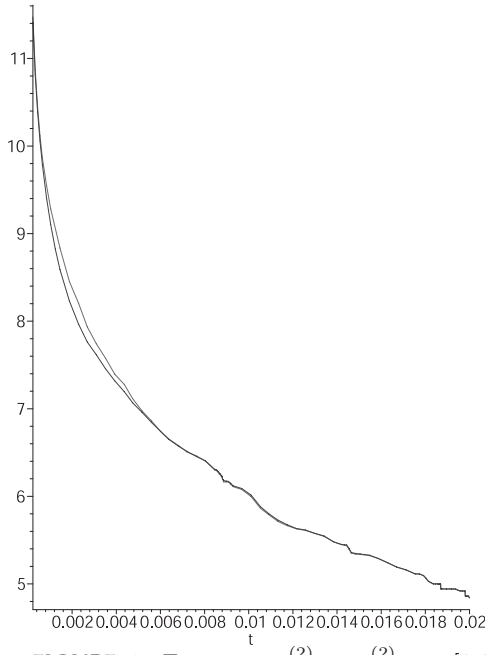

FIGURE 5. Fonctions $g_{600}^{(2)}$ et $g_{200}^{(2)}$ sur [0.0002,0.02] (sur [0.02,1.5] on observe une presque parfaite superposition des deux graphes).

l'expression de l'opérateur Γ_N^* (voir §4) sont rentrées dans les programmes sous forme de spline linéaire.

Tout se passe alors avec des fonctions polynomiales par morceaux. Pour éviter les saturations il est nécessaire d'introduire des procédures de calcul explicite de la liste des noeuds des fonctions composées.

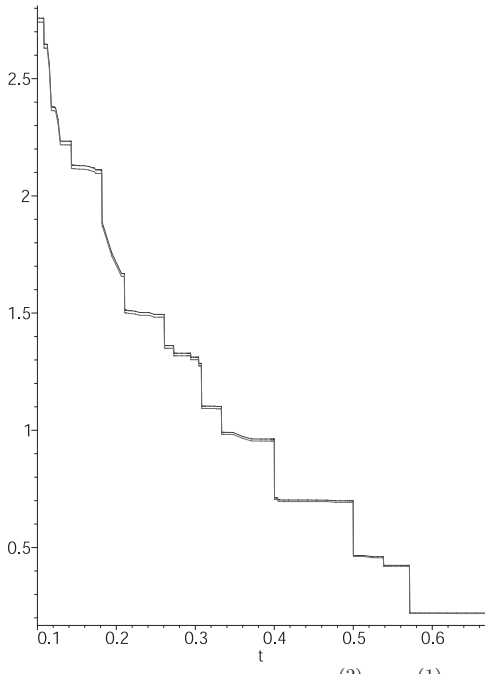

FIGURE 6. Fonctions $g_{600}^{(2)}$ et $g_{600}^{(1)}$ sur [0.1,1.5].

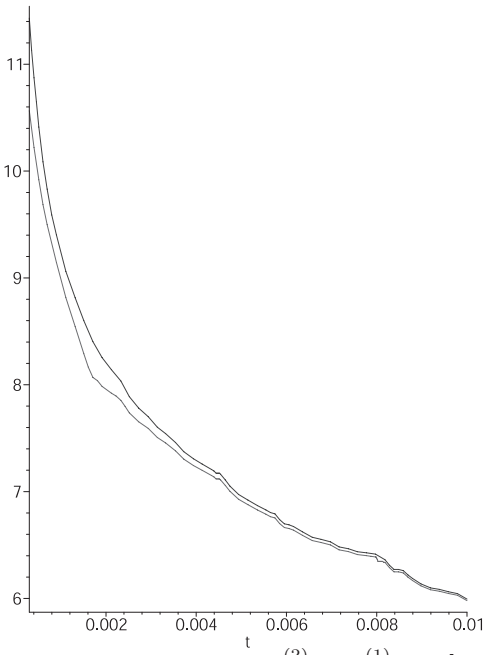

FIGURE 7. Fonctions $g_{600}^{(2)}$ et $g_{600}^{(1)}$ sur [0.0003,0.01]. On voit l'effet régularisant de la seconde méthode. La jonction entre les parties étagées et continues des approximations de g (formules (4-3) et (4-9)) se fait de manière plus progressive pour $g_{600}^{(2)}$.

D'autre part afin d'alléger les calculs, à chaque itération de l'opérateur, seules les N composantes d'indices $(p_n, 1)$ du produit matriciel $\mathcal{A}_N V$ sont effectivement déterminées, les autres composantes sont complétées en accord avec les relations du §3.3.

5.3 Comparaison du graphe de la densité stationnaire avec les histogammes des orbites issues de certains nombres

A l'occasion du calcul du développement des nombres $\pi - 3, \sqrt{2} - 1, (\sqrt{5} - 1)/2$ nous avons déterminé la liste des moyennes temporelles de $\Phi^n(x)$ pour les intervalles $[k/1000, (k+1)/1000]$ $(0 < k < 20)$, $[k/100, (k+1)/100]$ $(1 < k < 66)$ et $[66/100, 2/3]$. La concordance observée entre les histogrammes construits à partir de ces listes et le graphe de $g_{600}^{(2)}$ montre l'efficacité de la seconde méthode d'approximation de g et confirme l'hypothèse d'ergodicité de Φ.

5.4 Paramètres associés à g

Liste des dix premiers $\alpha_{p,1}$.
calculée à partir de $g_{200}^{(2)}$:

[2, 0.0364] [3, 0.11562] [5, 0.14578] [7, 0.09960] [11, 0.11531] [13, 0.04095] [17 0.05587] [19, 0.021406] [23, 0.03258] [29, 0.03453]

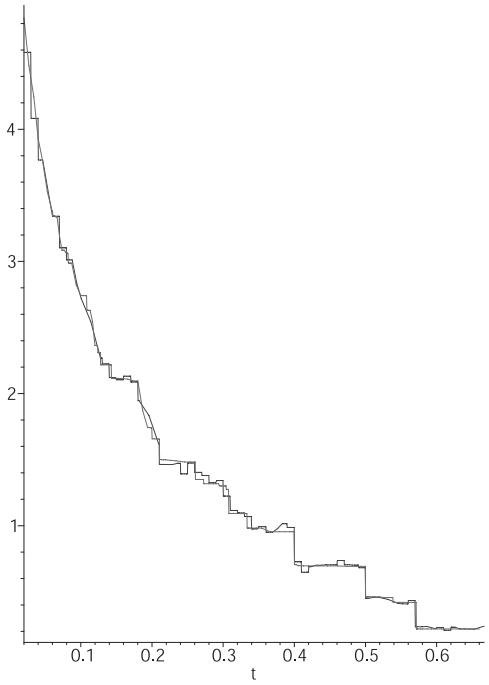

FIGURE 8. Histogramme de l'orbite de $\pi - 3$ et $g_{600}^{(2)}$ sur $[0.02, 1.5]$.

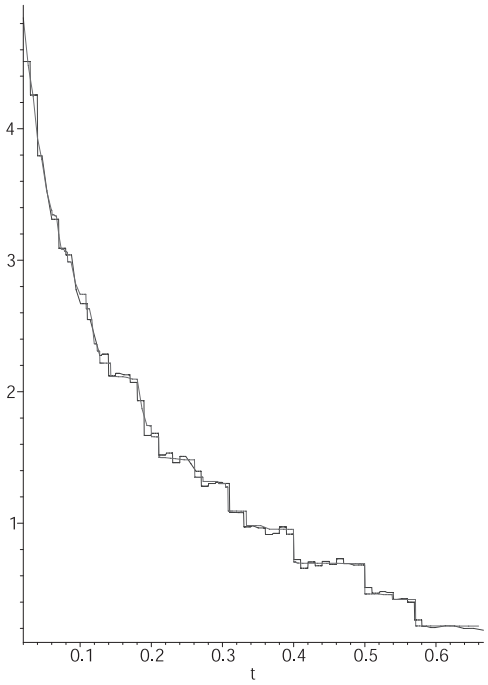

FIGURE 10. Histogramme de l'orbite de $(\sqrt{5} - 1)/2$ et $g_{600}^{(2)}$ sur $[0.02, 1.5]$.

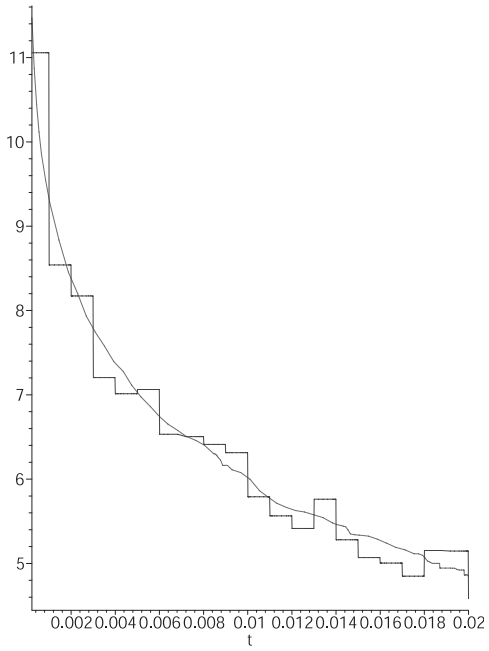

FIGURE 9. Histogramme de l'orbite de $\pi - 3$ et $g_{600}^{(2)}$ sur $[0, 0.02]$.

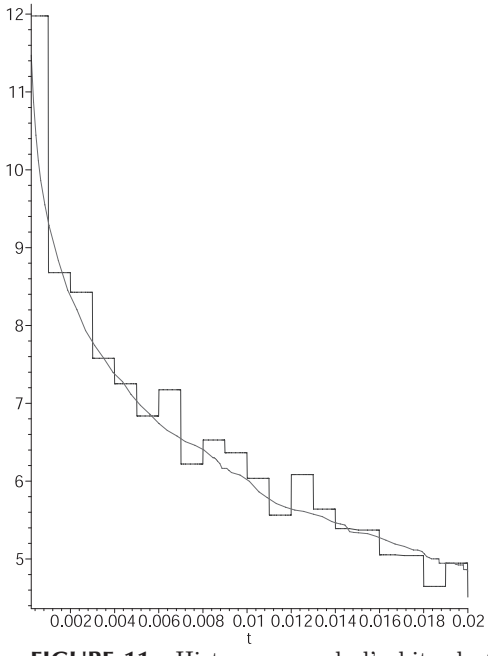

FIGURE 11. Histogramme de l'orbite de $(\sqrt{5} - 1)/2$ et $g_{600}^{(2)}$ sur $[0, 0.02]$.

calculée à partir de $g_{600}^{(2)}$:

[2, 0.03638] [3, 0.11544] [5, 0.14556] [7, 0.09948] [11, 0.11518] [13, 0.04092] [17, 0.05582] [19, 0.02139] [23, 0.03256] [29, 0.03452]

Liste des dix premiers coefficients de la loi de la variable P (c'est à dire des probabilités d'apparition des dix premiers nombres premiers)
calculée à partir de $g_{200}^{(2)}$:

[2, 0.0524] [3, 0.1340] [5, 0.1824] [7, 0.1154] [11, 0.1299] [13, 0.0421] [17, 0.0589] [19, 0.0216] [23, 0.0339] [29, 0.0360]

calculée à partir de $g_{600}^{(2)}$:

[2, 0.0523] [3, 0.1338] [5, 0.1822] [7, 0.1152] [11, 0.1298] [13, 0.0421] [17, 0.0588] [19, 0.0216] [23, 0.0339] [29, 0.0359]

Limite en 0 de g(t)/ log(1/t):
avec $g_{200}^{(2)}$: $1,19$, avec $g_{600}^{(2)}$: $1,28$

Liste pour n de 5 à 9 des rangs moyens des premières apparitions d'un nombre supérieur à 10^n
calculée à partir de $g_{200}^{(2)}$:

[5645, 47025, 402673, 3489183, 28490028]

calculée à partir de $g_{600}^{(2)}$:

[5610, 46622, 398311, 3456619, 29239766]

REMERCIEMENTS

Je remercie Christian Ballot de m'avoir suggéré d'introduire la notion d'indice de bifurcation afin de clarifier l'exposé.

BIBLIOGRAPHIE

[Cramér 36] H. Cramér. "On the order of magnitude of the difference between consecutive primes." *Acta. Arith.* **2** (1936), 396–403.

[Ellison et Mendès France 75] W. J. Ellison et M. Mendès France. *Les nombres premiers* Hermann, Paris, 1975.

[Feller 68] W. Feller. *An introduction to Probability Theory and Its Applications,* Vol. I, John Wiley and Sons, New York, 1968.

[Lasota and Mackey 94] A. Lasota and M. C. Mackey. *Chaos, Fractals, and Noise,* Springer, Berlin-Heidelberg-New York, 1994.

[Schweiger 95] F. Schweiger.*Ergodic theory of fibered systems and metric number theory,* Oxford Press, London, 1995.

[Ulam 60] S. M. Ulam. *A collection of Mathematical problems*, Interscience Tracts in Pure and Applied Math., vol. 8, Interscience, New York, 1960.

Alain Costé, LMNO, B.P. 5186, Université campus 2, 14032 Caen cedex, France (Coste@math.unicaen.fr)

Received April 30, 2001; accepted in revised form October 10, 2001.

Hecke Eigenvalues for Real Quadratic Fields

Kaoru Okada

CONTENTS

We describe an algorithm to compute the trace of Hecke operators acting on the space of Hilbert cusp forms defined relative to a real quadratic field with class number greater than one. Using this algorithm, we obtain numerical data for eigenvalues and characteristic polynomials of the Hecke operators. Within the limit of our computation, the conductors of the orders spanned by the Hecke eigenvalue for any principal split prime ideal contain a nontrivial common factor, which is equal to a Hecke L-value.

1. INTRODUCTION

Let F be a totally real algebraic number field with nontrivial class group. We shall study the space $\mathcal{S}_k(\mathfrak{c}, \psi)$ of Hilbert cusp forms (relative to F) and the Hecke operators $T(\mathfrak{a})$ acting on it. We shall describe our result using the framework first introduced in [Shimura 78]. Following Shimura's work, the trace formula (whose origin goes back to fundamental work of Eichler, Selberg, and Shimizu) was made more explicit in [Saito 84]. Saito's formula gives us a method for computing Hecke eigenvalues as long as the dimension of the space remains reasonably small. It is then natural to expect Hecke eigenvalues for prime ideals $\mathfrak{p}$ in a given ideal class to have a new feature specific to the ideal class. Such a new feature can be detected only by computing Hecke eigenvalues for the base field with nonprincipal ideal classes. The purpose of this paper is to compute examples of such Hecke eigenvalues for real quadratic fields with class number greater than 1 and to present a new phenomenon that we have discovered through our numerical examples.

We summarize our observations for the data of Hecke eigenvalues when the weight is parallel (k_1, k_1), the level $\mathfrak{c}$ is the maximal order $\mathfrak{o}_F$ of F, and the Hecke character ψ is the identity *1*. Let $\mathbf{f}$ be a primitive form contained in $\mathcal{S}_{(k_1,k_1)}(\mathfrak{o}_F, 1)$ that is orthogonal to any base change lift from $\mathbf{Q}$ (that is, $\mathbf{f}$ is a primitive form in the "F-proper" subspace of $\mathcal{S}_{(k_1,k_1)}(\mathfrak{o}_F, 1)$ as defined in [Doi et al. 98]). We denote by $C_{\mathbf{f}}(\mathfrak{p})$ the eigenvalue of $T(\mathfrak{p})$ satisfying $\mathbf{f}|T(\mathfrak{p}) = C_{\mathbf{f}}(\mathfrak{p})\mathbf{f}$, by $K_{\mathbf{f}}^{+}$ the subfield of the

2000 AMS Subject Classification: Primary 11F41; Secondary 11F60, 11F72, 11R42

Keywords: Hilbert cusp form, Hecke operator, eigenvalue, trace formula L-value

1058-6458/2001 $0.50 per page

Hecke field $K_{\mathbf{f}}$ of $\mathbf{f}$ generated by $C_{\mathbf{f}}((p)_F)$ for all *rational* primes p, and by $\mathfrak{o}_{K_{\mathbf{f}}^+}$ the maximal order of $K_{\mathbf{f}}^+$. For split prime ideals $\mathfrak{p}$, we computed the discriminant of the order $\Lambda_{\mathbf{f}}(\mathfrak{p})$ spanned by the eigenvalue of $T(\mathfrak{p})$ (that is, $\Lambda_{\mathbf{f}}(\mathfrak{p}) = \mathfrak{o}_{K_{\mathbf{f}}^+} + C_{\mathbf{f}}(\mathfrak{p})N(\mathfrak{p})^{(k_1-2)/2}\mathfrak{o}_{K_{\mathbf{f}}^+}$) to see whether it has extra factors outside the discriminant of the maximal order. Extra factors show up as the conductor of the order (for the definition of the conductor, see just above Lemma 2.3); so, we write $c(\Lambda_{\mathbf{f}}(\mathfrak{p}))$ for the conductor. Surprisingly enough, *as long as the prime ideals* $\mathfrak{p}$ *are principal and split, the conductors* $c(\Lambda_{\mathbf{f}}(\mathfrak{p}))$ *contain a nontrivial common factor* $\mathfrak{F}_{\mathbf{f}}$*, at least within the limit of our computations.* (see Sections 4.1 and 4.2).

In Section 2, we recall the space of Hilbert cusp forms for totally real number fields and Hecke operators. We then reformulate Saito's formula into a more computable one. The notion of the conductor of an order plays an important role in this process. In Section 3, we give an algorithm to compute the trace of Hecke operators for a real quadratic field F. Key points of the computation are the determination of the relative discriminant $D_{K/F}$, the character $\left(\frac{K}{\mathfrak{p}}\right)$, and the Hecke L-value $L_F(0, \chi_{K/F})$ for any totally imaginary quadratic extension K over F. In particular, the computation of the Hecke L-value by Shintani's method [Shintani 76] has been reduced to that of Hilbert symbols (cf. [Okazaki 91]). In Section 4, we give examples of eigenvalues and characteristic polynomials of Hecke operators restricted to the case where the weight is $(2,2)$ and F is $\mathbf{Q}(\sqrt{257})$ or $\mathbf{Q}(\sqrt{401})$, and we describe our analysis of the data to convince the reader of the conclusion we have already described.

While I was preparing the revision of this paper following the request of the referee to provide a more detailed study of $\mathfrak{F}_{\mathbf{f}}$, Professor Haruzo Hida provided the following crucial suggestion:

(1) Within the limits of the computations carried out, check that $\mathfrak{F}_{\mathbf{f}}\mathfrak{o}_{K_{\mathbf{f}}}$ is divisible by the common factor of $1 + N(\mathfrak{p}) - C_{\mathbf{f}}(\mathfrak{p})$ for the principal primes $\mathfrak{p}$.

(2) As is well known, several outstanding mathematicians have worked out the congruence primes between a primitive cusp form and an Eisenstein series, which are essentially given by the value at the weight of a Hecke L-function of the base field. Notably, A. Wiles studied in depth such an Eisenstein congruence, which is a key step in his proof of the Iwasawa conjecture for totally real fields. Therefore, if (1) is affirmative, his result presumably implies that $\mathfrak{F}_{\mathbf{f}}$ is divisible by the congruence primes. Here the congruence prime can be found in the prime factors of the numerator of the algebraic part of the Hecke L-value $L_F(2, \chi)$ associated with a nontrivial class character χ .

(3) Moreover, it is expected that the set of the primes of $\mathfrak{F}_{\mathbf{f}}$ *coincides with* that of the congruence primes between the F-proper cusp form $\mathbf{f}$ and the Eisenstein series of weight (k_1, k_1) with Mellin transform $L_F(s, \chi)L_F(s + 1 - k_1, \chi^{-1})$ for a nontrivial class character χ.

We shall give affirmative numerical evidence for (1) and (2) in Section 4.3. As for (3), we hope to discuss this property in a subsequent paper.

Notation

For an associative ring R with identity element, we denote by $R^\times$ the group of invertible elements of R. We write $M_2(R)$ for the ring of 2×2 matrices over R, and 1_2 for the identity element of $M_2(R)$.

For an algebraic number field F of finite degree, we denote by $\mathfrak{o}_F$, $\mathfrak{d}_F$, and D_F the maximal order of F, the different of F over $\mathbf{Q}$, and the discriminant of F over $\mathbf{Q}$. We write $I(F)$ for the ideal group of F, and $P(F)$, Cl(F), and h_F (respectively $P^+(F)$, Cl^+(F), and h_F^+) for the principal ideal group of F, the ideal class group of F, and the class number of F (respectively those in the narrow sense). For $\alpha \in F^\times$, we put $(\alpha)_F = \alpha\mathfrak{o}_F$. For a prime ideal $\mathfrak{p}$ of F and $\mathfrak{m} \in I(F)$, we denote by $\mathrm{ord}_{\mathfrak{p}}(\mathfrak{m})$ the order of $\mathfrak{m}$ at $\mathfrak{p}$. For $\alpha \in F$, we set $\alpha \gg 0$ if α is totally positive. We define $\mathfrak{o}_{F+}^\times = \{a \in \mathfrak{o}_F^\times \mid a \gg 0\}$. For integral ideals $\mathfrak{a}, \mathfrak{b}$ of F, we write $\mathfrak{a} \mid \mathfrak{b}$ if $\mathfrak{b}\mathfrak{a}^{-1} \subset \mathfrak{o}_F$; for elements $\alpha(\neq 0), \beta$ of $\mathfrak{o}_F$, we write $\alpha \mid \beta$ if $\beta\alpha^{-1} \in \mathfrak{o}_F$. For $\alpha_1, \ldots, \alpha_r \in F$, we write $[\alpha_1, \ldots, \alpha_r]$ for the $\mathbf{Z}$-submodule of F generated by $\alpha_1, \ldots, \alpha_r$. We denote by ζ_F the Dedekind zeta function of F.

For an extension K of F of finite degree, we denote by $D_{K/F}$ the relative discriminant of K over F. For an element α of K, we denote by $D_{K/F}(\alpha)$, $N_{K/F}(\alpha)$, and $\mathrm{Tr}_{K/F}(\alpha)$ the relative discriminant, the norm, and the trace of α in K over F. We denote by $N(\mathfrak{a})$ the norm of an ideal $\mathfrak{a}$ of F. (We also use the symbols $N_{K/F}(\alpha)$, $\mathrm{Tr}_{K/F}(\alpha)$, and $N(\mathfrak{a})$ when K and F are local fields.)

For $a \in \mathbf{R}$, we denote by $[a]$ the greatest integer not greater than a. Let $\left(\frac{a}{p}\right)$ be the Legendre symbol for $a \in \mathbf{Z}$ and a prime number p. For a set X, we denote the cardinality of X by $|X|$ and also by $\sharp X$. For a subgroup H of a group G, we write $[G : H] = |G/H|$. For a subfield F of a field K, the symbol $[K : F]$ means the degree of K

over F. The disjoint union of sets $Y_1, \dots, Y_s$ is denoted by $\bigsqcup_{i=1}^{s} Y_i$.

2. THE TRACE FORMULA FOR TOTALLY REAL NUMBER FIELDS

In this section, we first recall the definition of Hecke operators acting on the space of Hilbert cusp forms as given in [Shimura 78, §2]. (Cf. also [Shimura 91].)

2.1 Hilbert Cusp Forms and Hecke Operators

Let F be a totally real algebraic number field of degree g, and denote by $\mathbf{a}$ and $\mathbf{h}$ the sets of archimedean primes and nonarchimedean primes of F. For $\mathfrak{p} \in \mathbf{h}$, we also denote by $\mathfrak{p}$ the corresponding prime ideal of F. For any set X, we write $X^{\mathbf{a}}$ for the set of all indexed elements $(x_v)_{v\in\mathbf{a}}$ with $x_v \in X$. Let $F_{\mathbf{A}}$ be the ring of adeles of F, and $F_{\mathbf{A}}^{\times}$ the group of ideles of F. For $v \in \mathbf{a}\cup\mathbf{h}$ and $x \in F_{\mathbf{A}}$, let F_v be the v-completion of F, and x_v its v-component. We write $F_{\mathbf{a}}$ and $F_{\mathbf{h}}$ for the archimedean and nonarchimedean factors of $F_{\mathbf{A}}$, and identify $F_{\mathbf{a}}$ with $\mathbf{R}^{\mathbf{a}}$. For $\mathfrak{a} \in I(F)$ and $\mathfrak{p} \in \mathbf{h}$, we denote by $\mathfrak{a}_{\mathfrak{p}}$ the topological closure of $\mathfrak{a}$ in $F_{\mathfrak{p}}$. We abbreviate $(\mathfrak{o}_F)_{\mathfrak{p}}$ and $(\mathfrak{d}_F)_{\mathfrak{p}}$ by $\mathfrak{o}_{\mathfrak{p}}$ and $\mathfrak{d}_{\mathfrak{p}}$, for short. We then set $\mathfrak{o}_{\mathbf{h}} = \prod_{\mathfrak{p}\in\mathbf{h}} \mathfrak{o}_{\mathfrak{p}}$ and $\mathfrak{d}_{\mathbf{h}} = \prod_{\mathfrak{p}\in\mathbf{h}} \mathfrak{d}_{\mathfrak{p}}$. For $a \in F_{\mathbf{A}}^{\times}$, we denote by $a\mathfrak{o}_F$ the fractional ideal of F such that $(a\mathfrak{o}_F)_{\mathfrak{p}} = a_{\mathfrak{p}}\mathfrak{o}_{\mathfrak{p}}$ for every $\mathfrak{p} \in \mathbf{h}$ (i.e., $a\mathfrak{o}_F = F \cap F_{\mathbf{a}} \prod_{\mathfrak{p}\in\mathbf{h}} a_{\mathfrak{p}}\mathfrak{o}_{\mathfrak{p}}$). For $a \in F_{\mathbf{A}}^{\times}$, we set $\mathrm{ord}_{\mathfrak{p}}(a) = \mathrm{ord}_{\mathfrak{p}}(a\mathfrak{o}_F)$. We denote by $\pi_{\mathfrak{p}}$ a prime element of $F_{\mathfrak{p}}$. By a *Hecke character* of F, we understand a character of $F_{\mathbf{A}}^{\times}$ with values in $\mathbf{T} = \{z \in \mathbf{C} \mid |z| = 1\}$ that is trivial on $F^{\times}$.

Let $G = \mathrm{GL}_2(F)$. We set $G_v = \mathrm{GL}_2(F_v)$ for every $v \in \mathbf{a}\cup\mathbf{h}$. We consider the adelization $G_{\mathbf{A}}$ of G, and denote by $G_{\mathbf{a}}$ and $G_{\mathbf{h}}$ its archimedean and nonarchimedean factors. We set $G_{\mathbf{a}+} = \{x \in G_{\mathbf{a}} \mid \det(x) \gg 0\}$ and $G_+ = G \cap G_{\mathbf{a}+}G_{\mathbf{h}}$. For an element x of $G_{\mathbf{A}}$, we denote by $x_{\mathbf{a}}$ its $\mathbf{a}$-component. For $x \in G_{\mathbf{A}}$, we set $x^{\iota} = \det(x)x^{-1}$ and $x^{-\iota} = (x^{\iota})^{-1}$. We take an element $\delta_{\mathbf{h}}$ of $F_{\mathbf{h}}$ such that $\delta_{\mathbf{h}}\mathfrak{o}_{\mathbf{h}} = \mathfrak{d}_{\mathbf{h}}$, define subsets $Y_{\mathbf{h}}$ and $W_{\mathbf{h}}$ of $G_{\mathbf{h}}$ by

$$Y_{\mathbf{h}} = \begin{pmatrix} 1 & 0 \\ 0 & \delta_{\mathbf{h}} \end{pmatrix} M_2(\mathfrak{o}_{\mathbf{h}}) \begin{pmatrix} 1 & 0 \\ 0 & \delta_{\mathbf{h}} \end{pmatrix}^{-1} \cap G_{\mathbf{h}},$$

$$W_{\mathbf{h}} = \begin{pmatrix} 1 & 0 \\ 0 & \delta_{\mathbf{h}} \end{pmatrix} GL_2(\mathfrak{o}_{\mathbf{h}}) \begin{pmatrix} 1 & 0 \\ 0 & \delta_{\mathbf{h}} \end{pmatrix}^{-1},$$

and set

$$Y = G_{\mathbf{a}+}Y_{\mathbf{h}}, \qquad W = G_{\mathbf{a}+}W_{\mathbf{h}}.$$

We denote by H the complex upper half-plane. For $\alpha = (\alpha_v)_{v\in\mathbf{a}} = \left(\left(\begin{smallmatrix} a_v & b_v \\ c_v & d_v \end{smallmatrix}\right)\right)_{v\in\mathbf{a}} \in G_{\mathbf{a}+}$, $z = (z_v)_{v\in\mathbf{a}} \in H^{\mathbf{a}}$, $k = (k_v)_{v\in\mathbf{a}} \in \mathbf{Z}^{\mathbf{a}}$, and a $\mathbf{C}$-valued function f on $H^{\mathbf{a}}$, we set

$$\alpha(z) = \big((a_v z_v + b_v)/(c_v z_v + d_v)\big)_{v\in\mathbf{a}},$$

$$J_k(\alpha, z) = \prod_{v\in\mathbf{a}} \big(\det(\alpha_v)^{-k_v/2}(c_v z_v + d_v)^{k_v}\big),$$

$$(f\|_k\alpha)(z) = J_k(\alpha, z)^{-1} f(\alpha(z)),$$

and denote by $\tilde{\mathcal{S}}_k$ the space of all holomorphic functions f on $H^{\mathbf{a}}$ satisfying the following two conditions:

($\mathrm{i_a}$) There exists $0 < N \in \mathbf{Z}$ such that $f\|_k\gamma = f$ for all $\gamma \in SL_2(\mathfrak{o}_F) \cap (1_2 + N \cdot M_2(\mathfrak{o}_F))$.

($\mathrm{i_b}$) For every $\alpha \in G_+$, one has

$$(f\|_k\alpha)(z) = \sum_{0\ll\xi\in L_\alpha} c_\alpha(\xi) e_F(\xi z)$$

with a lattice L_α of F and $c_\alpha(\xi) \in \mathbf{C}$, where $e_F(\xi z) = \exp(2\pi\sqrt{-1}\sum_{v\in\mathbf{a}} \xi_v z_v)$.

Let ψ be a Hecke character of F of finite order such that *the nonarchimedean part of its conductor is equal to* $\mathfrak{o}_F$ (i.e. $\psi(\mathfrak{o}_{\mathbf{h}}^{\times}) = \{1\}$). We denote by $\mathcal{S}_k(\mathfrak{o}_F, \psi)$ the space of all $\mathbf{C}$-valued functions $\mathbf{f}$ on $G_{\mathbf{A}}$ satisfying the following two conditions:

($\mathrm{ii_a}$) $\mathbf{f}(s\alpha x w) = \psi(s)\mathbf{f}(x)$ for $s \in F_{\mathbf{A}}^{\times}$, $\alpha \in G$, and $w \in W_{\mathbf{h}}$ $(x \in G_{\mathbf{A}})$.

($\mathrm{ii_b}$) For every $x \in G_{\mathbf{h}}$, there exists an element f_x of $\tilde{\mathcal{S}}_k$ such that $\mathbf{f}(x^{-\iota}u) = (f_x\|_k u)(\mathbf{i})$ for all $u \in G_{\mathbf{a}+}$, where $\mathbf{i} = (\sqrt{-1}, \dots, \sqrt{-1}) \in H^{\mathbf{a}}$.

The elements of $\mathcal{S}_k(\mathfrak{o}_F, \psi)$ are called (adelic) *Hilbert cusp forms of weight k, level $\mathfrak{o}_F$, and character ψ*. We note that if $\mathcal{S}_k(\mathfrak{o}_F, \psi) \neq \{0\}$, then $\psi_v(-1) = (-1)^{k_v}$ for all $v \in \mathbf{a}$; moreover, $k_v > 0$ for all $v \in \mathbf{a}$ (cf. [Shimura 78, Proposition 1.1]).

Let $R_{\mathbf{C}}(W, Y)$ be the free $\mathbf{C}$-module generated by the double cosets $W\backslash Y/W$. For $WyW, WzW, WwW \in W\backslash Y/W$, we take coset decompositions $WyW = \bigsqcup_{i=1}^{m} Wy_i$ and $WzW = \bigsqcup_{j=1}^{n} Wz_j$, and set

$$m(WyW, WzW; WwW) = \sharp\{(i, j) \mid Wy_iz_j = Ww\}.$$

We then define the product $(WyW)(WzW)$ by

$$(WyW)(WzW) = \sum_{WwW\in W\backslash Y/W} m(WyW, WzW; WwW)WwW.$$

Note that the above sum is finite. We extend this product $\mathbf{C}$-linearly on $R_{\mathbf{C}}(W, Y)$. Then $R_{\mathbf{C}}(W, Y)$ becomes a $\mathbf{C}$-algebra, which is called the *Hecke algebra for W and Y*.

For every $y \in Y$, we may assume that $WyW = \bigsqcup_{i=1}^{m} Wy_i$ and $(y_i)_{\mathbf{a}} = 1$ $(\in G_{\mathbf{a}})$. For $\mathbf{f} \in \mathcal{S}_k(\mathfrak{o}_F, \psi)$, we define a function $\mathbf{f}|WyW$ on $G_{\mathbf{A}}$ by

$$(\mathbf{f}|WyW)(x) = \sum_{i=1}^{m} \mathbf{f}(xy_i^{\iota}) \quad (x \in G_{\mathbf{A}}).$$

Then, for $s \in F_{\mathbf{A}}^{\times}$, $\alpha \in G$, $w \in W_{\mathbf{h}}$, and $x \in G_{\mathbf{A}}$, we have $(\mathbf{f}|WyW)(s\alpha xw) = \psi(s)\sum_{i=1}^{m} \mathbf{f}(x(y_i w^{\iota})^{\iota}) = \psi(s)(\mathbf{f}|WyW)(x)$; moreover, for $x \in G_{\mathbf{h}}$, we have

$$\begin{aligned}(\mathbf{f}|WyW)(x^{-\iota}u) &= \sum_{i=1}^{m} \mathbf{f}((xy_i^{-1})^{-\iota}u) \\ &= (\sum_{i=1}^{m} (f_{xy_i^{-1}})\|_k u)(\mathbf{i})\end{aligned}$$

for all $u \in G_{\mathbf{a}+}$, where $f_{xy_i^{-1}}$ is as in $(\mathrm{ii}_{\mathrm{b}})$. Thus $\mathbf{f}|WyW \in \mathcal{S}_k(\mathfrak{o}_F, \psi)$. Extending this action $\mathbf{C}$-linearly to the whole of $R_{\mathbf{C}}(W, Y)$, we have a ring homomorphism ϕ of $R_{\mathbf{C}}(W, Y)$ into the $\mathbf{C}$-linear endomorphism algebra $\mathrm{End}_{\mathbf{C}}\big(\mathcal{S}_k(\mathfrak{o}_F, \psi)\big)$. We call an element of $\phi(R_{\mathbf{C}}(W, Y))$ a *Hecke operator*.

We now determine the generators of $R_{\mathbf{C}}(W, Y)$. For each integral ideal $\mathfrak{a}$ of F, we define elements $T(\mathfrak{a})$ and $S(\mathfrak{a})$ of $R_{\mathbf{C}}(W, Y)$ by

$$T(\mathfrak{a}) = \sum_{\substack{WyW \in W\backslash Y/W \\ \det(y)\mathfrak{o}_F = \mathfrak{a}}} WyW, \qquad S(\mathfrak{a}) = W\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} W,$$

where $a = (\pi_{\mathfrak{p}}^{\mathrm{ord}_{\mathfrak{p}}(\mathfrak{a})})_{\mathfrak{p} \in \mathbf{h}} \in F_{\mathbf{h}}^{\times}$ $(\subset F_{\mathbf{A}}^{\times})$. Now we set $T(\pi_{\mathfrak{p}}^{l}, \pi_{\mathfrak{p}}^{l'}) = W\left(\begin{smallmatrix} \pi_{\mathfrak{p}}^{l} & 0 \\ 0 & \pi_{\mathfrak{p}}^{l'} \end{smallmatrix}\right) W$ for $l, l' \in \mathbf{Z}$. (Note that $\pi_{\mathfrak{p}} \in F_{\mathfrak{p}}^{\times}$ $(\subset F_{\mathbf{A}}^{\times})$.) For $y, z \in Y$, we have

$$(WyW)(WzW) = WyzW \quad \text{if } \gcd(\det(y)\mathfrak{o}_F, \det(z)\mathfrak{o}_F) = \mathfrak{o}_F. \quad (2\text{–}1)$$

Thus we have

$$T(\mathfrak{a}) = \prod_{\mathfrak{p}|\mathfrak{a}} \Big(\sum_{l_{\mathfrak{p}}=0}^{[\mathrm{ord}_{\mathfrak{p}}(\mathfrak{a})/2]} T(\pi_{\mathfrak{p}}^{l_{\mathfrak{p}}}, \pi_{\mathfrak{p}}^{\mathrm{ord}_{\mathfrak{p}}(\mathfrak{a})-l_{\mathfrak{p}}}) \Big),$$

and hence

$$T(\mathfrak{a}\mathfrak{b}) = T(\mathfrak{a})T(\mathfrak{b}) \quad \text{if } \gcd(\mathfrak{a}, \mathfrak{b}) = \mathfrak{o}_F. \qquad (2\text{–}2)$$

For any integer $e \geq 0$, we have

$$T(1, \pi_{\mathfrak{p}}^{e}) = \bigsqcup_{f=0}^{e} \bigsqcup_{\substack{1 \leq j \leq N(\mathfrak{p}^f) \\ \gcd(m_{fj}, \pi_{\mathfrak{p}}^{f}, \pi_{\mathfrak{p}}^{e-f}) = 1}} W\begin{pmatrix} \pi_{\mathfrak{p}}^{e-f} & m_{fj}\delta_{\mathfrak{p}}^{-1} \\ 0 & \pi_{\mathfrak{p}}^{f} \end{pmatrix},$$

where $\{m_{fj}\}_{j=1}^{N(\mathfrak{p}^f)}$ is a complete set of representatives of $\mathfrak{o}_{\mathfrak{p}}/\pi_{\mathfrak{p}}^{f}\mathfrak{o}_{\mathfrak{p}}$. Moreover, for $l, m, n \geq 0$, we have

$$T(\pi_{\mathfrak{p}}^{l}, \pi_{\mathfrak{p}}^{l})T(\pi_{\mathfrak{p}}^{m}, \pi_{\mathfrak{p}}^{n}) = T(\pi_{\mathfrak{p}}^{l+m}, \pi_{\mathfrak{p}}^{l+n}). \qquad (2\text{–}3)$$

Thus

$$\begin{aligned}&T(1, \pi_{\mathfrak{p}})T(1, \pi_{\mathfrak{p}}^{e}) \\ &\quad = \begin{cases} T(1, \pi_{\mathfrak{p}}^{e+1}) + N(\mathfrak{p})T(\pi_{\mathfrak{p}}, \pi_{\mathfrak{p}})T(1, \pi_{\mathfrak{p}})^{e-1} & \text{if } e \geq 2, \\ T(1, \pi_{\mathfrak{p}}^{2}) + (N(\mathfrak{p}) + 1)T(\pi_{\mathfrak{p}}, \pi_{\mathfrak{p}}) & \text{if } e = 1. \end{cases}\end{aligned} \qquad (2\text{–}4)$$

Therefore, we have

$$\begin{aligned}T(\mathfrak{p})T(\mathfrak{p}^e) &= T(\mathfrak{p}^{e+1}) \\ &\quad + N(\mathfrak{p})S(\mathfrak{p})T(\mathfrak{p}^{e-1}) \quad \text{for } \mathfrak{p} \in \mathbf{h} \text{ and } e \geq 1.\end{aligned} \qquad (2\text{–}5)$$

From (2–1), (2–3), and (2–4), we see that $R_{\mathbf{C}}(W, Y)$ is the commutative $\mathbf{C}$-algebra generated by $T(\mathfrak{p})$ and $S(\mathfrak{p})$ for all prime ideals $\mathfrak{p}$ of F. We also denote by $T(\mathfrak{a})$ the image $\phi(T(\mathfrak{a}))$ in $\mathrm{End}_{\mathbf{C}}\big(\mathcal{S}_k(\mathfrak{o}_F, \psi)\big)$.

An element $\mathbf{f}$ of $\mathcal{S}_k(\mathfrak{o}_F, \psi)$ is called a *primitive form* if $\mathbf{f}$ is a normalized common eigenfunction of $T(\mathfrak{p})$ for all prime ideals $\mathfrak{p}$. Here *normalized* means that the coefficient $c(1)$ of the Fourier expansion $f_x(z) = \sum_{\xi} c(\xi)e_F(\xi z)$ for $x = 1_2$ $(\in G_{\mathbf{h}})$ is equal to 1, where f_x is as in $(\mathrm{ii}_{\mathrm{b}})$. (Cf. [Shimura 78, p. 650].)

2.2 The Trace Formula

It is known that the characteristic polynomial of a Hecke operator can be obtained immediately from traces of Hecke operators by using (2–2), (2–5), and Newton's identities ([Miyake 89, pp. 266–267]). In particular if we take a prime ideal $\mathfrak{p}$ of F, we can obtain the characteristic polynomial $X^r + a_1X^{r-1} + \cdots + a_{r-1}X + a_r$ of $T(\mathfrak{p})$ as follows:

Let $c_1, \ldots, c_r$ be the eigenvalues of $T(\mathfrak{p})$, and set $b_l = c_1^l + \cdots + c_r^l = \mathrm{tr}\,(T(\mathfrak{p})^l)$. Then by (2–5), we have

$$\mathrm{tr}\,(T(\mathfrak{p})^l) = \sum_{i=0}^{[l/2]} \left(\binom{l}{i} - \binom{l}{i-1} \right) N(\mathfrak{p})^i \psi(\pi_{\mathfrak{p}})^i \mathrm{tr}\, T(\mathfrak{p}^{l-2i})$$

for $l = 1, \ldots, r$, where $\binom{l}{-1} = 0$. Therefore, we can obtain b_l from $\mathrm{tr}\, T(\mathfrak{p}^{l-2i})$ $(i = 0, \ldots, [l/2])$. By Newton's formula, we have

$$b_l + b_{l-1}a_1 + b_{l-2}a_2 + \cdots + b_1a_{l-1} + la_l = 0$$

for $l = 1, \ldots, r$. Thus we can obtain $a_1, \ldots, a_r$ from $b_1, \ldots, b_r$.

Now we describe the trace formula of a Hecke operator $T(\mathfrak{a})$ on $\mathcal{S}_k(\mathfrak{o}_F, \psi)$ given by [Saito 84, Theorem 2.1]. But first, we introduce the following notation.

Let K be a quadratic extension of F. We denote by $\mathcal{O}_{K/F}$ the set of all orders in K containing $\mathfrak{o}_F$. Let $\Lambda \in \mathcal{O}_{K/F}$. Since Λ is an F-lattice, we can take $x_1, x_2 \in K$ and $\mathfrak{a}_1, \mathfrak{a}_2 \in I(F)$ such that $\Lambda = \mathfrak{a}_1 x_1 + \mathfrak{a}_2 x_2$. Then we define the integral ideal $D_{K/F}(\Lambda)$ of F by

$$D_{K/F}(\Lambda) = (\mathfrak{a}_1\mathfrak{a}_2)^2 \begin{vmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \end{vmatrix}^2,$$

where $x_j^{(1)}, x_j^{(2)}$ are the conjugates of x_j over F. We call $D_{K/F}(\Lambda)$ the *relative discriminant of* Λ *with respect to* K/F.

Theorem 2.1. *Let F ($\neq \mathbf{Q}$) be a totally real algebraic number field of degree g, ψ a Hecke character of F of finite order such that the nonarchimedean part of its conductor is equal to $\mathfrak{o}_F$, and $k = (k_1, \ldots, k_g) \in \mathbf{Z}^{\mathbf{a}}$ such that $k_j \geq 2$ and $\psi_{v_j}(-1) = (-1)^{k_j}$ for each $v_j \in \mathbf{a}$. For every element $\mathfrak{b}P(F) \in \mathrm{Cl(F)}$, we define a mapping η of $\mathrm{Cl(F)}$ into $\mathrm{Cl}^+\mathrm{(F)}$ by $\eta(\mathfrak{b}P(F)) = \mathfrak{b}^2P^+(F)$. Then, for any integral ideal $\mathfrak{a}$ of F, we have*

$$\begin{aligned} \operatorname{tr} T(\mathfrak{a}) = {} & \varepsilon(\mathfrak{a})\delta(\mathfrak{a})\frac{2\zeta_F(2)|D_F|^{3/2}}{(2\pi)^{2g}}\psi\Big(\big(\pi_{\mathfrak{p}}^{\mathrm{ord}_{\mathfrak{p}}(\mathfrak{a})/2}\big)_{\mathfrak{p}\in\mathbf{h}}\Big) \\ & \cdot \Big(\prod_{j=1}^{g}(k_j-1)\Big) \quad + \varepsilon(\mathfrak{a})(-1)^g 2^{-1} \\ & \cdot \sum_{\mathfrak{m}\in M_{\mathfrak{a}}} \psi\Big(\big(\pi_{\mathfrak{p}}^{\mathrm{ord}_{\mathfrak{p}}(\mathfrak{m})}\big)_{\mathfrak{p}\in\mathbf{h}}\Big)^{-1} \\ & \cdot \sum_{n\in N_{\mathfrak{m}}} \sum_{s\in S_n} \Big(\prod_{j=1}^{g}\Phi(s_j, n_j, k_j)\Big) \sum_{\Lambda\in R_{sn}} \frac{h(\Lambda)}{h_F[\Lambda^\times : \mathfrak{o}_F^\times]} \\ & + (-1)^{g-1}b(k) \sum_{\lambda\in C(\psi)} \lambda\Big(\big(\pi_{\mathfrak{p}}^{\mathrm{ord}_{\mathfrak{p}}(\mathfrak{a})}\big)_{\mathfrak{p}\in\mathbf{h}}\Big) \sum_{\substack{\mathfrak{b}\mid\mathfrak{a} \\ \mathfrak{b}\subset\mathfrak{o}_F \\ \mathfrak{b}\in I(F)}} N(\mathfrak{b}). \end{aligned} \tag{2–6}$$

Here

- $\varepsilon(\mathfrak{a}) = 1$ *or* 0 *depending on whether* $\mathfrak{a}P^+(F) \in \eta(\mathrm{Cl(F)})$ *or not;*
- $\delta(\mathfrak{a}) = 1$ *or* 0 *depending on whether* $\mathfrak{a}$ *is a square or not;*
- $M_{\mathfrak{a}} = \{\mathfrak{m}\}$ *is the set of all representatives of* $\{\mathfrak{m}P(F) \in \mathrm{Cl(F)} \mid \mathfrak{m}^2\mathfrak{a} \in \mathrm{P}^+\mathrm{(F)}\}$ *such that* $\mathfrak{m} \subset \mathfrak{o}_F$ *and* $\gcd(\mathfrak{m}, \mathfrak{a}) = \mathfrak{o}_F$*;*
- *for every* $\mathfrak{m} \in M_{\mathfrak{a}}$*, we take an element* $n_{\mathfrak{m}}$ *of* $\mathfrak{o}_F$ *such that* $(n_{\mathfrak{m}})_F = \mathfrak{m}^2\mathfrak{a}$ *and* $n_{\mathfrak{m}} \gg 0$*, and we set* $N_{\mathfrak{m}} = n_{\mathfrak{m}}E_F$*, where* E_F *is a complete set of representatives of* $\mathfrak{o}_{F+}^\times/(\mathfrak{o}_F^\times)^2$*;*
- *for* $n \in N_{\mathfrak{m}}$*, we set* $S_n = \{s \in \mathfrak{m} \mid s^2 - 4n \ll 0\}$*;*
- *let* s_j*,* n_j *be the* v_j*-components of* s*,* n *in* $F_{\mathbf{a}}$*, and* α_j*,* β_j *the roots of* $X^2 - s_jX + n_j$*; then we set*

$$\Phi(s_j, n_j, k_j) = \frac{\alpha_j^{k_j-1} - \beta_j^{k_j-1}}{\alpha_j - \beta_j} n_j^{-(k_j-2)/2};$$

- $K_{sn} = F(\sqrt{s^2-4n})$*, and* R_{sn} *is the set of all distinct orders* Λ *in* $\mathcal{O}_{K_{sn}/F}$ *satisfying* $D_{K_{sn}/F}(\Lambda) \mid (s^2-4n)_F\mathfrak{m}^{-2}$*;*
- $h(\Lambda)$ *is the class number of* Λ*; that is,* $h(\Lambda) = |(K_{sn}\otimes_F F_{\mathbf{h}})^\times/K_{sn}^\times \prod_{\mathfrak{p}\in\mathbf{h}}\Lambda_{\mathfrak{p}}^\times|$*, where* $\Lambda_{\mathfrak{p}}$ *is the topological closure of* Λ *in* $K_{sn}\otimes_F F_{\mathfrak{p}}$*;*
- $b(k) = 1$ *or* 0 *depending on whether* $k = (2, \ldots, 2)$ *or not;*
- $C(\psi)$ *is the set of all unramified Hecke characters* λ *of* F *such that* $\lambda^2 = \psi$*.*

Note that the second sum of the right-hand side of (2–6) is independent of the choice of $M_{\mathfrak{a}}$, $n_{\mathfrak{m}}$, and E_F. We remark also that (2–6) is shortened and corrected from the original formula which appeared in [Saito 84].

2.3 Preliminary Lemmas

We now present five lemmas for transforming (2–6) into a more computable form.

Lemma 2.2. *Let F be an algebraic number field of finite degree, and K a quadratic extension of F. For an integral ideal $\mathfrak{c}$ of F, we put $\rho(\mathfrak{c}) = \mathfrak{o}_F + \mathfrak{c}\mathfrak{o}_K$. Then ρ is a bijection of the set of all integral ideals of F onto $\mathcal{O}_{K/F}$.*

Proof: Thie result follows immediately from [Shimura 71, Proposition 4.11] when $F = \mathbf{Q}$, and we prove our assertion in a similar fashion. It is well known that there exist $\theta \in K$ and $\mathfrak{a} \in I(F)$ such that $\mathfrak{o}_K = \mathfrak{o}_F + \theta\mathfrak{a}$. Let $\mathfrak{c}$ be an integral ideal of F. Since $\mathfrak{c}\mathfrak{o}_K \subset \mathfrak{o}_F + \mathfrak{c}\mathfrak{o}_K \subset \mathfrak{o}_K$, we see that $\mathfrak{o}_F + \mathfrak{c}\mathfrak{o}_K$ is a $\mathbf{Q}$-lattice in K. Moreover, $\mathfrak{o}_F + \mathfrak{c}\mathfrak{o}_K$ is a subring of K containing $\mathfrak{o}_F$. Thus $\mathfrak{o}_F + \mathfrak{c}\mathfrak{o}_K \in \mathcal{O}_{K/F}$, and hence ρ is a mapping. If $\mathfrak{o}_F + \mathfrak{c}\mathfrak{o}_K = \mathfrak{o}_F + \mathfrak{c}'\mathfrak{o}_K$ with integral ideals $\mathfrak{c}$, $\mathfrak{c}'$ of F, then $\mathfrak{o}_F + \theta\mathfrak{a}\mathfrak{c} = \mathfrak{o}_F + \mathfrak{c}\mathfrak{o}_K = \mathfrak{o}_F + \mathfrak{c}'\mathfrak{o}_K = \mathfrak{o}_F + \theta\mathfrak{a}\mathfrak{c}'$. Since $\{1, \theta\}$ is a basis of K over F, we have $\mathfrak{c} = \mathfrak{c}'$. Thus ρ is injective. Let Λ be any order in $\mathcal{O}_{K/F}$. Since $\mathfrak{o}_K$ is the unique maximal order of K, we have $\Lambda \subset \mathfrak{o}_K$. Set $\mathfrak{b} = \{c \in \mathfrak{a} \mid \theta c \in \Lambda\}$. Since

$\mathfrak{o}_F \subsetneq \Lambda \subset \mathfrak{o}_K = \mathfrak{o}_F + \theta\mathfrak{a}$, we have $\{0\} \subsetneq \mathfrak{b} \subset \mathfrak{a}$. Moreover, $\mathfrak{b}$ is an $\mathfrak{o}_F$-module. Thus $\mathfrak{b} \in I(F)$. For any $x \in \Lambda$, we have $x = r + \theta s$ with $r \in \mathfrak{o}_F$ and $s \in \mathfrak{a}$, since $\Lambda \subset \mathfrak{o}_K$. Then $\theta s = x - r \in \Lambda$, and hence $s \in \mathfrak{b}$. Therefore, $\Lambda = \mathfrak{o}_F + \theta\mathfrak{b}$. Now we set $\mathfrak{c} = \mathfrak{b}\mathfrak{a}^{-1}$. Then $\mathfrak{c}$ is an integral ideal of F, and $\Lambda = \mathfrak{o}_F + \theta\mathfrak{b} = \mathfrak{o}_F + \theta\mathfrak{a}\mathfrak{c} = \mathfrak{o}_F + \mathfrak{c}\mathfrak{o}_K$. Thus ρ is surjective. □

We denote the mapping ρ^{-1} by c, and we call $c(\Lambda)$ the *conductor of* Λ for $\Lambda \in \mathcal{O}_{K/F}$.

Lemma 2.3. *Let F and K be as in Lemma* 2.2. *Then for $\Lambda \in \mathcal{O}_{K/F}$, we have*

$$D_{K/F}(\Lambda) \cdot {D_{K/F}}^{-1} = c(\Lambda)^2.$$

Proof: By Lemma 2.2, we have $\mathfrak{o}_K = \mathfrak{o}_F + \theta\mathfrak{a}$ and $\Lambda = \mathfrak{o}_F + \theta\mathfrak{a}c(\Lambda)$ with $\theta \in K$ and $\mathfrak{a} \in I(F)$. Now let $\theta^{(1)}, \theta^{(2)}$ be the conjugates of θ over F. Then we have $D_{K/F} = D_{K/F}(\mathfrak{o}_K) = \mathfrak{a}^2(\theta^{(2)} - \theta^{(1)})^2$ and $D_{K/F}(\Lambda) = (\mathfrak{a}c(\Lambda))^2(\theta^{(2)} - \theta^{(1)})^2$. □

Let F be an algebraic number field of finite degree, and K a quadratic extension of F. For $\mathfrak{p} \in \mathbf{h}$, we define

$$\left(\frac{K}{\mathfrak{p}}\right) = \begin{cases} 1 & \text{if } \mathfrak{p} \text{ splits in } K, \\ -1 & \text{if } \mathfrak{p} \text{ remains prime in } K, \\ 0 & \text{if } \mathfrak{p} \text{ ramifies in } K. \end{cases}$$

Lemma 2.4. *Let F be a totally real algebraic number field of finite degree, and K a totally imaginary quadratic extension of F. Then for $\Lambda \in \mathcal{O}_{K/F}$, we have*

$$h(\Lambda) = h_K \left[\mathfrak{o}_K^\times : \Lambda^\times\right]^{-1} N(c(\Lambda)) \prod_{\substack{\mathfrak{p} \mid c(\Lambda) \\ \mathfrak{p} \in \mathbf{h}}} \left(1 - \left(\frac{K}{\mathfrak{p}}\right) N(\mathfrak{p})^{-1}\right).$$

Proof: This can be proved in exactly the same way as in [Miyake 89, Theorem 6.7.2], which deals with the case $F = \mathbf{Q}$. For any lattice L in K and $\mathfrak{p} \in \mathbf{h}$, we write $L_\mathfrak{p}$ for the topological closure of L in $K \otimes_F F_\mathfrak{p}$. For $\Lambda \in \mathcal{O}_{K/F}$, we have

$$\begin{aligned} h(\Lambda) &= \Big|(K \otimes_F F_{\mathbf{h}})^\times \;/\; K^\times \prod_{\mathfrak{p} \in \mathbf{h}} \Lambda_\mathfrak{p}^\times\Big| \\ &= \Big|(K \otimes_F F_{\mathbf{h}})^\times \;/\; K^\times \prod_{\mathfrak{p} \in \mathbf{h}} (\mathfrak{o}_K)_\mathfrak{p}^\times\Big| \\ &\quad \cdot \Big|K^\times \prod_{\mathfrak{p} \in \mathbf{h}} (\mathfrak{o}_K)_\mathfrak{p}^\times \;/\; K^\times \prod_{\mathfrak{p} \in \mathbf{h}} \Lambda_\mathfrak{p}^\times\Big|. \end{aligned}$$

Generally for an abelian group G and subgroups H, I, and J satisfying $I \supset J$, the sequence

$$1 \to (H \cap I)/(H \cap J) \to I/J \to HI/HJ \to 1$$

is exact. Thus

$$\begin{aligned} h(\Lambda) &= h_K \cdot \Big|\prod_{\mathfrak{p} \in \mathbf{h}} (\mathfrak{o}_K)_\mathfrak{p}^\times \;/\; \prod_{\mathfrak{p} \in \mathbf{h}} \Lambda_\mathfrak{p}^\times\Big| \\ &\quad \cdot \Big|\big(K^\times \cap \prod_{\mathfrak{p} \in \mathbf{h}} (\mathfrak{o}_K)_\mathfrak{p}^\times\big) \;/\; \big(K^\times \cap \prod_{\mathfrak{p} \in \mathbf{h}} \Lambda_\mathfrak{p}^\times\big)\Big|^{-1} \\ &= h_K \cdot \Big(\prod_{\mathfrak{p} \in \mathbf{h}} \big|(\mathfrak{o}_K)_\mathfrak{p}^\times / \Lambda_\mathfrak{p}^\times\big|\Big) \cdot \big|\mathfrak{o}_K^\times / \Lambda^\times\big|^{-1}. \end{aligned}$$

Since $(\mathfrak{o}_K)_\mathfrak{p} \neq \Lambda_\mathfrak{p}$ if and only if $\mathfrak{p} \mid c(\Lambda)$, we need to show only that

$$|(\mathfrak{o}_K)_\mathfrak{p}^\times / \Lambda_\mathfrak{p}^\times| = N(c(\Lambda)_\mathfrak{p}) \left(1 - \left(\tfrac{K}{\mathfrak{p}}\right) N(\mathfrak{p})^{-1}\right) \qquad (2\text{–}7)$$

for $(\mathfrak{o}_K)_\mathfrak{p} \neq \Lambda_\mathfrak{p}$. We denote an element $\alpha \otimes \beta$ of $K \otimes_F F_\mathfrak{p}$ simply by $\alpha\beta$. Let $\mathfrak{p}$ satisfy $(\mathfrak{o}_K)_\mathfrak{p} \neq \Lambda_\mathfrak{p}$. Assume first that $\mathfrak{p}$ splits in K. Set $\mathfrak{o}_K = \mathfrak{o}_F + \theta\mathfrak{a}$ with $\theta \in K$ and $\mathfrak{a} \in I(F)$, and take $\alpha_\mathfrak{p} \in F_\mathfrak{p}$ such that $\alpha_\mathfrak{p}\mathfrak{o}_\mathfrak{p} = \mathfrak{a}_\mathfrak{p}$. Let f be the minimal polynomial of θ over F, and let θ_1, θ_2 be the two roots of f in $F_\mathfrak{p}$. For $a, b \in F_\mathfrak{p}$, we set $\tau(a + b\alpha_\mathfrak{p}\theta) = (a + b\alpha_\mathfrak{p}\theta_1, a + b\alpha_\mathfrak{p}\theta_2)$. Then τ is a topological $F_\mathfrak{p}$-algebra isomorphism of $K \otimes_F F_\mathfrak{p}$ onto $F_\mathfrak{p} \times F_\mathfrak{p}$. (cf. [Weil 67, Chapter III, Theorem 4]). Since $(\alpha_\mathfrak{p}(\theta_2 - \theta_1))^2\mathfrak{o}_\mathfrak{p} = (D_{K/F})_\mathfrak{p} = \mathfrak{o}_\mathfrak{p}$, we have $\tau((\mathfrak{o}_K)_\mathfrak{p}) = \mathfrak{o}_\mathfrak{p} \times \mathfrak{o}_\mathfrak{p}$ and $\tau(\Lambda_\mathfrak{p}) = \{(\alpha, \beta) \in \mathfrak{o}_\mathfrak{p} \times \mathfrak{o}_\mathfrak{p} \mid \alpha - \beta \in c(\Lambda)_\mathfrak{p}\}$. Hence $\tau(\Lambda_\mathfrak{p})^\times = \{(\alpha, \beta) \in \mathfrak{o}_\mathfrak{p}^\times \times \mathfrak{o}_\mathfrak{p}^\times \mid \alpha - \beta \in c(\Lambda)_\mathfrak{p}\}$. For $(\alpha, \beta) \in \mathfrak{o}_\mathfrak{p}^\times \times \mathfrak{o}_\mathfrak{p}^\times$, we set $\rho((\alpha, \beta)) = \alpha\beta^{-1}(1 + c(\Lambda)_\mathfrak{p})$. Then ρ is a group homomorphism of $\mathfrak{o}_\mathfrak{p}^\times \times \mathfrak{o}_\mathfrak{p}^\times$ onto $\mathfrak{o}_\mathfrak{p}^\times/(1 + c(\Lambda)_\mathfrak{p})$. Since $\mathrm{Ker}(\rho) = \tau(\Lambda_\mathfrak{p})^\times$, we have $\tau((\mathfrak{o}_K)_\mathfrak{p}^\times)/\tau(\Lambda_\mathfrak{p}^\times) = (\mathfrak{o}_\mathfrak{p}^\times \times \mathfrak{o}_\mathfrak{p}^\times)/\tau(\Lambda_\mathfrak{p})^\times \cong \mathfrak{o}_\mathfrak{p}^\times/(1 + c(\Lambda)_\mathfrak{p})$. Therefore,

$$|(\mathfrak{o}_K)_\mathfrak{p}^\times / \Lambda_\mathfrak{p}^\times| = \big|\mathfrak{o}_\mathfrak{p}^\times/(1 + c(\Lambda)_\mathfrak{p})\big| = N(c(\Lambda)_\mathfrak{p})(1 - N(\mathfrak{p})^{-1}).$$

Thus we obtain (2–7) in this case. Now assume that $\mathfrak{p}$ remains prime or ramifies in K. Then $K \otimes_F F_\mathfrak{p}$ is a field. For $\beta \in \mathfrak{o}_\mathfrak{p}^\times$ and $\gamma \in (\mathfrak{o}_K)_\mathfrak{p}^\times$, we set $\mu_1(\beta(1 + c(\Lambda)_\mathfrak{p})) = \beta(1 + c(\Lambda)_\mathfrak{p}(\mathfrak{o}_K)_\mathfrak{p})$ and $\mu_2(\gamma(1 + c(\Lambda)_\mathfrak{p}(\mathfrak{o}_K)_\mathfrak{p})) = \gamma\Lambda_\mathfrak{p}^\times$. Since $(1 + c(\Lambda)_\mathfrak{p}(\mathfrak{o}_K)_\mathfrak{p}) \cap \mathfrak{o}_\mathfrak{p}^\times = (1 + c(\Lambda)_\mathfrak{p} + \theta c(\Lambda)_\mathfrak{p}\mathfrak{a}_\mathfrak{p}) \cap \mathfrak{o}_\mathfrak{p}^\times = 1 + c(\Lambda)_\mathfrak{p}$ and $\Lambda_\mathfrak{p}^\times = \mathfrak{o}_\mathfrak{p}^\times + c(\Lambda)_\mathfrak{p}(\mathfrak{o}_K)_\mathfrak{p} = \mathfrak{o}_\mathfrak{p}^\times(1 + c(\Lambda)_\mathfrak{p}(\mathfrak{o}_K)_\mathfrak{p})$, the sequence

$$\begin{aligned} 1 \to \mathfrak{o}_\mathfrak{p}^\times/(1 + c(\Lambda)_\mathfrak{p}) &\xrightarrow{\mu_1} (\mathfrak{o}_K)_\mathfrak{p}^\times/(1 + c(\Lambda)_\mathfrak{p}(\mathfrak{o}_K)_\mathfrak{p}) \\ &\xrightarrow{\mu_2} (\mathfrak{o}_K)_\mathfrak{p}^\times/\Lambda_\mathfrak{p}^\times \to 1 \end{aligned}$$

is exact. Therefore,

$$\big|(\mathfrak{o}_K)_\mathfrak{p}^\times/\Lambda_\mathfrak{p}^\times\big| = \big|(\mathfrak{o}_K)_\mathfrak{p}^\times/(1 + c(\Lambda)_\mathfrak{p}(\mathfrak{o}_K)_\mathfrak{p})\big| \cdot \big|\mathfrak{o}_\mathfrak{p}^\times/(1 + c(\Lambda)_\mathfrak{p})\big|^{-1}.$$

Here we have

$$\begin{aligned} &\big|(\mathfrak{o}_K)_\mathfrak{p}^\times/(1 + c(\Lambda)_\mathfrak{p}(\mathfrak{o}_K)_\mathfrak{p})\big| = \big|((\mathfrak{o}_K)_\mathfrak{p}/c(\Lambda)_\mathfrak{p}(\mathfrak{o}_K)_\mathfrak{p})^\times\big| \\ &= \begin{cases} N(c(\Lambda)_\mathfrak{p})^2(1 - N(\mathfrak{p})^{-2}) & \text{if } \mathfrak{p} \text{ remains prime in } K, \\ N(c(\Lambda)_\mathfrak{p})^2(1 - N(\mathfrak{p})^{-1}) & \text{if } \mathfrak{p} \text{ ramifies in } K, \end{cases} \end{aligned}$$

and $\left|\mathfrak{o}_\mathfrak{p}^\times/(1+c(\Lambda)_\mathfrak{p})\right| = \left|(\mathfrak{o}_\mathfrak{p}/c(\Lambda)_\mathfrak{p})^\times\right| = N(c(\Lambda)_\mathfrak{p})(1-N(\mathfrak{p})^{-1})$, which proves (2–7) in this case. □

Lemma 2.5. *Let F and K be as in Lemma* 2.4. *Let g be the degree of F over $\mathbf{Q}$. Let $\chi_{K/F}$ be the ideal character corresponding to the extension K/F (by means of class field theory). Then*

$$L_F(0,\chi_{K/F}) = 2^{g-1}\frac{h_K}{h_F\left[\mathfrak{o}_K^\times:\mathfrak{o}_F^\times\right]},$$

where $L_F(s,\chi_{K/F})$ is the Hecke L-function associated with $\chi_{K/F}$.

Proof: Let W_K (respectively W_F) be the group of the roots of 1 in K (respectively F), and R_K (respectively R_F) the regulator of K (respectively F). Set $w_K = |W_K|$ and $w_F = |W_F|$. Let $Z_K(s) = \left((2\pi)^{1-s}\Gamma(s)\right)^g\zeta_K(s)$ and $Z_F(s) = \left(\pi^{-s/2}\Gamma(s/2)\right)^g\zeta_F(s)$. Then we have $\operatorname{Res}_{s=0}Z_K(s) = -(2\pi)^g h_K R_K w_K^{-1}$ and $\operatorname{Res}_{s=0}Z_F(s) = -2^g h_F R_F w_F^{-1}$. By $\zeta_K(s) = \zeta_F(s)L_F(s,\chi_{K/F})$, we have $Z_K(s) = \pi^{g(1-s)/2}\Gamma\left((s+1)/2\right)^g Z_F(s)L_F(s,\chi_{K/F})$, and hence $\operatorname{Res}_{s=0}Z_K(s) = \pi^g\operatorname{Res}_{s=0}Z_F(s)\cdot L_F(0,\chi_{K/F})$. Therefore,

$$L_F(0,\chi_{K/F}) = \frac{h_K R_K w_K^{-1}}{h_F R_F w_F^{-1}}.$$

Thus we need to show that

$$[\mathfrak{o}_K^\times:\mathfrak{o}_F^\times] = 2^{g-1}w_K w_F^{-1}R_K^{-1}R_F. \tag{2–8}$$

Let l be the mapping from $\mathfrak{o}_K^\times$ to $\mathbf{R}^g$ defined by $l(\delta) = (\log|\delta^{(1)}|,\dots,\log|\delta^{(g)}|)$, where $\delta^{(1)},\dots,\delta^{(g)}$ are the conjugates of δ over F. Then $\mathfrak{o}_K^\times/W_K\mathfrak{o}_F^\times \cong l(\mathfrak{o}_K^\times)/l(\mathfrak{o}_F^\times)$. Since $[l(\mathfrak{o}_K^\times):l(\mathfrak{o}_F^\times)] = 2^{g-1}R_K^{-1}R_F$, we have $[\mathfrak{o}_K^\times : W_K\mathfrak{o}_F^\times] = 2^{g-1}R_K^{-1}R_F$. On the other hand, we have $[W_K\mathfrak{o}_F^\times:\mathfrak{o}_F^\times] = w_F^{-1}w_K$. Thus we obtain (2–8). □

Lemma 2.6. *Let F and K be as in Lemma* 2.4. *Then, for any integral ideal $\mathfrak{f}$ of F, we have*

$$\begin{aligned}
&\sum_{\substack{\mathfrak{c}\mid\mathfrak{f}\\ \mathfrak{c}\subset\mathfrak{o}_F\\ \mathfrak{c}\in I(F)}}\left(N(\mathfrak{c})\prod_{\substack{\mathfrak{p}\mid\mathfrak{c}\\ \mathfrak{p}\in\mathbf{h}}}\left(1-\left(\tfrac{K}{\mathfrak{p}}\right)N(\mathfrak{p})^{-1}\right)\right)\\
&\quad=\left(\prod_{\substack{\mathfrak{p}\mid\mathfrak{f}\\ \left(\frac{K}{\mathfrak{p}}\right)=-1\\ \mathfrak{p}\in\mathbf{h}}}\frac{N(\mathfrak{p})^{\mathrm{ord}_\mathfrak{p}(\mathfrak{f})+1}+N(\mathfrak{p})^{\mathrm{ord}_\mathfrak{p}(\mathfrak{f})}-2}{N(\mathfrak{p})-1}\right)\\
&\qquad\cdot\left(\prod_{\substack{\mathfrak{p}\mid\mathfrak{f}\\ \left(\frac{K}{\mathfrak{p}}\right)=1\\ \mathfrak{p}\in\mathbf{h}}}N(\mathfrak{p})^{\mathrm{ord}_\mathfrak{p}(\mathfrak{f})}\right)\left(\prod_{\substack{\mathfrak{p}\mid\mathfrak{f}\\ \left(\frac{K}{\mathfrak{p}}\right)=0\\ \mathfrak{p}\in\mathbf{h}}}\frac{N(\mathfrak{p})^{\mathrm{ord}_\mathfrak{p}(\mathfrak{f})+1}-1}{N(\mathfrak{p})-1}\right).
\end{aligned}$$

Proof: For every prime ideal $\mathfrak{p}$ of F and $0\le s\in\mathbf{Z}$, we set

$$\varphi(\mathfrak{p}^s) = \begin{cases}1 & \text{if } s=0,\\ N(\mathfrak{p})^s\left(1-\left(\frac{K}{\mathfrak{p}}\right)N(\mathfrak{p})^{-1}\right) & \text{if } s\ge 1.\end{cases}$$

Now let $\mathfrak{f} = \mathfrak{p}_1^{e_1}\cdots\mathfrak{p}_r^{e_r}$ be the factorization of $\mathfrak{f}$ into prime factors. Then

$$\begin{aligned}
&\sum_{\mathfrak{c}\mid\mathfrak{f}}\left(N(\mathfrak{c})\prod_{\mathfrak{p}\mid\mathfrak{c}}\left(1-\left(\tfrac{K}{\mathfrak{p}}\right)N(\mathfrak{p})^{-1}\right)\right)\\
&\quad=\sum_{s_1=0}^{e_1}\cdots\sum_{s_r=0}^{e_r}\left(\varphi(\mathfrak{p}_1^{s_1})\cdots\varphi(\mathfrak{p}_r^{s_r})\right)\\
&\quad=\prod_{j=1}^{r}\Big(\sum_{s_j=0}^{e_j}\varphi(\mathfrak{p}_j^{s_j})\Big).
\end{aligned}$$

Here we have

$$\sum_{s_j=0}^{e_j}\varphi(\mathfrak{p}_j^{s_j}) = \begin{cases}N(\mathfrak{p}_j)^{e_j} & \text{if } \left(\frac{K}{\mathfrak{p}_j}\right)=1,\\ \left(N(\mathfrak{p}_j)^{e_j+1}+N(\mathfrak{p}_j)^{e_j}-2\right)\\ \quad\cdot\left(N(\mathfrak{p}_j)-1\right)^{-1} & \text{if } \left(\frac{K}{\mathfrak{p}_j}\right)=-1,\\ \left(N(\mathfrak{p}_j)^{e_j+1}-1\right)\left(N(\mathfrak{p}_j)-1\right)^{-1} & \text{if } \left(\frac{K}{\mathfrak{p}_j}\right)=0.\end{cases}$$

Therefore, we obtain our lemma. □

2.4 Formula for Computation

From the above lemmas, we obtain the following result:

Proposition 2.7. *With the notation of Theorem* 2.1, *we have*

$$\begin{aligned}
\operatorname{tr}T(\mathfrak{a}) &= \varepsilon(\mathfrak{a})\delta(\mathfrak{a})(-1)^g 2^{1-g}\zeta_F(-1)\psi\Big(\left(\pi_\mathfrak{p}^{\mathrm{ord}_\mathfrak{p}(\mathfrak{a})/2}\right)_{\mathfrak{p}\in\mathbf{h}}\Big)\\
&\quad\cdot\Big(\prod_{j=1}^{g}(k_j-1)\Big)+\varepsilon(\mathfrak{a})(-1)^g 2^{-g}\sum_{\mathfrak{m}\in M_\mathfrak{a}}\psi\Big(\left(\pi_\mathfrak{p}^{\mathrm{ord}_\mathfrak{p}(\mathfrak{m})}\right)_{\mathfrak{p}\in\mathbf{h}}\Big)^{-1}\\
&\quad\cdot\sum_{n\in N_\mathfrak{m}}\sum_{s\in S_n}\Big(\prod_{j=1}^{g}\Phi(s_j,n_j,k_j)\Big)\cdot L_F(0,\chi_{K_{sn}/F})\\
&\quad\cdot\Bigg(\prod_{\substack{\mathfrak{p}\mid\mathfrak{f}_{sn}\\ \left(\frac{K_{sn}}{\mathfrak{p}}\right)=-1\\ \mathfrak{p}\in\mathbf{h}}}\frac{N(\mathfrak{p})^{\mathrm{ord}_\mathfrak{p}(\mathfrak{f}_{sn})+1}+N(\mathfrak{p})^{\mathrm{ord}_\mathfrak{p}(\mathfrak{f}_{sn})}-2}{N(\mathfrak{p})-1}\Bigg)\\
&\quad\cdot\Bigg(\prod_{\substack{\mathfrak{p}\mid\mathfrak{f}_{sn}\\ \left(\frac{K_{sn}}{\mathfrak{p}}\right)=1\\ \mathfrak{p}\in\mathbf{h}}}N(\mathfrak{p})^{\mathrm{ord}_\mathfrak{p}(\mathfrak{f}_{sn})}\Bigg)\Bigg(\prod_{\substack{\mathfrak{p}\mid\mathfrak{f}_{sn}\\ \left(\frac{K_{sn}}{\mathfrak{p}}\right)=0\\ \mathfrak{p}\in\mathbf{h}}}\frac{N(\mathfrak{p})^{\mathrm{ord}_\mathfrak{p}(\mathfrak{f}_{sn})+1}-1}{N(\mathfrak{p})-1}\Bigg)\\
&\quad+(-1)^{g-1}b(k)\sum_{\lambda\in C(\psi)}\lambda\Big(\left(\pi_\mathfrak{p}^{\mathrm{ord}_\mathfrak{p}(\mathfrak{a})}\right)_{\mathfrak{p}\in\mathbf{h}}\Big)\sum_{\substack{\mathfrak{b}\mid\mathfrak{a}\\ \mathfrak{b}\subset\mathfrak{o}_F\\ \mathfrak{b}\in I(F)}}N(\mathfrak{b}),
\end{aligned}\tag{2–9}$$

where $\mathfrak{f}_{sn} = c\big(\mathfrak{o}_F + ((s+\sqrt{s^2-4n})/2)\mathfrak{o}_F\big)\mathfrak{m}^{-1} = \big((s^2-4n)_F D_{K_{sn}/F}{}^{-1}\big)^{1/2}\mathfrak{m}^{-1}$, *and* $\chi_{K_{sn}/F}$ *is the ideal character corresponding to the extension* K_{sn}/F. *(We note that* $(s^2-4n)_F D_{K_{sn}/F}{}^{-1}$ *is a square by Lemma* 2.3.*)*

Proof: In view of Theorem 2.1, we only need to show that

$$\frac{2\zeta_F(2)|D_F|^{3/2}}{(2\pi)^{2g}} = (-1)^g 2^{1-g}\zeta_F(-1), \tag{2–10}$$

$$\begin{aligned}\sum_{\Lambda\in R_{sn}} \frac{h(\Lambda)}{h_F[\Lambda^\times : \mathfrak{o}_F^\times]} &= 2^{1-g}L_F(0,\chi_{K_{sn}/F}) \\ &\cdot\Bigg(\prod_{\substack{\mathfrak{p}\mid\mathfrak{f}_{sn}\\ (\frac{K_{sn}}{\mathfrak{p}})=-1}} \frac{N(\mathfrak{p})^{\mathrm{ord}_\mathfrak{p}(\mathfrak{f}_{sn})+1}+N(\mathfrak{p})^{\mathrm{ord}_\mathfrak{p}(\mathfrak{f}_{sn})}-2}{N(\mathfrak{p})-1}\Bigg) \\ &\cdot\Bigg(\prod_{\substack{\mathfrak{p}\mid\mathfrak{f}_{sn}\\ (\frac{K_{sn}}{\mathfrak{p}})=1}} N(\mathfrak{p})^{\mathrm{ord}_\mathfrak{p}(\mathfrak{f}_{sn})}\Bigg)\Bigg(\prod_{\substack{\mathfrak{p}\mid\mathfrak{f}_{sn}\\ (\frac{K_{sn}}{\mathfrak{p}})=0}} \frac{N(\mathfrak{p})^{\mathrm{ord}_\mathfrak{p}(\mathfrak{f}_{sn})+1}-1}{N(\mathfrak{p})-1}\Bigg).\end{aligned} \tag{2–11}$$

By the functional equation of ζ_F, we have $\zeta_F(2) = |D_F|^{-3/2}(-2\pi^2)^g\zeta_F(-1)$, and hence we obtain (2–10). Next, by Lemma 2.4 and Lemma 2.5, we have

$$\begin{aligned}&\sum_{\Lambda\in R_{sn}} \frac{h(\Lambda)}{h_F[\Lambda^\times : \mathfrak{o}_F^\times]} \\ &= \sum_{\Lambda\in R_{sn}} \frac{h_{K_{sn}}}{h_F\big[\mathfrak{o}_{K_{sn}}^\times : \mathfrak{o}_F^\times\big]} N(c(\Lambda)) \prod_{\mathfrak{p}\mid c(\Lambda)} \Big(1-\big(\tfrac{K_{sn}}{\mathfrak{p}}\big)N(\mathfrak{p})^{-1}\Big) \\ &= 2^{1-g}L_F(0,\chi_{K_{sn}/F}) \sum_{\Lambda\in R_{sn}} N(c(\Lambda)) \\ &\quad\cdot \prod_{\mathfrak{p}\mid c(\Lambda)} \Big(1-\big(\tfrac{K_{sn}}{\mathfrak{p}}\big)N(\mathfrak{p})^{-1}\Big).\end{aligned}$$

Now, from Lemma 2.3, we have

$$\begin{aligned}R_{sn} &= \big\{\Lambda\in\mathcal{O}_{K_{sn}/F} \mid D_{K_{sn}/F}(\Lambda) \mid (s^2-4n)_F\mathfrak{m}^{-2}\big\} \\ &= \big\{\Lambda\in\mathcal{O}_{K_{sn}/F} \mid c(\Lambda)^2 \mid (s^2-4n)_F D_{K_{sn}/F}{}^{-1}\mathfrak{m}^{-2}\big\} \\ &= \big\{\Lambda\in\mathcal{O}_{K_{sn}/F} \mid c(\Lambda) \mid \mathfrak{f}_{sn}\big\}.\end{aligned}$$

Thus we have

$$\begin{aligned}&\sum_{\Lambda\in R_{sn}} N(c(\Lambda)) \prod_{\mathfrak{p}\mid c(\Lambda)} \Big(1-\big(\tfrac{K_{sn}}{\mathfrak{p}}\big)N(\mathfrak{p})^{-1}\Big) \\ &\qquad = \sum_{\substack{\mathfrak{c}\mid\mathfrak{f}_{sn}\\ \mathfrak{c}\subset\mathfrak{o}_F\\ \mathfrak{c}\in I(F)}} N(\mathfrak{c}) \prod_{\mathfrak{p}\mid\mathfrak{c}} \Big(1-\big(\tfrac{K_{sn}}{\mathfrak{p}}\big)N(\mathfrak{p})^{-1}\Big),\end{aligned}$$

since the mapping c is bijective by Lemma 2.2. Therefore we obtain (2–11) by Lemma 2.6. □

3. COMPUTATION FOR REAL QUADRATIC FIELDS

In this section, we give an algorithm to compute Formula (2–9) for a real quadratic field F. In particular, *we assume* $F = \mathbf{Q}(\sqrt{m})$ *with a square-free integer* m *satisfying* $m \equiv 1 \pmod{4\mathbf{Z}}$ exclusively, though the case $m \not\equiv 1 \pmod{4\mathbf{Z}}$ can be handled by a similar consideration below. Throughout this section, we let $\omega = (1+\sqrt{m})/2$, and denote by σ the nontrivial automorphism of F. We note that for every integral ideal $\mathfrak{a}$ of F there exist rational integers $a > 0$ and b such that $\mathfrak{a} = [a, b+\omega]$; moreover, it is well known how to check whether $\mathfrak{a} \in P(F)$ (respectively $\mathfrak{a} \in P^+(F)$) and to find an explicit generator of $\mathfrak{a}$ when $\mathfrak{a} \in P(F)$ (respectively $\mathfrak{a} \in P^+(F)$) by the theory of continued fractions (cf. [Dirichlet 1894], for example). For $\mathfrak{p} \in \mathbf{h}$, we call $\mathfrak{p}$ *odd* (respectively *even*) if $\mathfrak{p} \nmid (2)_F$ (respectively $\mathfrak{p} \mid (2)_F$).

3.1 Preliminaries

We note that $\zeta_F(-1) = \zeta(-1)\cdot L\big(-1, \big(\frac{m}{\cdot}\big)\big) = 24^{-1}B_{2,\left(\frac{m}{\cdot}\right)}$, where $B_{2,\left(\frac{m}{\cdot}\right)}$ is the second generalized Bernoulli number associated with $\big(\frac{m}{\cdot}\big)$. Here $\big(\frac{m}{\cdot}\big)$ is the character corresponding to $\mathbf{Q}(\sqrt{m})/\mathbf{Q}$. It is known that

$$B_{2,\left(\frac{m}{\cdot}\right)} = (6m)^{-1}\sum_{a=1}^{m}\big(\tfrac{m}{a}\big)(6a^2-6am+m^2)$$

(cf. [Iwasawa 72, §2]). Hence

$$\zeta_F(-1) = (144m)^{-1}\sum_{a=1}^{m}\big(\tfrac{m}{a}\big)(6a^2-6am+m^2).$$

Thus the first and third sums of the right-hand side of (2–9) are easily computable.

Hereinafter, we consider the second sum for the case $\varepsilon(\mathfrak{a}) = 1$, which implies $M_\mathfrak{a} \neq \emptyset$. We first explain a method for choosing $M_\mathfrak{a}$, $N_\mathfrak{m}$, and S_n in (2–9). Choose an arbitrary complete set of integral representatives C of Cl(F). Take the set of all elements $\mathfrak{b}_1, \ldots, \mathfrak{b}_u$ of C such that $\mathfrak{b}_j^2\mathfrak{a} \in P^+(F)$. If $\gcd(\mathfrak{b}_j, \mathfrak{a}) = \mathfrak{o}_F$, then set $\mathfrak{m}_j = \mathfrak{b}_j$; if $\gcd(\mathfrak{b}_j, \mathfrak{a}) \neq \mathfrak{o}_F$, then take a prime ideal $\mathfrak{p}_j$ of F satisfying $\mathfrak{p}_j \nmid \mathfrak{a}$ and $\mathfrak{p}_j^\sigma\mathfrak{b}_j \in P(F)$ (i.e., $\mathfrak{p}_j P(F) = \mathfrak{b}_j P(F)$), and set $\mathfrak{m}_j = \mathfrak{p}_j$. Then the set $M_\mathfrak{a}$ is given by

$$M_\mathfrak{a} = \{\mathfrak{m}_1, \ldots, \mathfrak{m}_u\}.$$

We fix $\mathfrak{m} \in M_\mathfrak{a}$. Then we can take $n_\mathfrak{m}$ satisfying $0 \ll n_\mathfrak{m} \in \mathfrak{o}_F$ and $(n_\mathfrak{m})_F = \mathfrak{m}^2\mathfrak{a}$. Now we can choose $\{1\}$ as E_F when $h_F = h_F^+$, and $\{1, \varepsilon\}$ as E_F when $h_F \neq h_F^+$, where ε is the fundamental unit of F satisfying $\varepsilon > 1$. Thus we can take $N_\mathfrak{m} = n_\mathfrak{m}E_F$. (Note that the choice of

$M_{\mathfrak{a}}$ and $N_{\mathfrak{m}}$ has no effect on $\operatorname{tr} T(\mathfrak{a})$, as remarked after Theorem 2.1.) We also fix $n \in N_{\mathfrak{m}}$. Now we set $\mathfrak{m} = [l_1, l_2+\omega]$ with $l_1, l_2 \in \mathbf{Z}$ satisfying $1 \le l_1$ and $0 \le l_2 < l_1$. For $s \in \mathfrak{m}$, we can set $s = yl_1 + x(l_2+\omega)$ with $y, x \in \mathbf{Z}$. Then we have

$$\begin{aligned} s^2 \ll 4n &\iff s^2 < 4n,\ (s^\sigma)^2 < 4n^\sigma \\ &\iff -2\sqrt{n} < s < 2\sqrt{n},\ -2\sqrt{n^\sigma} < s^\sigma < 2\sqrt{n^\sigma} \\ &\iff \begin{cases} y < -\dfrac{2l_2+1+\sqrt{m}}{2l_1}x + \dfrac{2\sqrt{n}}{l_1}, \\ y > -\dfrac{2l_2+1+\sqrt{m}}{2l_1}x - \dfrac{2\sqrt{n}}{l_1}, \\ y < -\dfrac{2l_2+1-\sqrt{m}}{2l_1}x + \dfrac{2\sqrt{n^\sigma}}{l_1}, \\ y > -\dfrac{2l_2+1-\sqrt{m}}{2l_1}x - \dfrac{2\sqrt{n^\sigma}}{l_1}. \end{cases} \end{aligned} \tag{3–1}$$

Thus we have

$$S_n = \{yl_1 + x(l_2+\omega) \mid y, x \in \mathbf{Z} \text{ satisfying (3–1)}\}.$$

Set $K = F(\sqrt{\alpha})$ with $\alpha \in \mathfrak{o}_F$ satisfying $\alpha \ll 0$. Then our study is reduced to the computation of the following :

(i) $D_{K/F} = \prod_{\mathfrak{p}\in\mathbf{h}} D_{\mathfrak{p}}$,

(ii) $\left(\frac{K}{\mathfrak{p}}\right)$ for $\mathfrak{p} \in \mathbf{h}$,

(iii) $L_F(0, \chi_{K/F})$,

where $\prod_{\mathfrak{p}\in\mathbf{h}} D_{\mathfrak{p}}$ is the prime factorization of $D_{K/F}$.

3.2 Determination of $D_{\mathfrak{p}}$ and $\left(\frac{K}{\mathfrak{p}}\right)$ for an Odd Prime $\mathfrak{p}$

First we explain a way to determine $D_{\mathfrak{p}}$ and $\left(\frac{K}{\mathfrak{p}}\right)$ for an *odd* prime ideal $\mathfrak{p}$ of F.

Proposition 3.1. *Let F be an algebraic number field of finite degree, and $K = F(\sqrt{\alpha})$ a quadratic extension of F with $\alpha \in \mathfrak{o}_F$. Let $\mathfrak{p}$ be an odd prime ideal of F. Then*

$$D_{\mathfrak{p}} = \begin{cases} \mathfrak{o}_F & \text{if } 2 \mid \operatorname{ord}_{\mathfrak{p}}(\alpha), \\ \mathfrak{p} & \text{otherwise.} \end{cases}$$

Proof: If $2 \mid \operatorname{ord}_{\mathfrak{p}}(\alpha)$, then we can find $\alpha_1 \in \mathfrak{o}_F$ such that $K = F(\sqrt{\alpha_1})$ and $\operatorname{ord}_{\mathfrak{p}}(\alpha_1) = 0$. Since $D_{K/F} \mid D_{K/F}(\sqrt{\alpha_1})$ and $D_{K/F}(\sqrt{\alpha_1}) = (4\alpha_1)_F$, we have $D_{\mathfrak{p}} = \mathfrak{o}_F$. Now assume $2 \nmid \operatorname{ord}_{\mathfrak{p}}(\alpha)$, take a prime ideal $\mathfrak{P}$ of K that lies above $\mathfrak{p}$, and let e be the ramification index of $\mathfrak{P}$ in K/F. Then $2\cdot\operatorname{ord}_{\mathfrak{P}}(\sqrt{\alpha}) = \operatorname{ord}_{\mathfrak{P}}(\alpha) = e\cdot\operatorname{ord}_{\mathfrak{p}}(\alpha)$, and hence $e = 2$. Since $[K_{\mathfrak{P}} : F_{\mathfrak{p}}] = 2$ and $\mathfrak{p} \nmid (2)_F$, we have $D_{\mathfrak{p}} = \mathfrak{p}^{e-1} = \mathfrak{p}$ (cf. [Weil 67, Chapter VIII, Corollary 3 of Proposition 7]). □

We can determine $D_{\mathfrak{p}}$ from this proposition.

Let F, K, and $\mathfrak{p}$ be as in Proposition 3.1. By Dedekind's discriminant theorem, we know that

$$\left(\tfrac{K}{\mathfrak{p}}\right) = 0 \iff \mathfrak{p} \mid D_{K/F}. \tag{3–2}$$

For an explicit determination of $\left(\frac{K}{\mathfrak{p}}\right)$ for $\mathfrak{p} \nmid D_{K/F}$, we start with the following lemma.

Lemma 3.2. *Let F be a Galois extension of $\mathbf{Q}$ of prime degree, and $K = F(\sqrt{\alpha})$ a quadratic extension of F with $\alpha \in \mathfrak{o}_F$. Let $\mathfrak{p}$ be an odd prime ideal of F satisfying $\mathfrak{p} \nmid D_{K/F}$, and p the prime number in $\mathbf{Q}$ that lies below $\mathfrak{p}$ (i.e., $\mathfrak{p} \mid (p)_F$, which means that $\mathfrak{p} \cap \mathbf{Z} = p\mathbf{Z}$). If p remains prime in $F/\mathbf{Q}$, then we set $a = N_{F/\mathbf{Q}}(\alpha p^{-\operatorname{ord}_{\mathfrak{p}}(\alpha)})$; if p ramifies in $F/\mathbf{Q}$, then we take $a \in (\alpha\pi_{\mathfrak{p}}^{-\operatorname{ord}_{\mathfrak{p}}(\alpha)} + \pi_{\mathfrak{p}}\mathfrak{o}_{\mathfrak{p}}) \cap \mathbf{Z}$; if p splits in $F/\mathbf{Q}$, then we take $a_0 \in (\alpha + p^{\operatorname{ord}_{\mathfrak{p}}(\alpha)+1}\mathfrak{o}_{\mathfrak{p}}) \cap \mathbf{Z}$, and set $a = a_0 p^{-\operatorname{ord}_{\mathfrak{p}}(\alpha)}$. Then we have*

$$\left(\tfrac{K}{\mathfrak{p}}\right) = \left(\tfrac{a}{p}\right).$$

Note that this criterion does not depend on the choices of α and a from the proof below.

Proof: Put $[F : \mathbf{Q}] = g$. Now, $\mathfrak{p}$ splits or remains prime in K/F by (3–2), and $2 \mid \operatorname{ord}_{\mathfrak{p}}(\alpha)$ by Proposition 3.1. Since $\mathfrak{p}$ is odd, we have

$$(1+\pi_{\mathfrak{p}}\mathfrak{o}_{\mathfrak{p}})^2 = 1 + \pi_{\mathfrak{p}}\mathfrak{o}_{\mathfrak{p}};$$

indeed, for any element $1+\pi_{\mathfrak{p}}y \in 1+\pi_{\mathfrak{p}}\mathfrak{o}_{\mathfrak{p}}$, the polynomial $\pi_{\mathfrak{p}}X^2 + 2X - y$ has a root $x \in \mathfrak{o}_{\mathfrak{p}}$ by Hensel's lemma,and thus $(1+\pi_{\mathfrak{p}}x)^2 = 1 + \pi_{\mathfrak{p}}y$. Note that $\mathfrak{p}$ splits in K/F if and only if the polynomial $X^2 - \alpha$ is reducible over $F_{\mathfrak{p}}$; that is,

$$\left(\tfrac{K}{\mathfrak{p}}\right) = 1 \iff \alpha \in (F_{\mathfrak{p}}^{\times})^2.$$

Assume first that p remains prime in $F/\mathbf{Q}$. Then $[F_{\mathfrak{p}} : \mathbf{Q}_p] = g$ and $\operatorname{ord}_{\mathfrak{p}}(p) = 1$. Write $\operatorname{Gal}(F/\mathbf{Q}) = \{\sigma_1, \dots, \sigma_g\}$ and $\alpha_0 = \alpha p^{-\operatorname{ord}_{\mathfrak{p}}(\alpha)} (\in \mathfrak{o}_{\mathfrak{p}}^{\times} \cap \mathfrak{o}_F)$. Then

$$\{\alpha_0^{p^j} + \mathfrak{o}_{\mathfrak{p}} \mid j = 0, \dots, g-1\} = \{\alpha_0^{\sigma_j} + \mathfrak{o}_{\mathfrak{p}} \mid j = 1, \dots, g\}$$

(cf. [Weil 67, Chapter I, Corollary 2 of Theorem 7]). Thus

$$\begin{aligned} \alpha \in (F_{\mathfrak{p}}^{\times})^2 &\iff \alpha_0 \in (\mathfrak{o}_{\mathfrak{p}}^{\times})^2 \\ &\iff (\alpha_0\alpha_0^p \cdots \alpha_0^{p^{g-1}})^{(p-1)/2} \\ &\qquad = \alpha_0^{(N(\mathfrak{p})-1)/2} \in 1 + p\mathfrak{o}_{\mathfrak{p}} \\ &\iff (\alpha_0^{\sigma_1}\alpha_0^{\sigma_2} \cdots \alpha_0^{\sigma_g})^{(p-1)/2} \\ &\qquad = N_{F/\mathbf{Q}}(\alpha_0)^{(p-1)/2} \in 1 + p\mathfrak{o}_{\mathfrak{p}} \\ &\iff a^{(p-1)/2} = N_{F/\mathbf{Q}}(\alpha_0)^{(p-1)/2} \in 1 + p\mathbf{Z} \\ &\iff \left(\tfrac{a}{p}\right) = 1. \end{aligned}$$

Next assume that p ramifies in $F/\mathbf{Q}$. Since $\alpha\pi_{\mathfrak{p}}^{-\operatorname{ord}_{\mathfrak{p}}(\alpha)} \in \mathfrak{o}_{\mathfrak{p}}^{\times}$ and $\mathfrak{o}_{\mathfrak{p}}/\pi_{\mathfrak{p}}\mathfrak{o}_{\mathfrak{p}} \cong \mathbf{Z}/p\mathbf{Z}$, we can take $a \in (\alpha\pi_{\mathfrak{p}}^{-\operatorname{ord}_{\mathfrak{p}}(\alpha)} + \pi_{\mathfrak{p}}\mathfrak{o}_{\mathfrak{p}}) \cap \mathbf{Z}$. Thus

$$\begin{aligned}\alpha \in (F_{\mathfrak{p}}^{\times})^2 &\Longleftrightarrow \alpha\pi_{\mathfrak{p}}^{-\operatorname{ord}_{\mathfrak{p}}(\alpha)} \in (\mathfrak{o}_{\mathfrak{p}}^{\times})^2\\ &\Longleftrightarrow a \in (\mathfrak{o}_{\mathfrak{p}}^{\times})^2\\ &\Longleftrightarrow a^{(p-1)/2} \in 1+\pi_{\mathfrak{p}}\mathfrak{o}_{\mathfrak{p}}\\ &\qquad \text{(i.e. } a^{(p-1)/2} \in 1+p\mathbf{Z})\\ &\Longleftrightarrow \left(\tfrac{a}{p}\right) = 1.\end{aligned}$$

Finally, assume that p splits in $F/\mathbf{Q}$. Then $F_{\mathfrak{p}} \cong \mathbf{Q}_p$ and $\operatorname{ord}_{\mathfrak{p}}(p) = 1$. Since $\mathbf{Z}$ is dense in $\mathfrak{o}_{\mathfrak{p}}$, we can take $a_0 \in (\alpha + p^{\operatorname{ord}_{\mathfrak{p}}(\alpha)+1}\mathfrak{o}_{\mathfrak{p}}) \cap \mathbf{Z}$. Then $\alpha(1+p\mathfrak{o}_{\mathfrak{p}}) = \alpha + p^{\operatorname{ord}_{\mathfrak{p}}(\alpha)+1}\mathfrak{o}_{\mathfrak{p}} = a_0(1+p\mathfrak{o}_{\mathfrak{p}})$. Thus

$$\begin{aligned}\alpha \in (F_{\mathfrak{p}}^{\times})^2 &\Longleftrightarrow a_0 \in (F_{\mathfrak{p}}^{\times})^2\\ &\qquad \text{(i.e. } a = a_0 p^{-\operatorname{ord}_{\mathfrak{p}}(\alpha)} \in (\mathfrak{o}_{\mathfrak{p}}^{\times})^2)\\ &\Longleftrightarrow a^{(p-1)/2} \in 1+\pi_{\mathfrak{p}}\mathfrak{o}_{\mathfrak{p}}\\ &\qquad \text{(i.e. } a^{(p-1)/2} \in 1+p\mathbf{Z})\\ &\Longleftrightarrow \left(\tfrac{a}{p}\right) = 1.\end{aligned}$$

This completes the proof. □

Lemma 3.3. *Let $F = \mathbf{Q}(\sqrt{m})$ be a real quadratic field with a square-free integer m satisfying $m \equiv 1 \pmod{4\mathbf{Z}}$. Let p be an odd prime number in $\mathbf{Q}$ such that $\left(\frac{m}{p}\right) = 1$, and $\mathfrak{p}$ the prime ideal of F that lies above p. Let r be an integer such that $\mathfrak{p} = [p, r+\omega]$. Then, for $u \in \mathbf{Z}$ and $1 \le j \in \mathbf{Z}$, $u \equiv \sqrt{m} \pmod{p^j\mathfrak{o}_{\mathfrak{p}}}$ if and only if $u^2 \equiv m \pmod{p^j\mathbf{Z}}$ and $u \equiv -2r-1 \pmod{p\mathbf{Z}}$.*

Proof: We first prove that $u \equiv \sqrt{m}$ or $-\sqrt{m} \pmod{p^j\mathfrak{o}_{\mathfrak{p}}}$ if and only if $u^2 \equiv m \pmod{p^j\mathbf{Z}}$ for $u \in \mathbf{Z}$ and $1 \le j \in \mathbf{Z}$. If $u - \sqrt{m} \in p^j\mathfrak{o}_{\mathfrak{p}}$ or $u + \sqrt{m} \in p^j\mathfrak{o}_{\mathfrak{p}}$, then $u^2 - m = (u-\sqrt{m})(u+\sqrt{m}) \in p^j\mathfrak{o}_{\mathfrak{p}}$, and hence $u^2 - m \in p^j\mathfrak{o}_{\mathfrak{p}} \cap \mathbf{Z} = p^j\mathbf{Z}$. Conversely, $X^2 \equiv m \pmod{p^j\mathbf{Z}}$ has exactly two solutions modulo $p^j\mathbf{Z}$, since $p \ne 2$ and $\left(\frac{m}{p}\right) = 1$. Since p is odd and $\sqrt{m} \in \mathfrak{o}_{\mathfrak{p}}^{\times}$, we have $\sqrt{m} \not\equiv -\sqrt{m} \pmod{p^j\mathfrak{o}_{\mathfrak{p}}}$. Thus if $u \in \mathbf{Z}$ satisfies the condition $u^2 \equiv m \pmod{p^j\mathbf{Z}}$, then $u \equiv \sqrt{m}$ or $-\sqrt{m} \pmod{p^j\mathfrak{o}_{\mathfrak{p}}}$. Now, $2^{-1}(2r+1+\sqrt{m}) = r+\omega \in \mathfrak{p} \subset p\mathfrak{o}_{\mathfrak{p}}$. Thus $2r+1+\sqrt{m} \in p\mathfrak{o}_{\mathfrak{p}}$, that is, $\sqrt{m} \equiv -2r-1 \pmod{p\mathfrak{o}_{\mathfrak{p}}}$. Therefore, we obtain our assertion. □

From the two lemmas above, we obtain the following proposition:

Proposition 3.4. *Let F and m be as in Lemma 3.3, and $0 \gg \alpha \in \mathfrak{o}_F$. Set $K = F(\sqrt{\alpha})$, and $\alpha = a_1 + a_2\omega$ with $a_1, a_2 \in \mathbf{Z}$. Let $\mathfrak{p}$ be an odd prime ideal of F satisfying $\mathfrak{p} \nmid D_{K/F}$, and p the prime number in $\mathbf{Q}$ that lies below $\mathfrak{p}$. Put $t = \operatorname{ord}_{\mathfrak{p}}(\alpha)$. In particular, if p splits in $F/\mathbf{Q}$, we set $\mathfrak{p} = [p, r+\omega]$ with $r \in \mathbf{Z}$, $l = \operatorname{ord}_p(\gcd(a_1, a_2))$, and take $u \in \mathbf{Z}$ such that $u^2 \equiv m \pmod{p^{t-l+1}\mathbf{Z}}$ and $u \equiv -2r-1 \pmod{p\mathbf{Z}}$. Set*

$$a = \begin{cases} p^{-2t}\left(a_1^2 + a_1a_2 + a_2^2(1-m)/4\right) \\ \qquad\qquad \text{if } p \text{ remains prime in } F/\mathbf{Q},\\ (mp^{-2})^{t/2}\left(a_1 - a_2(p-1)/2\right) \\ \qquad\qquad \text{if } p \text{ ramifies in } F/\mathbf{Q},\\ p^{-t}\left((p+1)/2\right)\left(2a_1 + a_2(1+u)\right) \\ \qquad\qquad \text{if } p \text{ splits in } F/\mathbf{Q}.\end{cases}$$

Then we have

$$\left(\tfrac{K}{\mathfrak{p}}\right) = \left(\tfrac{a}{p}\right).$$

Proof: If p remains prime in $F/\mathbf{Q}$, our assertion follows immediately from Lemma 3.2. Next we assume that p ramifies in $F/\mathbf{Q}$. Then $\mathfrak{p} = [p, (p-1)/2+\omega]$. Since $\sqrt{m}p^{-1} \in F^{\times}$, $\operatorname{ord}_{\mathfrak{p}}(\sqrt{m}p^{-1}) = -1$, and $\operatorname{ord}_{\mathfrak{q}}(\sqrt{m}p^{-1}) \ge 0$ for any $\mathfrak{q} \in \mathbf{h} - \{\mathfrak{p}\}$, we see that $\pi_{\mathfrak{p}} = \sqrt{m}^{-1}p$ is a prime element of $F_{\mathfrak{p}}$ and $\alpha\pi_{\mathfrak{p}}^{-t} = \alpha(mp^{-2})^{t/2} \in \mathfrak{o}_F$. Thus $a_1(mp^{-2})^{t/2}, a_2(mp^{-2})^{t/2} \in \mathbf{Z}$, and hence $\alpha\pi_{\mathfrak{p}}^{-t} + \pi_{\mathfrak{p}}\mathfrak{o}_{\mathfrak{p}} \supset \alpha(mp^{-2})^{t/2} + \mathfrak{p} \ni (mp^{-2})^{t/2}(a_1 - a_2(p-1)/2)$. Therefore, we obtain a in Lemma 3.2 in this case. Now assume that p splits in $F/\mathbf{Q}$. By Lemma 3.3, we have $u \equiv \sqrt{m} \pmod{p^{t-l+1}\mathfrak{o}_{\mathfrak{p}}}$. Since $p^l \mid a_2$, we have $a_2u \equiv a_2\sqrt{m} \pmod{p^{t+1}\mathfrak{o}_{\mathfrak{p}}}$, and hence

$$\begin{aligned}\left((p+1)/2\right)\left(2a_1 + a_2(1+u)\right) &\equiv \left((p+1)/2\right)\\ &\qquad \cdot \left(2a_1 + a_2(1+\sqrt{m})\right)\\ &= (p+1)\alpha\\ &\equiv \alpha \pmod{p^{t+1}\mathfrak{o}_{\mathfrak{p}}}.\end{aligned}$$

Thus $((p+1)/2)(2a_1 + a_2(1+u)) \in (\alpha + p^{t+1}\mathfrak{o}_{\mathfrak{p}}) \cap \mathbf{Z}$. Therefore, our assertion follows from Lemma 3.2. □

Remark 3.5. Note that we can find $u \in \mathbf{Z}$ satisfying

$$u^2 \equiv m \pmod{p^{t-l+1}\mathbf{Z}} \tag{3–3}$$

when $\left(\frac{m}{p}\right) = 1$. Then we have $u \equiv -2r-1$ or $2r+1 \pmod{p\mathbf{Z}}$, as we see in the proof of Lemma 3.3. Finding a solution u of (3–3) can be reduced to mod p calculation by the following procedure: Let $u = c_0 + c_1p + \cdots + c_{t-l}p^{t-l}$ with $0 \le c_j \le p-1$, and set $u_j = c_0 + c_1p + \cdots + c_jp^j$ for $0 \le j \le t-l$. Since $m \equiv u_j^2 = (u_{j-1} + c_jp^j)^2 \equiv u_{j-1}^2 + 2u_{j-1}c_jp^j \pmod{p^{j+1}\mathbf{Z}}$, we have

$$m \equiv c_0^2 \pmod{p\mathbf{Z}}, \tag{3–4a}$$

$$p^{-j}(m - u_{j-1}^2) \equiv 2u_{j-1}c_j \pmod{p\mathbf{Z}} \tag{3–4b}$$

for $1 \le j \le t-l$. Thus we can determine $c_0, \dots, c_{t-l}$ inductively by (3–4a, b).

From Proposition 3.4 and Remark 3.5, we can immediately determine $\left(\frac{K}{\mathfrak{p}}\right)$ for every *odd* prime ideal $\mathfrak{p}$ satisfying $\mathfrak{p} \nmid D_{K/F}$.

3.3 Determination of $D_{\mathfrak{p}}$ and $\left(\frac{K}{\mathfrak{p}}\right)$ for an Even Prime $\mathfrak{p}$

Next we give a method for determining $D_{\mathfrak{p}}$ and $\left(\frac{K}{\mathfrak{p}}\right)$ for an *even* prime ideal $\mathfrak{p}$ of F.

Proposition 3.6. *Let F, m, α, K, a_1, and a_2 be as in Proposition* 3.4. *Let $\mathfrak{p}$ be an even prime ideal of F. Assume $\mathfrak{p}^2 \nmid (\alpha)_F$. If $m \equiv 1 \pmod{8\mathbf{Z}}$, we set $l = (m-1)/8$ and $\mathfrak{p} = [2, r+\omega]$, where $r = 0$ or 1; we also set*

$$A_m = \{(a,b) \in \mathbf{Z}^2 \mid a + b(2l(-1)^r + r) - 1 \in 8\mathbf{Z}\},$$
$$A'_m = \{(a,b) \in \mathbf{Z}^2 \mid a - b(2l - r) - 1 \in 4\mathbf{Z}\}.$$

If $m \equiv 5 \pmod{8\mathbf{Z}}$, we set $l = (m-5)/8$,

$$\begin{aligned} A_m = \{(a,b) \in \mathbf{Z}^2 \mid\ & (a-1,b) \in 4\mathbf{Z} \times 8\mathbf{Z} \\ & \text{or } (a-2l-2, b-3) \in 8\mathbf{Z} \times 4\mathbf{Z} \\ & \text{or } (a-2l-1, b-1) \in (8\mathbf{Z})^2 \\ & \text{or } (a-2l-5, b-5) \in (8\mathbf{Z})^2\}, \\ A'_m = \{(a,b) \in \mathbf{Z}^2 \mid\ & (a-1,b) \in (4\mathbf{Z})^2 \\ & \text{or } (a-2l-2, b-3) \in (4\mathbf{Z})^2 \\ & \text{or } (a-2l-1, b-1) \in (4\mathbf{Z})^2\}. \end{aligned}$$

Then we have

$$\left(\frac{K}{\mathfrak{p}}\right) = \begin{cases} 1 & \text{if } \mathfrak{p} \nmid (\alpha)_F \text{ and } (a_1, a_2) \in A_m, \\ -1 & \text{if } \mathfrak{p} \nmid (\alpha)_F,\ (a_1, a_2) \notin A_m, \\ & \quad \text{and } (a_1, a_2) \in A'_m, \\ 0 & \text{otherwise}, \end{cases}$$

and

$$D_{\mathfrak{p}} = \begin{cases} \mathfrak{o}_F & \text{if } \left(\frac{K}{\mathfrak{p}}\right) \neq 0, \\ \mathfrak{p}^2 & \text{if } \left(\frac{K}{\mathfrak{p}}\right) = 0 \text{ and } \mathfrak{p} \nmid (\alpha)_F, \\ \mathfrak{p}^3 & \text{if } \mathfrak{p} \mid (\alpha)_F. \end{cases}$$

Proof: We set

$$\begin{aligned} f = \max\{j \in \mathbf{Z} \mid\ & 0 \le j \le 3, \\ & \text{there exists } \gamma \in \mathfrak{o}_F \text{ such that } \gamma^2 \equiv \alpha \pmod{\mathfrak{p}^j}\}. \end{aligned}$$

Then, by [Okazaki 91, Proposition 3], we have

$$\left(\frac{K}{\mathfrak{p}}\right) = \begin{cases} 1 & \text{if } f = 3, \\ -1 & \text{if } f = 2, \\ 0 & \text{if } f \le 1, \end{cases} \qquad D_{\mathfrak{p}} = \begin{cases} \mathfrak{p}^{2-2[f/2]} & \text{if } \mathfrak{p} \nmid (\alpha)_F, \\ \mathfrak{p}^3 & \text{if } \mathfrak{p} \mid (\alpha)_F. \end{cases}$$

When $\mathfrak{p} \mid (\alpha)_F$, we have $\alpha + \mathfrak{p}^2 \subset \pi_{\mathfrak{p}} \mathfrak{o}_{\mathfrak{p}}^{\times}$, thus $f = 1$, and hence $\left(\frac{K}{\mathfrak{p}}\right) = 0$; moreover, $D_{\mathfrak{p}} = \mathfrak{p}^3$. Thus our proposition holds in this case. Therefore, we may assume $\mathfrak{p} \nmid (\alpha)_F$. To prove our proposition, we need to show only that

$$f = 3 \Longleftrightarrow (a_1, a_2) \in A_m,$$
$$f \ge 2 \Longleftrightarrow (a_1, a_2) \in A'_m.$$

We first assume that $m \equiv 1 \pmod{8\mathbf{Z}}$. Then $\mathrm{ord}_{\mathfrak{p}}(\alpha) = 0$, and $\{t \in \mathbf{Z} \mid 0 \le t < 2^j,\ 2 \nmid t\}$ is a complete set of representatives of $(\mathfrak{o}_F/\mathfrak{p}^j)^{\times}$ for $j \ge 1$. Therefore, for $j \ge 1$, there exists $\gamma \in \mathfrak{o}_F$ such that $\gamma^2 \equiv \alpha \pmod{\mathfrak{p}^j}$ if and only if there exists $t \in \{t \in \mathbf{Z} \mid 0 \le t < 2^j,\ 2 \nmid t\}$ such that $\alpha - t^2 \in \mathfrak{p}^j$. Now we have

$$\mathfrak{p}^2 = [4, 2l - r + \omega], \qquad \mathfrak{p}^3 = [8, -2l(-1)^r - r + \omega].$$

Since $\alpha - t^2 = \big(a_1 + a_2(2l(-1)^r + r) - t^2\big) + a_2\big(-2l(-1)^r - r + \omega\big)$ and $1 \equiv 1^2 \equiv 3^2 \equiv 5^2 \equiv 7^2 \pmod{8\mathbf{Z}}$, we have

$$\begin{aligned} f = 3 &\Longleftrightarrow \text{there exists } t \in \{1,3,5,7\} \\ &\qquad \text{such that } \alpha - t^2 \in \mathfrak{p}^3 \\ &\Longleftrightarrow (a_1, a_2) \in A_m. \end{aligned}$$

Moreover, since $\alpha - t^2 = (a_1 - a_2(2l-r) - t^2) + a_2(2l - r + \omega)$ and $1 \equiv 1^2 \equiv 3^2 \pmod{4\mathbf{Z}}$, we have

$$\begin{aligned} f \ge 2 &\Longleftrightarrow \text{there exists } t \in \{1,3\} \text{ such that } \alpha - t^2 \in \mathfrak{p}^2 \\ &\Longleftrightarrow (a_1, a_2) \in A'_m. \end{aligned}$$

Thus the assertion is proved in this case. Now assume that $m \equiv 5 \pmod{8\mathbf{Z}}$. Then $\mathrm{ord}_{\mathfrak{p}}(\alpha) = 0$, and $\{u + v\omega \mid 0 \le u, v < 2^j,\ (u,v) \notin (2\mathbf{Z})^2\}$ is a complete set of representatives of $(\mathfrak{o}_F/\mathfrak{p}^j)^{\times}$ for $j \ge 1$. Thus for $j \ge 1$, there exists $\gamma \in \mathfrak{o}_F$ such that $\gamma^2 \equiv \alpha \pmod{\mathfrak{p}^j}$ if and only if there exists $(u,v) \in \{(u,v) \in \mathbf{Z}^2 \mid 0 \le u, v < 2^j,\ (u,v) \notin (2\mathbf{Z})^2\}$ such that $\alpha - (u + v\omega)^2 \in \mathfrak{p}^j = [2^j, 2^j\omega]$. Since $\alpha - (u+v\omega)^2 = (a_1 - u^2 - (2l+1)v^2) + (a_2 - 2uv - v^2)\omega$ and

$$\begin{aligned} &(a_1 - u^2 - (2l+1)v^2,\ a_2 - 2uv - v^2) \\ &\quad \equiv \begin{cases} (a_1 - 1, a_2) & \text{if } 2 \nmid u \text{ and } 4 \mid v, \\ (a_1 - 5, a_2) & \text{if } 2 \nmid u,\ 2 \mid v, \text{ and } 4 \nmid v, \\ (a_1 - 2l - 2, a_2 - 3) & \text{if } 2 \nmid u,\ 2 \nmid v, \text{ and } 4 \mid u - v, \\ (a_1 - 2l - 2, a_2 - 7) & \text{if } 2 \nmid u,\ 2 \nmid v, \text{ and } 4 \nmid u - v, \\ (a_1 - 2l - 1, a_2 - 1) & \text{if } 4 \mid u \text{ and } 2 \nmid v, \\ (a_1 - 2l - 5, a_2 - 5) & \text{if } 2 \mid u,\ 4 \nmid u, \text{ and } 2 \nmid v, \end{cases} \\ &\qquad\qquad \pmod{8\mathbf{Z}}, \end{aligned}$$

we have

$$\begin{aligned} f = 3 &\Longleftrightarrow \text{there exists } (u,v) \in \{(u,v) \mid 0 \le u, v \le 7, \\ &\qquad\qquad (u,v) \notin (2\mathbf{Z})^2\} \\ &\qquad \text{such that } \alpha - (u + v\omega)^2 \in [8, 8\omega] \\ &\Longleftrightarrow (a_1, a_2) \in A_m. \end{aligned}$$

Moreover, since

$$(a_1 - u^2 - (2l+1)v^2, a_2 - 2uv - v^2) \equiv \begin{cases} (a_1 - 1, a_2) & \text{if } 2 \nmid u \text{ and } 2 \mid v, \\ (a_1 - 2l - 2, a_2 - 3) & \text{if } 2 \nmid u \text{ and } 2 \nmid v, \\ (a_1 - 2l - 1, a_2 - 1) & \text{if } 2 \mid u \text{ and } 2 \nmid v, \end{cases} \pmod{4\mathbf{Z}},$$

we have

$$\begin{aligned} f \geq 2 &\Longleftrightarrow \text{there exists } (u,v) \in \{(u,v) \mid 0 \leq u, v \leq 3, \\ &\qquad (u,v) \notin (2\mathbf{Z})^2\} \\ &\quad \text{such that } \alpha - (u + v\omega)^2 \in [4, 4\omega] \\ &\Longleftrightarrow (a_1, a_2) \in A'_m. \end{aligned}$$

This completes the proof. □

Remark 3.7. Let F, m, α, and K be as in Proposition 3.4. Let $\mathfrak{p}$ be an *even* prime ideal of F. When $\mathfrak{p}^2 \mid (\alpha)_F$, we cannot apply this case to Proposition 3.6. However we can take α_1, instead of α, satisfying

$$K = F(\sqrt{\alpha_1}), \quad \alpha_1 \in \mathfrak{o}_F, \quad \mathfrak{p}^2 \nmid (\alpha_1)_F \tag{3–5}$$

as follows:

(i) If $m \equiv 1 \pmod{8\mathbf{Z}}$, then we take $\gamma \in \mathfrak{o}_F$ such that $(\gamma)_F = (\mathfrak{p}^\sigma)^{h_F} \in P_F$, and put

$$\alpha_1 = \alpha(2^{-1}\gamma)^{2[\operatorname{ord}_\mathfrak{p}(\alpha)/2]}.$$

Then α_1 satisfies (3–5), since $(2^{-1}\gamma)_F = \mathfrak{p}^{-1}(\mathfrak{p}^\sigma)^{h_F - 1}$.

(ii) If $m \equiv 5 \pmod{8\mathbf{Z}}$, set

$$\alpha_1 = \alpha \cdot 2^{-2[\operatorname{ord}_\mathfrak{p}(\alpha)/2]}.$$

Then α_1 satisfies (3–5), since $\mathfrak{p} = (2)_F$.

Thus, by Proposition 3.6 and Remark 3.7, we can immediately determine $D_\mathfrak{p}$ and $\left(\frac{K}{\mathfrak{p}}\right)$ for an *even* prime ideal $\mathfrak{p}$.

3.4 Hecke L-Values

Finally, we explain the method for computing $L_F(0, \chi_{K/F})$ that was established by [Shintani 76] for totally real algebraic number fields F. [Okazaki 91] deals with Shintani's formula for the case of real quadratic fields F, and we observe that the ideal character corresponding to K/F is expressed by the Legendre symbols and the Hilbert symbols. Applying this result to Shintani's formula, we obtain Formula (3–6) with simple calculation. We note that the conductor of $\chi_{K/F}$ is equal to $D_{K/F}$ (cf. [Weil 67, Chapter XIII, Theorem 9]), and we can determine $D_{K/F}$ by Proposition 3.1 and Proposition 3.6.

Let F, m, α, K, a_1, and a_2 be as in Proposition 3.4. Let ε be the fundamental unit of F that is greater than 1, and set

$$\varepsilon_+ = \begin{cases} \varepsilon & \text{if } \varepsilon \gg 0, \\ \varepsilon^2 & \text{otherwise.} \end{cases}$$

We take $e, e' \in \mathbf{Z}$ such that $\varepsilon_+ = e + e'\omega$. Let $\mathfrak{a}_1, \ldots, \mathfrak{a}_{h_F^+}$ be a complete set of representatives of $\mathrm{Cl}^+(\mathrm{F})$ such that $\mathfrak{a}_\mu \subset \mathfrak{o}_F$ for all μ. For $1 \leq \mu \leq h_F^+$, we can determine uniquely integers d_μ, d'_μ, and d''_μ such that

$$d_\mu[d'_\mu, d''_\mu + \omega] = \mathfrak{a}_\mu D_{K/F},$$

$d_\mu, d'_\mu > 0$, and $0 \leq d''_\mu < d'_\mu$; we take integers s_μ, s'_μ, and Q_μ as follows:

(i) If $\mathfrak{a}_\mu D_{K/F} \in P(F)$, then take s_μ, s'_μ such that

$$\left(s_\mu + s'_\mu\omega\right)_F = \mathfrak{a}_\mu D_{K/F},$$

and set

$$Q_\mu = 1.$$

(ii) If $\mathfrak{a}_\mu D_{K/F} \notin P(F)$, then we can take an odd prime ideal $\mathfrak{q}_\mu$ of F such that $\mathfrak{q}_\mu$ splits in $F/\mathbf{Q}$, $\gcd(\mathfrak{q}_\mu, (\alpha)_F) = \mathfrak{o}_F$, and $\mathfrak{q}_\mu\mathfrak{a}_\mu D_{K/F} \in P(F)$. Then $\mathfrak{q}_\mu = [q_\mu, r_\mu + \omega]$ with $q_\mu, r_\mu \in \mathbf{Z}$. We take s_μ, s'_μ such that

$$\left(s_\mu + s'_\mu\omega\right)_F = \mathfrak{q}_\mu\mathfrak{a}_\mu D_{K/F},$$

and set

$$Q_\mu = \left(\tfrac{a_1 - a_2 r_\mu}{q_\mu}\right).$$

Moreover, for $1 \leq i \leq d_\mu$ and $1 \leq j \leq e' d_\mu d'_\mu$, we take the integer $1 \leq r_{\mu ij} \leq e' d_\mu d'_\mu$ such that

$$r_{\mu ij} \equiv e' d'_\mu i - (e + e'(d''_\mu + 1))j \pmod{e' d_\mu d'_\mu \mathbf{Z}},$$

and we set

$$\begin{aligned} B_{\mu ij} &= 4^{-1}(e' d_\mu d'_\mu)^{-2}\left((2e + e')(r_{\mu ij}^2 + j^2) + 4 r_{\mu ij} j\right) \\ &\quad - 4^{-1}(e' d_\mu d'_\mu)^{-1}(2e + e' + 2)(r_{\mu ij} + j) \\ &\quad + 12^{-1}(2e + e' + 3), \\ u_{\mu ij} &= (e' d_\mu d'_\mu)^{-1}\left(r_{\mu ij} s_\mu + j\left(e s_\mu + e' s'_\mu \frac{m-1}{4}\right)\right), \\ v_{\mu ij} &= (e' d_\mu d'_\mu)^{-1}\left(r_{\mu ij} s'_\mu + j(e s'_\mu + e' s_\mu + e' s'_\mu)\right). \end{aligned}$$

Note that $u_{\mu ij}, v_{\mu ij} \in \mathbf{Z}$. Now, for an *odd* prime ideal $\mathfrak{p}$ of F and $u+v\omega \in \mathfrak{o}_F$, we set

$$\chi_{\mathfrak{p}}(u+v\omega) = \begin{cases} \left(\frac{N_{F/\mathbf{Q}}(u+v\omega)}{p}\right) & \text{if } \mathfrak{p} \text{ remains prime in } F/\mathbf{Q}, \\ \left(\frac{u-vr}{p}\right) & \text{otherwise,} \end{cases}$$

where p is the prime number in $\mathbf{Q}$ that lies below $\mathfrak{p}$, and r an integer satisfying $\mathfrak{p} = [p, r+\omega]$. For an *even* prime ideal $\mathfrak{p}$ of F dividing $D_{K/F}$ and $\beta \in \mathfrak{o}_F$, we set

$$\chi_{\mathfrak{p}}(\beta) = \begin{cases} (\beta, \alpha)_{F_{\mathfrak{p}}} & \text{if } \mathfrak{p} \nmid (\beta)_F, \\ 0 & \text{otherwise,} \end{cases}$$

where $(\ ,\ )_{F_{\mathfrak{p}}}$ is the Hilbert symbol (and α is retained as in Proposition 3.4). Then we obtain

$$\begin{aligned} L_F(0, \chi_{K/F}) = \sum_{\mu=1}^{h_F^+} \operatorname{sgn}\bigl(N_{F/\mathbf{Q}}(s_\mu + s'_\mu\omega)\bigr) Q_\mu \\ \cdot \Biggl(\sum_{i=1}^{d_\mu} \sum_{j=1}^{e' d_\mu d'_\mu} B_{\mu ij} \prod_{\substack{\mathfrak{p} \mid D_{K/F} \\ \mathfrak{p} \in \mathbf{h}}} \chi_{\mathfrak{p}}(u_{\mu ij} + v_{\mu ij}\omega) \\ + \sum_{l=1}^{d_\mu} \frac{2l - d_\mu}{2d_\mu} \prod_{\substack{\mathfrak{p} \mid D_{K/F} \\ \mathfrak{p} \in \mathbf{h}}} \chi_{\mathfrak{p}}\bigl(l d_\mu^{-1}(s_\mu + s'_\mu\omega)\bigr) \Biggr). \end{aligned} \tag{3–6}$$

Thus, if we can compute the Hilbert symbol $(\beta, \alpha)_{F_{\mathfrak{p}}}$ for an *even* prime $\mathfrak{p}$ and $\alpha, \beta \in \mathfrak{o}_F - \{0\}$ satisfying $\mathfrak{p} \nmid (\beta)_F$, we can determine $L_F(0, \chi_{K/F})$ by (3–6). We note that the Hilbert symbol

$$(\ ,\ )_{F_{\mathfrak{p}}} : F_{\mathfrak{p}}^\times \times F_{\mathfrak{p}}^\times \longrightarrow \{\pm 1\}$$

satisfies

$$(a, b)_{F_{\mathfrak{p}}} = (b, a)_{F_{\mathfrak{p}}}, \tag{3–7a}$$

$$(a, bc)_{F_{\mathfrak{p}}} = (a, b)_{F_{\mathfrak{p}}} (a, c)_{F_{\mathfrak{p}}}, \tag{3–7b}$$

for $a, b, c \in F_{\mathfrak{p}}^\times$ and naturally induces the mapping

$$F_{\mathfrak{p}}^\times / (F_{\mathfrak{p}}^\times)^2 \times F_{\mathfrak{p}}^\times / (F_{\mathfrak{p}}^\times)^2 \longrightarrow \{\pm 1\} \tag{3–8}$$

(for the Hilbert symbol, see [Neukirch 86], for example).

3.5 Hilbert Symbol for $m \equiv 1 \pmod 8$

We first give a method for computing the Hilbert symbol when $m \equiv 1 \pmod{8\mathbf{Z}}$.

Lemma 3.8. *Let $\mu \in \mathbf{Z}_2^\times$ satisfying $\mu^2 \in \mathbf{Z}$. Then $u \equiv \mu \pmod{2^j \mathbf{Z}_2}$ if and only if $u^2 \equiv \mu^2 \pmod{2^{j+1}\mathbf{Z}}$ and $u \equiv \mu \pmod{4\mathbf{Z}_2}$ for $u \in \mathbf{Z}$ and $2 \le j \in \mathbf{Z}$.*

Proof: Suppose $u \equiv \mu \pmod{2^j \mathbf{Z}_2}$ with $u \in \mathbf{Z}$ and $2 \le j \in \mathbf{Z}$. Then $u \equiv \mu \pmod{4\mathbf{Z}_2}$. Since $u + \mu \equiv 2\mu \pmod{4\mathbf{Z}_2}$ and $\mu \in \mathbf{Z}_2^\times$, we have $2^{-1}(u+\mu) \in \mu + 2\mathbf{Z}_2 = \mathbf{Z}_2^\times$, and hence $\operatorname{ord}_2(u+\mu) = 1$. Thus $u^2 - \mu^2 = (u-\mu)(u+\mu) \in 2^{j+1}\mathbf{Z}_2$, that is, $u^2 - \mu^2 \in 2^{j+1}\mathbf{Z}$. Conversely, if $u^2 \equiv \mu^2 \pmod{2^{j+1}\mathbf{Z}}$ and $u \equiv \mu \pmod{4\mathbf{Z}_2}$, then $\operatorname{ord}_2(u+\mu) = 1$, and hence $u - \mu = (u^2 - \mu^2)(u+\mu)^{-1} \in 2^j \mathbf{Z}_2$. □

Lemma 3.9. *Let $F = \mathbf{Q}(\sqrt{m})$ be a real quadratic field with a square-free integer m satisfying $m \equiv 1 \pmod{8\mathbf{Z}}$. Let $\mathfrak{p} = [2, r+\omega]$ be an even prime ideal of F, where $r = 0$ or 1. Then, for $u \in \mathbf{Z}$ and $2 \le j \in \mathbf{Z}$, it follows that $u \equiv \sqrt{m} \pmod{2^j \mathfrak{o}_{\mathfrak{p}}}$ if and only if $u^2 \equiv m \pmod{2^{j+1}\mathbf{Z}}$ and $u \equiv -2r - 1 \pmod{4\mathbf{Z}}$.*

Proof: Since $m \equiv 1 \pmod{8\mathbf{Z}}$, we have $F_{\mathfrak{p}} \cong \mathbf{Q}_2$. Thus, by Lemma 3.8, we see that $u \equiv \sqrt{m} \pmod{2^j \mathfrak{o}_{\mathfrak{p}}}$ if and only if $u^2 \equiv m \pmod{2^{j+1}\mathbf{Z}}$ and $u \equiv \sqrt{m} \pmod{4\mathfrak{o}_{\mathfrak{p}}}$ for $u \in \mathbf{Z}$ and $2 \le j \in \mathbf{Z}$. Since $2^{-1}(2r + 1 + \sqrt{m}) = r + \omega \in \mathfrak{p} \subset 2\mathfrak{o}_{\mathfrak{p}}$, we have $\sqrt{m} \equiv -2r - 1 \pmod{4\mathfrak{o}_{\mathfrak{p}}}$. Thus $u \equiv \sqrt{m} \pmod{4\mathfrak{o}_{\mathfrak{p}}}$ if and only if $u \equiv -2r - 1 \pmod{4\mathbf{Z}}$. This completes the proof. □

From the two lemmas above, we obtain the following result:

Proposition 3.10. *Let F, m, $\mathfrak{p}$, and r be as in Lemma 3.9. Let $\beta_1, \beta_2 \in \mathfrak{o}_F - \{0\}$, and set $\beta_j = c_j + d_j\omega$ with $c_j, d_j \in \mathbf{Z}$. We take $u_j \in \mathbf{Z}$ such that $u_j^2 \equiv m \pmod{2^{\operatorname{ord}_{\mathfrak{p}}(\beta_j) - l_j + 5}\mathbf{Z}}$ and $u_j \equiv -2r - 1 \pmod{4\mathbf{Z}}$, where $l_j = \min\{\operatorname{ord}_2(2c_j + d_j), \operatorname{ord}_2(d_j)\}$. We set $t_j = 2^{-\operatorname{ord}_{\mathfrak{p}}(\beta_j)}(c_j + d_j(1+u_j)/2)$. Then we have*

$$\begin{aligned} &(\beta_1, \beta_2)_{F_{\mathfrak{p}}} \\ &= (-1)^{(t_1-1)(t_2-1)/4 + \operatorname{ord}_{\mathfrak{p}}(\beta_1)(t_2^2-1)/8 + \operatorname{ord}_{\mathfrak{p}}(\beta_2)(t_1^2-1)/8}. \end{aligned}$$

Proof: Since $\mathfrak{o}_{\mathfrak{p}}^\times = (1 + 2^3\mathfrak{o}_{\mathfrak{p}}) \cup (3 + 2^3\mathfrak{o}_{\mathfrak{p}}) \cup (5 + 2^3\mathfrak{o}_{\mathfrak{p}}) \cup (7 + 2^3\mathfrak{o}_{\mathfrak{p}})$, we have $(\mathfrak{o}_{\mathfrak{p}}^\times)^2 \subset 1 + 2^3\mathfrak{o}_{\mathfrak{p}}$. Conversely, for any element $1 + 2^3 y \in 1 + 2^3\mathfrak{o}_{\mathfrak{p}}$, we can take a root $x \in \mathfrak{o}_{\mathfrak{p}}$ of the polynomial $2X^2 + X - y$ by Hensel's lemma, then $1 + 2^3 y = (1 + 2^2 x)^2$, and hence $1 + 2^3\mathfrak{o}_{\mathfrak{p}} \subset (\mathfrak{o}_{\mathfrak{p}}^\times)^2$. Therefore,

$$(\mathfrak{o}_{\mathfrak{p}}^\times)^2 = 1 + 2^3\mathfrak{o}_{\mathfrak{p}}. \tag{3–9}$$

Now set $e_j = \operatorname{ord}_{\mathfrak{p}}(\beta_j)$. Since $\beta_j = 2^{l_j - 1}(2^{-l_j}(2c_j + d_j) + 2^{-l_j} d_j \sqrt{m})$, we have $e_j \ge l_j - 1$. Thus $u_j \equiv \sqrt{m} \pmod{2^{e_j - l_j + 4}\mathfrak{o}_{\mathfrak{p}}}$ by Lemma 3.9, and hence $d_j(1 + u_j)/2 \equiv d_j\omega \pmod{2^{e_j+3}\mathfrak{o}_{\mathfrak{p}}}$. Therefore, $t_j = 2^{-e_j}(c_j + d_j(1+u_j)/2) \equiv 2^{-e_j}\beta_j \pmod{2^3\mathfrak{o}_{\mathfrak{p}}}$, that is, $t_j(\mathfrak{o}_{\mathfrak{p}}^\times)^2 =$

$2^{-e_j}\beta_j(\mathfrak{o}_\mathfrak{p}^\times)^2$. It follows that $(\beta_1,\beta_2)_{F_\mathfrak{p}} = (2^{e_1}t_1, 2^{e_2}t_2)_{F_\mathfrak{p}}$. Since $F_\mathfrak{p} \cong \mathbf{Q}_2$, we have

$$(2^{e_1}t_1, 2^{e_2}t_2)_{F_\mathfrak{p}} = (-1)^{(t_1-1)(t_2-1)/4+e_1(t_2^2-1)/8+e_2(t_1^2-1)/8}$$

(cf. [Neukirch 86, Chapter III, Theorem 5.6], for example). This completes the proof. □

Remark 3.11. It is well known that $X^2 \equiv a \pmod{2^j\mathbf{Z}}$ has exactly four solutions modulo $2^j\mathbf{Z}$ for $a \in 1+8\mathbf{Z}$ and $3 \le j \in \mathbf{Z}$. If x_j is one of the solutions, then all the solutions are x_j, $x_j + 2^{j-1}$, $-x_j$, and $-x_j + 2^{j-1}$, since x_j is odd. Now we can inductively determine x_j by

$$x_3 = 1,$$
$$x_j = \begin{cases} x_{j-1} & \text{if } 2 \mid 2^{-j+1}(x_{j-1}^2 - a), \\ x_{j-1} + 2^{j-2} & \text{otherwise,} \end{cases}$$

for $4 \le j \in \mathbf{Z}$. To prove this, we assume that x_{j-1} is a solution of $X^2 \equiv a \pmod{2^{j-1}\mathbf{Z}}$ with $j \ge 4$, and set

$$c = \begin{cases} 0 & \text{if } 2 \mid 2^{-j+1}(x_{j-1}^2 - a), \\ 1 & \text{otherwise,} \end{cases}$$

and $x_j = x_{j-1} + c2^{j-2}$. Since $2^{-j+1}(x_{j-1}^2 - a) \equiv c \equiv -x_{j-1}c \pmod{2\mathbf{Z}}$, we have $x_{j-1}^2 - a \equiv -x_{j-1}c2^{j-1} \pmod{2^j\mathbf{Z}}$. Thus

$$\begin{aligned} x_j^2 &= x_{j-1}^2 + x_{j-1}c2^{j-1} + c^2 2^{j+j-4} \\ &\equiv x_{j-1}^2 - (x_{j-1}^2 - a) = a \pmod{2^j\mathbf{Z}}. \end{aligned}$$

From this, we can easily find u_j of Proposition 3.10 and u of Proposition 3.14 below.

From Proposition 3.10 and Remark 3.11, we can immediately compute the Hilbert symbol when $m \equiv 1 \pmod{8\mathbf{Z}}$.

3.6 Hilbert Symbol for $m \equiv 5 \pmod{8\mathbf{Z}}$

Now we explain a method for computing the Hilbert symbol when $m \equiv 5 \pmod{8\mathbf{Z}}$.

By [Okazaki 91, Proposition 4], we have

Proposition 3.12. *Let $F = \mathbf{Q}(\sqrt{m})$ be a real quadratic field with a square-free integer m satisfying $m \equiv 5 \pmod{8\mathbf{Z}}$. Let $\mathfrak{p}$ be an even prime ideal of F. Put $\tau = 1 + 4\omega$ and $\xi = (m-9)/4 + \omega$. Then $F_\mathfrak{p}^\times/(F_\mathfrak{p}^\times)^2$ is generated by $-1, \tau, \xi, 2$, and we have*

$$(b,a)_{F_\mathfrak{p}} = \begin{cases} 1 & \text{if } (b,a) = (-1,-1) \\ & \quad \text{or } (-1,\tau) \text{ or } (-1,2) \\ & \quad \text{or } (\tau,\tau) \text{ or } (\tau,\xi) \\ & \quad \text{or } (\xi,2) \text{ or } (2,2), \\ -1 & \text{if } (b,a) = (-1,\xi) \text{ or } (\tau,2) \text{ or } (\xi,\xi). \end{cases}$$

From this proposition, we can take

$$\left\{(-1)^{i_1}\tau^{i_2}\xi^{i_3}2^{i_4} \;\middle|\; i_1,i_2,i_3,i_4 \in \{0,1\}\right\}$$

as a complete set of representatives of $F_\mathfrak{p}^\times/(F_\mathfrak{p}^\times)^2$. Thus for any β_1, β_2 in $\mathfrak{o}_F - \{0\}$, we can compute $(\beta_1,\beta_2)_{F_\mathfrak{p}}$ by (3–7a), (3–7b), (3–8), and Proposition 3.12, if we can find $l_{j1}, l_{j2}, l_{j3}, l_{j4} \in \{0,1\}$ such that $\beta_j(F_\mathfrak{p}^\times)^2 = (-1)^{l_{j1}}\tau^{l_{j2}}\xi^{l_{j3}}2^{l_{j4}}(F_\mathfrak{p}^\times)^2$ for $j = 0, 1$. Now we have $\beta_j(-1)^{i_{j1}}\tau^{i_{j2}}\xi^{i_{j3}}2^{-\operatorname{ord}_\mathfrak{p}(\beta_j)} \in \mathfrak{o}_F \cap \mathfrak{o}_\mathfrak{p}^\times$ for any $i_{j1}, i_{j2}, i_{j3} \in \{0,1\}$, and

$$\begin{aligned} &\beta_j(-1)^{i_{j1}}\tau^{i_{j2}}\xi^{i_{j3}}2^{-\operatorname{ord}_\mathfrak{p}(\beta_j)} \in (\mathfrak{o}_\mathfrak{p}^\times)^2 \\ &\qquad \Longleftrightarrow \beta_j(F_\mathfrak{p}^\times)^2 = (-1)^{i_{j1}}\tau^{i_{j2}}\xi^{i_{j3}}2^{i_{j4}}(F_\mathfrak{p}^\times)^2, \end{aligned}$$

where $i_{j4} = 0$ or 1 according as $2 \mid \operatorname{ord}_\mathfrak{p}(\beta_j)$ or $2 \nmid \operatorname{ord}_\mathfrak{p}(\beta_j)$. Thus, for an element δ of $\mathfrak{o}_F$, we wish to give an effective method to see that $\delta \in (\mathfrak{o}_\mathfrak{p}^\times)^2$. (In fact, this will be given in Proposition 3.14 below.)

For $\gamma \in (\mathbf{Q}_2)^2$, we denote by $\gamma^{1/2}$ the root of $X^2 - \gamma$ contained in $(\bigsqcup_{j=-\infty}^{\infty} 2^j(1+4\mathbf{Z}_2)) \cup \{0\}$. Note that ${\gamma_1}^{1/2}{\gamma_2}^{1/2} = (\gamma_1\gamma_2)^{1/2}$ for $\gamma_1, \gamma_2 \in (\mathbf{Q}_2)^2$.

Lemma 3.13. *Let F, m, and $\mathfrak{p}$ be as in Proposition 3.12. Let $\delta \in F_\mathfrak{p}$. Then $\delta \in (\mathfrak{o}_\mathfrak{p}^\times)^2$ if and only if*

(i) $N_{F_\mathfrak{p}/\mathbf{Q}_2}(\delta) \in (\mathbf{Z}_2^\times)^2$; *and*

(ii) $\operatorname{Tr}_{F_\mathfrak{p}/\mathbf{Q}_2}(\delta) + 2N_{F_\mathfrak{p}/\mathbf{Q}_2}(\delta)^{1/2}$, $\operatorname{Tr}_{F_\mathfrak{p}/\mathbf{Q}_2}(\delta) - 2N_{F_\mathfrak{p}/\mathbf{Q}_2}(\delta)^{1/2} \in (\mathbf{Z}_2)^2 \cup m(\mathbf{Z}_2)^2$.

Proof: In this proof, we abbreviate $\operatorname{Tr}_{F_\mathfrak{p}/\mathbf{Q}_2}$ and $N_{F_\mathfrak{p}/\mathbf{Q}_2}$ by Tr and N. Since $\mathfrak{o}_\mathfrak{p}$ is the topological closure of $\mathfrak{o}_F$ in $F_\mathfrak{p}$, we have

$$\mathfrak{o}_\mathfrak{p} = \{2^{-1}(x + y\sqrt{m}) \mid x, y \in \mathbf{Z}_2,\ x - y \in 2\mathbf{Z}_2\}.$$

We first prove the "only if" part. Since $\delta \in (\mathfrak{o}_\mathfrak{p}^\times)^2$, there exist $x, y \in \mathbf{Z}_2$ such that $\delta = (2^{-1}(x+y\sqrt{m}))^2$, and hence $\operatorname{Tr}(\delta) = 2^{-1}(x^2 + y^2m)$ and $N(\delta) = (4^{-1}(x^2 - y^2m))^2$. Since $\operatorname{ord}_\mathfrak{p}(\delta) = 0$, we have $N(\delta) \in (\mathbf{Z}_2^\times)^2$. Now we have $N(\delta)^{1/2} = (-1)^l 4^{-1}(x^2 - y^2m)$ with $l = 0$ or 1. Therefore,

$$\begin{aligned} \operatorname{Tr}(\delta) + (-1)^l 2N(\delta)^{1/2} &= x^2 \in (\mathbf{Z}_2)^2, \\ \operatorname{Tr}(\delta) - (-1)^l 2N(\delta)^{1/2} &= y^2 m \in m(\mathbf{Z}_2)^2. \end{aligned}$$

Now we prove the "if" part. Since $N(\delta) \in (\mathbf{Z}_2^\times)^2$, we have $\delta \in \mathfrak{o}_\mathfrak{p}^\times$. Set $\delta = 2^{-1}(a + b\sqrt{m})$ with $a, b \in \mathbf{Z}_2$. Since $\operatorname{Tr}(\delta)^2 - 4N(\delta) = b^2m$, we have

$$a = \operatorname{Tr}(\delta), \quad b = (-1)^j(m^{-1}(\operatorname{Tr}(\delta)^2 - 4N(\delta)))^{1/2}$$

with $j = 0$ or 1. On the other hand, since $(\mathrm{Tr}(\delta)^2 - 4N(\delta)) \in m(\mathbf{Z}_2)^2$, we have $\mathrm{Tr}(\delta)+(-1)^l 2N(\delta)^{1/2} \in (\mathbf{Z}_2)^2$ and $\mathrm{Tr}(\delta) - (-1)^l 2N(\delta)^{1/2} \in m(\mathbf{Z}_2)^2$ with $l = 0$ or 1. Thus we set

$$x = (-1)^j \big(\mathrm{Tr}(\delta) + (-1)^l 2N(\delta)^{1/2}\big)^{1/2} \quad (\in \mathbf{Z}_2),$$
$$y = \big(m^{-1}(\mathrm{Tr}(\delta) - (-1)^l 2N(\delta)^{1/2})\big)^{1/2} \quad (\in \mathbf{Z}_2).$$

Then

$$\begin{aligned}&\big(2^{-1}(x + y\sqrt{m})\big)^2\\ &= 2^{-1}\Big(\mathrm{Tr}(\delta) + (-1)^j \big(m^{-1}(\mathrm{Tr}(\delta)^2 - 4N(\delta))\big)^{1/2}\sqrt{m}\Big)\\ &= 2^{-1}(a + b\sqrt{m}) = \delta.\end{aligned}$$

Since $\delta \in \mathfrak{o}_\mathfrak{p}^\times$, we have $\delta \in (\mathfrak{o}_\mathfrak{p}^\times)^2$. This completes the proof. □

From the above lemma, we obtain the following result:

Proposition 3.14. *Let F, m, and $\mathfrak{p}$ be as in Proposition 3.12. Let $\delta \in \mathfrak{o}_F$, and set $\delta = 2^{-1}(a+b\sqrt{m})$ with $a, b \in \mathbf{Z}$. Then $\delta \in (\mathfrak{o}_\mathfrak{p}^\times)^2$ if and only if*

(i) $8 \mid 4^{-1}(a^2 - b^2 m) - 1$*; and*

(ii) $b = 0$ *and* $4 \mid 2^{-1}a - 1$*; or*
$b \neq 0$ *and* $4 \mid a + 1$*; or*
$b \neq 0$, $\mathrm{ord}_2(a) = 1$, $2 \mid r$, *and* $4 \mid 2^{-r}(a + 2u) - 1$,

where u is a rational integer such that $u^2 \equiv 4^{-1}(a^2-b^2m) \pmod{2^{2\,\mathrm{ord}_2(b)}\mathbf{Z}}$ and $u \equiv 1 \pmod{4\mathbf{Z}}$, and $r = \mathrm{ord}_2(a+2u)$.

Proof: In this proof, again abbreviate $\mathrm{Tr}_{F_\mathfrak{p}/\mathbf{Q}_2}$ and $N_{F_\mathfrak{p}/\mathbf{Q}_2}$ by Tr and N. Then $\mathrm{Tr}(\delta) = a$ and $N(\delta) = 4^{-1}(a^2 - b^2 m) \in \mathbf{Z}$. By Lemma 3.13, we have

$$\begin{aligned}\delta \in (\mathfrak{o}_\mathfrak{p}^\times)^2 \Longleftrightarrow\ & N(\delta) \in (\mathbf{Z}_2^\times)^2,\\ & \mathrm{Tr}(\delta) + 2N(\delta)^{1/2},\\ & \mathrm{Tr}(\delta) - 2N(\delta)^{1/2} \in (\mathbf{Z}_2)^2 \cup m(\mathbf{Z}_2)^2.\end{aligned} \tag{3–10}$$

Here, by (3–9), we have $(\mathbf{Z}_2^\times)^2 = 1 + 8\mathbf{Z}_2$, and hence

$$(\mathbf{Z}_2)^2 \cup m(\mathbf{Z}_2)^2 = \Big(\bigsqcup_{j=0}^{\infty} 4^j(1 + 4\mathbf{Z}_2)\Big) \cup \{0\}.$$

Since $N(\delta) \in \mathbf{Z}$, we have

$$N(\delta) \in (\mathbf{Z}_2^\times)^2 \Longleftrightarrow 8 \mid 4^{-1}(a^2 - b^2 m) - 1. \tag{3–11}$$

Henceforth, until the end of this proof, we may assume $N(\delta) \in (\mathbf{Z}_2^\times)^2$. When $b = 0$, we have $N(\delta)^{1/2} = (-1)^l 2^{-1}a \in \mathbf{Z}_2^\times$ with $l = 0$ or 1, and hence

$$\begin{aligned}&\mathrm{Tr}(\delta) + (-1)^l 2N(\delta)^{1/2} = 4 \cdot 2^{-1}a,\\ &\mathrm{Tr}(\delta) - (-1)^l 2N(\delta)^{1/2} = 0.\end{aligned}$$

Note that $2^{-1}a \in \mathbf{Z}_2^\times \cap \mathbf{Z} = 1 + 2\mathbf{Z}$. If $4 \mid 2^{-1}a - 1$, then $4 \cdot 2^{-1}a \in 4(1+4\mathbf{Z}) \subset (\mathbf{Z}_2)^2 \cup m(\mathbf{Z}_2)^2$; if $4 \mid 2^{-1}a - 3$, then $4 \cdot 2^{-1}a \notin (\bigsqcup_{j=0}^{\infty} 4^j(1 + 4\mathbf{Z}_2)) \cup \{0\} = (\mathbf{Z}_2)^2 \cup m(\mathbf{Z}_2)^2$. Therefore, by (3–10), we have

$$\delta \in (\mathfrak{o}_\mathfrak{p}^\times)^2 \Longleftrightarrow 4 \mid 2^{-1}a - 1.$$

Now we also assume $b \neq 0$. Since $(\mathrm{Tr}(\delta) + 2N(\delta)^{1/2})(\mathrm{Tr}(\delta) - 2N(\delta)^{1/2}) = b^2 m \neq 0$, we have $\mathrm{Tr}(\delta) + 2N(\delta)^{1/2} \neq 0$. Hence, if $\mathrm{Tr}(\delta) + 2N(\delta)^{1/2} \in (\mathbf{Z}_2)^2 \cup m(\mathbf{Z}_2)^2$, then

$$\begin{aligned}\mathrm{Tr}(\delta) - 2N(\delta)^{1/2} &= b^2 m(\mathrm{Tr}(\delta) + 2N(\delta)^{1/2})^{-1}\\ &\in ((\mathbf{Q}_2^\times)^2 \cup m(\mathbf{Q}_2^\times)^2) \cap \mathbf{Z}_2\\ &\subset (\mathbf{Z}_2)^2 \cup m(\mathbf{Z}_2)^2.\end{aligned}$$

Thus

$$\begin{aligned}&\mathrm{Tr}(\delta) + 2N(\delta)^{1/2}, \mathrm{Tr}(\delta) - 2N(\delta)^{1/2} \in (\mathbf{Z}_2)^2 \cup m(\mathbf{Z}_2)^2\\ &\quad\Longleftrightarrow \mathrm{Tr}(\delta) + 2N(\delta)^{1/2} \in (\mathbf{Z}_2)^2 \cup m(\mathbf{Z}_2)^2.\end{aligned} \tag{3–12}$$

When $\mathrm{ord}_2(a) \geq 2$, we have $\mathrm{Tr}(\delta) = a \in 4\mathbf{Z}_2$. Since $N(\delta)^{1/2} \in 1 + 2\mathbf{Z}_2$, we have $\mathrm{Tr}(\delta) + 2N(\delta)^{1/2} \in 2(1 + 2\mathbf{Z}_2) = 2\mathbf{Z}_2^\times$. Thus $\mathrm{Tr}(\delta) + 2N(\delta)^{1/2} \notin (\mathbf{Z}_2)^2 \cup m(\mathbf{Z}_2)^2$, and hence $\delta \notin (\mathfrak{o}_\mathfrak{p}^\times)^2$ by (3–10). When $\mathrm{ord}_2(a) = 0$, we have $\mathrm{Tr}(\delta) + 2N(\delta)^{1/2} \in a + 2 + 4\mathbf{Z}_2\ (\subset \mathbf{Z}_2^\times)$. Thus, by (3–10) and (3–12), we have

$$\delta \in (\mathfrak{o}_\mathfrak{p}^\times)^2 \Longleftrightarrow 4 \mid a + 1.$$

When $\mathrm{ord}_2(a) = 1$, we have $\mathrm{ord}_2(b) \geq 2$ by Equation (3–11). Thus, by Lemma 3.8, we have $u \equiv N(\delta)^{1/2} \pmod{2^{2\,\mathrm{ord}_2(b)-1}\mathbf{Z}_2}$, and hence $a + 2u \equiv \mathrm{Tr}(\delta) + 2N(\delta)^{1/2} \pmod{2^{2\,\mathrm{ord}_2(b)}\mathbf{Z}_2}$. Since $\mathrm{Tr}(\delta), 2N(\delta)^{1/2} \in 2(1 + 2\mathbf{Z}_2)$, we have $\mathrm{Tr}(\delta) - 2N(\delta)^{1/2} \in 4\mathbf{Z}_2$. Thus, since $(\mathrm{Tr}(\delta) + 2N(\delta)^{1/2})(\mathrm{Tr}(\delta) - 2N(\delta)^{1/2}) = mb^2$, we have $\mathrm{ord}_2(\mathrm{Tr}(\delta) + 2N(\delta)^{1/2}) \leq 2\,\mathrm{ord}_2(b) - 2$, and hence

$$r = \mathrm{ord}_2(a + 2u) = \mathrm{ord}_2(\mathrm{Tr}(\delta) + 2N(\delta)^{1/2}).$$

It follows that $a + 2u \equiv \mathrm{Tr}(\delta) + 2N(\delta)^{1/2} \pmod{2^{r+2}\mathbf{Z}_2}$, that is,

$$2^{-r}(a + 2u) \equiv 2^{-r}(\mathrm{Tr}(\delta) + 2N(\delta)^{1/2}) \pmod{4\mathbf{Z}_2}.$$

$\mathfrak{p}$	$C_{\mathbf{f}}(\mathfrak{p})$	$N((C_{\mathbf{f}}(\mathfrak{p}))_{K_{\mathbf{f}}})$	$\mathfrak{p}$	$C_{\mathbf{f}}(\mathfrak{p})$	$N((C_{\mathbf{f}}(\mathfrak{p}))_{K_{\mathbf{f}}})$
$[2,\ \omega]$	$(1\pm\sqrt{13})/2$	3	$[349,\ 206+\omega]$	$-18\pm\sqrt{13}$	311
$[11,\ 4+\omega]$	1	1	$[373,\ 233+\omega]$	27	3^6
$[13,\ 9+\omega]$	$\pm\sqrt{13}$	13	$[379,\ 153+\omega]$	$15\pm2\sqrt{13}$	173
$[17,\ 11+\omega]$	$4\pm\sqrt{13}$	3	$[397,\ 164+\omega]$	$-20\pm3\sqrt{13}$	283
$[23,\ 10+\omega]$	$4\pm\sqrt{13}$	3	$[401,\ 80+\omega]$	$14\pm\sqrt{13}$	$3\cdot61$
$[29,\ 2+\omega]$	$2\pm\sqrt{13}$	3^2	$[419,\ 222+\omega]*$	$-19\pm\sqrt{13}$	$2^2\cdot3\cdot29$
$[31,\ 29+\omega]$	$1\pm2\sqrt{13}$	$3\cdot17$	$[433,\ 30+\omega]$	$-24\pm\sqrt{13}$	563
$[59,\ 45+\omega]$	$\pm\sqrt{13}$	13	$[457,\ 43+\omega]$	-15	$3^2\cdot5^2$
$[61,\ 23+\omega]*$	$-1\pm\sqrt{13}$	$2^2\cdot3$	$[479,\ 434+\omega]$	$5\pm6\sqrt{13}$	443
$[67,\ 24+\omega]*$	$-3\pm3\sqrt{13}$	$2^2\cdot3^3$	$[491,\ 340+\omega]$	$14\pm5\sqrt{13}$	$3\cdot43$
$[73,\ 14+\omega]$	$4\pm3\sqrt{13}$	101	$[499,\ 130+\omega]*$	$9\pm7\sqrt{13}$	$2^2\cdot139$
$[79,\ 19+\omega]$	$12\pm\sqrt{13}$	131	$[503,\ 105+\omega]$	$9\pm6\sqrt{13}$	$3^2\cdot43$
$[89,\ 68+\omega]$	$-1\pm2\sqrt{13}$	$3\cdot17$	$[523,\ 303+\omega]$	21	$3^2\cdot7^2$
$[113,\ 62+\omega]*$	$5\pm3\sqrt{13}$	$2^2\cdot23$	$[563,\ 34+\omega]$	$17\pm4\sqrt{13}$	3^4
$[137,\ 89+\omega]$	$14\pm\sqrt{13}$	$3\cdot61$	$[571,\ 172+\omega]$	$21\pm2\sqrt{13}$	389
$[139,\ 18+\omega]$	$-12\pm\sqrt{13}$	131	$[587,\ 445+\omega]*$	$-2\pm4\sqrt{13}$	$2^2\cdot3\cdot17$
$[157,\ 73+\omega]*$	$\pm2\sqrt{13}$	$2^2\cdot13$	$[593,\ 478+\omega]$	$-16\pm\sqrt{13}$	3^5
$[173,\ 27+\omega]$	$3\pm6\sqrt{13}$	$3^3\cdot17$	$[613,\ 486+\omega]$	-19	19^2
$[193,\ 100+\omega]*$	$-3\pm\sqrt{13}$	2^2	$[631,\ 559+\omega]*$	$-18\pm8\sqrt{13}$	$2^2\cdot127$
$[197,\ 40+\omega]*$	$-1\pm3\sqrt{13}$	$2^2\cdot29$	$[643,\ 336+\omega]*$	$12\pm4\sqrt{13}$	2^6
$[199,\ 177+\omega]$	$-18\pm\sqrt{13}$	311	$[647,\ 51+\omega]$	39	$3^2\cdot13^2$
$[211,\ 136+\omega]$	$-4\pm3\sqrt{13}$	101	$[653,\ 173+\omega]*$	$10\pm8\sqrt{13}$	$2^2\cdot3\cdot61$
$[223,\ 152+\omega]$	$-8\pm3\sqrt{13}$	53	$[673,\ 620+\omega]$	$-21\pm6\sqrt{13}$	3^3
$[227,\ 102+\omega]*$	$2\pm2\sqrt{13}$	$2^4\cdot3$	$[683,\ 133+\omega]$	$-16\pm\sqrt{13}$	3^5
$[239,\ 148+\omega]$	$-5\pm6\sqrt{13}$	443	$[701,\ 452+\omega]$	$6\pm3\sqrt{13}$	3^4
$[241,\ 122+\omega]*$	$-19\pm3\sqrt{13}$	$2^2\cdot61$	$[709,\ 38+\omega]$	$27\pm4\sqrt{13}$	521
$[283,\ 95+\omega]$	$-16\pm3\sqrt{13}$	139	$[719,\ 310+\omega]$	$-14\pm7\sqrt{13}$	$3^2\cdot7^2$
$[293,\ 267+\omega]$	1	1	$[727,\ 205+\omega]$	$-6\pm13\sqrt{13}$	2161
$[307,\ 116+\omega]$	$-3\pm4\sqrt{13}$	199	$[739,\ 558+\omega]*$	14	$2^2\cdot7^2$
$[317,\ 280+\omega]$	$\pm3\sqrt{13}$	$3^2\cdot13$	$[769,\ 550+\omega]$	$-17\pm2\sqrt{13}$	$3\cdot79$

$*$: principal prime ideal

TABLE 1. The eigenvalues $C_{\mathbf{f}}(\mathfrak{p})$ of $T(\mathfrak{p})|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_{\mathbf{Q}(\sqrt{257})},1)}$ and their norm for a split prime $\mathfrak{p}$.

Therefore, by (3–10) and (3–12), we have

$$\begin{aligned}\delta\in(\mathfrak{o}_{\mathfrak{p}}^{\times})^2 &\Longleftrightarrow \mathrm{Tr}(\delta)+2N(\delta)^{1/2}\in\bigsqcup_{j=0}^{\infty}4^j(1+4\mathbf{Z}_2)\\ &\Longleftrightarrow 2\mid \mathrm{ord}_2(\mathrm{Tr}(\delta)+2N(\delta)^{1/2}),\\ &\qquad 2^{-\mathrm{ord}_2(\mathrm{Tr}(\delta)+2N(\delta)^{1/2})}(\mathrm{Tr}(\delta)+2N(\delta)^{1/2})\\ &\qquad\qquad\in 1+4\mathbf{Z}_2\\ &\Longleftrightarrow 2\mid r,\ 4\mid 2^{-r}(a+2u)-1.\end{aligned}$$

This completes the proof. □

Note that we can easily find u of Proposition 3.14 by Remark 3.11, since $4^{-1}(a^2-b^2m)=N_{F_{\mathfrak{p}}/\mathbf{Q}_2}(\delta)\in1+8\mathbf{Z}$.

4. NUMERICAL EXAMPLES FOR $\mathbf{Q}(\sqrt{257})$ AND $\mathbf{Q}(\sqrt{401})$

In this section, we shall give numerical examples of eigenvalues and characteristic polynomials of Hecke operators for real quadratic fields $\mathbf{Q}(\sqrt{257})$ and $\mathbf{Q}(\sqrt{401})$, whose class numbers are three and five, respectively.

Let F and m be as in Section 3. *We treat only the case* $k=(2,2)$ *and* ψ *is the identity* (i.e., $\psi(F_{\mathbf{A}}^{\times})=\{1\}$). We denote this character by *1*. Let $S_2(\Gamma_0(m),(\frac{m}{\cdot}))$ be the space of elliptic cusp forms of "Neben"-type of level m, and $\mathcal{S}^N_{(2,2)}(\mathfrak{o}_F,1)$ the subspace of $\mathcal{S}_{(2,2)}(\mathfrak{o}_F,1)$ that consists of Hilbert cusp forms coming from $S_2(\Gamma_0(m),(\frac{m}{\cdot}))$ through the Doi–Naganuma lifting (cf. [Doi and Naganuma 69] and [Naganuma 73]). We denote by $\mathcal{S}^0_{(2,2)}(\mathfrak{o}_F,1)$ the "F-proper" subspace of $\mathcal{S}_{(2,2)}(\mathfrak{o}_F,1)$, that is, the orthogonal complement of $\mathcal{S}^N_{(2,2)}(\mathfrak{o}_F,1)$ with respect to the standard inner product. It is known that $\mathcal{S}^0_{(2,2)}(\mathfrak{o}_F,1)$ and $\mathcal{S}^N_{(2,2)}(\mathfrak{o}_F,1)$ are stable under the action of $T(\mathfrak{p})$ for all prime ideals $\mathfrak{p}$ of F. In the following, we shall determine eigenvalues and characteristic polynomials of $T(\mathfrak{p})|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_F,1)}$ for several prime ideals $\mathfrak{p}$.

We denote by $\Psi_{\mathfrak{p}}(X)$ the characteristic polynomial of $T(\mathfrak{p})|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_F,1)}$. For a primitive form $\mathbf{f}$ in $\mathcal{S}^0_{(2,2)}(\mathfrak{o}_F,1)$, we denote by $C_{\mathbf{f}}(\mathfrak{p})$ the eigenvalue of $T(\mathfrak{p})$ satisfying

$\mathbf{f}|T(\mathfrak{p}) = C_{\mathbf{f}}(\mathfrak{p})\mathbf{f}$, and then denote by $K_{\mathbf{f}}$ the Hecke field of $\mathbf{f}$, that is, the field generated over $\mathbf{Q}$ by $C_{\mathbf{f}}(\mathfrak{p})$ for all prime ideals $\mathfrak{p}$. Let $K_{\mathbf{f}}^+$ be the subfield of $K_{\mathbf{f}}$ generated by $C_{\mathbf{f}}((p)_F)$ for all rational primes p. We note that $[K_{\mathbf{f}} : K_{\mathbf{f}}^+] = 2$.

Now we set $\Lambda_{\mathbf{f}}(\mathfrak{p}) = \mathfrak{o}_{K_{\mathbf{f}}^+} + C_{\mathbf{f}}(\mathfrak{p})\mathfrak{o}_{K_{\mathbf{f}}^+}$ for a split prime $\mathfrak{p}$. Then $\Lambda_{\mathbf{f}}(\mathfrak{p})$ is an order in $\mathcal{O}_{K_{\mathbf{f}}/K_{\mathbf{f}}^+}$ in the sense of Section 2.2, and the conductor $c(\Lambda_{\mathbf{f}}(\mathfrak{p}))$ is given by Lemma 2.3 as follows:

$$c(\Lambda_{\mathbf{f}}(\mathfrak{p})) = \left(D_{K_{\mathbf{f}}/K_{\mathbf{f}}^+}(C_{\mathbf{f}}(\mathfrak{p})) \cdot D_{K_{\mathbf{f}}/K_{\mathbf{f}}^+}{}^{-1}\right)^{1/2}.$$

4.1 Example for $\mathbf{Q}(\sqrt{257})$

Let $F = \mathbf{Q}(\sqrt{257})$. Then $h_F = h_F^+ = 3$. We have

$$\begin{aligned}\dim_{\mathbf{C}} \mathcal{S}^0_{(2,2)}(\mathfrak{o}_F, 1) &= \dim_{\mathbf{C}} \mathcal{S}_{(2,2)}(\mathfrak{o}_F, 1) \\ &\quad - \frac{1}{2}\dim_{\mathbf{C}} S_2(\Gamma_0(257), \left(\tfrac{257}{}\right)) \\ &= \operatorname{tr} T(\mathfrak{o}_F) - \frac{1}{2}\cdot 20 = 2.\end{aligned}$$

- Table 1 gives numerical data for the eigenvalues of $T(\mathfrak{p})|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_F,1)}$ and their norms for split primes $\mathfrak{p}$ satisfying $N(\mathfrak{p}) \le 769$.
- Table 2 gives numerical data for the eigenvalues of $T((p)_F)|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_F,1)}$ for rational primes p such that p remains prime in F and $p \le 97$. (Note that the characteristic polynomial of $T((p)_F)|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_F,1)}$ has a double root.)

$\mathfrak{p}$	$C_{\mathbf{f}}(\mathfrak{p})$	$\mathfrak{p}$	$C_{\mathbf{f}}(\mathfrak{p})$	$\mathfrak{p}$	$C_{\mathbf{f}}(\mathfrak{p})$
$(3)_F$	-4	$(37)_F$	52	$(53)_F$	-8
$(5)_F$	-2	$(41)_F$	-18	$(71)_F$	-30
$(7)_F$	0	$(43)_F$	30	$(83)_F$	50
$(19)_F$	-18	$(47)_F$	46	$(97)_F$	90

TABLE 2. The eigenvalue $C_{\mathbf{f}}(\mathfrak{p})$ of $T(\mathfrak{p})|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_{\mathbf{Q}(\sqrt{257})},1)}$ for $\mathfrak{p} = (p)_F$.

For a primitive form $\mathbf{f}$ in $\mathcal{S}^0_{(2,2)}(\mathfrak{o}_F, 1)$, we have

$$K_{\mathbf{f}} = \mathbf{Q}(\sqrt{13}), \quad K_{\mathbf{f}}^+ = \mathbf{Q}.$$

Within the limit of Table 1, *we observe that*

$$\mathfrak{p} \textit{ is principal} \iff (2)_{K_{\mathbf{f}}} \mid C_{\mathbf{f}}(\mathfrak{p})$$

for split primes $\mathfrak{p}$*; in particular, we remark that*

$$2 \mid c(\Lambda_{\mathbf{f}}(\mathfrak{p}))$$

for all principal split primes $\mathfrak{p}$ in the table.

4.2 Example for $\mathbf{Q}(\sqrt{401})$

Let $F = \mathbf{Q}(\sqrt{401})$. Then $h_F = h_F^+ = 5$. We have

$$\dim_{\mathbf{C}} \mathcal{S}^0_{(2,2)}(\mathfrak{o}_F, 1) = 24 - \frac{1}{2}\cdot 32 = 8.$$

- Table 3 gives numerical data for the characteristic polynomials $\Psi_{(p)_F}(X)$ of $T((p)_F)|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_F,1)}$ for rational primes p such that p remains prime in F and $p \le 23$.

$\mathfrak{p}$	$\Psi_{\mathfrak{p}}(X)$
$(3)_F$	$(X^4 + 7X^3 + 4X^2 - 32X + 1)^2$
$(13)_F$	$(X^4 + 24X^3 + 120X^2 - 113X - 571)^2$
$(17)_F$	$(X^4 + 2X^3 - 110X^2 - 111X + 3019)^2$
$(19)_F$	$(X^4 + 10X^3 - 339X^2 - 1360X + 22759)^2$
$(23)_F$	$(X^4 - 16X^3 - 495X^2 + 8532X - 11671)^2$

TABLE 3. The characteristic polynomial $\Psi_{\mathfrak{p}}(X)$ of $T(\mathfrak{p})|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_{\mathbf{Q}(\sqrt{401})},1)}$ for $\mathfrak{p} = (p)_F$.

- Table 4 gives numerical data for the coefficients of the characteristic polynomials $\Psi_{\mathfrak{p}}(X) = X^8 + a_1X^7 + \cdots + a_7X + a_8$ of $T(\mathfrak{p})|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_F,1)}$ for principal split primes $\mathfrak{p}$ satisfying $N(\mathfrak{p}) \le 643$ and nonprincipal split primes $\mathfrak{p}$ satisfying $N(\mathfrak{p}) \le 263$.

The characteristic polynomial $\Psi_{[2,\omega]}(X)$ of $T([2,\omega])|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_F,1)}$ is irreducible over $\mathbf{Q}$, and the roots of $\Psi_{[2,\omega]}(X)$ are

$$c_{ijl} = \frac{1}{40}\Bigg(15 - (-1)^i 5\sqrt{5} + (-1)^{i+j}\sqrt{5}\sqrt{110 + 10\sqrt{5}} + (-1)^l\sqrt{4900 - (-1)^i 100\sqrt{5} + (-1)^j\left(150 - (-1)^i 10\sqrt{5}\right)\sqrt{110 + 10\sqrt{5}}}\Bigg),$$

where $0 \le i, j, l \le 1$. Now we take the primitive form $\mathbf{f}$ such that $\mathbf{f}|T([2,\omega]) = c_{000}\mathbf{f}$. Then we have

$$K_{\mathbf{f}} = \mathbf{Q}\left(\sqrt{4900 - 100\sqrt{5} + (150 - 10\sqrt{5})\sqrt{110 + 10\sqrt{5}}}\right),$$
$$K_{\mathbf{f}}^+ = \mathbf{Q}\left(\sqrt{110 + 10\sqrt{5}}\right).$$

(The degree of the Galois closure of $K_{\mathbf{f}}$ over $\mathbf{Q}$ is $128 = 2^7$.) Then we have

$$\begin{aligned} D_{K_{\mathbf{f}}/\mathbf{Q}} &= 5^4 \cdot 29^2 \cdot 131 \cdot 139, \\ D_{K_{\mathbf{f}}^+/\mathbf{Q}} &= 5^2 \cdot 29, \\ N(D_{K_{\mathbf{f}}/K_{\mathbf{f}}^+}) &= 131 \cdot 139. \end{aligned}$$

- Table 5 gives numerical data for the norm of $c(\Lambda_{\mathbf{f}}(\mathfrak{p}))$.

$\mathfrak{p}$		a_1,	a_2,	a_3,	a_4,	a_5,	a_6,	a_7,	a_8
$[2, \quad \omega]$		−3,	−10,	28,	37,	−78,	−58,	53,	19
$[5, \quad \omega]$		−3,	−18,	41,	111,	−163,	−234,	155,	1
$[7, \ 5+\omega]$		−6,	−12,	120,	−175,	−42,	175,	−83,	11
$[11, \ 7+\omega]$		9,	10,	−101,	−253,	149,	918,	809,	179
$[29, \ 22+\omega]$		13,	−5,	−523,	−408,	8053,	1917,	−48078,	38359
$[41, \ 27+\omega]$		−16,	−90,	2332,	−2437,	−86432,	249407,	263905,	43909
$[43, \ 26+\omega]$		−32,	309,	−148,	−11780,	43156,	18606,	−138350,	−42131
$[47, \ 44+\omega]$		15,	−46,	−1114,	861,	24682,	−30282,	−133239,	211541
$[73, \ 39+\omega]$		4,	−362,	−1343,	36721,	107356,	−1265768,	−1867352,	14282224
$[83, \ 30+\omega]$	∗	−32,	92,	5761,	−41264,	−341588,	3084758,	6681459,	−69332531
$[89, \ 60+\omega]$		−10,	−278,	3166,	13505,	−241434,	703443,	−114403,	−611281
$[103, \ 85+\omega]$		27,	−118,	−6456,	−10369,	372198,	1371298,	−3479619,	−15228421
$[109, \ 74+\omega]$		29,	−186,	−10994,	−61417,	448682,	4111874,	7230513,	−1231091
$[113, \ 23+\omega]$		−34,	88,	7570,	−77675,	−156608,	5544655,	−26618987,	40066931
$[149, \ 55+\omega]$		−68,	1775,	−22208,	131704,	−237922,	−810800,	2670928,	1079011
$[151, \ 92+\omega]$		−22,	−151,	5415,	−7430,	−329865,	1334049,	−291087,	−2791879
$[173, \ 110+\omega]$		−11,	−381,	7795,	−37942,	−157229,	2060155,	−6349795,	5629151
$[179, \ 107+\omega]$		−88,	2950,	−45092,	245937,	1261812,	−21112768,	83231088,	−104306576
$[181, \ 21+\omega]$		−7,	−310,	2548,	18965,	−210804,	594832,	−493872,	−100624
$[197, \ 45+\omega]$		40,	168,	−9515,	−110360,	121710,	6701416,	29340685,	31232399
$[223, \ 31+\omega]$		−68,	1453,	−2076,	−317709,	3811290,	−9269721,	−49448362,	130826831
$[229, \ 57+\omega]$		33,	−67,	−8851,	−51342,	253023,	1988777,	−1662381,	−18676169
$[239, \ 206+\omega]$		−53,	1066,	−10574,	54335,	−132516,	96784,	72897,	−69191
$[241, \ 167+\omega]$		−35,	218,	2705,	−15695,	−84085,	59836,	48125,	−30371
$[257, \ 122+\omega]$	∗	10,	−1119,	−14842,	336025,	6080144,	−7871441,	−548680256,	−2339785241
$[263, \ 201+\omega]$		−36,	−295,	12183,	91010,	−605219,	−5055489,	−2431643,	17176609
$[337, \ 172+\omega]$	∗	−47,	105,	24343,	−327430,	−1820659,	57625757,	−337202462,	582948571
$[379, \ 103+\omega]$	∗	−25,	−1214,	26233,	458881,	−6336001,	−81101528,	302094243,	3696976091
$[383, \ 205+\omega]$	∗	−63,	286,	61218,	−1739923,	17945914,	−49335124,	−306466744,	1582083824
$[397, \ 197+\omega]$	∗	−77,	690,	64185,	−1089907,	−15159265,	255263350,	1030269191,	−910014589
$[421, \ 304+\omega]$	∗	48,	−494,	−47086,	−216435,	13127696,	123014259,	−842081917,	−9542329681
$[487, \ 70+\omega]$	∗	−25,	−1553,	38867,	743048,	−20170693,	−84273537,	3477790992,	−11951208719
$[499, \ 264+\omega]$	∗	−22,	−1402,	39386,	246139,	−13711178,	90023337,	71238749,	−717378001
$[643, \ 474+\omega]$	∗	−72,	−896,	147914,	−1370115,	−68019308,	937509055,	3125818491,	−4860460921

∗ : principal prime ideal

TABLE 4. The coefficients of the characteristic polynomial $X^8+a_1X^7+\cdots+a_7X+a_8$ of $T(\mathfrak{p})|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_{\mathbf{Q}(\sqrt{401})},1)}$ for a split prime $\mathfrak{p}$.

From Table 5, *we immediately observe that*

$$19 \mid N\big(c(\Lambda_{\mathbf{f}}(\mathfrak{p}))\big)$$

for all principal split primes $\mathfrak{p}$ in the table. Moreover, if we set

$$\mathfrak{P}_{19} = c(\Lambda_{\mathbf{f}}([83, 30+\omega])),$$

then $\mathfrak{P}_{19}$ is a prime ideal of $K_{\mathbf{f}}^+$, and there exist prime ideals $\mathfrak{P}'_{19}, \mathfrak{P}''_{19}$ of $K_{\mathbf{f}}^+$ such that

$$\begin{aligned}(19)_{K_{\mathbf{f}}^+} &= \mathfrak{P}_{19}\mathfrak{P}'_{19}\mathfrak{P}''_{19},\\ \mathfrak{P}_{19}\mathfrak{P}'_{19} &= [19, 4+(1+\sqrt{5})/2]\cdot\mathfrak{o}_{K_{\mathbf{f}}^+},\\ \mathfrak{P}''_{19} &= [19, 14+(1+\sqrt{5})/2]\cdot\mathfrak{o}_{K_{\mathbf{f}}^+}.\end{aligned}$$

Then *we can observe that*

$$\mathfrak{P}_{19} \mid c(\Lambda_{\mathbf{f}}(\mathfrak{p}))$$

for all principal split primes $\mathfrak{p}$ in the table.

4.3 Calculation Based on Hida's Suggestion

We check (1), (2), and (3) of the Introduction.

When $F=\mathbf{Q}(\sqrt{257})$, the common factor of $N_{K_{\mathbf{f}}/\mathbf{Q}}(1+N(\mathfrak{p})-C_{\mathbf{f}}(\mathfrak{p}))=\Psi_{\mathfrak{p}}(1+N(\mathfrak{p}))$ is 2^2 from Table 6. Moreover, it follows immediately from Table 1 that the common factor of $1+N(\mathfrak{p})-C_{\mathbf{f}}(\mathfrak{p})$ is $(2)_{K_{\mathbf{f}}}=\mathfrak{F}_{\mathbf{f}}\mathfrak{o}_{K_{\mathbf{f}}}$. When $F=\mathbf{Q}(\sqrt{401})$, the common factor of $N_{K_{\mathbf{f}}/\mathbf{Q}}(1+N(\mathfrak{p})-C_{\mathbf{f}}(\mathfrak{p}))=\Psi_{\mathfrak{p}}(1+N(\mathfrak{p}))$ is 19^2 from Table 7.

$\mathfrak{p}$	$N(c(\Lambda_{\mathbf{f}}(\mathfrak{p})))$
$[2,\ \omega]$	1
$[5,\ \omega]$	1
$[7,\ 5+\omega]$	1
$[11,\ 7+\omega]$	1
$[29,\ 22+\omega]$	1
$[41,\ 27+\omega]$	31
$[43,\ 26+\omega]$	31
$[47,\ 44+\omega]$	31
$[73,\ 39+\omega]$	3^4
$[83,\ 30+\omega]$ $*$	19
$[89,\ 60+\omega]$	41
$[103,\ 85+\omega]$	23^2
$[109,\ 74+\omega]$	$19\cdot 29$
$[113,\ 23+\omega]$	1
$[149,\ 55+\omega]$	5^2
$[151,\ 92+\omega]$	29
$[173,\ 110+\omega]$	41
$[179,\ 107+\omega]$	19
$[181,\ 21+\omega]$	11
$[197,\ 45+\omega]$	379
$[223,\ 31+\omega]$	61
$[229,\ 57+\omega]$	19
$[239,\ 206+\omega]$	1
$[241,\ 167+\omega]$	7^2
$[257,\ 122+\omega]$ $*$	$19\cdot 139$
$[263,\ 201+\omega]$	409
$[337,\ 172+\omega]$ $*$	$19\cdot 41$
$[379,\ 103+\omega]$ $*$	$19\cdot 31$
$[383,\ 205+\omega]$ $*$	$7^2\cdot 19$
$[397,\ 197+\omega]$ $*$	$19\cdot 41$
$[421,\ 304+\omega]$ $*$	$11\cdot 19^2$
$[487,\ 70+\omega]$ $*$	$19^2\cdot 31$
$[499,\ 264+\omega]$ $*$	$11^2\cdot 19$
$[643,\ 474+\omega]$ $*$	$19\cdot 79$

$*$: principal prime ideal

TABLE 5. The norm of $c(\Lambda_{\mathbf{f}}(\mathfrak{p}))$ for the primitive form $\mathbf{f}$ in $\mathcal{S}^0_{(2,2)}(\mathfrak{o}_{\mathbf{Q}(\sqrt{401})}, 1)$ and a split prime $\mathfrak{p}$.

Moreover, we can observe that the common factor of $1+N(\mathfrak{p})-C_{\mathbf{f}}(\mathfrak{p})$ is $\mathfrak{P}_{19}\mathfrak{o}_{K_{\mathbf{f}}} = \mathfrak{F}_{\mathbf{f}}\mathfrak{o}_{K_{\mathbf{f}}}$. Thus (1) and (2) are affirmative, and (3) is correct in this case.

By using [Siegel 69, (22)], we have calculated the value at 2 of the Hecke L-function associated with a a nontrivial class character χ. In the case $F=\mathbf{Q}(\sqrt{257})$, we have

$$\prod_{\chi:\text{nontrivial}} \frac{D_F^{3/2}}{(2\pi)^4} L_F(2,\chi) = 2^2.$$

In the case $F=\mathbf{Q}(\sqrt{401})$, we have

$$\prod_{\chi:\text{nontrivial}} \frac{D_F^{3/2}}{(2\pi)^4} L_F(2,\chi) = 19^2.$$

$\mathfrak{p}$	$N_{K_{\mathbf{f}}/\mathbf{Q}}(1+N(\mathfrak{p})-C_{\mathbf{f}}(\mathfrak{p}))$
$(3)_F$	$2^2\cdot 7^2$
$(5)_F$	$2^4\cdot 7^2$
$(7)_F$	$2^2\cdot 5^4$
$(19)_F$	$2^4\cdot 5^2\cdot 19^2$
$(37)_F$	$2^2\cdot 659^2$
$(41)_F$	$2^4\cdot 5^4\cdot 17^2$
$(43)_F$	$2^4\cdot 5^2\cdot 7^2\cdot 13^2$
$(47)_F$	$2^4\cdot 541^2$
$(53)_F$	$2^2\cdot 1409^2$
$(71)_F$	$2^8\cdot 317^2$
$(83)_F$	$2^6\cdot 3^4\cdot 5^2\cdot 19^2$
$(97)_F$	$2^6\cdot 5^2\cdot 233^2$
$[61,\ 23+\omega]$	$2^2\cdot 23\cdot 43$
$[67,\ 24+\omega]$	$2^2\cdot 1231$
$[113,\ 62+\omega]$	$2^2\cdot 17\cdot 173$
$[157,\ 73+\omega]$	$2^4\cdot 3^2\cdot 173$
$[193,\ 100+\omega]$	$2^2\cdot 3\cdot 53\cdot 61$
$[197,\ 40+\omega]$	$2^2\cdot 9871$
$[227,\ 102+\omega]$	$2^4\cdot 3\cdot 1063$
$[241,\ 122+\omega]$	$2^2\cdot 3^2\cdot 1889$
$[419,\ 222+\omega]$	$2^2\cdot 3^2\cdot 53\cdot 101$
$[499,\ 130+\omega]$	$2^2\cdot 3^2\cdot 6679$
$[587,\ 445+\omega]$	$2^2\cdot 3\cdot 53\cdot 547$
$[631,\ 559+\omega]$	$2^2\cdot 3^2\cdot 13\cdot 17\cdot 53$
$[643,\ 336+\omega]$	$2^4\cdot 3\cdot 8317$
$[653,\ 173+\omega]$	$2^4\cdot 3\cdot 8623$
$[739,\ 558+\omega]$	$2^2\cdot 3^2\cdot 11^4$

TABLE 6. $N_{K_{\mathbf{f}}/\mathbf{Q}}(1+N(\mathfrak{p})-C_{\mathbf{f}}(\mathfrak{p}))$ for a principal prime $\mathfrak{p}$, where $C_{\mathbf{f}}(\mathfrak{p})$ is an eigenvalue of $T(\mathfrak{p})|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_{\mathbf{Q}(\sqrt{257})},1)}$.

$\mathfrak{p}$	$N_{K_{\mathbf{f}}/\mathbf{Q}}(1+N(\mathfrak{p})-C_{\mathbf{f}}(\mathfrak{p}))$
$(3)_F$	$19^2\cdot 29^2\cdot 31^2$
$(13)_F$	$11^4\cdot 19^4\cdot 61^2\cdot 359^2$
$(17)_F$	$11^2\cdot 19^2\cdot 34030181^2$
$(19)_F$	$7^4\cdot 11^2\cdot 19^2\cdot 521^2\cdot 3299^2$
$(23)_F$	$19^2\cdot 4020433831^2$
$[83,\ 30+\omega]$	$19^2\cdot 31\cdot 11059\cdot 12836389$
$[257,\ 122+\omega]$	$11\cdot 19^2\cdot 1279\cdot 3191\cdot 1236962761$
$[337,\ 172+\omega]$	$19^2\cdot 641\cdot 1109\cdot 2011\cdot 8111\cdot 35099$
$[379,\ 103+\omega]$	$11^3\cdot 19^2\cdot 1638061\cdot 511689281$
$[383,\ 205+\omega]$	$2^4\cdot 11\cdot 19^2\cdot 1621\cdot 1790869\cdot 2150221$
$[397,\ 197+\omega]$	$11^2\cdot 19^2\cdot 131\cdot 94781\cdot 942456979$
$[421,\ 304+\omega]$	$19^2\cdot 60830069\cdot 50854477409$
$[487,\ 70+\omega]$	$11\cdot 19^2\cdot 2699\cdot 17191\cdot 16454332679$
$[499,\ 264+\omega]$	$11^2\cdot 19^2\cdot 61\cdot 829\cdot 1681246642091$
$[643,\ 474+\omega]$	$19^2\cdot 421\cdot 334889\cdot 515366804791$

TABLE 7. $N_{K_{\mathbf{f}}/\mathbf{Q}}(1+N(\mathfrak{p})-C_{\mathbf{f}}(\mathfrak{p}))$ for a principal prime $\mathfrak{p}$, where $C_{\mathbf{f}}(\mathfrak{p})$ is an eigenvalue of $T(\mathfrak{p})|_{\mathcal{S}^0_{(2,2)}(\mathfrak{o}_{\mathbf{Q}(\sqrt{401})},1)}$.

ACKNOWLEDGMENTS

I would like to express my deep gratitude to Professor Haruzo Hida for giving an introductory lecture on Hilbert modular forms in the summer of 1994 at Ritsumeikan University and for suggesting the explicit goal as described above. Also, I wish to thank Dr. Keiji Goto, who read the original draft with deep interest and offered both mathematical and linguistic comments. Finally, I would like to thank one of the referees for encouraging further study of the observed phenomenon, which gave me the chance to investigate Professor Hida's suggestion.

REFERENCES

[Dirichlet 1894] P. G. L. Dirichlet. *Vorlesungen über Zahlentheorie*, Herausgegeben und mit Zusätzen versehen von R. Dedekind, Vierte, Braunschweig, 1894.

[Doi and Naganuma 69] K. Doi and H. Naganuma. "On the functional equation of certain Dirichlet series." *Invent. Math.* **9** (1969), 1–14.

[Doi et al. 98] K. Doi, H. Hida, and H. Ishii. "Discriminant of Hecke fields and twisted adjoint L-values for $GL(2)$." *Invent. Math.* **134** (1998), 547–577.

[Iwasawa 72] K. Iwasawa. *Lectures on p-Adic L-Functions,* Ann. of Math. Studies 74, Princeton Univ. Press, 1972.

[Miyake 89] T. Miyake. *Modular Forms*, Springer-Verlag, Berlin, Heidelberg, and New York, 1989.

[Naganuma 73] H. Naganuma. "On the coincidence of two Dirichlet series associated with cusp forms of Hecke's "Neben"-type and Hilbert modular forms over a real quadratic field." *J. Math. Soc. Japan* **25** (1973), 547–555.

[Neukirch 86] J. Neukirch. *Class Field Theory*, Springer-Verlag, Berlin, Heidelberg, and New York, 1986.

[Okazaki 91] R. Okazaki. "On evaluation of L-functions over real quadratic fields." *J. Math. Kyoto Univ.* **31** (1991), 1125–1153.

[Saito 84] H. Saito. "On an operator U_χ acting on the space of Hilbert cusp forms." *J. Math. Kyoto Univ.* **24** (1984), 285–303.

[Shimura 71] G. Shimura. *Introduction to the Arithmetic Theory of Automorphic Functions*, Iwanami Shoten and Princeton Univ. Press, 1971.

[Shimura 78] G. Shimura. "The special values of the zeta functions associated with Hilbert modular forms." *Duke Math. J.* **45** (1978), 637–679.

[Shimura 91] G. Shimura. "The critical values of certain Dirichlet series attached to Hilbert modular forms." *Duke Math. J.* **63** (1991), 557–613.

[Shintani 76] T. Shintani. "On evaluation of zeta functions of totally real algebraic number fields at non-positive integers." *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* **23** (1976), 393–417.

[Siegel 69] C. L. Siegel. "Berechnung von Zetafunktionen an ganzzahligen Stellen." *Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl.* (1969), 87–102.

[Weil 67] A. Weil. *Basic Number Theory*, Springer-Verlag, Berlin, Heidelberg, and New York, 1967.

Kaoru Okada, Department of Mathematical Sciences, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan
(okada-k@se.ritsumei.ac.jp)

Received May 3, 2000; accepted in revised form October 12, 2001.

Kashaev's Conjecture and the Chern-Simons Invariants of Knots and Links

Hitoshi Murakami, Jun Murakami, Miyuki Okamoto, Toshie Takata, and Yoshiyuki Yokota

CONTENTS

2000 AMS Subject Classification: Primary 57M27, 57M25, 57M50; Secondary 41A60, 17B37, 81R50

Keywords: Volume conjecture, Kashaev's conjecture, colored Jones polynomial, Chern-Simons invariant, volume

R. M. Kashaev conjectured that the asymptotic behavior of the link invariant he introduced [Kashaev 95], which equals the colored Jones polynomial evaluated at a root of unity, determines the hyperbolic volume of any hyperbolic link complement. We observe numerically that for knots 6_3, 8_9 and 8_{20} and for the Whitehead link, the colored Jones polynomials are related to the hyperbolic volumes and the Chern–Simons invariants and propose a complexification of Kashaev's conjecture.

1. INTRODUCTION

In [Kashaev 95], R.M. Kashaev defined a link invariant associated with the quantum dilogarithm, depending on a positive integer N, which is denoted by $\langle L\rangle_N$ for a link L. Moreover, in [Kashaev 97], he conjectured that for any hyperbolic link L, the asymptotics at $N\to\infty$ of $|\langle L\rangle_N|$ gives its volume, that is

$$\mathrm{vol}(L)=2\pi\lim_{N\to\infty}\frac{\log|\langle L\rangle_N|}{N}$$

with $\mathrm{vol}(L)$ the hyperbolic volume of the complement of L. He showed that this conjecture is true for three doubled knots 4_1, 5_2, and 6_1. Unfortunately, his proof is not mathematically rigorous.

Afterwards, in [Murakami and Murakami 01], the first two authors proved that for any link L, Kashaev's invariant $\langle L\rangle_N$ is equal to the colored Jones polynomial evaluated at $\exp\left(2\pi\sqrt{-1}/N\right)$, which is written by $J_N(L)$, and extended Kashaev's conjecture as follows.

Conjecture 1.1. (Volume conjecture.)

$$\|L\|=\frac{2\pi}{v_3}\lim_{N\to\infty}\frac{\log|J_N(L)|}{N},$$

where $\|L\|$ is the simplicial volume of the complement of L and v_3 is the volume of the ideal regular tetrahedron.

Note that the hyperbolic volume $\mathrm{vol}(L)$ of a hyperbolic link L is equal to $\|L\|$ multiplied by v_3. This conjecture is not true for links in general, as $J_N(L)$ vanishes for a split link L. It is shown by Kashaev and O. Tirkkonen in [Kashaev and Tirkkonen 00] that the volume conjecture holds for torus knots. See [Thurston 99] and [Yokota 00, Yokota 02] for discussions about Kashaev's conjecture for hyperbolic knots from the viewpoint of tetrahedron decomposition.

In this paper, following Kashaev's way to analyze the asymptotic behavior of the invariant, we observe numerically, by using MAPLE V (a product of Waterloo Maple Inc.) and SnapPea [Weeks 02], that for the hyperbolic knots 6_3, 8_9, 8_{20}, and for the Whitehead link, the colored Jones polynomials are related to the hyperbolic volumes and the Chern–Simons invariants. Note that the knots 6_3 and 8_9 are not doubles of the unknot.

We also discuss a relation between the asymptotic behavior of $J_N(L)$ and the Chern–Simons invariant of the complement of the above-mentioned links L, and propose the following conjecture.

Conjecture 1.2. (Complexification of Kashaev's conjecture.) *Let L be a hyperbolic link. Then the following formula holds.*

$$J_N(L) \sim \exp \frac{N}{2\pi}(\mathrm{vol}(L) + \sqrt{-1}\,\mathrm{CS}(L)) \quad (N \to \infty)$$

where $\mathrm{CS}(L)$ is the Chern–Simons invariant of L [Chern and Simons 74, Meyerhoff 86]*. Note that the complement of L is a hyperbolic manifold with cusps.*

The statement of this conjecture will be given more properly in the last section.

2. PRELIMINARIES

First, we will briefly review the colored Jones polynomials of links following [Kirby and Melvin 91]. It is obtained from the quantum group $U_q(sl(2,\mathbb{C}))$ and its N-dimensional irreducible representation.

Let L be an oriented link. We consider a $(1,1)$-tangle presentation of L, obtained by cutting a component of the link. We assume that all crossing and local extreme points are as in Figure 1. We can calculate the N-colored Jones polynomial $J_L(N)$ evaluated at the N-th root of unity for L in the following way. We start with a labeling of the edges of the $(1,1)$-tangle presentation with labels $\{0,1,\ldots,N-1\}$. Here we label the two edges containing the end points of the tangle by 0. Following the labeling,

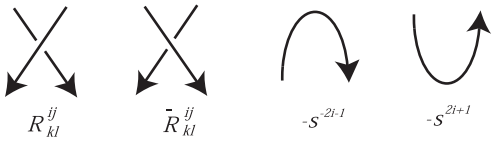

FIGURE 1. Crossings and local extrema in a link diagram.

we associate a positive (respectively, negative) crossing with the element R^{ij}_{kl} (respectively, $\bar{R}^{ij}_{kl}$), a maximal point $\cap$ labeled by i with the element $-s^{-2i-1}$, and a minimal point $\cup$ labeled by i with the element $-s^{2i+1}$ with $s = \exp\left(\frac{\pi\sqrt{-1}}{N}\right)$ as in Figure 1.

Here R^{ij}_{kl} and $\bar{R}^{ij}_{kl}$ are given by

$$R^{ij}_{kl} = \sum_{n=0}^{\min(N-1-i,j)} \delta_{l,i+n}\delta_{k,j-n} \frac{(i+n)!(N-1+n-j)!}{(i)!(N-1-j)!(n)!} \times s^{2(i-\frac{N-1}{2})(j-\frac{N-1}{2})-n(i-j)-\frac{n(n+1)}{2}},$$

$$\bar{R}^{ij}_{kl} = \sum_{n=0}^{\min(N-1-j,i)} \delta_{l,i-n}\delta_{k,j+n} \frac{(j+n)!(N-1+n-i)!}{(j)!(N-1-i)!(n)!}(-1)^n \times s^{-2(i-\frac{N-1}{2})(j-\frac{N-1}{2})-n(i-j)+\frac{n(n+1)}{2}}$$

with $(n)! = (s-s^{-1})(s^2-s^{-2})\cdots(s^n-s^{-n})$.

After multiplying all elements associated with the critical points, we sum over all indices, ignoring framings of links.

We calculate the colored Jones polynomial of the Whitehead link as an example. We can label each edge in the following way, noting Kronecker's deltas in R^{ij}_{kl} and $\bar{R}^{ij}_{kl}$.

We have to rotate a crossing where edges go up. In that case, we use $\cup$ and/or $\cap$ to calculate the invariant.

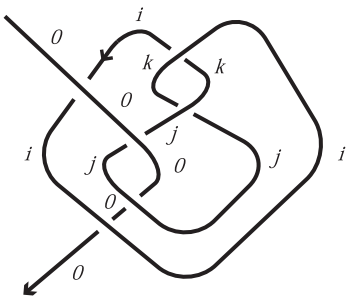

FIGURE 2. Labeling of the Whitehead link.

$$J_N(6_3) = \sum_{\substack{0\le k,l,m\\ k+l+m\le N-1}} (-1)^{k+l} s^{\frac{(l+k)(l+k+1)}{2}-\frac{(m+k)(m+k+1)}{2}+\frac{k(k+1)}{2}+2(m-l)(k+1)+N(m-l+k)} \times \frac{(N-1-l)!(N-1-m)!(l+m+k)!(N-1)!(1-s^{-2N-2})\cdots(1-s^{-2N-2k})}{(N-1-l-m-k)!(N-1-l-k)!(N-1-m-k)!(l)!(m)!(k)!}.$$

FIGURE 3. Calculation of $J_N(6_3)$.

Then we calculate the formula

$$J_N(L) = \sum_{\substack{0\le i,j,k\le N-1\\ i,j\ge k}} \frac{(q)_i(q)_j\{(q)_{N-1-k}\}^2}{\{(q)_k\}^2(q)_{N-1-i}(q)_{N-1-j}(q)_{i-k}(q)_{j-k}} q^{-k(i+j+1)}, \tag{2–1}$$

where $q = s^2 = \exp\left(\frac{2\pi\sqrt{-1}}{N}\right)$. Here $(x)_k = (1-x)(1-x^2)\cdots(1-x^k)$.

Next, the Chern–Simons invariant of a link is defined. Let $\mathcal{A}$ be the set of all $SO(3)$-connections of the trivial $SO(3)$-bundle of a closed three-manifold M and $\mathrm{cs}\colon \mathcal{A}\to\mathbb{R}$ the Chern–Simons functional defined by

$$\mathrm{cs}(A) = \frac{1}{8\pi^2}\,\mathrm{Tr}\left(A\wedge d\,A + \frac{2}{3}A\wedge A\wedge A\right).$$

The Chern–Simons invariant of the connection A is then defined to be the integral

$$\mathrm{cs}_M(A) = \int_{s(M)} \mathrm{cs}(A) \in \mathbb{R}/\mathbb{Z},$$

where the integral is over a section s of the $SO(3)$-bundle (i.e., an orthonormal frame field on M) [Chern and Simons 74]. If M is hyperbolic, we define $\mathrm{cs}(M)$ to be the Chern–Simons invariant of the connection defined by the hyperbolic metric.

The definition of the Chern–Simons invariant for hyperbolic three-manifolds with cusps is due to R. Meyerhoff [Meyerhoff 86]. It is defined modulo 1/2 by using a special singular frame field which is linear near the cusps. See [Meyerhoff 86] for details. See [Coulson et al. 00] to examine how it is computed by SnapPea [Weeks 02]. Throughout this paper, we use another normalization $\mathrm{CS}(M) = -2\pi^2\,\mathrm{cs}(M)$ so that $\mathrm{vol}(M)+\sqrt{-1}\,\mathrm{CS}(M)$ is a natural complexification of the hyperbolic volume $\mathrm{vol}(M)$ (see [Neumann and Zagier 85, Yoshida 85]).

3. KNOT 6_3

Let us calculate the colored Jones polynomial of the knot 6_3 using the labeling as in Figure 4.

Putting $k = n_1+n_2$ and using the formula in [Murakami and Murakami 01]

$$\sum_{i=0}^{N-1}(-1)^i s^{\beta i}\begin{bmatrix}\alpha\\ i\end{bmatrix} = \prod_{j=1}^{\alpha}(1-s^{\beta+\alpha+1-2j}) \tag{3–1}$$

with $\alpha = k$, $i = n_1$, and $\beta = -k-1-2N$, we calculate $J_N(6_3)$ as shown in Figure 3.

The colored Jones polynomial of the knot 6_3 is given by

$$J_N(6_3) = \sum_{\substack{k,l,m\ge 0\\ k+l+m\le N-1}} \left|\frac{(q)_{k+l+m}}{(q)_l(q)_m}\right|^2 (q)_{k+l}(\bar{q})_{m+k}\, q^{(m-l)(k+1)}. \tag{3–2}$$

We review the technique in [Kashaev 97]. For a complex number p and a positive real number γ with $|\,\mathrm{Re}\,p| < \pi+\gamma$, we define

$$S_\gamma(p) = \exp\frac{1}{4}\int_{-\infty}^{\infty}\frac{e^{px}}{\sinh(\pi x)\sinh(\gamma x)}\frac{dx}{x}.$$

Here Re denotes the real part. This function has two properties:

(a) $\left(1+\exp(\sqrt{-1}p)\right) S_\gamma(p+\gamma) = S_\gamma(p-\gamma)$,

(b) $S_\gamma(p) \sim \exp\left(\frac{1}{2\gamma\sqrt{-1}}\mathrm{Li}_2\left(-\exp(\sqrt{-1}p)\right)\right)$ $(\gamma\to 0)$,

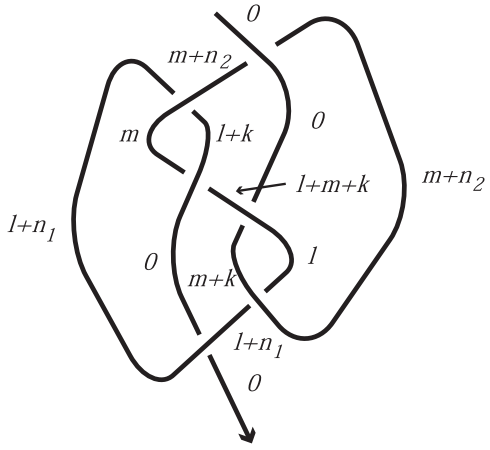

FIGURE 4. Labeling of 6_3.

where

$$\mathrm{Li}_2(z) = -\int_0^z \frac{\log(1-u)}{u}du.$$

We put

$$f_\gamma(p) = \frac{S_\gamma(\gamma-\pi)}{S_\gamma(p)}, \quad \bar{f}_\gamma(p) = \frac{S_\gamma(-p)}{S_\gamma(\pi-\gamma)},$$

so that

$$(q)_k = f_\gamma(-\pi+(2k+1)\gamma), \quad (\bar{q})_k = \bar{f}_\gamma(-\pi+(2k+1)\gamma).$$

Following Kashaev's analysis, we rewrite the formula (3–2) as a multiple integral with appropriately chosen contours. (Note that there is considerable doubt as to the contours.) By using the property (b), it can be asymptotically approximated by

$$\iiint \exp\frac{\sqrt{-1}}{2\gamma}V_{6_3}(z,u,v)\,dz\,du\,dv$$

with $\gamma = \pi/N$. Here z, u, and v correspond to q^k, q^m, and q^l, respectively, and

$$\begin{aligned}V_{6_3}(z,u,v) &= \mathrm{Li}_2(zuv) - \mathrm{Li}_2\left(\frac{1}{zuv}\right) + \mathrm{Li}_2(zv)\\ &\quad - \mathrm{Li}_2\left(\frac{1}{zu}\right) - \mathrm{Li}_2(u) + \mathrm{Li}_2\left(\frac{1}{u}\right)\\ &\quad - \mathrm{Li}_2(v) + \mathrm{Li}_2\left(\frac{1}{v}\right) - \log z \log\frac{u}{v}.\end{aligned}$$

Then there exists a stationary point $(z_0,u_0,v_0) = (0.204323 - 0.978904\sqrt{-1}, 1.60838 + 0.558752\sqrt{-1}, 0.554788 + 0.192734\sqrt{-1})$ of V_{6_3} with

$$\mathrm{Im}\,V_{6_3}(z_0,u_0,v_0) < 0, \quad \arg z_0 + \arg u_0 + \arg v_0 \le 2\pi,$$

and we have

$$\begin{aligned}&-\mathrm{Im}\,V_{6_3}(z_0,u_0,v_0) = 5.693021\ldots,\\ &\mathrm{Re}\,V_{6_3}(z_0,u_0,v_0) = 0.\end{aligned}$$

From values of $\mathrm{vol}(6_3)$ and $\mathrm{CS}(6_3)$ given by SnapPea, we see that the equation

$$\exp\frac{\sqrt{-1}}{2\gamma}V_{6_3}(z_0,u_0,v_0) = \exp\frac{\mathrm{vol}(6_3)+\sqrt{-1}\,\mathrm{CS}(6_3)}{2\gamma}$$

holds up to digits shown above.

4. KNOT 8_9

We label the edges of the $(1,1)$-tangle presentation of the knot 8_9 as in Figure 5.

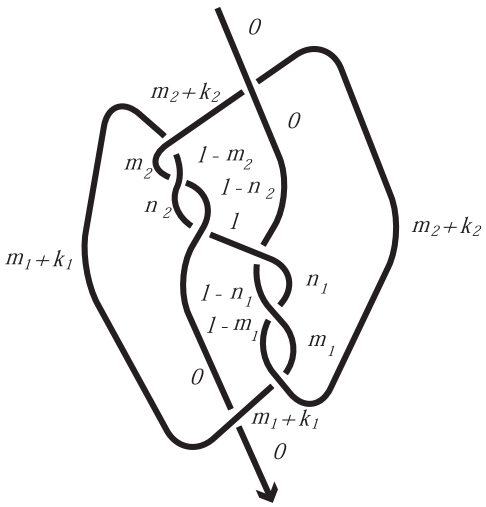

FIGURE 5. Labeling of 8_9.

We obtain the following formula of the colored Jones polynomial of the knot 8_9, where we put $l = m_1+m_2+k_1+k_2$ and use the formula (3–1):

$$\begin{aligned}J_N(8_9) &= \sum_{\substack{0\le l,m_1,m_2,n_1,n_2\le N-1\\ m_1+n_1,m_2+n_2\le l\\ m_1+m_2\le l}} \left|\frac{(q)_{l-m_1}(q)_l(q)_{l-m_2}}{(q)_{m_1}(q)_{m_2}(q)_{n_1}(q)_{n_2}}\right|^2\\ &\quad\times\frac{(\bar{q})_{l-n_1}(q)_{l-n_2}}{(q)_{l-m_1-n_1}(\bar{q})_{l-m_2-n_2}}\\ &\quad\times q^{(m_2-m_1)(l-m_1-m_2)+(n_2-n_1)(l-n_1-n_2)+m_2-m_1+n_2-n_1},\end{aligned}$$

which can be asymptotically approximated by

$$\int\cdots\int \exp\frac{\sqrt{-1}}{2\gamma}V_{8_9}(x,y,z,u,v)\,dx\,dy\,dz\,du\,dv,$$

where x, y, z, u, and v correspond to q^{-l}, q^{m_1}, q^{m_2}, q^{n_1}, and q^{n_2} respectively, and

$$\begin{aligned}&V_{8_9}(x,y,z,u,v)\\ &\quad= -\mathrm{Li}_2(xy) + \mathrm{Li}_2\left(\frac{1}{xy}\right) - \mathrm{Li}_2(xz) + \mathrm{Li}_2\left(\frac{1}{xz}\right)\\ &\quad- \mathrm{Li}_2(xu) + \mathrm{Li}_2\left(\frac{1}{xv}\right) - \mathrm{Li}_2(x) + \mathrm{Li}_2\left(\frac{1}{x}\right) - \mathrm{Li}_2(y)\\ &\quad+ \mathrm{Li}_2\left(\frac{1}{y}\right) - \mathrm{Li}_2(z) + \mathrm{Li}_2\left(\frac{1}{z}\right) - \mathrm{Li}_2(u) + \mathrm{Li}_2\left(\frac{1}{u}\right)\\ &\quad- \mathrm{Li}_2(v) + \mathrm{Li}_2\left(\frac{1}{v}\right) + \mathrm{Li}_2(xzv) - \mathrm{Li}_2\left(\frac{1}{xyu}\right)\\ &\quad- \log\frac{y}{z}\log(xzv) - \log\frac{u}{v}\log(xyu).\end{aligned}$$

Thus, we have

$$\begin{aligned}&-\mathrm{Im}\,V_{8_9}(x_0,y_0,z_0,u_0,v_0) = 7.5881802\ldots,\\ &\mathrm{Re}\,V_{8_9}(x_0,y_0,z_0,u_0,v_0) = 0\end{aligned}$$

for

$$\begin{aligned}
x_0 &= 0.7366011609 - 0.6763273835\sqrt{-1},\\
y_0 &= 0.4472176075 - 0.1647027124\sqrt{-1},\\
z_0 &= 1.968989044 - 0.7251455025\sqrt{-1},\\
u_0 &= 0.3859112582 - 0.0202712198\sqrt{-1},\\
v_0 &= 2.584139126 - 0.1357401508\sqrt{-1}
\end{aligned}$$

satisfying

$$\begin{aligned}
&\operatorname{Im} V_{8_9}(x_0, y_0, z_0, u_0, v_0) < 0,\\
&\arg x_0 + \arg y_0 + \arg u_0 \le 2\pi,\\
&\arg x_0 + \arg z_0 + \arg v_0 \le 2\pi,\\
&\arg x_0 + \arg u_0 + \arg v_0 \le 2\pi.
\end{aligned}$$

It follows from the calculation by SnapPea that

$$\begin{aligned}
&\exp\frac{\sqrt{-1}}{2\gamma} V_{8_9}(x_0, y_0, z_0, u_0, v_0)\\
&\quad = \exp\frac{\operatorname{vol}(8_9) + \sqrt{-1}\,\operatorname{CS}(8_9)}{2\gamma},
\end{aligned}$$

up to digits shown above.

5. KNOT 8_{20}

In this section, we discuss a relation between the asymptotic behavior of the colored Jones polynomial and the Chern–Simons invariant for the knot 8_{20}. We label each edge in the diagram of the knot in Figure 6.

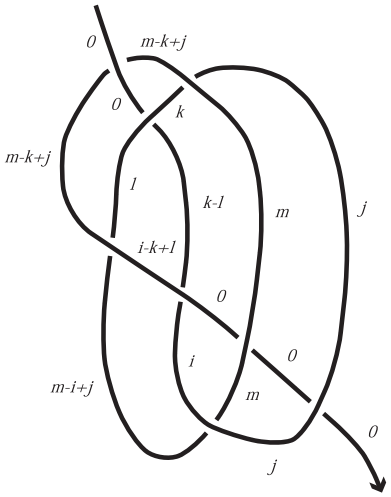

FIGURE 6. Labeling of 8_{20}.

The N-colored Jones polynomial of the knot 8_{20} is given by

$$\sum_{\substack{j,l\le k\le i+l\le j+m\\ j\le i\\ 0\le i,j,k,l,m\le N-1}} \frac{\{(\bar{q})_i (q)_k (\bar{q})_m\}^2}{\{(\bar{q})_j (q)_l\}^2 (q)_{k-l} (\bar{q})_{i-k+l} (\bar{q})_{j+m-i-l} (q)_{i-j} (q)_{k-j}} \times q^{k+m+im+km-il}, \quad (5\text{–}1)$$

which can be rewritten in the integral

$$\int\cdots\int \exp\frac{\sqrt{-1}}{2\gamma} V_{8_{20}}(x, y, z, u, v)\, dx\, dy\, dz\, du\, dv$$

with

$$\begin{aligned}
V_{8_{20}}(x, y, z, u, v) = {}& -2\operatorname{Li}_2(x) + 2\operatorname{Li}_2\left(\frac{1}{y}\right) + 2\operatorname{Li}_2(z)\\
&- 2\operatorname{Li}_2\left(\frac{1}{u}\right) - 2\operatorname{Li}_2\left(\frac{1}{x}\right) - \operatorname{Li}_2\left(\frac{1}{xy}\right)\\
&- \operatorname{Li}_2\left(\frac{z}{y}\right) - \operatorname{Li}_2(zu) + \operatorname{Li}_2(xzu)\\
&+ \operatorname{Li}_2\left(\frac{1}{xyuv}\right) + \log x \log u\\
&+ \log x \log v - \log z \log v + \frac{\pi^2}{2}.
\end{aligned}$$

Here x, y, z, u, and v correspond to q^{-i}, q^j, q^k, q^{-l}, and q^m, respectively.

Stationary points are solutions to partial differential equations,

$$\frac{\partial V_{8_{20}}}{\partial x} = \frac{\partial V_{8_{20}}}{\partial y} = \frac{\partial V_{8_{20}}}{\partial z} = \frac{\partial V_{8_{20}}}{\partial u} = \frac{\partial V_{8_{20}}}{\partial v} = 0.$$

From these equations, we have the following system of algebraic equations:

$$\begin{aligned}
&(1-x)^2\left(1-\frac{1}{xyuv}\right)uv = \left(1-\frac{1}{xy}\right)(1-xzu),\\
&\left(1-\frac{1}{xy}\right)\left(1-\frac{z}{y}\right) = \left(1-\frac{1}{y}\right)^2\left(1-\frac{1}{xyuv}\right),\\
&(1-z)^2(1-xzu)\,v = (1-zu)\left(1-\frac{z}{y}\right),\\
&(1-zu)\left(1-\frac{1}{xyuv}\right)x = \left(1-\frac{1}{u}\right)^2(1-xzu),\\
&\left(1-\frac{1}{v}\right)^2 z = \left(1-\frac{1}{xyuv}\right)x.
\end{aligned}$$

Using MAPLE V, we get a stationary point $(x_0, y_0, z_0, u_0, v_0)$ which satisfies the conditions

$$\arg\frac{1}{u_0} \le \arg z_0, \quad \arg z_0 \le \arg\frac{1}{x_0} + \arg\frac{1}{u_0}$$

from the range in the summation in (5–1), and

$$\operatorname{Im} V_{8_{20}}(x_0, y_0, z_0, u_0, v_0) < 0,$$

where Im denotes the imaginary part. Note that the range of (5–1) can be read as

$$\begin{gathered}\arg\frac{1}{u} \le \arg z \le \arg\frac{1}{x} + \arg\frac{1}{u},\\ \arg\frac{1}{x} + \arg\frac{1}{y} + \arg\frac{1}{u} \le \arg v,\\ 0 \le \arg\frac{1}{x} + \arg\frac{1}{y},\\ 0 \le \arg\frac{1}{x}, \arg z, \arg\frac{1}{u}, \arg v \le 2\pi.\end{gathered}$$

To put it concretely,

$$\begin{aligned}x_0 &= 2.878599677 + 2.657408013\sqrt{-1},\\ y_0 &= \infty,\\ z_0 &= -0.4425377456 - 0.4544788919\sqrt{-1},\\ u_0 &= 0.3542198353 - 0.02180673815\sqrt{-1},\\ v_0 &= 0.1458832937 - 0.3399257634\sqrt{-1}.\end{aligned}$$

Then we obtain

$$\begin{aligned}&-\operatorname{Im} V_{8_{20}}(x_0, y_0, z_0, u_0, v_0) = 4.1249032\ldots,\\ &-\frac{\operatorname{Re} V_{8_{20}}(x_0, y_0, z_0, u_0, v_0) + \pi^2}{2\pi^2} = 0.1033634\ldots.\end{aligned}$$

Applying values of $\mathrm{vol}(8_{20})$ and $\mathrm{CS}(8_{20})$ given by SnapPea [Weeks 02], we see that the following equation holds up to digits shown above.

$$\begin{aligned}&\exp\frac{\sqrt{-1}}{2\gamma} V_{8_{20}}(x_0, y_0, z_0, u_0, v_0)\\ &\quad = \exp\frac{\mathrm{vol}(8_{20}) + \sqrt{-1}\,\mathrm{CS}(8_{20})}{2\gamma}.\end{aligned}$$

Note that $\mathrm{CS}(8_{20})$ is defined modulo π^2.

6. WHITEHEAD LINK

For the final example, we calculate the limit of the colored Jones polynomial of the Whitehead link given by (2–1), which can be changed to the formula

$$J_N(L) = \sum_{\substack{0\le i,j,k\le N-1\\ k\le i,j}} \frac{\{(\bar{q})_i(\bar{q})_j\}^2}{(q)_k^4(\bar{q})_{i-k}(\bar{q})_{j-k}} q^{-(N-1)N/2}.$$

This can be asymptotically approximated by

$$\iiint \exp\frac{\sqrt{-1}}{2\gamma} V_L(x, y, z)\, dx\, dy\, dz,$$

where

$$\begin{aligned}V_L(x, y, z) = &-2\mathrm{Li}_2\left(\frac{1}{x}\right) - 2\mathrm{Li}_2\left(\frac{1}{y}\right) - 4\mathrm{Li}_2(z)\\ &+ \mathrm{Li}_2\left(\frac{z}{x}\right) + \mathrm{Li}_2\left(\frac{z}{y}\right) + \pi^2,\end{aligned}$$

and x, y, and z correspond to q^i, q^j, and q^k respectively. For a stationary point $(x_0, y_0, z_0) = (\infty, \infty, 1+\sqrt{-1})$, we obtain

$$\begin{aligned}&-\operatorname{Im} V_L(x_0, y_0, z_0) = 3.663862\ldots,\\ &-\frac{\operatorname{Re} V_L(x_0, y_0, z_0)}{2\pi^2} = -0.1250000\ldots.\end{aligned}$$

Since these values agree with SnapPea, the equation

$$\exp\frac{\sqrt{-1}}{2\gamma} V_L(x_0, y_0, z_0) = \exp\frac{\mathrm{vol}(L) + \sqrt{-1}\,\mathrm{CS}(L)}{2\gamma}$$

holds up to digits shown above.

7. TOPOLOGICAL CHERN-SIMONS INVARIANT AND SOME EXAMPLES

We propose a topological definition of the Chern–Simons invariant for links.

For a link L, if there exists the limit

$$2\pi \operatorname{Im} \lim_{N\to\infty} \log\frac{J_{N+1}(L)}{J_N(L)} \mod \pi^2,$$

then we denote it by $\mathrm{CS}_{\mathrm{TOP}}(L)$ and call it the *topological Chern–Simons invariant* of L.

Let us give some numerical examples.

For the knot 5_2, we list some values of $(N, 2\pi\log(J_{N+1}(5_2)/J_N(5_2)))$ by Pari-Gp in Table 1.

By fitting the above data to quadratic functions on $1/N$, we can obtain the limit value

$$2.82813 - 3.02414\sqrt{-1}$$

of $2\pi\log(J_{N+1}(5_2)/J_N(5_2))$ as $N \to \infty$ numerically, which agrees with the value

$$2.8281220 - 3.02412837\sqrt{-1}$$

by SnapPea. We display our data graphically in Figures 7 and 8, which help us to see the limit.

Similarly, for the Whitehead link L, we illustrate our numerical check in Table 2, Figure 9, and Figure 10.

Fitting, we get the numerical limit value $3.66386 + 2.46742\sqrt{-1}$ of $2\pi\log(J_{N+1}(L)/J_N(L))$ as $N \to \infty$, which agrees with our result in Section 6.

$(40, 3.05822372126184272261388 5956 - 3.022924613281720287391974968\sqrt{-1})$
$(50, 3.01308150853018835357385 4822 - 3.023340368517507069134855780\sqrt{-1})$
$(60, 2.98274431875358069682177 2299 - 3.023574042878935429645720640\sqrt{-1})$
$(70, 2.96095540496173917074911 4151 - 3.023717381786374852930574631\sqrt{-1})$
$(80, 2.94454826917045011244696 6301 - 3.023811574968472287718611711\sqrt{-1})$
$(100, 2.92148390610822899301846 9212 - 3.023923719027833555669502480\sqrt{-1})$
$(120, 2.90604642138866600028254 2398 - 3.023985374930307234443986632\sqrt{-1})$
$(150, 2.89055988190753712837200 1511 - 3.024036295143969179028770901\sqrt{-1})$
$(200, 2.87502423422694162032715 6350 - 3.024076266558545340852410631\sqrt{-1})$
$(250, 2.86567925096953853156209 9056 - 3.024094905811349375139149331\sqrt{-1})$

TABLE 1.

$(40, 3.89292035910181109780952 5583 + 2.457483997330866045812504703\sqrt{-1})$
$(50, 3.84816146640291422515453 0180 + 2.461039474018016569869745301\sqrt{-1})$
$(60, 3.81802901334949931270823 6153 + 2.462976748675980254703390855\sqrt{-1})$
$(70, 3.79636250120953769107894 4556 + 2.464147191795881614582476451\sqrt{-1})$
$(80, 3.78003432756002219508201 5385 + 2.464907923404764622274395868\sqrt{-1})$
$(100, 3.75706225898547785724799 1239 + 2.465803785962819679236327339\sqrt{-1})$
$(120, 3.74167460817902367315914 4258 + 2.466291085896660260688606142\sqrt{-1})$
$(150, 3.72622864972655859050782 8429 + 2.466690204011030007962113880\sqrt{-1})$

TABLE 2. $(N, 2\pi \log(J_{N+1}(L)/J_N(L)))$ for the Whitehead link L

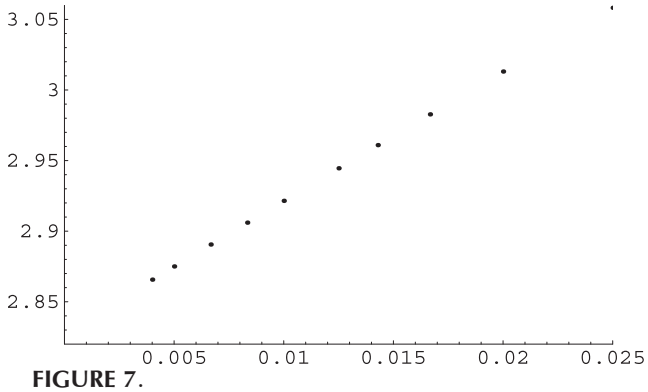

FIGURE 7.
Dots indicate $(1/N, 2\pi \operatorname{Re} \log(J_{N+1}(5_2)/J_N(5_2)))$ for $N = 40, 50, 60, 70, 80, 100, 120, 150, 200, 250$. The origin corresponds to $(0, 2.82)$.

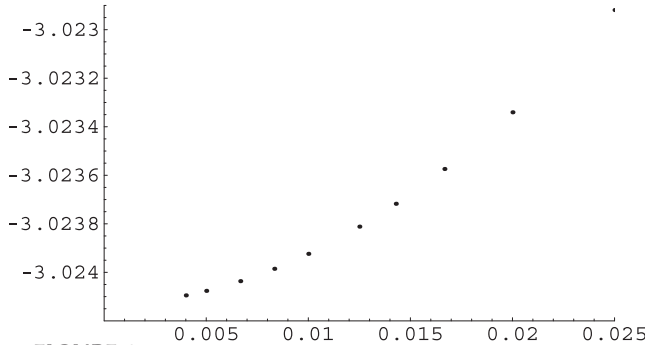

FIGURE 8.
Dots indicate $(1/N,\ 2\pi \operatorname{Im} \log(J_{N+1}(5_2)/J_N(5_2)))$ for $N = 40, 50, 60, 70, 80, 100, 120, 150, 200, 250$. The origin corresponds to $(0, -3.0242)$.

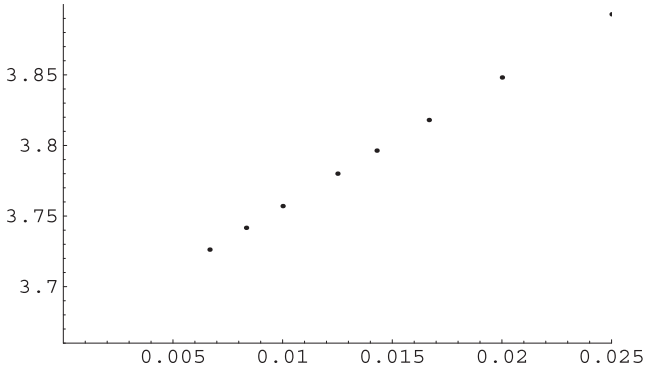

FIGURE 9.
Dots indicate $(1/N, 2\pi \operatorname{Re}\log(J_{N+1}(L)/J_N(L)))$ for $N = 40, 50, 60, 70, 80, 100, 120, 150$. The origin corresponds to $(0, 3.66)$.

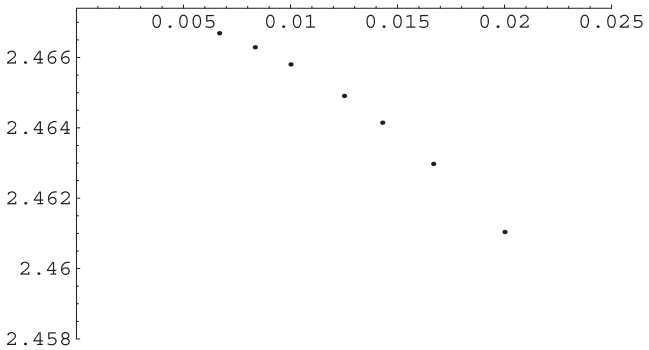

FIGURE 10.
Dots indicate $(1/N, 2\pi \operatorname{Im}\log(J_{N+1}(L)/J_N(L)))$ for $N = 40, 50, 60, 70, 80, 100, 120, 150$. The origin corresponds to $(0, 2.4674)$.

8. CONCLUSION

We have shown the following by direct calculation.

Observation 8.1. Let L be one of the hyperbolic knots 6_3, 8_9, and 8_{20}, or the Whitehead link. Following Kashaev's way, we approximate the colored Jones polynomial $J_N(L)$ of L asymptotically by

$$\int \cdots \int \exp \frac{N\sqrt{-1}}{2\pi} V_L(\mathbf{x}) d\mathbf{x}.$$

Then there exists a stationary point $\mathbf{x}_0$ of V_L such that the formula

$$\exp \frac{N\sqrt{-1}}{2\pi} V_L(\mathbf{x}_0) = \exp \frac{N}{2\pi}(\operatorname{vol}(L) + \sqrt{-1}\,\mathrm{CS}(L))$$

holds up to 6 digits.

Conjecture 8.2. (Complexification of Kashaev's conjecture.) *Let L be a hyperbolic link. Then, it holds that*

$$\operatorname{vol}(L) = 2\pi \lim_{N\to\infty} \frac{\log|\langle L\rangle_N|}{N}$$

with $\operatorname{vol}(L)$ *the hyperbolic volume of the complement of L. Moreover, there exists the topological Chern–Simons invariant* $\mathrm{CS}_{TOP}(L)$ *of L*

$$\mathrm{CS}_{TOP}(L) = 2\pi \operatorname{Im} \lim_{N\to\infty} \log \frac{J_{N+1}(L)}{J_N(L)} \mod \pi^2,$$

and $\mathrm{CS}_{TOP}(L)$ *equals to* $\mathrm{CS}(L)$ *modulo π^2. Here* $\mathrm{CS}(L)$ *is the Chern–Simons invariant of L* [Chern and Simons 74, Meyerhoff 86]. *Note that the complement of L is a hyperbolic manifold with cusps.*

We note that Observation 8.1 also holds for the knots 4_1, 5_2 and 6_1 by calculating Kashaev's examples in [Kashaev 97] using MAPLE V and SnapPea.

Therefore we conclude that the complexified Kashaev conjecture is true, up to several digits, up to choices of contours when we change summations into integrals, and up to choices of saddle (stationary) points when we approximate integrals by the saddle point method, for the six hyperbolic knots above and for the Whitehead link.

Note that if the complexified Kashaev conjecture is true then the topological Chern–Simons invariant of a hyperbolic link coincides with its Chern–Simons invariant associated with the hyperbolic metric. Moreover if the volume conjecture is true then the colored Jones polynomial would give both the simplicial volume and the topological Chern–Simons invariant for any knot.

ACKNOWLEDGMENTS

We thank the participants in the meeting "Volume conjecture," October 1999 and those in the workshop "Recent Progress toward the Volume Conjecture," March 2000, both of which were held at the International Institute for Advanced Study. The latter was financially supported by the Research Institute for Mathematical Sciences, Kyoto University. We are grateful to both of the institutes.

H. M., J. M. and M. O. express their gratitude to the Graduate School of Mathematics, Kyushu University, where the essential part of this work was carried out in December 1999.

Thanks are also due to S. Kojima for his suggestion of the Chern–Simons invariant, to K. Mimachi for valuable discussions, and to K. Hikami for information on Pari-Gp [Cohen 02] and fitting.

This research is partially supported by Grant-in-Aid for Scientific Research, The Ministry of Education, Science, Sports and Culture.

REFERENCES

[Chern and Simons 74] S.-S. Chern and J. Simons. "Characteristic forms and geometric invariants." *Ann. of Math. (2)* **99** (1974) 48–69.

[Cohen 02] H. Cohen. *Pari-Gp*: *a computer program for number theory.* available at http://www.parigp-home.de/.

[Coulson et al. 00] D. Coulson, O. A. Goodman, C. D Hodgson, and W. D. Neumann. "Computing arithmetic invariants of 3-manifolds." *Experiment. Math.* **9**:1 (2000) 127–152.

[Kashaev and Tirkkonen 00] R. M. Kashaev and O. Tirkkonen. "A proof of the volume conjecture on torus knots." *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 269(Vopr. Kvant. Teor. Polya i Stat. Fiz. 16) (2000) 262–268, 370.

[Kashaev 95] R. M. Kashaev. "A link invariant from quantum dilogarithm." *Modern Phys. Lett. A* **10**:19 (1995), 1409–1418.

[Kashaev 97] R. M. Kashaev. "The hyperbolic volume of knots from the quantum dilogarithm." *Lett. Math. Phys.* **39**:3 (1997) 269–275.

[Kirby and Melvin 91] R. Kirby and P. Melvin. "The 3-manifold invariants of Witten and Reshetikhin–Turaev for $sl(2,\mathbb{C})$." 105: (1991) 473–545.

[Meyerhoff 86] R. Meyerhoff. "Density of the Chern–Simons invariant for hyperbolic 3-manifolds." In *Low-dimensional topology and Kleinian groups (Coventry/Durham, 1984)*, pp. 217–239, Cambridge Univ. Press, Cambridge, 1986.

[Murakami and Murakami 01] H. Murakami and J. Murakami. "The colored Jones polynomials and the simplicial volume of a knot." *Acta Math.* **186**:1 (2001) 85–104.

[Neumann and Zagier 85] W. D. Neumann and D. Zagier. "Volumes of hyperbolic three-manifolds." *Topology* **24**:3 (1985) 307–332.

[Thurston 99] D. Thurston. *Hyperbolic volume and the Jones polynomial.* Lecture notes, École d'été de Mathématiques 'Invariants de nœuds et de variétés de dimension 3', Institut Fourier - UMR 5582 du CNRS et de l'UJF Grenoble (France) du 21 juin au 9 juillet 1999.

[Weeks 02] J. Weeks. *SnapPea*: *a computer program for creating and studying hyperbolic* 3*-manifolds.* available at http://www.northnet.org/weeks/index/SnapPea.html.

[Yokota 00] Y. Yokota. "On the volume conjecture of hyperbolic knots." In *Knot Theory – dedicated to Professor Kunio Murasugi for his 70th birthday*, M. Sakuma, editor, pp 362–367, March 2000.

[Yokota 02] Y. Yokota. "On the volume conjecture for hyperbolic knots." preprint, available at www.comp.metro-u.ac.jp/˜jojo/volume_conjecture.ps.

[Yoshida 85] T. Yoshida. "The η-invariant of hyperbolic 3-manifolds." *Invent. Math.* **81**:3 (1985) 473–514.

Hitoshi Murakami, Department of Mathematics, Tokyo Institute of Technology, Oh-okayama, Meguro, Tokyo 152-8551, Japan (starshea@tky3.3web.ne.jp)

Jun Murakami, Department of Mathematical Sciences, School of Science and Engineering, Waseda University, 3-4-1, Ohkubo, Shinjuku-ku, Tokyo, 169-8555 Japan (murakami@waseda.jp)

Miyuki Okamoto, University of the Sacred Heart, Hiroo, Shibuya, Tokyo 150-8938, Japan (miyuki3@hh.iij4u.or.jp)

Toshie Takata, Department of Mathematics, Faculty of Science, Niigata University, Niigata 950-2181, Japan (takata@math.sc.niigata-u.ac.jp)

Yoshiyuki Yokota, Department of Mathematics, Tokyo Metropolitan University, Tokyo 192-0397, Japan (jojo@math.metro-u.ac.jp)

Received April 30, 2001; accepted in revised form October 24, 2002.

The EKG Sequence

J. C. Lagarias, E. M. Rains and N. J. A. Sloane

CONTENTS

The EKG or electrocardiogram sequence is defined by $a(1) = 1$, $a(2) = 2$ and, for $n \geq 3$, $a(n)$ is the smallest natural number not already in the sequence with the property that $\gcd\{a(n-1), a(n)\} > 1$. In spite of its erratic local behavior, which when plotted resembles an electrocardiogram, its global behavior appears quite regular. We conjecture that almost all $a(n)$ satisfy the asymptotic formula $a(n) = n(1 + 1/(3 \log n)) + o(n/\log n)$ as $n \to \infty$; and that the exceptional values $a(n) = p$ and $a(n) = 3p$, for p a prime, produce the spikes in the EKG sequence. We prove that $\{a(n) : n \geq 1\}$ is a permutation of the natural numbers and that $c_1 n \leq a(n) \leq c_2 n$ for constants c_1, c_2. There remains a large gap between what is conjectured and what is proved.

1. INTRODUCTION

Consider the sequence defined by $a(1) = 1$, $a(2) = 2$ and, for $n \geq 3$, $a(n)$ is the smallest natural number not in $\{a(k) : 1 \leq k \leq n-1\}$ with the property that $\gcd\{a(n-1), a(n)\} \geq 2$. This sequence might be called a greedy gcd sequence, but because of its striking appearance when plotted, we will name it the *EKG* (or *electrocardiogram*) *sequence*—see Figures 1, 2. It was apparently first discovered by Jonathan Ayres [Ayres 01] and appears as sequence A064413 in [Sloane 01]. The first 30 terms are

1	2	4	6	3	9	12	8	10	5
15	18	14	7	21	24	16	20	22	11
33	27	30	25	35	28	26	13	39	36	⋯

Although the local behavior is erratic, plots of the first 1000 or 10000 terms show considerable regularity (see Figures 3 and 4 in Section 4).

The EKG sequence has a simple recursive definition, yet seems surprisingly difficult to analyze. Its definition combines both additive and multiplicative aspects of the integers, and the greedy property of its definition produces a complicated dependence on the earlier terms of the sequence. Indeed, it is not immediately obvious whether it contains all positive integers, but we show

2000 AMS Subject Classification: Primary 11Bxx, 11B83, 11B75; Secondary 11N36

Keywords: Electrocardiagram sequence, EKG sequence

1058-6458/2001 $0.50 per page

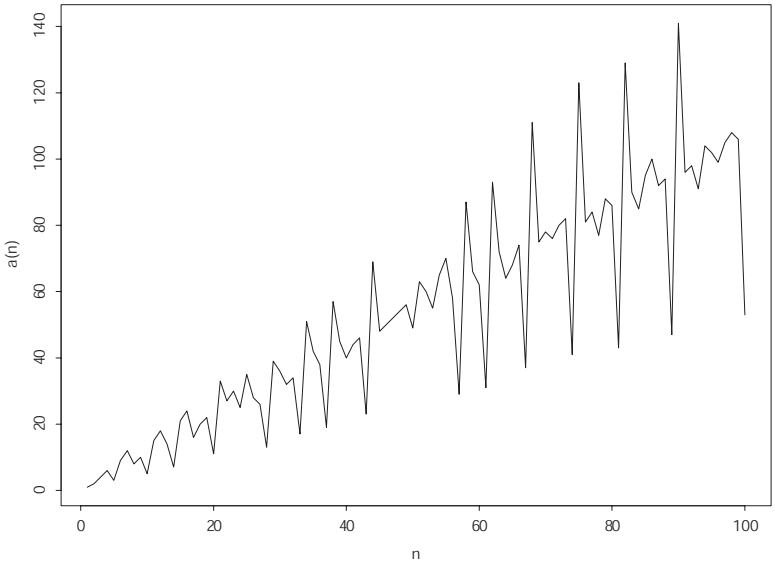

FIGURE 1. Plot of $a(1)$ to $a(100)$, with successive points joined by lines.

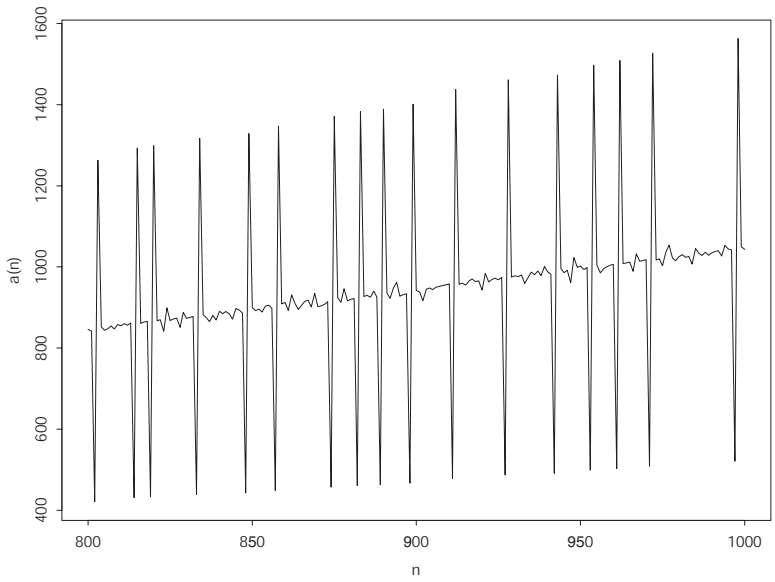

FIGURE 2. Terms 800 to 1000, with successive points joined by lines.

this is the case—the EKG sequence is a permutation of the positive integers.

In comparing Figures 1 and 3, one is reminded of the contrast between the irregular plot of $\pi(x)$ (the number of primes $\leq x$) for $x \leq 100$ and the very smooth plot for $x \leq 50000$ as shown in Don Zagier's lecture on "The first 50 million prime numbers" [Zagier 77]. This is not a coincidence, as we will see, because (experimentally) the spikes in the EKG sequence are associated with the primes arranged in increasing order. However, the spacings between spikes are not the same as the spacings between consecutive primes.

Although the EKG sequence itself seems only to have been proposed recently, in the early 1980s Erdős, Freud, and Hegyvari [Erdős et al. 83] studied properties of integer permutations with restrictions placed on allowed values of greatest common divisors of consecutive terms.

In Section 2 we derive a number of basic properties of the EKG sequence, and prove that it is a permutation of the natural numbers $\mathbb{N}$. An efficient algorithm for computing the sequence is given in Section 3. Using this algorithm we computed 10^7 terms; this led us to conjectured asymptotic formulae given in Section 4. We give a heuristic argument why these formulae may be true,

but it seems likely they will be very hard to prove. We are able to rigorously extablish linear upper and lower bounds on the EKG sequence, namely $\frac{1}{260}n \le a(n) \le 14n$ (Sections 5 and 6); the proofs use sieving ideas. Section 7 discusses experimental results concerning the cycle structure of the associated permutation of $\mathbb{N}$. The final section discusses generalizations to other sorts of integer permutations resulting from greedy constructions with restrictions on gcd's of consecutive terms.

2. THE SEQUENCE IS A PERMUTATION

We begin with some general remarks about the sequence.

For $n \ge 2$, let $g = \gcd\{a(n-1), a(n)\}$. For some prime p dividing $a(n-1)$, $a(n)$ is the smallest multiple of p not yet seen (otherwise, the smaller multiple of p would be a better candidate for $a(n)$). We call such primes p the *controlling primes* for $a(n)$. There may be more than one, and their product divides g.

For any prime p and $n \ge 2$, let $B_p(n)$ be the smallest multiple of p that is not in $\{a(1), \ldots, a(n-1)\}$. For example, the sequence $\{B_2(n) : n \ge 2\}$ begins 2, 4, 6, 8, 8, 8, 8, 10, 14, ... Clearly

$$B_p(n) \le B_p(n+1) \le pn \tag{2–1}$$

for all $p, n \ge 2$. Then we have $a(1) = 1$,

$$a(n) = \min\{B_p(n) : \ p \text{ divides } a(n-1)\} \ , \tag{2–2}$$

for $n \ge 2$, which provides an alternative definition of the sequence.

Lemma 2.1. *Let p be a prime > 2 that divides some term of the sequence. If p first divides $a(n)$ then $a(n) = qp$ where q is the smallest prime dividing $a(n-1)$, q is less than p, $a(n+1) = p$, and either $a(n)$ or $a(n+2)$ is equal to $2p$. The new primes that divide the terms of the sequence appear in increasing order.*

Proof: Let $a(n)$ be the first term divisible by p. The numbers pq where q is a prime dividing $a(n-1)$ are all candidates for $a(n)$, and so $a(n) = pq$ where q is the smallest such prime. Also p must be the smallest prime that has not appeared as a divisor of $\{a(1), \ldots, a(n-1)\}$ (for if p' were a smaller such prime, then $p'q$ would be a better candidate for $a(n)$). In particular, the primes that divide the terms of the sequence must appear in increasing order, and $q < p$. Then p is a candidate for $a(n+1)$, and is less than $B_q(n+1) \ge B_q(n) = pq$, so $a(n+1) = p$. Finally, either $a(n) = 2p$ or else $2p$ is the winning candidate for $a(n+2)$ □

Lemma 2.2. *The primes that appear in the sequence occur in increasing order.*

Proof: This follows from Lemma 2.1, since the first time p divides a term of the sequence the next term is p itself. □

Lemma 2.3. *If infinitely many multiples of a prime p appear in the sequence, then all multiples of p appear.*

Proof: We argue by contradiction, and let kp be the first multiple of p that is missed. Choose n_0 so that $a(n) > kp$ for all $n \ge n_0$. Since infinitely many multiples of p occur, there exists $n > n_0$ with $a(n) = lp$ for some l. But now we must have $a(n+1) = kp$, because $\gcd\{a(n), kp\} \ge p$ is allowed, and all smaller possible values which are ever going to appear in the sequence have already appeared. This is a contradiction. □

Lemma 2.4. *If all multiples of a prime p appear in the sequence, then all positive integers appear.*

Proof: Again we argue by contradiction and let $k \ge 2$ be the first integer that is missed. Since infinitely many multiples of k occur among all the multiples of p, we get a contradiction just as in Lemma 2.3. Namely, there exists for $n > n_0$ a value $a(n) = klp$ for some l, and $\gcd\{a(n), k\} \ge k$ is allowed, and all smaller possible values have already been used. Thus $a(n+1) = k$, a contradiction. □

Theorem 2.5. *$\{a(n) : n \ge 0\}$ is a permutation of the natural numbers.*

Proof: No number can appear twice, by construction, so it suffices to show that every number appears. Suppose only finitely many different primes divide the terms of the sequence. Then one of them would appear infinitely many times, and Lemmas 2.3 and 2.4 would imply that all integers occur, which is a contradiction.

Therefore, infinitely many different primes p divide the terms of the sequence. Then by Lemma 2.1 infinitely many even numbers $2p$ occur, by Lemma 2.3, all even numbers occur, and by Lemma 2.4, all positive integers occur. □

Remark 2.6. As will be discussed in Section 8, the principle of this proof generalizes to a wide variety of other integer sequences defined by restrictions on the gcd's of consecutive terms.

3. NUMERICAL INVESTIGATIONS

To compute the EKG sequence, it is better not to use the original definition, but to use (2–2) and to store the current values of $B_p(n)$ for primes p. An efficient way to arrange the computation is to maintain four tables:

$\text{hit}(m)$ = 0 if m has not yet appeared, otherwise 1;

$\text{gap}(m)$ = current value of $B_m(n)$ if m is a prime, otherwise m;

$\text{small}(m)$ = smallest prime factor of m;

$\text{quot}(m)$ = largest factor of m not divisible by $\text{small}(m)$.

Combining these tables in a C "struct" minimizes memory access.

Suppose we wish to compute the sequence until $a(n)$ reaches or exceeds N. The first step is to precompute $\text{small}(m)$ and $\text{quot}(m)$ for $m \le N$. Since it is only necessary to consider primes $\le \sqrt{N}$, this takes about $\Sigma_{p\le\sqrt{N}} N/p = O(N \log\log N)$ steps.

In the main loop, let $a(n)$ be the current value. Set $k = a(n)$, $B = N$ and repeat until k reaches 1:

$$\begin{aligned} p &= \text{small}(k)\,, \\ B &= \min\{B, \text{gap}(p)\}\,, \\ k &= \text{quot}(k)\,. \end{aligned}$$

Then we set $a(n+1) = B$, $\text{hit}(B) = 1$, and update $\text{gap}(q)$ for primes q dividing B.

We have not analyzed the complexity of the main loop in detail, but it also appears to take roughly $O(N \log\log N)$ steps, comparable to and not much greater than the number of steps needed for the precomputation part of the calculation. The program computed 10^7 terms of the sequence in less than a minute. For example, $a(10954982) = 11184814$.

4. A CONJECTURED ASYMPTOTIC FORMULA

The results from the experimental data suggest that whenever a prime p occurs in the sequence, it is preceded by $2p$ and (consequently) followed by $3p$. Although it is theoretically possible that some other multiple of p occurs before p, for example, we might have seen $\ldots, 3p, p, 2p, \ldots$, this does not happen in the first 10^7 terms.

Conjecture 4.1. *Whenever a prime p occurs in the sequence, it is immediately preceded by $2p$ (and hence followed by $3p$).*

The numerical results also strongly suggest that the terms of the sequence fall close to three lines (see Figures 3 and 4).

- if $a(n) = m$ and m is neither a prime nor three times a prime, then $a(n) \approx n$;
- if $a(n) = p$, p prime, then $a(n) \approx n/2$;
- if $a(n) = 3p$, p prime, then $a(n) \approx 3n/2$.

This was also observed by Ayres [Ayres 01]. In fact, if we smooth the sequence by replacing every term $a(n) = p$ or $3p$, p prime > 2, by $a(n) = 2p$, the terms of the sequence lie close to a single line (see Figure 5).

A plausible but nonrigorous argument suggests a more precise conjecture.

Conjecture 4.2. *Let $f(n) \sim g(n)$ mean that the ratio of the two sides approaches 1 as $n \to \infty$.*

(i) If $a(n) = m$, $m \ne p$ or $3p$ for p prime, then

$$a(n) \sim n\left(1 + \frac{1}{3\log n}\right)\,; \tag{4–1}$$

(ii) If $a(n) = p$, p prime, then

$$a(n) \sim \frac{1}{2}n\left(1 + \frac{1}{3\log n}\right)\,; \tag{4–2}$$

(iii) If $a(n) = 3p$, p prime, then

$$a(n) \sim \frac{3}{2}n\left(1 + \frac{1}{3\log n}\right)\,. \tag{4–3}$$

To see why this conjecture might be true, consider a term $a(n) = m$ of the smoothed sequence. Examination of Figure 5 suggests that the smoothed sequence has hit all the numbers from 1 to m at least once (the numbers occur a little out of order, but never by much). However, we have smoothed away the numbers p and $3p$ that are $\le m$, while picking up the primes p and $3p$ for $2p \le m$. Therefore one expects

$$n \sim m - \pi(m) - \pi\left(\frac{m}{3}\right) + 2\pi\left(\frac{m}{2}\right), \tag{4–4}$$

where $\pi(x) =$ number of primes $\le x$. Then (4–1) follows at once from the asymptotic formula $\pi(x) \sim x/\log x$. Equations (4–2) and (4–3) are based on (4–1) and the observations made at the beginning of this section.

Although we are unable to prove this conjecture, it is an excellent fit to the data.

If we try to write

$$a(n) \approx n\left(1 + \frac{1}{3\log n} + \frac{c}{(\log n)^2}\right) \quad (?) \tag{4–5}$$

then the values of c do not appear to converge to a single value (see Figure 6), although c is very often close

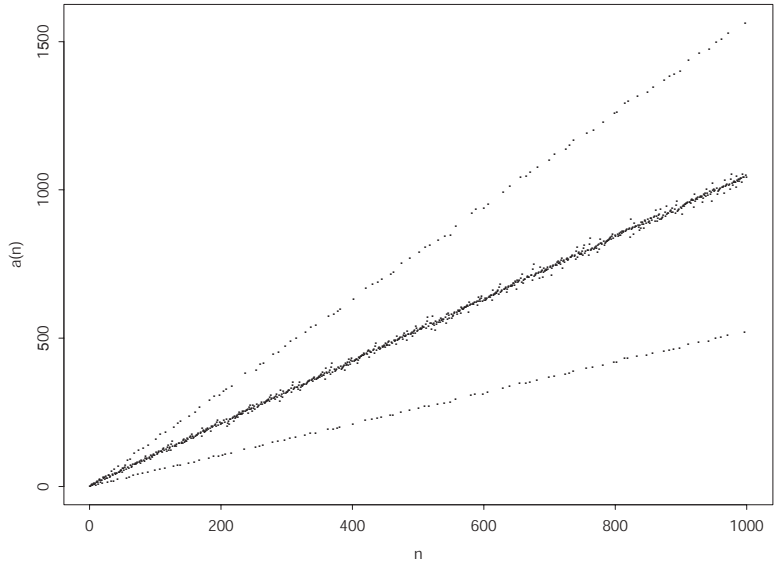

FIGURE 3. The first 1000 terms (represented by dots), successive points not joined.

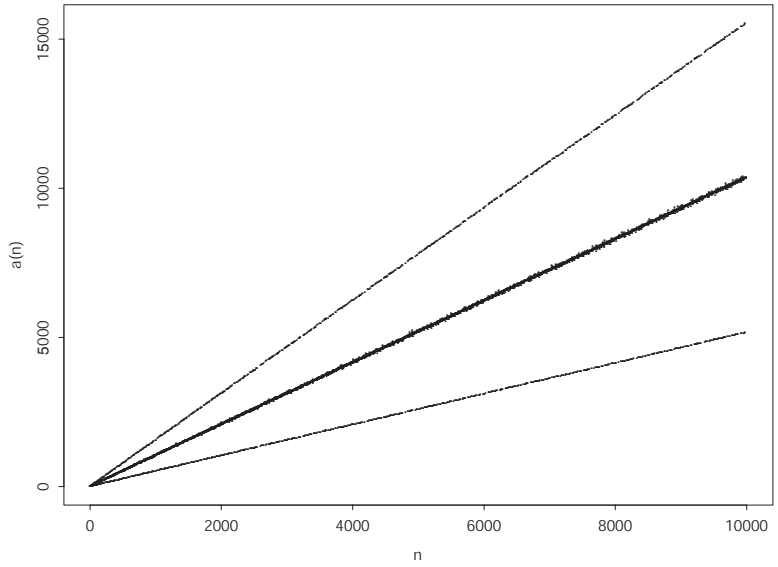

FIGURE 4. The first 10000 terms (represented by dots), successive points not joined.

to 0.11. It seems conceivable that c might converge in distribution to a limiting distribution. Using two terms of the asymptotic expansion of $\pi(x)$ would give

$$a(n) \sim n\left(1 + \frac{1}{3\log n} + \frac{c'}{(\log n)^2}\right) \quad (?) \qquad (4\text{–}6)$$

where $c' = 4/9 + (\log 3)/3 - \log 2 = 0.1175\ldots$.

Conjectures 4.1 and 4.2 predict that the k-th prime p_k will occur in the pattern $a(n) = 2p_k$, $a(n+1) = p_k$, $a(n+2) = 3p_k$, where

$$n \sim \frac{2p_k}{1 + \frac{1}{3\log(2p_k)}} \ .$$

These conjectures may be hard to settle, because the permutation $a(n)$ encodes an intricate interaction between additive and multiplicative properties of integers, which by the "greedy" property of the definition depends on all the earlier terms of the sequence.

In the next two sections we establish linear upper and lower bounds for the sequence, namely

$$\frac{1}{260}n \le a(n) \le 14n.$$

The numerical evidence supports the following conjectural bounds.

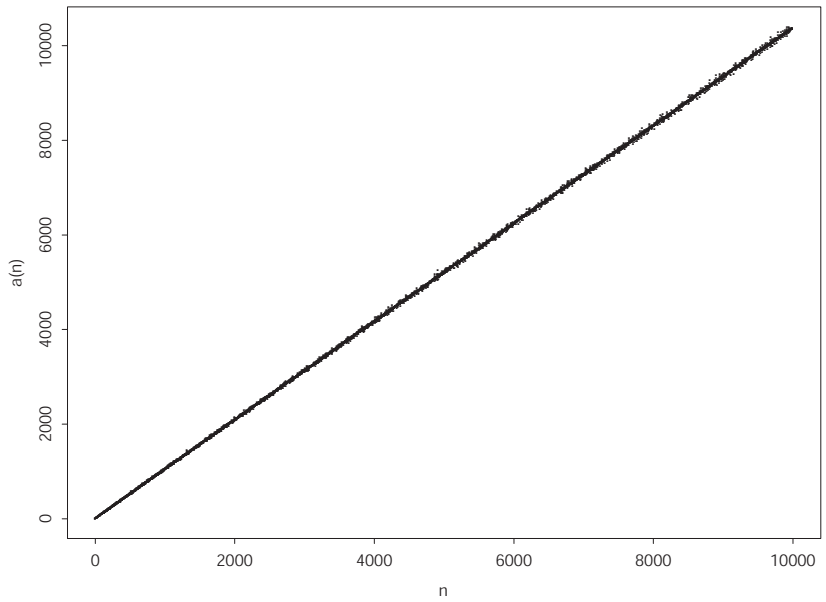

FIGURE 5. The sequence smoothed by replacing $a(n) = p$ or $3p$, p prime > 2, by $a(n) = 2p$.

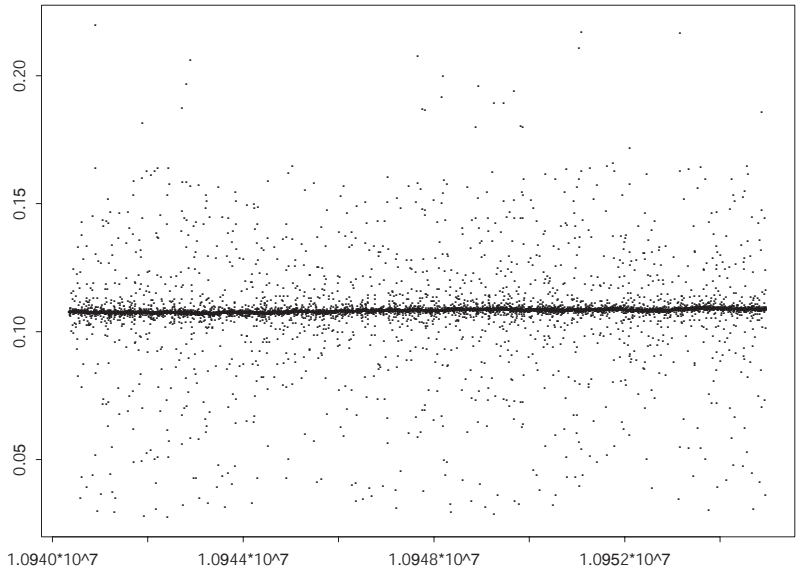

FIGURE 6. Values of c in (4–5) for n near 10^7.

Conjecture 4.3. *The sequence $a(n)$ satisfies $a(n) \geq \frac{13}{28}n$, with equality if and only if $n = 28$, and $a(n) \leq \frac{12}{7}n$, with equality if and only if $n = 7$.*

For large n the asymptotic lower and upper bounds would be $n/2$ and $3n/2$, by Conjecture 4.2.

5. A LINEAR UPPER BOUND

Theorem 5.1. *$a(n) \leq 14n$, for $n \geq 1$.*

Remark 5.2. The proof is by contradiction. The basic idea of the proof uses the fact that the function values $\{a(k) : 1 \leq k \leq n\}$ have a sandpile-like structure, in which each element can be built only at the top of a ladder of all smaller multiples of a controlling prime p dividing it. To reach a number at height exceeding $14n$ via a ladder of multiples of a controlling prime p, one repeatedly falls off this ladder while building it, at various smaller multiples kp of p of size between $2n$ and $14n$ (see property (P4) below). To get back on this ladder, one must use ladders of other controlling primes q which

reach to some currently omitted multiple of p at the top of their ladder. The total number of elements in such ladders is shown to be large using a combinatorial sieve argument (compare [Hooley 76], pp. 4-5); a contradiction results by showing that the sandpile contains more than n elements.

We begin with a preliminary lemma.

Lemma 5.3. *If $a(n)$ is divisible by a prime p, then $p \le n$. If $p \neq 2$, $p < n$.*

Proof: The result is true if $p = 2$ or $n \le 3$, so we may assume $p \ge 3$ and $n \ge 4$. We may also assume $a(n)$ is the first term divisible by p. By Lemma 2.1, $a(n-1) = pq$ where q is the controlling prime for $a(n)$. Then $pq = a(n) = B_q(n) \le q(n-1)$ by (2–1). □

Proof of Theorem 5.1: We argue by contradiction, and let n be the smallest number such that $a(n) > 14n$. By direct verification, we know n is large (in fact, $n > 10^7$, although we will not use that in the proof). Let p be the smallest controlling prime for $a(n)$, say $a(n) = lp > 14n$, so $l \ge 15$. Also $a(n) = B_p(n)$, so from (2–1) and Lemma 5.3, $17 \le p < n$.

Let $l' = \lceil l/2 \rceil \ge 8$, and consider the "window"

$$\mathcal{W} = \{l'p, (l'+1)p, \ldots, (l-1)p\}\,.$$

Since p is a controlling prime for $a(n)$, every element of $\mathcal{W}$ has already appeared in the sequence: $a(t) \in \mathcal{W}$ implies $t \le n-1$. Call $a(t) \in \mathcal{W}$ an *entry point* if $a(t-1) \notin \mathcal{W}$, an *exit point* if $a(t+1) \notin \mathcal{W}$. The following properties hold:

(P1) The number of entry points is equal to the number of exit points.

(P2) At most one entry point $a(t)$ has p as a controlling prime. This can only happen if $a(t-1) = (l'-1)p$, $a(t) = l'p$, for some $t \le n-1$.

(P3) At most one exit point $a(t)$ has p as a controlling prime for $a(t+1)$. This happens just if $a(n-1) \in \mathcal{W}$. Furthermore, if this is not the case, then $a(n-1) = ip$ with $i < l'$, and hence there is no entry point with p as a controlling prime.

(P4) Let $a(t) = \alpha p \in \mathcal{W}$, $a \in [l', l-1]$. If α is a multiple of a prime $q \le 7$, then $a(t)$ is an exit point. (For $a(t+1) \le B_q(t+1) \le 7t \le 7(n-1) < l'p$.)

For a set of primes $\mathcal{P}$, let $D_{\mathcal{P}}(a,b)$ denote the number of integers $a \le i \le b$ such that i is a multiple of some element of $\mathcal{P}$. Let

$$\theta = D_{\{2,3,5,7\}}(l', l-1)\,.$$

By (P4), the number of exit points is at least θ, while if there exists an entry point controlled by p, then there are, in fact, at least $\theta + 1$ exit points. By (P1), we conclude that the number of entry points not controlled by p is at least θ.

Let $\mathcal{S} = \{q_1, q_2, \ldots\}$ (with $q_1 < q_2 < \cdots$) denote the set of controlling primes for entry points not controlled by p. Note that $2, 3, 5, 7, p$ are not in $\mathcal{S}$ (by the same argument as in (P4)), and $|\mathcal{S}| \le \pi(l-1) - 4$. Then

$$\begin{aligned}
\theta &= D_{\{2,3,5,7\}}(l', l-1) \\
&\le \text{number of entry points not controlled by } p \\
&\le D_{\mathcal{S}}(l', l-1) \\
&\le \sum_{q \in \mathcal{S}} \left\lceil \frac{l-l'}{q} \right\rceil \\
&\le (l-l'-1) \sum_{q \in \mathcal{S}} \frac{1}{q} + |\mathcal{S}|\,. \qquad (5\text{–}1)
\end{aligned}$$

Setting $\phi = \sum_{q \in \mathcal{S}} \frac{1}{q}$, we have

$$\phi \ge \frac{\theta - \pi(l-1) + 4}{\lfloor \frac{l}{2} \rfloor - 1}\,. \qquad (5\text{–}2)$$

The right-hand side of (5–2) is a function of the single variable l, and is $\ge 2/9$ for all $l \ge 15$, with equality if and only if l is 20 or 21. (This is easily verified by computer for small l, say $l \le 1000$, and analytically for larger l.) In other words,

$$\sum_{i \ge 1} \frac{1}{q_i} \ge \frac{2}{9}\,.$$

Define k by

$$\sum_{i=1}^{k-1} \frac{1}{q_i} < \frac{2}{9} \le \sum_{i=1}^{k} \frac{1}{q_i}\,, \qquad (5\text{–}3)$$

and let $\mathcal{S}' = \{q_1, \ldots, q_k\} \subseteq \mathcal{S}$. Note that $q_k \ge 17$, since $1/11 + 1/13 + 1/17 \ge 2/9$, but no proper subset of $\{1/11, 1/13, 1/17\}$ has this property.

Every multiple of any element $q \in \mathcal{S}$ that is $\le l'p$ must have already occurred in the sequence, by definition. We obtain a contradiction by showing that there are more than n different multiples of elements of $\mathcal{S}$ that are $\le l'p$. By inclusion-exclusion. we have

$$D_{\mathcal{S}}(1, l'p) \ge D_{\mathcal{S}'}(1, l'p) \ge \sum_{i=1}^{k} \left\lfloor \frac{l'p}{q_i} \right\rfloor - \sum_{1 \le i < j \le k} \left\lfloor \frac{l'p}{q_i q_j} \right\rfloor\,. \qquad (5\text{–}4)$$

To bound the first term in (5–4), observe that for $q_i \in \mathcal{S}$ there is a multiple of $q_i p$ in $\mathcal{W}$, so

$$q_i < \frac{14n}{17} < \frac{lp}{17},$$

$$\begin{aligned}
\left\lfloor \frac{l'p}{q_i} \right\rfloor &\geq 8, \\
\left\lfloor \frac{l'p}{q_i} \right\rfloor &\geq \frac{8}{9}\frac{l'p}{q_i}, \\
\sum_{i=1}^{k} \left\lfloor \frac{l'p}{q_i} \right\rfloor &\geq \frac{8}{9}\sum_{i=1}^{k} \frac{l'p}{q_i}.
\end{aligned}$$

To bound the second term in (5–4), we use

$$\begin{aligned}
&\sum_{i\leq i<j\leq k} \frac{1}{q_i q_j} \\
&\quad = \frac{1}{2}\left(\sum_{i=1}^{k-1}\frac{1}{q_i}\right)^2 - \frac{1}{2}\sum_{i=1}^{k-1}\frac{1}{q_i^2} + \frac{1}{q_k}\sum_{i=1}^{k-1}\frac{1}{q_i} \\
&\quad < \frac{1}{2}\left(\sum_{i=1}^{k-1}\frac{1}{q_i}\right)^2 + \frac{1}{2q_k}\sum_{i=1}^{k-1}\frac{1}{q_i}.
\end{aligned}$$

Then, since $q_k \geq 17$, we have

$$\begin{aligned}
D_{\mathcal{S}}(1, l'p) &> l'p\left\{\frac{8}{9}\left(\sum_{i=1}^{k}\frac{1}{q_i}\right) - \frac{1}{2}\left(\sum_{i=1}^{k-1}\frac{1}{q_i}\right)^2 \right. \\
&\qquad \left. - \frac{1}{2q_k}\sum_{i=1}^{k-1}\frac{1}{q_i}\right\} \\
&\geq l'p\left\{\frac{8}{9}\cdot\frac{2}{9} - \frac{1}{2}\left(\frac{2}{9}\right)^2 - \frac{1}{34}\cdot\frac{2}{9}\right\} \\
&> 7n\frac{229}{1377} = \frac{1603}{1377}n > n,
\end{aligned}$$

which is the desired contradiction. □

6. A LINEAR LOWER BOUND

In this section we show:

Theorem 6.1. *$a(n) \geq \lceil \frac{1}{260} n \rceil$, for $n \geq 1$.*

Remark 6.2. The proof is a modification of that of Theorem 2.5. It aims to show that if some number less than $n/260$ is missed in $\{a(k) : 1 \leq k \leq n\}$, then there are at least $n/65$ numbers in this set that are even numbers, and Lemma 6.4 below provides the mechanism to get a contradiction. The method of Theorem 2.5 seems inherently weaker when used for a lower bound, so we have not attempted to streamline the proof. It would certainly be possible to reduce the constant 260, but not to anything close to 14.

We begin with three lemmas.

Lemma 6.3. *For a prime p, if $a(n) = kp$ for some k, then $a(j) = k$ for some $j \leq n+1$.*

Proof: We argue by induction on k. The result is true for $k = 1$ since $a(1) = 1$.

Let q be a controlling prime for $a(n)$, so $q|a(n-1)$ and $q|a(n) = kp$.

Case (i): $q \neq p$. Then $q|k$, say $k = mq$, $a(n) = mqp$. Hence all multiples iq with $i < mp$ have already appeared, and in particular $a(j) = mq = k$ for some $j < n$.

Case (ii): $q = p$. All multiples ip with $i < k$ have already occurred. By the induction hypothesis, i has occurred for all $i < k$. If k has occurred, then $a(j) = k$ with $j < n$, and otherwise $a(n+1) = k$. □

Lemma 6.4. *If at least $4k$ even numbers occur in $\{a(1), \ldots, a(n)\}$, then all numbers $\{1, \ldots, k\}$ occur in $\{a(1), \ldots, a(n+1)\}$.*

Proof: In view of Lemma 6.3 (taking $p = 2$), it is enough to show that $\{2, 4, \ldots, 2k\}$ are in $\{a(1), \ldots, a(n)\}$.

Suppose not, and let $2m$ be the largest even number $\leq 2k$ not in $\{a(1), \ldots, a(n)\}$. Every even number $a(i) > 2m$ with $i \leq n$ will be followed by $a(i+1) \leq 2k$ (since $2m$ is always available). But in $\{a(1), \ldots, a(n)\}$ we have at least $4k$ even numbers, and so at least $2k$ even numbers $> 2k$. Therefore in $\{a(1), \ldots, a(n+1)\}$, we see all the numbers from 1 to $2k$, including $2m$, a contradiction. □

Lemma 6.5. *If $a(n) = p$, a prime, then all numbers $\{1, \ldots, p-1\}$ occur in $\{a(1), \ldots, a(n-1)\}$.*

Proof: By Lemma 2.1, $a(n-1) = qp$, q prime, $q < p$; so q, $2q, \ldots, (p-1)q$ have already appeared in $\{a(1), \ldots, a(n-2)\}$. The result now follows by Lemma 6.3. □

Proof of Theorem 6.1: By direct verification, we may assume $n \geq 260^2$. Let $m = \lceil n/260 \rceil$, and suppose, seeking a contradiction, that $a(n) < m$. Note that the lower bound on n implies that $n/m > 259$.

We will show that at least $4m$ even numbers have occurred in $\{a(1), \ldots, a(n-2)\}$, which gives a contradiction

by Lemma 6.4. No primes greater than m can occur in this interval, or we get a contradiction by Lemma 6.5.

Some number $\geq n$ must occur among $\{a(1), \ldots, a(n-1)\}$, since there are $n-1$ numbers and $a(n) < m$ is missing. Suppose $a(j) \geq n$, with controlling prime $p \leq m$, say $a(j) = lp$, and $j \leq n-1$. Since $lp \geq n$, $l \geq 260$. Let $R \leq m$ be the smallest missing number among $\{a(1), \ldots, a(n-1)\}$. Then $l \leq R$, for if $l > R$, then we have seen Rp at time $j-1$, and by Lemma 6.3 we have seen R by time j, a contradiction. Therefore, $l \leq m$ and so $p \geq n/l > 259$. Both l and p are in the range $[260, m]$.

We consider the "window"

$$\mathcal{W} = \left\{ \left\lfloor \frac{lp}{4} \right\rfloor, \left\lfloor \frac{lp}{4} \right\rfloor + p, \ldots, lp \right\},$$

and define "entry" and "exit" point as in the proof of Theorem 5.1.

There are at least $\lfloor 3l/4 \rfloor$ multiples of p in $\mathcal{W}$, and at least $\lfloor 3l/8 \rfloor - 1$ of them are even. Any such even multiple of p is an exit point (from Lemma 6.4). There must therefore be at least $\lfloor 3l/8 \rfloor - 2$ entry points which are not controlled by p. Let $\mathcal{S} = \{q_1, q_2, \ldots\}$ (with $q_1 < q_2 < \cdots \leq l$) be the set of controlling primes for these entry points. Then

$$D_{\mathcal{S}}\left(\left\lfloor \frac{l}{4} \right\rfloor, l \right) \geq \lfloor 3l/8 \rfloor - 2 .$$

As in (5–1), (5–2), we get

$$\sum_{i \geq 1} \frac{1}{q_i} \geq \frac{\left\lfloor \frac{3l}{8} \right\rfloor - 2 - \pi(l)}{l - \left\lfloor \frac{l}{4} \right\rfloor} \tag{6–1}$$

which is $\geq 43/214$ for $l \geq 260$. We define k by

$$\sum_{i=1}^{k-1} \frac{1}{q_i} < \frac{43}{214} \leq \sum_{i=1}^{k} \frac{1}{q_i} . \tag{6–2}$$

Every multiple of any element $q \in \mathcal{S}$ that is $\leq \lfloor lp/4 \rfloor$ must have already occurred in $\{a(1), \ldots, a(n-2)\}$. Of these, at least $\lfloor lp/8q \rfloor$ are even. Therefore the number of distinct even numbers in the range $1, \ldots, \lfloor lp/4 \rfloor$ that have occurred is at least

$$\sum_{i \geq 1}^{k} \left\lfloor \frac{X}{2q_i} \right\rfloor - \sum_{1 \leq i < j \leq k} \left\lfloor \frac{X}{2q_i q_j} \right\rfloor$$

where $X = lp/4$. Proceeding as in the proof of Theorem 5.1, and using $q_1 \geq 2$, we find that this is at least $4.024m$. This exceeds $4m$ and provides the desired contradiction. □

7. CYCLE STRUCTURES

Since the sequence is a permutation of $\mathbb{N}$, it is also natural to investigate the cycle structure. The experimental evidence suggests that there are infinitely many finite cycles and infinitely many infinite cycles. It seems very likely to be hard to prove either of these observations, or even to prove that there is at least one infinite cycle.

The first few finite cycles start at the points

$$1, 2, 3, 8, 40, 64, 121, 149, 359, 2879, 5563, 28571, 251677,$$

and have lengths

$$1, 1, 6, 1, 1, 1, 2, 12, 11, 25, 8, 22, 11,$$

respectively. There is also a large number of apparently infinite cycles, of which the first two are

$$\cdots 229310, 117833, \ldots, 22, 27, 26, 28, 13, 14, \mathbf{7}, 12, 18, 20, 11, 15, 21, \ldots, 636551, 652766, \ldots$$

and

$$\ldots, 502008, 257519, \ldots, 248, 253, 131, 138, \mathbf{73}, 82, 129, 201, 212, \ldots, 645906, 662330, \ldots$$

with minimal representatives 7 and 73, respectively. The first 15 of these apparently distinct cycles have not coalesced in the first 700000 terms of the sequence. However, although it seems unlikely, it is theoretically possible that they could coalesce at some later point. It would be nice to know more!

8. GENERALIZATIONS

The EKG sequence can be generalized in various ways, while retaining the basic construction of a greedy sequence with a condition on gcd's of consecutive terms. For fixed $M \geq 2$, let $b(n) = n$ for $1 \leq n \leq M$, and for $n \geq M+1$, let $b(n)$ be the smallest natural number not already in the sequence with the property that $\gcd\{b(n-1), b(n)\} \geq M$. The proof of Theorem 2.5 easily extends to show that $(b(n) : n \geq 1)$ is also a permutation of $\mathbb{N}$. For the cases $M = 3, 4, 5$, see sequences A064417, A064418, A064418 in [Sloane 01].

The first 1000 terms for the case $M = 3$ are shown in Figure 7. This sequence appears to behave in a similar way to the EKG sequence: the spikes in the sequence are associated with the primes, occurring in the order $3p, p, 2p, 4p$. All points of the sequence seem to lie near lines of slope $1/3, 2/3, 1$ and $4/3$.

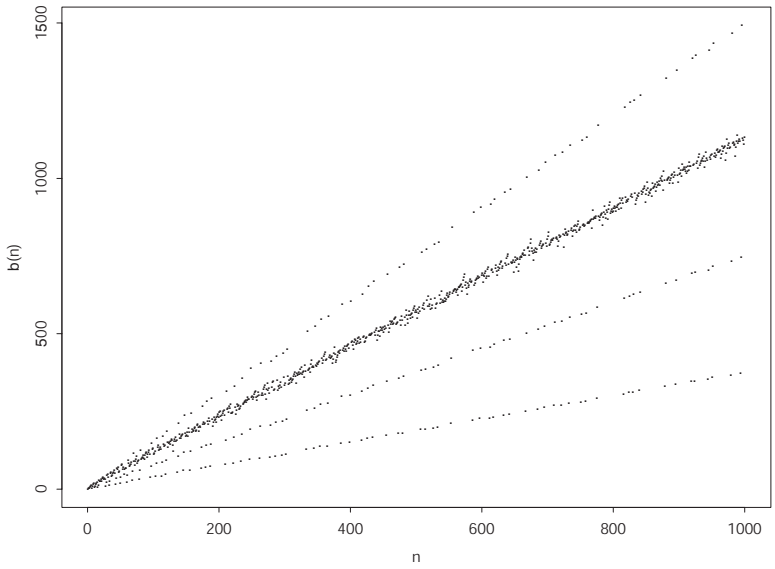

FIGURE 7. Plot of $b(1)$ to $b(1000)$ for the case $M = 3$.

More generally, we can allow an arbitrary finite prefix at the beginning of such sequences. Consider a sequence $c(n)$ with the generating rule that $\gcd\{c(n-1), c(n)\} \geq M$, for $n \geq N$, but with a finite prefix $c(n) = a_n : 1 \leq n \leq N$, such that all terms a_n are distinct natural numbers, and that the prefix includes values $a(n) = k$ for of $1 \leq k \leq M$. With a little work, the proof of Theorem 2.5 can be modified to apply to these sequences as well to show they are permutations.

If we use the greedy gcd rule with $M = 2$, starting with an arbitrary finite prefix, one obtains an infinite number of different EKG-like permutations. Many of these sequences, perhaps all, eventually coalesce with the original EKG sequence. In any case, the qualitative properties of the resulting permutations appear similar to the EKG sequence: the primes provably appear in consecutive order (excluding those in the prefix terms); the general plot of the permutations remains similar to Figure 3.

It appears that linear upper and lower bounds hold for all the permutations $c(n)$ of these types. Very likely such bounds can be established rigorously in any particular case using methods similar to those used in proving Theorems 5.1 and 6.1.

ACKNOWLEDGMENTS

We thank Jonathan Ayres for discovering this wonderful sequence. We also thank a referee for helpful comments.

REFERENCES

[Ayres 01] J. Ayres, personal communication, Sept. 30, 2001.

[Erdős et al. 83] P. Erdős, R. Freud and N. Hegyvari. "Arithmetical properties of partitions of integers." *Acta Math. Acad. Sci. Hungar.* **41** (1983), 169–176.

[Hooley 76] C. Hooley. *Applications of Sieve Methods to the Theory of Numbers*, Cambridge Tracts in Math. No. 70, Cambridge Univ. Press: Cambridge 1976.

[Sloane 01] N. J. A. Sloane. *The On-Line Encyclopedia of Integer Sequences*, published electronically at www.research.att.com/∼njas/sequences/

[Zagier 77] D. Zagier. "The first 50 million prime numbers." *Math. Intelligencer*, **0** (1977), 7–19.

J. C. Lagarias, Information Sciences Research Center, AT & T Shannon Lab, Florham Park, NJ 07932-0971 (jcl@research.att.com)

E. M. Rains, Information Sciences Research Center, AT & T Shannon Lab, Florham Park, NJ 07932-0971 Current address: Institute for Defense Analysis, Princeton, NJ, 08540 (rains@idaccr.org)

N. J. A. Sloane, Information Sciences Research Center, AT & T Shannon Lab, Florham Park, NJ 07932-0971 (njas@research.att.com)

Received December 12, 2001; accepted in revised form March 11, 2002.

Instructions for Authors

Experimental Mathematics is devoted to experimental aspects of mathematical research. It publishes results inspired by experimentation, conjectures suggested by experiments, surveys of certain areas from the experimental point of view, descriptions of algorithms and software for mathematical exploration, and general articles of interest to the community. A more detailed statement of philosophy and of the publishability criteria is available on the Web at http://www.expmath.org, or by request from the publisher (see address below, or send e-mail to editorial@expmath.org).

How to Submit an Article

To submit a contribution, you may either send e-mail to editorial@expmath.org with an attachment (pdf file) or address from which the paper (pdf file) can be downloaded, or send four printed copies of the material to

Experimental Mathematics
A K Peters, Ltd.
63 South Avenue
Natick, MA 01760-4626
phone: 508-655-9933
fax: 508-655-5847

In either case, you must include a note stating that the paper is intended for publication in *Experimental Mathematics* and contact information for each author, consisting of (at least) full name, postal address, electronic address and phone number.

Conditions of Submission

By submiting a paper, authors agree and confirm that: substantially the same work has not been published elsewhere (in a journal or proceedings, though it may have appeared in the form of an abstract or as part of a lecture, review, or thesis); substantially the same work is not under consideration for publication elsewhere; if and when the manuscript is accepted for publication, substantially the same work will not be published elsewhere, except that each author retains the right of republication in any book of which he/she is the author or editor; publication has been approved by all authors and, if required, by the institution where the work was carried out.

Submissions will be acknowledged, but not returned.

Charges

There are no page charges for publications, but authors are expected to contribute toward the cost of color illustrations in their articles. Rates will take into account funding available to authors and editorial necessity.

Offprints

Authors will receive 25 free offprints of their work. At production time authors may order up to 75 additional offprints at cost.

Manuscript Requirements

Manuscripts must be written clearly and concisely. We reserve the right to edit contributions for style and format, with changes subject to the authors' approval.

All submissions must include the following elements:

1. title and (if title exceeds 75 characters) alternative short title for running heads;

2. postal address, affiliation (if appropriate) and electronic address (if available) for each author;

3. an abstract of at most 150 words, in the same language as the article, and an English translation if the article is not in English.

References

References should include full information: author or institution; full title; publisher, city and year (for books, manuals, etc.), or full journal name, volume, year and page range (for papers). References to software should contain complete manufacturer's or distributor's names and addresses. All references in the bibliography should be cited in the text, or accompanied by comments stating their relevance. Reference tags in the text should include author's last name and year of publication, in brackets [Poincaré 1901]. Use a comma to separate a tag from a subsequent page or section number, and semicolons to separate several tags in the same brackets.

1058-6458/2001 $0.50 per page
Experimental Mathematics **11**:3, page 447

Code and Tables

Experimental Mathematics does not publish computer programs in printed form. You can include short illustrative excerpts from your programs, either within the text itself (if at most three lines) or as a separate display. Please supply a caption and a number for each displayed listing. Keep in mind that many readers will not be familiar with the programming language in which your program is written; it is almost always better to explain what a program does in words than to let the program speak for itself.

Similar considerations apply to program output and interactive sessions.

Tables should be kept to a minimum, and generally serve an illustrative rather than archival purpose. Very short tables can be embedded in the text; all others should be able to float and have a caption.

Figures

All figures should be available in electronic form. For electronically-generated figures, you can use photographs or printouts for hard-copy submission, but you must supply the electronic source files if your article is accepted for publication. Under no circumstances will we reproduce printouts, low-resolution scans, or screen photographs.

Figure source files should be in Encapsulated PostScript (EPS). All art files must be supplied in either grayscale, or in CMYK (if color will be included in your article); RGB files should not be used. Halftone images should have a resolution of 300 dpi for best printing quality. Line art (black and white with no grays) should ideally have a resolution of 1200 dpi, but definitely no less than 800 dpi. Please be sure to check and make sure that your line art is truly black and white and not RGB in disguise.

When in doubt whether your figure source is in an acceptable format, check with the editors by sending electronic mail to editorial@expmath.org.

For each figure, please supply a caption and a number by which the figure is referred to in the text. If possible, integrate the figures with the text; otherwise, indicate their optimal placement by means of a comment such as "Place Figure 1 here." In referring to the figure, avoid constructions ("the curve looks like this:") that require the exact placement to be known in advance.

See also the section on Charges on the preceeding page.

Electronic Text

If your article is accepted, we request that you submit the text in electronic form. You can transfer it by e-mail, ftp, or diskette. Send e-mail to editorial@expmath.org for details. We also require that you send us a printed copy of the final version of your article.

Experimental Mathematics is typeset in LaTeX. If you have prepared your manuscript in a form other than LaTeX or TeX, please save your files as text-only or ASCII.

- If possible, please use LaTeX article style.
- Do not redefine existing LaTeX commands.
- Do not embed any new definitions in your text.
- Avoid using explicit vertical spacing commands such as `\vskip`, `\medskip`, `\bigbreak`.
- All user-defined macros must be placed in a separate file which is input at the top of the document.
- Avoid using specialized style files which may work only on your installation—if you use other style files, they must be submitted with your article
- Avoid using explicit horizontal spacing commands. If you must use extra spacing, do it consistently, by means of macros.
- Do not, under any circumstances, insert forced line breaks or page breaks in your document.
- Use double-column format if possible; otherwise single column is acceptable. Since Experimental Mathematics is set in double-column format, your preparation of the electronic files in this way will ensure that your mathematical formulas are broken correctly. It will also help in the sizing of your illustrations.