INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

Approximation of Complex Numbers by Cyclotomic Integers

M.A. Shokrollahi and V. Stemann

TR-96-033

August 1996

Abstract

We present a new method of approximating complex numbers by cyclotomic integers in $\mathbb{Z}[e^{2\pi i/2^n}]$ whose coefficients with respect to the basis given by powers of $e^{2\pi i/2^n}$ are bounded in absolute value by a given integer M. It has been suggested by Cozzens and Finkelstein [5] that such approximations reduce the dynamic range requirements of the discrete Fourier transform. For fixed n our algorithm gives approximations with an error of $O(1/M^{2^{n-2}-1})$. This proves a heuristic formula of Cozzens and Finkelstein. We will also prove a matching lower bound for the worst case error of any approximation algorithm and hence show that our algorithm is essentially optimal. Further, we derive a slightly different and more efficient algorithm for approximation by 16th roots of unity. The basic ingredients of our algorithm are the explicit Galois theory of cyclotomic fields as well as cyclotomic units. We use a deep number theoretic property of these units related to the class number of the field. Various examples and running times for this case and that of approximation by 32nd roots of unity are included. Finally, we derive the algebraic and analytic foundations for the generalization of our results to arbitrary algebraic number fields.

1 Introduction

Numerical computations are vulnerable to basically two different types of errors: *quantization errors* which are caused, e.g., by inaccurate input, and *roundoff errors*, which falsify results of arithmetic operations. While quantization errors are inherent to computations run on inaccurate data, roundoff errors can be avoided if the input and all program constants are approximated by elements of an algebraic structure in which exact arithmetic is possible. If the approximation errors are not too large, this procedure yields outputs of guaranteed good precision.

An important class of computations for which this approach has been developed in detail is the fast Fourier transform [4], FFT for short. The basic idea is as follows (see [5, 6, 7] and the references therein): the input vector as well as the roots of unity involved are approximated by Gaussian integers, i.e., elements of the ring $\mathbb{Z}[i]$. This step is accomplished by scaling the complex numbers by a large number, and then rounding to the nearest Gaussian integer. Since the coefficients arising in the course of the computation may be large, arithmetic in $\mathbb{Z}[i]$ is performed using Chinese Remaindering Techniques modulo appropriate primes. Fixing the primes leads to working with elements of $\mathbb{Z}[i]$ with bounded coefficients. To obtain reasonable approximations of the input data however, one needs to work with large scaling factors. Hence, there is a fundamental tradeoff between the so-called dynamic range requirements, and the precision of the quantization. For example, approximation by Gaussian integers up to an error of $1/n$ would require integer coefficients in the range $[-n/\sqrt{2}, \dots, n/\sqrt{2}]$, see [8].

In a pioneering paper Cozzens and Finkelstein [5] suggested to replace the fourth root of unity i by a 2^nth root $\zeta = e^{2\pi i/2^n}$. While $\mathbb{Z}[i]$ forms a discrete lattice in $\mathbb{C}$, the ring $\mathbb{Z}[\zeta] := \left\{\alpha_0 + \alpha_1\zeta + \ldots + \alpha_{2^{n-1}-1}\zeta^{2^{n-1}-1} \;\middle|\; \alpha_i \in \mathbb{Z}\right\}$ is *dense* in $\mathbb{C}$ for $n \geq 3$. Hence for all $z \in \mathbb{C}$ and all $\varepsilon > 0$ there is an approximation $a \in \mathbb{Z}[\zeta]$ of z, such that $|a - z| \leq \varepsilon$. Since the set

$$\mathbb{Z}[\zeta]_M := \left\{\alpha_0 + \alpha_1\zeta + \ldots + \alpha_{2^{n-1}-1}\zeta^{2^{n-1}-1} \;\middle|\; \alpha_i \in \mathbb{Z}, |\alpha_i| \leq M\right\}$$

is finite, we will still be confronted with the tradeoff problem between the dynamic range requirements and the precision. However, we might hope for better approximations within the same range. Since any complex number can be represented by the sum of a Gaussian integer and a complex number of absolute value ≤ 1, we focus in the following *Approximation Problem* on complex numbers inside the unit circle:

> *Design an algorithm that approximates a given complex number of absolute value ≤ 1 by an element of the set $\mathbb{Z}[\zeta]_M$, where $\zeta = e^{2\pi i/2^n}$.*

The main result of this paper is a general solution to the above problem by exhibiting a meta-algorithm that produces an approximation algorithm on input M and n. Usually, we will consider n as being fixed and study algorithms solving the above Approximation Problem with respect to the following questions: (1) What is a worst case bound on the error of the approximation in terms of M? (2) What is the running time of the approximation algorithm in terms of M?

Only partial answers have been given to these questions in previous work. Cozzens and Finkelstein give in [5] a heuristic argument which suggests obtaining an approximation error of order $O(1/M^{2^{n-2}-1})$ for fixed n might be possible. Furthermore they describe a simple algorithm which, for fixed n, needs an $O(M\log(M))$ bound on the coefficients to achieve a precision of $O(1/M^{2^{n-2}-1})$. As this algorithm has a precomputation phase whose basic ingredient is exhaustive search, it is impractical for larger values of M and n.

Games [6, 7] develops a greedy algorithm for the special case of the 8th roots of unity. A rough sketch of the algorithm is as follows: in a first step a small element ε of $\mathbb{Z}[\zeta]_M$ is found. Now the algorithm starts at the origin and successively improves the approximation by adding an appropriate element of the form $\zeta^i\varepsilon$ without violating the bound on the coefficients. He shows that an extension of this algorithm achieves an approximation error $\leq |\varepsilon|/\sqrt{2-\sqrt{2}}$. Finally he gives an explicit algorithm, based on continued fractions, to find a ε of smallest absolute value. The final result is an algorithm with worst case approximation error of $O(1/M)$, which compares favorably with the heuristic formula in [5]. However, no attempt has been made to justify the algorithm for the case of sixteenth roots. Furthermore, due to the large amount of computations involved, this algorithm is not well suited for real time applications.

In this paper we give a complete answer to the above questions by presenting a general method to solve the Approximation Problem. For fixed n the algorithm approximates a complex number in the unit circle with an approximation error of $O(1/M^{2^{n-2}-1})$. This proves the heuristic formula by Cozzens and Finkelstein. The algorithm runs in time $O(\log(M))$ and uses $O(\log(M))$ additional memory. We prove a matching lower bound on the approximation error thereby showing the optimality of our algorithm. Due to its simple structure, it is also suitable for real time computations.

In a first step we reduce the complex approximation problem to the approximation problem of real numbers in the interval $[0,1]$. This is done by separately approximating the real and imaginary part. This only gives an additional factor of two in the bound on the coefficients, and a factor of $\sqrt{2}$ on the absolute error of the final approximation.

The main part of our meta-algorithm is the construction of a set E of small positive real elements of $\mathbb{Z}[\zeta]_M$. E has the property that for all real $a\in\mathbb{Z}[\zeta]_M$ there exists an $\varepsilon\in E$ such that $a+\varepsilon$ is still in $\mathbb{Z}[\zeta]_M$, i.e., the sum does not violate the bound on the coefficients.

Games' algorithm [7] for finding the elements of smallest absolute value in the set $\mathbb{Z}[e^{2\pi i/8}]_M$ shows that these always have absolute norm ± 1 or ± 2. Recall that the absolute norm of an algebraic number is the product of its Galois-conjugates. Inspired by this result we construct the elements of E as power products of certain units in the cyclotomic field $\mathbb{Q}(\zeta)$, the so-called cyclotomic units. The main property of this set of units is that their so-called regulator is nonzero. This means that the matrix whose (i,j)-entry is the logarithm of the absolute value of the ith conjugate of the jth unit is nonsingular. This is a deep property which follows from the nonvanishing of the Dirichlet L-series at $s=1$. The interested reader is referred to [10, Chapter 8].

To construct the sets E we start in Section 2 with establishing a fundamental relationship between the size of the coefficients of an algebraic integer and the size of its conjugates. After recalling some basic results about cyclotomic fields in Section 3, we proceed in Section 4 with lower bounds for the Approximation Problem. We prove that the approximation error of any algorithm solving this problem is $\Omega(1/M^{2^{n-2}-1})$ by showing that there exists a point z such that for all $a\in\mathbb{Z}[\zeta]$ $|z-a|=\Omega(1/M^{2^{n-2}-1})$. Further, we also show that the smallest element of $\mathbb{Z}[\zeta]_M$ has order $\Omega(1/M^{2^{n-2}-1})$. This result, though interesting in its own, will be used to prove that our approximation algorithm runs in time $O(\log(M))$.

Matching upper bounds are proved in Section 5 by constructing the set E. We derive certain conditions on the conjugates of units in E which guarantee for any $a \in \mathbb{Z}[\zeta]_M$ the existence of $\varepsilon \in E$ such that $a + \varepsilon \in \mathbb{Z}[\zeta]_M$. We then reduce the problem of finding power products of cyclotomic units which satisfy these conditions to an integer linear programming problem, and show how to find (asymptotically optimal) solutions by solving a system of linear equations.

Section 6 deals with a modification of the general algorithm for the case of 16th roots of unity. This modified algorithm uses a signature technique. It is asymptotically comparable to the general method, but is much more efficient in practice. The algorithm is inspired by a simple approximation algorithm in $\mathbb{Z}[e^{2\pi i/8}]$ found by the authors [9].

In Section 7 we report on the implementation of our algorithms, on their running times, and on strategies to enhance their performance. Finally, Section 8 prepares the ground for a far reaching generalization of the results of this paper. We have included a theorem obtained by Clausen and Shokrollahi [3] on the characterization of complex numbers w such that $\mathbb{Z}[w]$ is dense in $\mathbb{C}$. If w is an algebraic integer of degree > 2, we can design a meta-algorithm for approximation in $\mathbb{Z}[w]$ along the same lines as in this paper, provided we know a set of units in the algebraic number field $\mathbb{Q}(w)$ with nonvanishing regulator.

We would like to thank Michael Clausen for communicating the problem to us and for his permission to include Theorem 30.

2 Galois Extensions

Let $L \supseteq K$ be a Galois extension of fields with group $G := \mathrm{Gal}(L/K) = \{\tau_0, \ldots, \tau_{d-1}\}$. Further, let $\theta_0, \ldots, \theta_{d-1}$ be a K-basis of L, and $a := \sum_{i=0}^{d-1} \alpha_i \theta_i \in L$ with $\alpha_i \in K$. We wish to derive a relationship between the sets $\{\alpha_0, \ldots, \alpha_{d-1}\}$, $\{\theta_0, \ldots, \theta_{d-1}\}$, and $\{\tau_0, \ldots, \tau_{d-1}\}$. We start by noting that for any nonzero $a \in L^\times$ there exists an invertible matrix $R_a \in \mathrm{GL}(d, K)$, such that

$$\begin{pmatrix} a \cdot \theta_0 \\ \vdots \\ a \cdot \theta_{d-1} \end{pmatrix} = R_a \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_{d-1} \end{pmatrix}. \tag{1}$$

The map $a \mapsto R_a$ is an injective homomorphism of L into $\mathrm{GL}(d, k)$ (regular representation). Since the entries of R_a belong to K, they are invariant under G. Hence, taking the G-conjugates of (1) yields

$$T \cdot D_a = R_a \cdot T \tag{2}$$

where

$$T := \begin{pmatrix} \tau_0(\theta_0) & \tau_1(\theta_0) & \cdots & \tau_{d-1}(\theta_0) \\ \tau_0(\theta_1) & \tau_1(\theta_1) & \cdots & \tau_{d-1}(\theta_1) \\ \vdots & \vdots & \ddots & \vdots \\ \tau_0(\theta_{d-1}) & \tau_1(\theta_{d-1}) & \cdots & \tau_{d-1}(\theta_{d-1}) \end{pmatrix},$$

and $D_a := \mathrm{diag}(\tau_0(a), \ldots, \tau_{d-1}(a))$ is the diagonal matrix with diagonal entries $\tau_0(a), \ldots, \tau_{d-1}(a)$. The following result is well-known.

Lemma 1. *T is invertible.*

Proof. Since L is a separable extension of K, there exists an irreducible separable polynomial $f \in K[X]$, such that $L = K(a)$ for some root a of f. Let S be the matrix $(\tau_j(a^i))_{0 \leq i,j < d}$ and let C be the transition matrix from the K-basis $(\theta_0, \ldots, \theta_{d-1})$ to the K-basis $(1, a, \ldots, a^{d-1})$ of L. S is van der Monde, hence $\det(S) \neq 0$ as $\{\tau_0(a), \ldots, \tau_{d-1}(a)\}$ are different roots of f. Furthermore, as $S = C \cdot T$, $\det(T) \neq 0$ □

In the following we will assume that $\theta_0 = 1$ and τ_0 is the identity map. The definition of R_a implies that $(1, 0, \ldots, 0) \cdot R_a = (\alpha_0, \ldots, \alpha_{d-1})$ (multiply both sides of (1) with $(1, 0, \ldots, 0)$ from the left). Hence $(1, 0, \ldots, 0) \cdot T D_a T^{-1} = (\alpha_0, \ldots, \alpha_{d-1})$. Noting that the first row of T is the all one vector, we obtain:

$$(\alpha_0, \ldots, \alpha_{d-1}) = (a, \tau_1(a), \ldots, \tau_{d-1}(a)) \cdot T^{-1}. \tag{3}$$

3 Cyclotomic Fields

In this section we review some basic and well-known facts about cyclotomic fields. We will later use the results stated here to derive our approximation algorithms in Sections 5 and 6. Proofs of the classic results not explicitly proved here can be found in, e.g., Washington's book [10].

3.1 Explicit Galois Theory

Let $\zeta_n := e^{\frac{2\pi i}{2^n}}$, and $K_n := \mathbb{Q}(\zeta_n)$ be the cyclotomic field generated over $\mathbb{Q}$ by ζ_n. K_n has a $\mathbb{Q}$-basis $(1, \zeta_n, \ldots, \zeta_n^{2^{n-1}-1})$ and its ring of integers is $\mathbb{Z}[\zeta_n] := \{\sum_{i=0}^{2^{n-1}-1} \alpha_i \zeta_n^i \mid \forall i\colon \quad \alpha_i \in \mathbb{Z}\}$. Recall that for a positive integer M we denote by $\mathbb{Z}[\zeta_n]_M$ the subset $\{\sum_{i=0}^{2^{n-1}-1} \alpha_i \zeta_n^i \mid \forall i\colon \quad \alpha_i \in \mathbb{Z}, |\alpha_i| \leq M\}$ of $\mathbb{Z}[\zeta_n]$.

K_n is a Galois extension of $\mathbb{Q}$ with Galois group isomorphic to $(\mathbb{Z}/2^n\mathbb{Z})^\times$ under the canonical isomorphism:

$$(\mathbb{Z}/2^n\mathbb{Z})^\times \ni c \mapsto (\sigma_c : \zeta_n \mapsto \zeta_n^c) \in \mathrm{Gal}(K_n/\mathbb{Q}).$$

It is well known that $(\mathbb{Z}/2^n\mathbb{Z})^\times = \langle 5 \bmod 2^n \rangle \times \langle -1 \bmod 2^n \rangle$, so $\mathrm{Gal}(K_n/\mathbb{Q}) = \langle \sigma_5 \rangle \times \langle \sigma_{-1} \rangle$. Note that σ_{-1} is the complex conjugation. Hence its fixed field, denoted by K_n^+, is a subfield of $\mathbb{R}$ and has index 2 in K_n.

Let $\theta_0 := 1$ and for $i \geq 1$ let $\theta_i := \zeta_n^i + \zeta_n^{-i}$. The elements $\theta_0, \ldots, \theta_{2^{n-2}-1}$ of K_n^+ form an integral basis of $K_n^+/\mathbb{Q}$, i.e., they are a $\mathbb{Q}$-basis of K_n^+ and $\mathbb{Z}\theta_0 + \ldots + \mathbb{Z}\theta_{2^{n-2}-1}$ is the ring of integers $\mathbb{Z}[\zeta_n]^+$ of K_n^+. For a positive integer M we define

$$\mathbb{Z}[\zeta_n]_M^+ := \left\{ \sum_{i=0}^{2^{n-2}-1} \alpha_i \theta_i \;\middle|\; \forall i\colon \quad \alpha_i \in \mathbb{Z}, |\alpha_i| \leq M \right\} \subset \mathbb{Z}[\zeta_n]_M.$$

By Galois theory we have $\mathrm{Gal}(K_n^+/\mathbb{Q}) \simeq (\mathbb{Z}/2^n\mathbb{Z})^\times / \langle -1 \bmod 2^n \rangle \simeq \langle 5 \bmod 2^n \rangle$. An explicit isomorphism is as follows: Let $\tau \in \mathrm{Gal}(K_n^+/\mathbb{Q})$ be defined by $\tau(\zeta_n^i + \zeta_n^{-i}) := \zeta_n^{5i} + \zeta_n^{-5i}$. Then $\mathrm{Gal}(K_n^+/\mathbb{Q}) = \langle \tau \rangle \simeq \langle 5 \bmod 2^n \rangle$.

Example 2. *Let $n = 4$ and $\zeta := \zeta_n$. The images of the θ_i under τ are given by:*

$$\begin{array}{lcl}
\theta_0 & \overset{\tau}{\longmapsto} & \theta_0 \\
\theta_1 & \overset{\tau}{\longmapsto} & \zeta^5 + \zeta^{-5} = \zeta^{8-3} + \zeta^{-(8-3)} = -\theta_3 \\
\theta_2 & \overset{\tau}{\longmapsto} & \zeta^{10} + \zeta^{-10} = \zeta^{8+2} + \zeta^{-(8+2)} = -\theta_2 \\
\theta_3 & \overset{\tau}{\longmapsto} & \zeta^{15} + \zeta^{-15} = \zeta^{-1} + \zeta^{1} = \theta_1
\end{array}$$

For $a \in K_n^+$ we set $a^{(i)} := \tau^i(a)$. In accordance with Section 2 we define $T = (\theta_i^{(j)})_{0\leq i,j<2^{n-2}}$. T is invertible by Lemma 1 and we can explicitly determine T^{-1} in this case.

Lemma 3. *With T as above we have $T^{-1} = \frac{1}{2^{n-1}}(\tilde{\theta}_j^{(i)})_{0\leq i,j<2^{n-2}}$ where $\tilde{\theta}_j := \zeta_n^j + \zeta_n^{-j}$.*

Proof. For $i \neq 0$ the (i,j)-entry of TT^{-1} equals:

$$\begin{aligned}
\frac{1}{2^{n-1}}\sum_{l=0}^{2^{n-2}-1}\theta_i^{(l)}\tilde{\theta}_j^{(l)} &= \frac{1}{2^{n-1}}\sum_{l=0}^{2^{n-2}-1}(\zeta_n^{5^l i} + \zeta_n^{-5^l i})(\zeta_n^{5^l j} + \zeta_n^{-5^l j}) \\
&= \frac{1}{2^{n-1}}\sum_{l=0}^{2^{n-2}-1}(\zeta_n^{5^l(i+j)} + \zeta_n^{5^l(i-j)} + \zeta_n^{5^l(j-i)} + \zeta_n^{-5^l(i+j)}) \\
&= \frac{1}{2^{n-1}}\left(\sum_{\sigma\in\mathrm{Gal}(K_n/\mathbb{Q})}\sigma(\zeta_n^{i-j}) + \sigma(\zeta_n^{i+j})\right) \\
&= \delta_{ij},
\end{aligned}$$

where δ is the Kronecker-function. (Note that if $\zeta_n^l \neq 1$, then the sum of the conjugates of ζ_n^l under $\mathrm{Gal}(K_n/\mathbb{Q})$ is zero.) Analogously one proves that the $(0,j)$-entry of TT^{-1} equals $\delta_{0,j}$. □

Example 4. *For $n = 4$ we have*

$$T = \begin{pmatrix} 1 & 1 & 1 & 1 \\ \theta_1 & -\theta_3 & -\theta_1 & \theta_3 \\ \theta_2 & -\theta_2 & \theta_2 & -\theta_2 \\ \theta_3 & \theta_1 & -\theta_3 & -\theta_1 \end{pmatrix}, \qquad T^{-1} = \frac{1}{8}\begin{pmatrix} 2 & \theta_1 & \theta_2 & \theta_3 \\ 2 & -\theta_3 & -\theta_2 & \theta_1 \\ 2 & -\theta_1 & \theta_2 & -\theta_3 \\ 2 & \theta_3 & -\theta_2 & -\theta_1 \end{pmatrix}.$$

For $a := \sum_{i=0}^{2^{n-2}-1}\alpha_i\theta_i \in K_n^+$ let $L_\infty(a) := \max\left\{|\alpha_i| \;\middle|\; i = 0,\ldots,2^{n-2}-1\right\}$. Stated in terms of this function, our problem is to approximate a given real number in $[0,1]$ by an element $a \in \mathbb{Z}[\zeta_n]^+$ with $L_\infty(a) \leq M$. $L_\infty(a)$ is related to the different conjugates of a in the following way.

Lemma 5. *For $a \in K_n^+$ we have*

$$L_\infty(a) \leq \frac{1}{2^{n-2}}\sum_{i=0}^{2^{n-2}-1}|a^{(i)}|.$$

Proof. By (3) we know that

$$(a^{(0)}, \ldots, a^{(2^{n-2}-1)}) \cdot T^{-1} = (\alpha_0, \ldots, \alpha_{2^{n-2}-1}).$$

Since $|\tilde{\theta}_j^{(i)}| \leq 2$ for all i, j, the absolute values of the entries of T^{-1} are $\leq 1/2^{n-2}$ by Lemma 3. □

We will use this lemma in the following form.

Corollary 6. *If $a \in \mathbb{Z}[\zeta_n]^+$ is such that $\sum_{i=0}^{2^{n-2}-1} |a^{(i)}| \leq 2^{n-2}M$ for some positive integer M, then $a \in \mathbb{Z}[\zeta_n]^+_M$.*

For $a \in K_n^+$ we define

$$\operatorname{maxconj}(a) := \max\Big\{|a^{(i)}| \;\Big|\; i = 0, \ldots, 2^{n-2} - 1\Big\}.$$

Lemma 5 explains the significance of this quantity: we always have $L_\infty(a) \leq \operatorname{maxconj}(a)$.

Proposition 7. *Let $0 \leq i < 2^{n-2}$ and λ, ν be such that $i = 2^\nu \lambda$ and λ odd. Then $\theta_{2^{n-\nu-3}}^{(i)} = -\operatorname{maxconj}(\theta_{2^{n-\nu-3}}) = -\theta_{2^{n-\nu-3}} < -1$.*

Proof. Let us first prove that $\theta_{2^{n-\nu-3}} > 1$: note that $\theta_i < \theta_j$ for $i > j > 0$. Hence, $\theta_{2^{n-\nu-3}} > \theta_{2^{n-3}} = \zeta_3 + \zeta_3^{-1} = \sqrt{2} > 1$.

Next we show for all $n \geq 3$ and all odd λ

$$\theta_1^{(2^{n-3}\lambda)} = -\theta_1 = -\operatorname{maxconj}(\theta_1).$$

Notice that $|\theta_1^{(l)}| \leq |\theta_1|$ for all l. Hence, we need to prove the left equality. Observe that $5^{2^{n-3}} \equiv 1 \bmod 2^{n-1}$ since $(\mathbb{Z}/2^{n-1}\mathbb{Z})^\times = \langle 5 \rangle \times \langle -1 \rangle$. Thus, $5^{2^{n-3}} \equiv 2^{n-1} + 1 \bmod 2^n$, and for odd λ we have $5^{\lambda 2^{n-3}} \equiv (2^{n-1}+1)^\lambda \equiv 1 + 2^{n-1} \bmod 2^n$. Hence,

$$\theta_1^{(2^{n-3}\lambda)} = \zeta_n^{5^{2^{n-3}}\lambda} + \zeta_n^{-5^{2^{n-3}}\lambda} = -(\zeta_n + \zeta_n^{-1}) = -\theta_1$$

Now note that $\theta_{2^{n-\nu-3}} = \zeta_n^{2^{n-\nu-3}} + \zeta_n^{-2^{n-\nu-3}} = \zeta_{\nu+3} + \zeta_{\nu+3}^{-1}$. By the above we have $\theta_{2^{n-\nu-3}}^{(2^\nu\lambda)} = -\theta_{2^{n-\nu-3}} = \operatorname{maxconj}(\theta_{2^{n-\nu-3}})$ (replace n by $\nu + 3$). □

3.2 Cyclotomic Units

The norm of an element $u \in \mathbb{Z}[\zeta_n]^+$ is defined as

$$\mathrm{N}(u) := \prod_{i=0}^{2^{n-2}-1} u^{(i)}.$$

As u is an algebraic integer, $\mathrm{N}(u) \in \mathbb{Z}$. u is called a unit if $\mathrm{N}(u) \in \{\pm 1\}$. A set $\{u_1, \ldots, u_l\}$ of units is called *independent* if $|\prod_{i=1}^l u_i^{x_i}| = 1$, $x_i \in \mathbb{Z}$, implies $x_1 = \cdots = x_l = 0$. For general number fields, it can be quite hard to find a maximal set of independent units. For cyclotomic fields and their maximal real subfields, however, the situation is quite different: the *cyclotomic* units form a maximal set of independent units. They are defined as follows: for $j = 1, \ldots, 2^{n-2} - 1$ we set $\eta_j := \zeta_n^j (1 - \zeta_n)/(1 - \zeta_n^{2j+1})$ ([10, Chapter 8]). The following important fact holds for the matrix $\Xi := (\log |\eta_j^{(i)}|)_{1 \leq i,j < 2^{n-1}}$.

Theorem 8. *The matrix Ξ is invertible.*

A proof of this classic result can be found in [10, Chap. 8.1]. The nonvanishing of the determinant of this matrix comes from the nonvanishing of the Dirichlet L-series $L(s,\chi)$ at $s=1$.

4 Lower Bounds

In this section we will prove by using a volume argument that any algorithm solving the Approximation Problem will have a worst case approximation error of $\Omega(1/M^{2^{n-2}-1})$ for fixed n, see Theorem 11. A matching upper bound will be derived in the next section. Furthermore, we will show that the order of the smallest absolute value of an element in $\mathbb{Z}[\zeta_n]_M$ is also $\Omega(1/M^{2^{n-2}-1})$ for fixed n. This result which is of independent interest will be applied in the next section to analyze the running time of our approximation algorithm.

Proposition 9. *Let $n \geq 3$, $M \in \mathbb{N}$, and $r := (2M+1)^{-(2^{n-2}-1)}$. Then there exists $z \in \mathbb{Z}[\zeta_n]_M$, $|z| < 1$, such that for all $w \in \mathbb{Z}[\zeta_n]_M \setminus \{z\}$ we have $|w-z| > r$.*

For the proof of this proposition we need the following result.

Lemma 10. *There are at most $(2M+1)^{2^{n-1}-2}$ elements of $\mathbb{Z}[\zeta_n]_M$ inside the unit circle.*

Proof. Let $\alpha_j \in \{-M,\dots,M\}$, $j \neq 0$, $j \neq 2^{n-2}$, be given. Then there exists exactly one pair $(\alpha_0, \alpha_{2^{n-2}}) \in \mathbb{Z}^2$ such that

$$\sum_{j=0}^{2^{n-1}-1} \alpha_j \zeta_n^j \in \Big\{ z \ \Big|\ \max(|\mathrm{Re}(z)|, |\mathrm{Im}(z)|) \leq 1 \Big\}.$$

Hence, there is at most one pair $(\alpha_0, \alpha_{2^{n-2}})$ with $|\alpha_0|, |\alpha_{2^{n-2}}| \leq M$, such that $|\sum_{j=0}^{2^{n-1}-1} \alpha_j \zeta_n^j| \leq 1$. Thus the assertion follows.

Proof of Proposition 9. Let

$$\delta_M := \max_{\{z \in \mathbb{Z}[\zeta_n]_M \mid |z|<1\}} \ \min_{\{z' \in \mathbb{Z}[\zeta_n]_M \mid |z'| \leq 1, z' \neq z\}} |z - z'|.$$

It suffices to prove that $r < \delta_M$. The circles of radius δ_M around the elements of $\Big\{ z \in \mathbb{Z}[\zeta_n]_M \ \Big|\ |z| < 1 \Big\}$ cover the unit circle. Hence by Lemma 10:

$$(2M+1)^{2^{n-1}-2} \cdot \delta_M^2 \pi > \pi \Rightarrow \delta_M > (2M+1)^{-(2^{n-2}-1)} = r. \ \square$$

Theorem 11. *Let $n \geq 3$. There exists a constant c_n depending on n and for any $M > 0$ there exists $z_M \in \mathbb{C}$, $|z_M| \leq 1$, such that*

$$\min_{z \in \mathbb{Z}[\zeta_n]_M} |z - z_M| \geq \frac{c_n}{M^{2^{n-2}-1}}.$$

PROOF. Let r and z be as in Lemma 10 and let z_M be such that $|z_M - z| = r$ and $|z_M| < 1$. □

The next theorem gives a lower bound on the size of the elements of $\mathbb{Z}[\zeta_n]_M$.

Theorem 12. *For fixed $n \geq 3$ there exists a constant γ_n depending on n, such that*

$$\min_{z \in \mathbb{Z}[\zeta_n]_M \setminus \{0\}} |z| \geq \frac{\gamma_n}{M^{2^{n-2}-1}}.$$

PROOF. We will only prove this statement for the elements of $\mathbb{Z}[\zeta_n]_M^+$. The corresponding assertion for $\mathbb{Z}[\zeta_n]_M$ can be proved exactly in the same way. During this proof we set $m := 2^{n-2}$. Let $\varepsilon \in \mathbb{Z}[\zeta_n]_M^+$ be the smallest element of this set. Multiplying both sides of equation (3) with T we obtain

$$(\varepsilon^{(0)}, \ldots, \varepsilon^{(m-1)}) = (e_0, \ldots, e_{m-1})T,$$

where $\varepsilon = \sum_i e_i \theta_i$. Let the L_1*-norm* $L_1(v)$ of a real valued vector v be the sum of the absolute values of its entries. Note that if A is a square matrix, then $L_1(vA) \leq L_1(A)L_1(v)$, where $L_1(A)$ is the maximum of the L_1-norms of the columns of A. Hence, taking L_1-norms of the above equation we obtain

$$\sum_{i=0}^{m-1} |\varepsilon^{(i)}| \leq m L_\infty(\varepsilon) L_1(T).$$

Let $\ell := \prod_{i=0}^{m-1} \varepsilon^{(i)}$ be the the norm of ε. We get

$$\frac{|\ell|}{|\varepsilon|} = \prod_{i=1}^{m-1} |\varepsilon^{(i)}| \leq L_\infty(\varepsilon)^{m-1} L_1(T)^{m-1} \left(\frac{m}{m-1}\right)^{m-1} \leq \frac{M^{m-1}}{\gamma_n},$$

where $\gamma_n := \Big(mL_1(T)/(m-1)\Big)^{-m+1}$ is a constant depending on n. Since ℓ is a nonzero integer, we have $|\ell| \geq 1$, which implies the assertion. □

5 The General Approximation Algorithm

In this section we describe a general procedure to approximate complex numbers by elements of $\mathbb{Z}[\zeta_n]$. We adopt the notation of the previous sections. In addition, we set $m := 2^{n-2}$.

5.1 Reduction to the Real Case

We start by reducing the problem to that of the approximation of real numbers.

Proposition 13. *Let $z \in \mathbb{C}$ have real part z_0 and imaginary part z_1. Suppose that $a_0, a_1 \in \mathbb{Z}[\zeta_n]_M^+$ are approximations of z_0 and z_1, respectively, and that $|z_j - a_j| \leq \delta$ for some $\delta \geq 0$. Then $a_0 + ia_1 \in \mathbb{Z}[\zeta_n]_{2M}$ is an approximation of z with error at most $\delta\sqrt{2}$.*

Proof. The bound on the error being obvious we focus on proving that $a_0 + ia_1$ has coefficients bounded by $2M$. But this is clear once noting that if $a_j = \sum_\ell \alpha_{j,\ell}\theta_\ell$, then $a_0+ia_1 = \alpha_{0,0}+\sum_{j=1}^{2^{n-2}-1}(\alpha_{0,j}+\alpha_{1,2^{n-2}-j})\zeta_n^j+\alpha_{1,0}\zeta_n^{2^{n-2}} +\sum_{j=1}^{2^{n-2}-1}(\alpha_{1,j}-\alpha_{0,2^{n-2}-j})\zeta_n^{2^{n-2}+j}$. □

Given $\omega \in [0,1]$ and some $M \in \mathbb{N}$, our aim is to find $a = \sum_i \alpha_i\theta_i \in \mathbb{Z}[\zeta_n]^+$ such that $L_\infty(a) \leq M$ and $|\omega - a| \leq \frac{c}{M^{m-1}}$ for some constant c (possibly depending on n). In view of Corollary 6 it suffices to find a such that $\sum_{i=0}^{m-1} |a^{(i)}| \leq mM$.

5.2 The Algorithm

The basic precomputation step is to find a subset $E = \{\varepsilon_1, \ldots, \varepsilon_{2m-2}\}$ of $\mathbb{Z}[\zeta_n]^+$ consisting of positive elements with the following properties:

(a) $\forall 1 \leq k < m$: $\mathrm{maxconj}(\varepsilon_k) = \varepsilon_k^{(k)}$ and $\mathrm{maxconj}(\varepsilon_{k+m-1}) = -\varepsilon_{k+m-1}^{(k)}$,

(b) $\forall \varepsilon \in E$: $\sum_{i=0}^{m-1} |\varepsilon^{(i)}| \leq 2\,\mathrm{maxconj}(\varepsilon)$,

(c) $\max_{l=1}^{2m-2} \mathrm{maxconj}(\varepsilon_l) \leq M$,

We first show that if E has the above properties, then for an arbitrary element a in $\mathbb{Z}[\zeta_n]_M^+ \cap [0,1]$ there always exists $\varepsilon \in E$ such that $a + \varepsilon \in \mathbb{Z}[\zeta_n]_M^+$. Roughly speaking, one has to choose from E an element ε that has its maximum conjugate at the same position as a, but with a different sign. Recall that the sign of a nonzero real number α, denoted by $\mathrm{sign}(\alpha)$ is $+1$ if α is positive, -1 if it is negative.

Lemma 14. *Let $a \in \mathbb{Z}[\zeta_n]^+ \cap (0,1)$, $\sum_{i=0}^{m-1} |a^{(i)}| \leq mM$, $\mathrm{maxconj}(a) =: |a^{(l)}|$, and $e := \mathrm{sign}(a^{(l)})$. Further let $b := a + \varepsilon_l$ if $e = -1$, and $b := a + \varepsilon_{l+m-1}$ if $e = 1$, where the ε_ℓ are defined as above. Then $\sum_{i=0}^{m-1} |b^{(i)}| \leq mM$.*

Proof. Let $\varepsilon := b - a$. Suppose first that $|a^{(l)}| \geq |\varepsilon^{(l)}|$. Then $\sum_{i=0}^{m-1} |b^{(i)}| \leq \sum_{i \neq l}(|a^{(i)}| + |\varepsilon^{(i)}|) + |a^{(l)}| - |\varepsilon^{(l)}| \leq \sum_{i\neq l} |a^{(i)}| + |a^{(l)}| \leq mM$, since by Condition (b) on E we have $|\varepsilon^{(l)}| \geq \sum_{i \neq l} |\varepsilon^{(i)}|$.

Suppose now that $|a^{(l)}| < |\varepsilon^{(l)}|$. Then $\sum_{i=0}^{m-1} |b^{(i)}| = \sum_{i=0}^{m-1} |a^{(i)} + \varepsilon^{(i)}| \leq \sum_{i\neq l} |a^{(i)} + \varepsilon^{(i)}| + |\varepsilon^{(l)}| - |a^{(l)}| \leq \sum_{i \neq l}(|a^{(l)}| + |\varepsilon^{(i)}|) + |\varepsilon^{(l)}| - |a^{(l)}| \leq (m-2)|a^{(l)}| + 2|\varepsilon^{(l)}| < m|\varepsilon^{(l)}| \leq mM$, where the last inequality follows from Condition (c) on E. □

The idea of the approximation algorithm is now fairly simple. We compute sets E_ℓ for different $\ell \leq M$ and starting from 0 we improve our approximation by adding elements from the sets E_ℓ until no further improvement is possible. Using Lemma 14 we know that we can always find such elements without violating the bound on the size of the coefficients. If $\ell < M$, we increase the value of k and start all over again.

Suppose that we have found sets $E_{2^k} := \{\varepsilon_{1,k}, \ldots, \varepsilon_{2m-2,k}\}$ for $0 \leq k \leq \lfloor \log M \rfloor$, such that Conditions (a)-(c) are satisfied. Let $\max(E_{2^k}) := \max_{i=1}^{2m-2} \varepsilon_{i,k}$ and $\min(E_{2^k}) := \min_{i=1}^{2m-2} \varepsilon_{i,k}$. The General Approximation Algorithm (**GAA**) is given in pseudocode in Figure 1.

```
Precomputation: Sets $E_{2^k}$ for $0 \le k \le \lfloor \log(L) \rfloor$ as described above.
Input: $\omega \in [0,1]$, $M \le L$, $M \in \mathbb{N}$.
Output: $a \in \mathbb{Z}[\zeta_n]_M^+$ such that $0 \le \omega - a \le \max(E_{2^{\lfloor \log(M) \rfloor}})$.
$a := 0$
for $k = 0$ to $\lfloor \log M \rfloor$ do
    while $a - \omega < 0$ do
        Compute maxconj$(a) =: |a^{(l)}|$
        $e := \mathrm{sign}(a^{(l)})$
        $\varepsilon_k := \begin{cases} \varepsilon_{l,k} & \text{if } e = -1 \\ \varepsilon_{l+m-1,k} & \text{if } e = 1 \end{cases}$
        $a := a + \varepsilon_k$
```

Figure 1: General Approximation Algorithm (**GAA**)

Theorem 15. *Algorithm* **GAA** *computes an element* $a \in \mathbb{Z}[\zeta_n]^+$ *such that* $|a - \omega| \le \max(E_{2^{\lfloor \log M \rfloor}})$ *and uses no more than*

$$\frac{1}{\min(E_1)} + \sum_{k=1}^{\lfloor \log M \rfloor} \frac{\max(E_{2^{k-1}})}{\min(E_{2^k})}$$

iterations.

Proof. At step $k \ge 1$ the difference between the a's obtained from two successive runs of the inner loop is at least $\min(E_{2^k})$ and the inner loop terminates iff $\omega - a < \varepsilon_k \le \max(E_{2^k})$. Hence, the inner loop is performed at most $\max(E_{2^{k-1}})/\min(E_{2^k})$ times. For $k = 0$ the inner loop is performed at most $\omega/\min(E_1) \le 1/\min(E_1)$ times. □

The remaining (and more difficult) problem is the design of the sets E_{2^k}. We will study the more general problem of constructing E_M for arbitrary M. In a first step we show how to construct E_M from a set E'_M with only half as many elements. The elements of E'_M have their conjugates at different positions. The idea is to construct from an element in E'_M another element having its maximal conjugate at the same position but with a different sign, by multiplying this element with an appropriate θ_i.

Lemma 16. *Let* $\varepsilon \in \mathbb{Z}[\zeta_n]^+$, $\varepsilon^{(l)} := \max|\varepsilon^{(i)}|$, *and suppose that* $\sum_i |\varepsilon^{(i)}| \le 2|\varepsilon^{(l)}|$. *Let* $\lambda, \nu \in \mathbb{N}$, λ *odd, be such that* $l = 2^\nu \lambda$,*and let let* $\chi := \varepsilon \theta_{2^{n-\nu-3}}$. *Then*

(a) $\max|\chi^{(i)}| = |\chi^{(l)}|$, $\chi^{(l)} \cdot \varepsilon^{(l)} < 0$

(b) $\sum_{i=0}^{m-1} |\chi^{(i)}| \le 2|\chi^{(l)}|$

(c) $|\chi^{(l)}| < 2|\varepsilon^{(l)}|$

Proof. Let $\theta := \theta_{2^{n-\nu-3}}$.
(a) By Proposition 7 we have $\theta^{(l)} = -\mathrm{maxconj}(\theta)$, hence (a) follows.
(b) We have

$$\sum_{i=0}^{m-1} |\chi^{(i)}| = \sum_{i=0}^{m-1} |\theta^{(i)} \varepsilon^{(i)}| \le \theta \sum_{i=0}^{m-1} |\varepsilon^{(i)}| \le 2|\varepsilon^{(l)}|\theta = 2|\chi^{(l)}|.$$

(c) By Proposition 7 we know that $|\theta| \geq |\theta^{(i)}|$ for all i. Hence we get $|\chi^{(l)}| = \theta|\varepsilon^{(l)}| < 2|\varepsilon^{(l)}|$. □

Using the above lemma we only need to construct a set E'_M described below, which only has half as many elements as the set E_M. The construction is summarized as follows.

Remark 17. *Let* $E'_M := \{\varepsilon_1, \dots, \varepsilon_{m-1}\} \subset \mathbb{Z}[\zeta_n]^+ \cap \mathbb{R}_+$ *be a subset such that*

$$\sum_{j=0}^{m-1} |\varepsilon_i^{(j)}| \leq 2|\varepsilon_i^{(i)}| \leq M$$

for all i. *For each* k *let* $t_k := \theta_{2^{n-\operatorname{ord}_2(k)-3}}$, *where* $\operatorname{ord}_2(k) := \max\{\ell \mid 2^\ell$ *divides* $k\}$. *Then by Lemma 16 the set* $E_M := \{\varepsilon_1, \dots, \varepsilon_{m-1}, t_1\varepsilon_1, \dots, t_{m-1}\varepsilon_{m-1}\}$ *satisfies Conditions (a)–(c) stated at the beginning of this section.*

Note that $\max(E_M) \leq 2\max(E'_M)$ and $\min(E_M) \geq \min(E'_M)$ since $1 \leq t_k \leq 2$ by Proposition 7.

5.3 Design of E'_M

To construct the sets E'_M we will use power products of the cyclotomic units introduced in Section 3.2. Recall that for $\ell = 1, \dots, m-1$ we defined $\eta_\ell := \zeta^\ell(1-\zeta)/(1-\zeta^{2\ell+1})$.

For given $i \in \{1, \dots, m-1\}$ we want to find $k_1, \dots, k_{m-1} \in \mathbb{Z}$ such that $\varepsilon_i := \prod_\ell \eta_\ell^{k_\ell}$ satisfies the inequalities given in Remark 17. (To keep the notation simple, we suppress the dependency of the k_ℓ on i.) In fact, we will find $k_1, \dots, k_{m-1}$ satisfying the stronger conditions

$$\forall j \neq i: \quad |\varepsilon_i^{(j)}| \leq \frac{1}{m-2}|\varepsilon_i^{(i)}|, \text{ and } |\varepsilon_i^{(i)}| \leq \frac{M}{2}.$$

Taking logarithms this gives

$$\begin{aligned} \forall j \neq i: \quad \sum_{\ell=1}^{m-1} k_\ell\Big(\log|\eta_\ell^{(j)}| - \log|\eta_\ell^{(i)}|\Big) &\leq -\log(m-2), \\ \sum_{\ell=1}^{m-1} k_\ell \log|\eta_\ell^{(i)}| &\leq \log(M/2). \end{aligned}$$

Recall the matrix $\Xi = (\log|\eta_j^{(i)}|)_{1 \leq i,j < 2^{n-1}}$ defined in Section 3.2. Let Ξ_i be the matrix obtained from Ξ by subtracting the ith row from the jth for $j \neq i$, and leaving the ith row unchanged. Obviously, Ξ_i is invertible by Theorem 8. In terms of this matrix the above inequalities can be summarized as

$$\Xi_i \begin{pmatrix} k_1 \\ \vdots \\ k_{m-1} \end{pmatrix} \leq L_i, \tag{4}$$

where L_i is the vector having entry $\log(M/2)$ at position i and entry $-\log(m-2)$ at positions $j \neq i$, and the inequalites are to hold component-wise. Our aim is to find a small

ε_i satisfying these inequalities. This gives us the following Integer Linear Programming problem:

ILP1: *Minimize* $\sum_{\ell=1}^{m-1} k_\ell \log|\eta_\ell|$ *subject to* (4).

While Branch and Bound methods easily yield optimal solutions of the above problem for small values of m (e.g., $m = 3$ or $m = 7$), the Integer Linear Programming approach is not feasible for larger m. Below we will give another approach to find (asymptotically optimal) solutions of **ILP1**, which will also prove the following.

Theorem 18. *The Integer Linear Programming problem* **ILP1** *is solvable and for fixed* m*, the optimal value of its objective function is* $O(M^{-(m-1)})$*.*

Let $\Xi_i = (x_{\ell,j})$, and for each $\ell = 1,\ldots,m-1$ let $y_\ell := \sum_{j,x_{\ell,j}<0} x_{\ell,j}$. Let Y be the vector having entry $-\log(m-2)+y_\ell$ at position $\ell \neq i$, and entry $\log(M/2)+y_i$ at position i. The following proposition shows that **ILP1** is always solvable.

Proposition 19. *Let* $\kappa := (\kappa_1,\ldots,\kappa_{m-1})^\top := \Xi_i^{-1}Y$*, and* $k_\ell := \lfloor \kappa_\ell \rfloor$*. Then* $k_1,\ldots,k_{m-1}$ *satisfy the inequalities* (4)*.*

Proof. Let $\delta_\ell := \kappa_\ell - k_\ell \in [0,1)$. The jth entry of $\Xi_i\kappa$ gives the equation $\sum_j x_{j,\ell}(k_\ell+\delta_\ell) = \lambda_j + y_j$, where λ_j is the jth entry of L_i. Hence,

$$\sum_\ell x_{j,\ell}k_\ell = \lambda_j + y_j - \sum_\ell x_{j,\ell}\delta_\ell \le \lambda_j$$

by the definition of y_j. □

The next proposition shows that the optimal value of the objective function of **ILP1** is $O(M^{-(m-1)})$ for fixed m, and hence completes the proof of Theorem 18.

Proposition 20. *With the notation of the previous lemma let* $\varepsilon_i := \prod_{\ell=1}^{m-1}\eta_\ell^{k_\ell}$*. Then for fixed* m *we have* $\varepsilon_i = O(M^{-(m-1)})$*.*

Proof. With the notation of the proof of the previous proposition we have for $j \neq i$

$$\log|\varepsilon_i^{(j)}| - \log|\varepsilon_i^{(i)}| = \sum_\ell x_{j,\ell}k_\ell = -\log(m-2) + y_j - \sum_\ell x_{j,\ell}\delta_\ell.$$

Hence, there exists a nonnegative d_j depending on m but not on M such that $\log|\varepsilon_i^{(j)}| - \log|\varepsilon_i^{(i)}| \le -\log(m-2) - d_j$. Analogously, there is a nonnegative d_i depending on m and not on M such that $\log|\varepsilon_i^{(i)}| \le \log(M/2) - d_i$. Since ε_i is a unit, we have $\log|\varepsilon_i| = -\sum_{\ell=1}^{m-1}\log|\varepsilon_i^{(\ell)}|$. Summing up, we obtain

$$\log|\varepsilon_i| = -(m-1)\log(M/2) + (m-1)\log(m-2) + (m-2)d_i + \sum_{\ell=1}^{m-1} d_\ell.$$

Hence, for fixed m we get the assertion. □

The last two propositions form the proof of one of the main theorems of this paper.

Theorem 21. *Let n be an integer, $m = 2^{n-2}$. For any $M \in \mathbb{N}$ one can construct in time $O(m^3 \log(M))$ an approximation algorithm with the following properties: for fixed m it computes in time $O(\log(M))$ on input $\omega \in [0,1]$ an element $a \in \mathbb{Z}[\zeta_n]_M^+$ such that $L_\infty(a) \leq M$ and $|\omega - a| = O(M^{-(m-1)})$.*

Proof. We first compute the sets E'_{2^k} for $k = 0, \ldots, \lfloor \log(M) \rfloor$. Using the previous propositions we see that each of these sets can be constructed in time $O(m^3)$ (solving linear equations). The computation of all the E_{2^k} takes $O(m^3 \log(M))$ time. From E'_{2^k} we construct E_{2^k} using Remark 17. This takes $O(m \log(M))$ time. We then incorporate these sets into Algorithm **GAA** given in Figure 1. Since the maximal and the minimal elements of the sets E_{2^k} are of order $\Theta(2^{-(m-1)k})$ for fixed m by the previous proposition and Theorem 12, Theorem 15 implies that Algorithm **GAA** computes a with $L_\infty(a) \leq M$ and $|\omega - a| = O(M^{-(m-1)})$ in time $O(\log(M))$. □

Theorem 11 in Section 4 shows that the approximation error given in Theorem 21 is essentially optimal.

It should be noted that restriction to $\omega \in [0,1]$ is not essential. Actually, for any $\omega \in [0, M/2]$ we can compute an approximation $a \in \mathbb{Z}[\zeta_n]_M^+$ in the following way. We first approximate $a - \lfloor a \rfloor$ with an element of $\mathbb{Z}[\zeta_n]_{M/2}^+$, and then add $\lfloor a \rfloor$ to this approximation to obtain an element in $\mathbb{Z}[\zeta_n]_M^+$. Furthermore, combining the approximation algorithm with scaling techniques, we can obtain approximations of order $O(1/M^{2^{n-2}})$ by elements of the set $\mathbb{Z}[\zeta_n]_M^+/(M/2)$, see Section 7.

6 Sixteenth Roots of Unity

In this section we describe in detail the approximation of real numbers in the interval $(0,1)$ by elements from $\mathbb{Z}[e^{2\pi i/16}]_M^+$. In this special case we use a slightly different algorithm based on the signature of an element. It is inspired by the approximation algorithm in $\mathbb{Z}[e^{2\pi i/8}]$ found by the authors in [9]. Since we will not deal with the conjugates of the elements explicitly, this algorithm runs faster in practice than the general one. We adopt the notation of the previous sections. In addition, throughout this section we set $\zeta := \zeta_4 = e^{2\pi i/16}$.

6.1 Signature

Roughly speaking the signature of an element of $\mathbb{Z}[\zeta]^+$ is the vector of signs of its coefficients. Unfortunately the sign of zero is not uniquely determined, so we have to use a more technical definition. For $\alpha \in \mathbb{Z}$ let

$$\mathrm{sgn}(\alpha) := \begin{cases} \{+\}, & \text{if} \quad \alpha > 0 \\ \{-\}, & \text{if} \quad \alpha < 0 \\ \{+,-\}, & \text{if} \quad \alpha = 0 \end{cases},$$

For $a = \sum_{i=0}^{3} \alpha_i \theta_i \in \mathbb{Z}[\zeta]^+$ let $\mathrm{sgn}(a) := \mathrm{sgn}(\alpha_0) \times \cdots \times \mathrm{sgn}(\alpha_3)$, and, $-\mathrm{sgn}(a) := \mathrm{sgn}(-a)$.

Examples are $\mathrm{sgn}(1 - \theta_3) = \{(+,+,+,-),(+,+,-,-),(+,-,+,-),(+,-,-,-)\}$ and $\mathrm{sgn}(1 + \theta_1 - \theta_2 + \theta_3) = \{(+,+,-,+)\}$.

For $\sigma = (\sigma_0, \ldots, \sigma_3) \in \{+,-\}^4$ let $\mathrm{wt}(\sigma) := |\{i \mid \sigma_i = +\}|$ and define $\mathrm{wt}(a) := \{\mathrm{wt}(\sigma) \mid \sigma \in \mathrm{sgn}(a)\}$. Hence $\mathrm{wt}(1-\theta_3) = \{3,2,1\}$. The following properties of the sgn-function can be easily verified.

Remark 22. *Let $a = \sum_{i=0}^{3} \alpha_i \theta_i \in \mathbb{Z}[\zeta]^+ \cap [0,1]$. Then*

(a) $|\mathrm{sgn}(a)| = 16 \Leftrightarrow a = 0 \Leftrightarrow 4 \in \mathrm{wt}(a)$,

(b) $|\mathrm{sgn}(a)| = 8 \Leftrightarrow a = 1$,

(c) $a \notin \{0,1\} \Leftrightarrow |\mathrm{sgn}(a)| \in \{1,2,4\}$.

(d) *If $k \in \mathrm{wt}(a)$ then $|\{i \mid \alpha_i \geq 0\}| \geq k$ and $|\{i \mid \alpha_i \leq 0\}| \geq 4-k$.*

An important relationship between the wt-function and the bound on the coefficients is given in the next lemma.

Lemma 23. *Let $M \geq 4$ and $a = \sum_{i=0}^{3} \alpha_i \theta_i \in \mathbb{Z}[\zeta]_M^+ \cap (0,1)$.*

(a) *If $3 \in \mathrm{wt}(a)$ then there exists i such that $0 \leq \alpha_i \leq \frac{2}{3}M$.*

(b) *If $1 \in \mathrm{wt}(a)$ then there exists i such that $0 \geq \alpha_i \geq -\frac{2}{3}M$.*

Proof. (a) There are three $\alpha_i \geq 0$ by Remark 22(d). Suppose that all of them are $> 2M/3$. Then $a > M(2(1+\theta_2+\theta_3)/3 - \theta_1) > 1$ for $M \geq 4$, a contradiction.
(b) There are three $\alpha_i \leq 0$ by Remark 22 (d). Suppose that all of them are $< -2M/3$. Then $a < M(-2(1+\theta_2+\theta_3)+\theta_1)/3 < 0$ for all M, a contradiction. □

6.2 The Main Idea

The main idea of the approximation algorithm is to start with zero and then to increase the value of the current approximation step by step by adding a small algebraic integer with inverse signature. This ensures that the coefficients of the sum are still bounded in absolute value by M. This idea is captured in the following result.

Lemma 24. *Suppose that $a, b \in \mathbb{Z}[\zeta]_M^+$ and $\mathrm{sgn}(a) \cap (-\mathrm{sgn}(b)) \neq \emptyset$. Then $a+b \in \mathbb{Z}[\zeta]_M^+$.*

Proof. Let $a = \sum_{i=0}^{3} \alpha_i \theta_i$ and $b = \sum_{i=0}^{3} \beta_i \theta_i$. The assumption implies that $\alpha_i \beta_i \leq 0$ for all $i = 0, \ldots, 3$. Since $|\alpha_i|, |\beta_i| \leq M$, we obtain $|\alpha_i + \beta_i| \leq M$ for all $i = 0, \ldots, 3$. □

We have thus to construct a set of small elements of $\mathbb{Z}[\zeta]_M^+$ having all possible signatures. The following lemma shows that it is enough to construct a set with six appropriate signatures, if we make a small sacrifice on the bound M. This new set is going to play the role of the set E_M from Section 5.

Let $\sigma_1 := (+,-,-,+)$, $\sigma_2 := (+,+,-,-)$, $\sigma_3 := (+,-,+,-)$, and for $M \in \mathbb{N}$ let $E_M := \{\varepsilon_1, \ldots, \varepsilon_6\} \subset \mathbb{Z}[\zeta]_M^+$ be such that $\sigma_i \in \mathrm{sgn}(\varepsilon_i)$ and $-\sigma_i \in \mathrm{sgn}(\varepsilon_{i+3})$, $i = 1,2,3$.

Lemma 25. *Let $a = \sum_{i=0}^{3} \alpha_i \theta_i \in \mathbb{Z}[\zeta]_M^+ \cap (0,1)$. Then there exists $\varepsilon \in E_{\lfloor M/3 \rfloor}$ such that $a + \varepsilon \in \mathbb{Z}[\zeta]_M^+$.*

PROOF. If $\mathrm{sgn}(a) \cap \{\pm\sigma_1, \pm\sigma_2, \pm\sigma_3\} \neq \emptyset$ then apply Lemma 24.

Suppose $\mathrm{sgn}(a) \cap \{\pm\sigma_1, \pm\sigma_2, \pm\sigma_3\} = \emptyset$. Then either $1 \in \mathrm{wt}(a)$ or $3 \in \mathrm{wt}(a)$. Suppose that $3 \in \mathrm{wt}(a)$, and let $\alpha_j := \min\{\alpha_i | \alpha_i \geq 0\}$. Since $2 \in \mathrm{wt}(a - \alpha_j\theta_j)$, there exists $\varepsilon \in E_{\lfloor M/3 \rfloor}$ such that $\mathrm{sgn}(a - \alpha_j\theta_j) \cap (-\mathrm{sgn}(\varepsilon)) \neq \emptyset$. Let $b := \sum_{i=0}^{3} \beta_i\theta_i := a - \alpha_j\theta_j + \varepsilon$. Then for $i \neq j$ we have $|\beta_i| \leq M$ and for $i = j$ we have $|\beta_i| \leq M/3$. Hence $b + \alpha_j\theta_j = a + \varepsilon \in \mathbb{Z}[\zeta]^+_M$, since $\alpha_j \leq 2/3M$ by Lemma 23. The case $1 \in \mathrm{wt}(a)$ is handled analogously. □

As in Section 5 we construct E_M from a smaller set E'_M having only three elements with disjoint signatures. The missing signatures can be obtained by multiplying with appropriate elements of $\mathbb{Z}[\zeta]^+$.

Lemma 26. *Suppose that* $E'_M = \{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$ *is such that* $\mathrm{sgn}(\varepsilon_i) \cap \{\sigma_i, -\sigma_i\} \neq \emptyset$, *and* $\varepsilon_i \in \mathbb{Z}[\zeta]^+_{M/3} \cap (0, 1/2)$ *for* $i = 1, 2, 3$. *Let* $E_M := \{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_1(\theta_2 - 1), \varepsilon_2\theta_3, \varepsilon_3\theta_1\}$. *Then*

(a) $E_M \subset \mathbb{Z}[\zeta]^+_M \cap (0, 1)$.

(b) $E_M = \{\varepsilon_1, \ldots, \varepsilon_6\}$ *with* $\sigma_i \in \mathrm{sgn}(\varepsilon_i) \cap (-\mathrm{sgn}(\varepsilon_{i+3}))$, $i = 1, 2, 3$.

(c) $\max(E_M) \leq 2\max(E'_M)$, $\min(E_M) = \theta_1 \min(E'_M)$.

PROOF. Part (c) is obvious, so we concentrate on the first two parts. Let $\varepsilon_1 = \sum_{i=0}^{3} \alpha_i\theta_i \in \mathbb{Z}[\zeta]^+_{\lfloor M/3 \rfloor}$ be such that $\sigma_1 \in \mathrm{sgn}(\varepsilon_1)$. We prove that $(\theta_2 - 1)\varepsilon_1 \in \mathbb{Z}[\zeta]^+_M$ and $\mathrm{sgn}((\theta_2 - 1)\varepsilon_1) \cup \mathrm{sgn}(\varepsilon_1) \supseteq \{\sigma_1, -\sigma_1\}$. Note first that we have

$$\begin{aligned} (\theta_2 - 1)\varepsilon_1 &=: \beta_0 + \beta_1\theta_1 + \beta_2\theta_2 + \beta_3\theta_3 \\ &= (2\alpha_2 - \alpha_0) + \theta_1\alpha_3 + \theta_2(\alpha_0 - \alpha_2) + \theta_3(\alpha_1 - 2\alpha_3) \in \mathbb{Z}[\zeta]^+_M, \end{aligned}$$

as $\alpha_i \leq M/3$. Furthermore, a direct calculation shows that $\mathrm{sgn}(\beta_i) = -\mathrm{sgn}(\alpha_i)$ for all i.

The other cases dealing with multiplication with θ_3 and θ_1 are handled analogously. □

The algorithm *Approximation By Signatures* (**ABS**) is given in Figure 2 and will be analyzed later in this section. The basic precomputation step of this algorithm is the construction of sets $E_{2^k} := \{\varepsilon_{1,k}, \ldots, \varepsilon_{6,k}\}$, $0 \leq k \leq \lfloor \log M \rfloor$, satisfying $\sigma_i \in \mathrm{sgn}(\varepsilon_{i,k})$ and $-\sigma_i \in \mathrm{sgn}(\varepsilon_{i+3,k})$, $i = 1, 2, 3$.

6.3 Construction of E'_M and Analysis of ABS

It remains to construct E'_M. Once again, E'_M will consist of cyclotomic units.

By (3) we know that for $a = \sum_{i=0}^{3} \alpha_i\theta_i \in \mathbb{Z}[\zeta]^+ \cap (0, 1)$ the equation

$$(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (a, a^{(1)}, a^{(2)}, a^{(3)}) \cdot T^{-1}$$

holds. Using Lemma 3 and Example 4, we get the following explicit form:

$$(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (a, a^{(1)}, a^{(2)}, a^{(3)}) \cdot \frac{1}{8} \begin{pmatrix} 2 & \theta_1 & \theta_2 & \theta_3 \\ 2 & -\theta_3 & -\theta_2 & \theta_1 \\ 2 & -\theta_1 & \theta_2 & -\theta_3 \\ 2 & \theta_3 & -\theta_2 & -\theta_1 \end{pmatrix}.$$

From this equation we can derive a relationship between the position of the maximal conjugate of an element and its signature.

```
Precomputation: Sets $E_{2^k}$ for $0 \le k \le \lfloor \log(L) \rfloor$ as described above.
Input: $\omega \in [0,1]$, $M \le L$, $M \in \mathbb{N}$.
Output: $a \in \mathbb{Z}[\zeta]^+_{9M} \cap (0,1)$, $|a - \omega| \le \max(E_{2^{\lfloor \log(M) \rfloor}})/M^3$
$a := 0$
for $l = 0$ to $\lfloor \log(M) \rfloor$ do
    while $\sum_{i=0}^{3} \alpha_i\theta_i < \omega$ do
        if there exists $\varepsilon \in E_{2^l}$ such that $\mathrm{sgn}(\varepsilon) \cap (-\mathrm{sgn}(a)) \ne \emptyset$ then
          $a := a + \varepsilon$
        else
            if $3 \in \mathrm{wt}(a)$ then
              $\alpha_j := \min\{\alpha_i \mid \alpha_i \ge 0\}$
              Find $\varepsilon \in E_{2^l}$ such that $\mathrm{sgn}(\varepsilon) \cap (-\mathrm{sgn}(a - \alpha_j\theta_j)) \ne \emptyset$
              $a := a + \varepsilon$
            else
                $\alpha_j := \min\{\alpha_i \mid \alpha_i \le 0\}$
                Find $\varepsilon \in E_{2^l}$ such that $\mathrm{sgn}(\varepsilon) \cap (-\mathrm{sgn}(a + \alpha_j\theta_j)) \ne \emptyset$
                $a := a + \varepsilon$
```

Figure 2: Approximation By Signatures (**ABS**) in $\mathbb{Z}[e^{2\pi i/16}]^+$

Proposition 27. *Suppose that $|a^{(i)}| \ge \theta_1\left(\sum_{j\ne i,0} |a^{(j)}|\right)/\theta_3$ for some $i \in \{1,2,3\}$. Then we have*

$$\mathrm{sgn}(a) \ni \begin{cases} \sigma_1 & \textit{if} \quad i=1 \quad a^{(1)} \ge 0 \\ -\sigma_1 & \textit{if} \quad i=1 \quad a^{(1)} \le 0 \\ \sigma_2 & \textit{if} \quad i=3 \quad a^{(3)} \ge 0 \\ -\sigma_2 & \textit{if} \quad i=3 \quad a^{(3)} \le 0 \\ \sigma_3 & \textit{if} \quad i=2 \quad a^{(2)} \ge 0 \\ -\sigma_3 & \textit{if} \quad i=2 \quad a^{(2)} \le 0 \end{cases}.$$

Proof. Suppose that $i = 1$, $a^{(1)} > 0$. Then

$$\begin{aligned} \alpha_0 &= \frac{1}{4}\sum_{k=0}^{3} a^{(k)} \ge \frac{1}{4}(a^{(1)} + a - |a^{(2)}| - |a^{(3)}|) \ge \frac{a}{4} \ge 0, \\ \alpha_1 &= \frac{1}{8}(\theta_1 a - \theta_3 a^{(1)} - \theta_1 a^{(2)} + \theta_3 a^{(3)}) \le \frac{\theta_1 a}{8} < 1, \\ \alpha_2 &\le \frac{\theta_2}{8}(a + |a^{(2)}| + |a^{(3)}| - a^{(1)}) \le \frac{\theta_2 a}{8} < 1, \\ \alpha_3 &\ge \frac{\theta_3 a}{8} \ge 0. \end{aligned}$$

The remaining cases can be proved analogously. □

The next theorem gives sufficient conditions for elements ε to belong to the set E'_M.

Theorem 28. *Suppose that $a \in \mathbb{Z}[\zeta]^+ \cap (0,1)$ is such that $|a^{(i)}| \ge \theta_1\left(\sum_{j\ne 0,i} |a^{(j)}|\right)/\theta_3$ for some $i \in \{1,2,3\}$, and $\sum_{j=1}^{3} |a^{(j)}| \le 4M - 1$. Then $a \in \mathbb{Z}[\zeta]^+_M$ and* $\mathrm{sgn}(a) \cap \{\sigma_i, -\sigma_i\} \ne \emptyset$.

PROOF. The assertion on $\mathrm{sgn}(a)$ has been proved in the previous proposition. The other claim is obtained using Lemma 5:

$$L_\infty(a) \;\leq\; \frac{1}{4}\sum_{i=0}^{3}|a^{(i)}| \leq \frac{1}{4}(4M-1+1) = M. \;\square$$

As in the general case in Section 5, we choose the elements of E'_M as power products of $\eta_i := \zeta^i(1-\zeta)/(1-\zeta^{2i+1})$, $i = 1,2,3$.

Suppose that $\varepsilon = \prod_{j=1}^{3}\eta_j^{k_j}$ and that i is such that $|\varepsilon^{(i)}| \geq 2\frac{\theta_1}{\theta_3}|\varepsilon^{(j)}|$ for $j \neq i, 0$, and $(1+\frac{\theta_3}{\theta_1})|\varepsilon^{(i)}| \leq 4M-1$. Then ε satisfies the assumptions of the above theorem. Taking logarithms we obtain;

$$\sum_{l=1}^{3} k_l(\log|\eta_l^{(j)}| - \log|\eta_l^{(i)}|) \leq -\log(2) - \log(\theta_1) + \log(\theta_3)$$

and

$$\sum_{l=1}^{3} k_l(\log|\eta_l^{(i)}| \leq \log(4M-1) + \log(\theta_1) - \log(\theta_1+\theta_3).$$

Minimizing $\sum_{l=1}^{3} k_l \log|\eta_l|$ subject to the above inequalities gives an Integer Linear Programming problem in three variables k_1, k_2, k_3. Exactly in the same way as in Section 5 one can prove that the optimal value of the objective function of this ILP is $O(M^{-3})$.

Now we can analyze Algorithm **ABS** along the same lines as Algorithm **GAA**.

Theorem 29. *Algorithm* **ABS** *computes the desired output in less than* $c\log(M)$ *iterations, where* c *is an absolute constant.*

PROOF. (Sketch) Let $k := \lfloor\log(M)\rfloor$. In the same way as we did in Theorem 15 we prove that **ABS** computes the desired output in no more than

$$\frac{1}{\min(E_1)} + \sum_{k=1}^{\lfloor\log M\rfloor}\frac{\max(E_{2^{k-1}})}{\min(E_{2^k})}$$

iterations. Hence, it suffices to show that $\max(E_{2^k})$ and $\min(E_{2^k})$ are $\Theta(M^{-3})$. But this follows from the above discussion and Theorem 12. $\square$

7 Implementations

In this section, we report on our implementations of the algorithms **GAA** (for $n = 5$) and **ABS**. We start with the approximation algorithm using 32nd roots of unity.

Table 1: Number of iterations of **GAA** for $n = 5$

M	est. worst case	impl.
45	723	26
58	1157	37
70	1302	39
80	1356	43
103	1405	48
118	1579	57
143	1866	91
182	2333	103
213	2392	110
238	2453	103
274	2511	104
355	2665	135
452	2695	117
509	2817	126
638	2955	131
653	3165	142
827	3324	156
1092	7504	185
1196	7815	174
1481	7899	166
1938	8145	158
2187	8211	204

7.1 32nd roots of unity

In a preprocessing phase we computed sets E_ℓ for a special sequence of ℓ's described below. The resulting ILP's were solved by the Integer Linear Programming Package `lp_solve` [2]. All the computations in the field $\mathbb{Q}(e^{2\pi i/32})$ including those of the units were done with the package `PARI` [1]. For each element found in this way, we also computed and stored all the conjugates. This data is used by the approximation algorithm to reduce computing the conjugates of the approximations to table-lookups and additions/subtractions.

To avoid a large number of iterations, we had to design the sets E_ℓ in such a way as to minimize the sum given in Theorem 15. Deviation from the sequence $\ell = 2^k$ to $\ell = \lfloor 1.2^k \rfloor$ resulted in a good tradeoff between the number of iterations and the amount of memory used to store all the E_ℓ. The theoretical upper bounds for the worst case running time (obtained via Theorem 15) for different M are compared in Table 1 with the maximum number of iterations performed for 1000 random numbers in the interval $[0, 1]$.

7.2 16th roots of unity

The units giving rise to the sets E_{2^k} were computed by solving the Integer Linear Programming problems of Section 6. Using these units we have computed upper bounds for theoretical worst case running times (in terms of the number of iterations) of the approximation algorithm as given in Theorems 15 and 29. The results are summarized in Table 2. The entry in the third column in that table gives the maximum number of iterations performed for 1000 random numbers in the interval $[0, 1]$. For the implementation of **ABS** we deviated a little from the algorithm given in Section 6 by replacing the single addition steps by multisteps obtained from multiplying the current unit by an appropriate multiple. This resulted in considerable savings of the running time, as is seen in the last column of Table 2.

7.3 Approximation of roots of unity

We used our algorithm to approximate $e^{2\pi i/1024}$ by elements of $\mathbb{Z}[e^{2\pi i/16}]_M$ and $\mathbb{Z}[e^{2\pi i/32}]_M$ for various M. Roots of unity of order a power of two are particularly im-

Table 2: Number of iterations of **ABS**

M	est. worst case	impl.	enhanced
48	23	9	8
96	134	17	11
192	221	23	16
384	332	42	31
768	406	37	26
1536	445	53	30
3072	484	52	24
6144	514	64	29
12288	645	74	33
24576	656	71	22
49152	697	86	43
98304	736	100	33
196608	767	111	33

Table 3: Approximation of $e^{2\pi i/1024}$ by elements of $\mathbb{Z}[e^{2\pi i/32}]_M$.

M	Coefficients																Absolute error
45	0	3	-1	2	-3	1	2	-4	3	-2	0	1	-1	4	-5	3	0.0064690975925
58	4	2	0	1	-3	-1	2	-4	3	-2	0	3	-1	5	-6	4	0.0055076967698
70	-4	3	-1	1	-3	2	3	4	-2	-2	-3	-2	3	-3	3	-1	0.0002526805055
80	-4	3	-1	1	-3	2	3	4	-2	-2	-3	-2	3	-3	3	-1	0.0002526805055
103	5	-1	0	-2	2	-2	3	2	5	-4	-9	2	0	6	-6	7	0.0000826104368
118	2	7	-6	2	-4	2	-2	1	5	-3	-4	-2	6	2	0	-1	0.0000591644332
143	2	7	-2	1	-10	6	-3	8	-7	4	-5	2	0	1	4	-1	0.0000363813739
182	3	-4	-1	4	-3	8	-4	8	-16	24	-26	18	-21	16	-11	10	0.0000050421465
213	-4	4	0	0	-13	0	6	7	1	3	-12	-2	7	6	-6	-6	0.0000009374573
238	14	-2	-12	3	3	3	-8	-4	1	14	2	-5	-9	3	6	0	0.0000002046956
274	14	-2	-12	3	3	3	-8	-4	1	14	2	-5	-9	3	6	0	0.0000002046956
355	-34	-14	-1	18	15	-6	6	18	1	-8	-12	4	-21	-12	-5	12	0.0000000335135
452	-2	21	10	13	-6	-43	-39	-1	1	11	33	41	0	-7	-16	-23	0.0000000275519
509	-2	29	14	16	-7	-38	-35	-6	-13	6	37	46	-1	-4	-12	-15	0.0000000020860
638	-2	29	14	16	-7	-38	-35	-6	-13	6	37	46	-1	-4	-12	-15	0.0000000020860
653	3	5	14	-21	54	-44	21	-60	40	-32	51	-14	-2	-5	4	9	0.0000000005213
827	3	5	14	-21	54	-44	21	-60	40	-32	51	-14	-2	-5	4	9	0.0000000005213
1092	-36	14	-62	-6	131	-46	-118	-58	159	36	-22	-26	59	-54	-86	24	0.0000000000594
1196	23	-65	56	-92	45	-151	33	-37	144	7	29	-51	69	46	-156	-41	0.0000000000152
1481	23	-19	-7	29	-68	9	-150	127	-50	171	-154	109	-44	167	-219	5	0.0000000000071
1938	54	8	19	-189	173	-27	75	-119	56	-17	-1	87	-61	-9	-135	168	0.0000000000004
2187	205	-115	119	-324	269	-89	159	-122	56	-14	-85	149	-157	126	-235	291	0.0000000000003

Table 4: Approximation of $e^{2\pi i/1024}$ by elements of $\mathbb{Z}[e^{2\pi i/16}]_M$.

M	Coefficients								Absolute error
96	0	-1	-6	5	1	1	0	-5	0.0164480060092
192	16	14	20	-12	-21	-6	4	40	0.0032793605419
384	30	22	-51	-5	-24	69	-3	-14	0.0001455843210
768	-49	98	33	-84	37	-108	119	-12	0.0000134058600
1536	-109	-80	-226	132	182	32	-40	-300	0.0000021850237
3072	-13	-14	-22	-123	216	-85	50	-100	0.0000000836765
6144	-279	817	354	-244	-867	-174	968	15	0.0000000236388
12288	-19	1287	-1446	1001	-971	83	708	-3	0.0000000013689
24576	2102	-3775	3899	-4872	4598	-2134	1035	-443	0.0000000003403
49152	3033	-2525	-16	465	-2921	5445	-2552	635	0.0000000000214
98304	-2606	2062	-2620	6018	-6008	8756	-12278	5498	0.0000000000062
196608	-1020	-10992	-9535	14226	9254	-12248	7467	-14144	0.0000000000013
393216	14984	-52173	2855	-2070	26775	24690	-30325	-21645	0.0000000000001

portant as they are used in radix 2 FFT-algorithms. Our approximation results are summarized in Tables 3 and 4. These tables clearly show the impressive change in dynamic range requirements when we switch from $\mathbb{Z}[e^{2\pi i/16}]$ to $\mathbb{Z}[e^{2\pi i/32}]$.

7.4 Scaling

We can approximate complex numbers in the unit circle by elements of $\mathbb{Z}[\zeta]_M/(M/4)$ rather than by those of $\mathbb{Z}[\zeta]_M$. This technique, called *scaling* is as follows: suppose that ω is a complex number inside the unit circle, and let ω_1 and ω_2 be its real and imaginary part respectively. Let $\tau_j := \omega_j M/4$. We first approximate $\tau_j - \lfloor\tau_j\rfloor$ by a number in $\mathbb{Z}[\zeta]^+_{M/4}$. Adding $\lfloor\tau_j\rfloor$ to this number gives an approximation α_j of $\omega_j M/4$ in $\mathbb{Z}[\zeta]^+_{M/2}$, and $\alpha := \alpha_1 + i\alpha_2$ gives an approximation of $\omega M/4$ in $\mathbb{Z}[\zeta]_M$. Hence, $\alpha/(M/4)$ is an approximation of ω in $\mathbb{Z}[\zeta]_M/(M/4)$, which is usually much better than a corresponding approximation in $\mathbb{Z}[\zeta]_M$. For instance, using this scaling technique we obtain an approximation of $e^{2\pi i/1024}$ as

$$\frac{1}{98304}(-1943\zeta^7 + 2840\zeta^6 - 6374\zeta^5 + 5136\zeta^4 + 362\zeta^3 + 386\zeta^2 - 1347\zeta + 96999)$$

with an absolute error of less than 1.3×10^{-15}.

7.5 Running times and Further Remarks

Compilation of Table 1 took 17 seconds on an ULTRASPARC-1. Note that it consists of 22000 approximations of real numbers in $[0, 1]$. Table 2 which accounts for 13000 approximations took 0.7 seconds on the same machine. Table 3 used 0.04 seconds, and Table 4 used under 0.01 seconds of CPU time on the same machine.

The tables also show that the bounds we have obtained on the range of the approximation is quite pessimistic. This experience is supported by a great many examples that we computed with our program.

8 Extension to other Fields

In this section we prepare the ground for a far reaching generalization of the main results of this paper. We will start with the characterization of those complex numbers w such that $\mathbb{Z}[w]$ is dense in $\mathbb{C}$, see Theorem 30. This theorem has been taken from Clausen and Shokrollahi [3]. It turns out that the necessary conditions for this to hold are also sufficient: if w is neither real nor an algebraic integer of degree 2 (in which case $\mathbb{Z}[w]$ is a lattice), then $\mathbb{Z}[w]$ is dense.

Theorem 30. *For a complex number w the following statements are equivalent.*

(a) $\mathbb{Z}[w]$ *is dense in* $\mathbb{C}$.

(b) $w \notin \mathbb{R}$ *and* $w^2 \notin \mathbb{Z} + \mathbb{Z}w$.

PROOF. It is obvious that (a) implies (b), hence we focus on the converse. The proof proceeds in several steps.
CLAIM 1: if $t \in \mathbb{C} \setminus \mathbb{R}$ and $|t| < 1$ then $\mathbb{Z}[t]$ is dense in $\mathbb{C}$.

To prove this note that, according to our assumptions, all $U_m := \mathbb{Z}t^m + \mathbb{Z}t^{m+1}$ are 2-dimensional lattices in $\mathbb{C}$. Hence, for every $z \in \mathbb{C}$ there exist uniquely determined $u, v \in \mathbb{Z}$ such that

$$z - (ut^m + vt^{m+1}) \in P_m := \{\alpha t^m + \beta t^{m+1} \mid 0 \leq \alpha, \beta < 1\}.$$

Since the diameter of P_m is smaller than $|t|^m$ and $\lim_{m\to\infty} |t|^m = 0$, our first claim follows.
CLAIM 2: If $w \in \mathbb{C} \setminus \mathbb{R}$ and $\mathbb{Z}[w]$ contains a non-zero t with $|t| < 1$ then $\mathbb{Z}[w]$ is dense in $\mathbb{C}$.

This is obvious by Claim 1, since $\mathbb{Z}[t^n w] \subseteq \mathbb{Z}[w]$ for all n.

Now let $w \in \mathbb{C} \setminus \mathbb{R}$ and $w^2 \notin \mathbb{Z} + \mathbb{Z}w$. If $\mathbb{Z}[w]$ contains a non-zero t with $|t| < 1$ then (a) follows by Claim 2. So we are left with the case

$$\mathbb{Z}[w] \cap \Big\{z \in \mathbb{C} \;\Big|\; |z| < 1\Big\} = \{0\}. \tag{5}$$

We are going to show that this case is impossible. To begin with, we note that $\mathbb{Z}[w]$ is discrete, since by (5) the difference of any two different elements in $\mathbb{Z}[w]$ has absolute value ≥ 1. Hence there exists $t \in \mathbb{Z}[w] \setminus \mathbb{R}$ satisfying

$$|t| = \min\Big\{|z| \;\Big|\; z \in \mathbb{Z}[w] \setminus \mathbb{R}\Big\}.$$

CLAIM 3: $\mathbb{Z}[w] = \mathbb{Z} + \mathbb{Z}t$.

To see this, let $z \in \mathbb{Z}[w]$. Since t is not real there exist $a, b \in \mathbb{Z}$ and $\alpha, \beta \in \mathbb{R}$ with $|\alpha|, |\beta| \leq 1/2$ such that $z = (a + bt) + (\alpha + \beta t)$. Since $t, z \in \mathbb{Z}[w]$ we have $\alpha + \beta t \in \mathbb{Z}[w]$. But then, using $t \notin \mathbb{R}$ and $|t| \geq 1$ (see (5)), we get

$$|\alpha + \beta t| \leq \frac{1}{2}\max(|1 + t|, |1 - t|) < \frac{1}{2}(1 + |t|) \leq |t|.$$

If $\alpha + \beta t \notin \mathbb{R}$ we get $\alpha + \beta t = 0$ by minimality of t. If $\alpha + \beta t \in \mathbb{R}$ then $\beta = 0$; hence $\alpha \in \mathbb{Z}[w]$. Combining this with $|\alpha| \leq 1/2$ we get $\alpha = 0$, by (5). In both cases $z \in \mathbb{Z} + \mathbb{Z}t$ and our claim follows.

To finish the proof we derive the contradiction $w^2 \in \mathbb{Z} + \mathbb{Z}w$. By Claim 3 we already know that $w = a + bt$ and $t^2 = c + dt$ for suitable $a, b, c, d \in \mathbb{Z}$. Hence

$$\begin{aligned} w^2 &= (a+bt)w = aw + b(at+bt^2) = aw + b(at + b(c+dt)) \\ &= aw + (a+bd)bt + b^2c = aw + (a+bd)(w-a) + b^2c \in \mathbb{Z} + \mathbb{Z}w. \end{aligned}$$

This completes the proof of the Theorem. □

In the same way as above we can also show that for real α the set $\mathbb{Z}[\alpha]$ is dense in $\mathbb{R}$ if and only if α is not an integer.

This result suggests the following generalization of the general approximation algorithm **GAA**: suppose that $\mathbb{Q}(w)$ is a *CM-field* [10, pp. 38], i.e., $\mathbb{Q}(w) = \mathbb{Q}(\alpha)(i)$, where α is a totally real element. Suppose further that we already know a set of units of the ring $\mathbb{Z}[w^i + \overline{w}^i \mid i \geq 0]$ with nonvanishing regulator, where bar means complex conjugation. Then in exactly the same way as in Section 5 we can construct an approximation algorithm in $\mathbb{Z}[w]$ whose worst case approximation error is $O(1/M^n)$ if the absolute values of the coefficients are bounded by M, where n is the maximal number of independent units of the ring. By Dirichlet's Unit Theorem [10] the number n equals $[\mathbb{Q}(w){:}\,\mathbb{Q}]/2$.

Abelian number fields satisfy the above assumptions. Hence, our algorithm is readily extendable to all these fields.

References

[1] C. Batut, D. Bernardi, H. Cohen, and M. Olivier. *User's Guide to PARI-GP.* Université Bordeaux, 351 Cours de la Libération, May 1995. Obtainable via anonymous ftp from `megrez.math.u-bordeaux.fr`.

[2] M. Berkelaar. `lp_solve 2.0 release`. `michel@es.ele.tue.nl`.

[3] M. Clausen and M. A. Shokrollahi. Dense $\mathbb{Z}$-modules in $\mathbb{C}$. Unpublished manuscript, 1988.

[4] J.W. Cooley and J.W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[5] J. H. Cozzens and L. A. Finkelstein. Computing the discrete Fourier transform using residue number systems in a ring of algebraic integers. *IEEE Transactions on Information Theory*, 31(5):580–588, 1985.

[6] R. A. Games. Complex approximations using algebraic integers. *IEEE Transactions on Information Theory*, 31(5):565–579, 1985.

[7] R. A. Games. An algorithm for complex approximation in $\mathbb{Z}[e^{2\pi i/8}]$. *IEEE Transactions on Information Theory*, 32(4):603–607, 1986.

[8] M. W. Marcellin and Th. R. Fischer. Encoding algorithms for complex approximation in $\mathbb{Z}[e^{2\pi i/8}]$. *IEEE Transactions on Information Theory*, 35(5):1133–1136, September 1989.

[9] M. A. Shokrollahi and V. Stemann. Approximation of complex numbers in $\mathbb{Z}[e^{2\pi i/8}]$. Technical Report TR–96–032, International Computer Science Institute, 1996.

[10] L. C. Washington. *Introduction to Cyclotomic Fields.* Springer-Verlag, New York, 1982.