Daniel KRENN

Digit Expansions with Applications in Cryptography

PHD THESIS

written to obtain the academic degree of a Doctor of Engineering Sciences

Doctoral studies of Engineering
at the doctoral school "Mathematics and Scientific Computing"

Graz University of Technology

Graz University of Technology

Supervisor:
Ao.Univ.-Prof. Dipl.-Ing. Dr.techn. Clemens HEUBERGER

Institute of Optimization and Discrete Mathematics (Math B)

Graz, December 2012

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

....

(date) (signature)

Contents

Contents iii

List of Figures v

Preface vii

Publication List ix

1 Introduction to Digit Expansions with Applications in Cryptography 1
- 1.1 Non-Adjacent Forms and Typical Questions that Arise . . . 2
- 1.2 Frobenius-and-add Methods . . . 2
- 1.3 Results on the Existence Question . . . 3
- 1.4 Results on the Optimality Question . . . 4
- 1.5 Results on the Analysis . . . 6

2 Optimality of the Width-w Non-adjacent Form 9
- 2.1 Expansions and Number Systems . . . 10
- 2.2 The Optimality Result . . . 13
- 2.3 Optimality for Integer Bases . . . 16
- 2.4 Voronoi Cells . . . 17
- 2.5 Digit Sets for Imaginary Quadratic Bases . . . 19
- 2.6 Optimality for Imaginary Quadratic Bases . . . 21
- 2.7 The p-is-3-q-is-3-Case . . . 27
- 2.8 The p-is-2-q-is-2-Case . . . 28
- 2.9 The p-is-0-Case . . . 32
- 2.10 Computational Results . . . 38

3 Existence and Optimality of w-Non-adjacent Forms 39
- 3.1 w-Non-Adjacent Forms and Digit Sets . . . 39
- 3.2 Lattices and $\mathcal{D}$-w-NAFs . . . 42
- 3.3 Tiling Based Digit Sets . . . 43

3.4 Minimal Norm Digit Set . . . 45
3.5 Optimality of $\mathcal{D}$-w-NAFs . . . 48

4 Analysis of the Width-w Non-Adjacent Form 51
4.1 Non-Adjacent Forms . . . 51
4.2 The Set-Up and Notations . . . 53
4.3 Some Basic Properties and some Remarks . . . 54
4.4 Bounds for the Value of Non-Adjacent Forms . . . 56
4.5 Right-infinite Expansions . . . 59
4.6 The Fundamental Domain . . . 60
4.7 Cell Rounding Operations . . . 65
4.8 The Characteristic Sets . . . 67
4.9 Counting the Occurrences of a non-zero Digit in a Region . . . 69
4.10 Counting Digits in Conjunction with Hyperelliptic Curve Cryptography . 82

5 On Linear Combinations of Units with Bounded Coefficients 85
5.1 Introduction . . . 85
5.2 Proof of Theorem 5.1.2 . . . 87
5.3 The Case of Simplest Cubic Fields . . . 94
5.4 Application to Signed Double-Base Expansions . . . 94

6 Sylow p-groups of Polynomial Permutations 101
6.1 Introduction . . . 101
6.2 Polynomial functions and permutations . . . 102
6.3 A group between G_1 and G_2 . . . 105
6.4 Sylow subgroups of H and G_n . . . 106

7 Analysis of Parameters of Trees Corresponding to Huffman Codes 109
7.1 Introduction . . . 109
7.2 The Generating Function . . . 114
7.3 The Height . . . 115
7.4 The Number of Distinct Depths of Leaves . . . 116
7.5 The Number of Leaves on the Last Level . . . 117
7.6 The Path Length . . . 118
7.7 The Width . . . 120
7.8 Supplement to Section 7.2, "The Generating Function" . . . 121
7.9 Supplement to Section 7.3, "The Height" . . . 123
7.10 Supplement to Section 7.4, "The Number of Distinct Depths of Leaves" . . 125
7.11 Supplement to Section 7.5, "The Number of Leaves on the Last Level" . . 127
7.12 Supplement to Section 7.6, "The Path Length" . . . 127
7.13 Supplement to Section 7.7, "The Width" . . . 136

Bibliography 145

List of Figures

2.4.1 Voronoi cell V for 0 . . . 18
2.4.2 Restricted Voronoi cell $\widetilde{V}$ for 0 . . . 18
2.5.1 Digit sets for different τ and w . . . 20
2.6.1 Bounds for the optimality of 2-NAFs . . . 22
2.6.2 Optimality of $w-1$-NAFs . . . 26
2.7.1 Digit sets for different τ and w . . . 27
2.8.1 Digit sets for different τ and w . . . 28
2.8.2 The w-is-even situation . . . 30
2.9.1 Digit sets for different τ and w . . . 33
2.9.2 The q-is-even situation . . . 34
2.9.3 The q-is-odd situation . . . 35
2.10.1 The optimality map including computational results . . . 37

4.6.1 Automaton recognising $\bigcup_{j\in\mathbb{N}} \widetilde{U}_j$ . . . 63
4.8.1 Characteristic sets $\mathcal{W}_\eta$ . . . 68

7.1.1 All elements of external size 5 . . . 111

Preface

This PhD thesis is, according to §5 Abs. 6 of the *Curriculum für das Doktoratsstudium der Technischen Wissenschaften* (Curriculum 2007 in the version 2012), a "Manteldissertation", this means, it contains a collection of publications of the author. A full list of those can be found in the chapter following this preface. This list also includes a publication of the author, which is not contained in this thesis.

The structure of this PhD thesis is as follows: The Chapters 2, 3 and 4 have a common introduction in Chapter 1. In contrast, the Chapters 5, 6 and 7 have their own introduction as first section in those chapters. Each chapter after the mentioned introduction corresponds to one publication. At the beginning of each of that chapters more informations about the publication, for example the current submission status, can be found.

Publication List

[i] Sophie Frisch and Daniel Krenn, *Sylow p-groups of polynomial permutations on the integers mod p^n*, arXiv:1112.1228v1 [math.NT], 2011.

[ii] Clemens Heuberger and Daniel Krenn, *Optimality of the width-w non-adjacent form: General characterisation and the case of imaginary quadratic bases*, arXiv:1110.0966v1 [math.NT], 2011.

[iii] ———, *Analysis of width-w non-adjacent forms to imaginary quadratic bases*, to appear in J. Number Theory (2012), earlier version available at arXiv:1009.0488v2 [math.NT].

[iv] Clemens Heuberger and Daniel Krenn, *Existence and optimality of w-non-adjacent forms with an algebraic integer base*, to appear in Acta Math. Hungar. (2012), earlier version available at arXiv:1205.4414v1 [math.NT].

[v] Clemens Heuberger, Daniel Krenn, and Stephan Wagner, *Analysis of parameters of trees corresponding to Huffman codes and sums of unit fractions*, 2013 Proceedings of the Tenth Workshop on Analytic Algorithmics and Combinatorics, to appear.

[vi] Daniel Krenn, *Analysis of the width-w non-adjacent form in conjunction with hyperelliptic curve cryptography and with lattices*, arXiv:1209.0618v1 [math.NT], 2012.

[vii] Daniel Krenn, Jörg Thuswaldner, and Volker Ziegler, *On linear combinations of units with bounded coefficients and double-base digit expansions*, to appear in Monatsh. Math., arXiv:1205.4833v1 [math.NT], 2012.

Chapter 1

Introduction to Digit Expansions with Applications in Cryptography

The RSA algorithm [88], the Diffie–Hellman key exchange [26] and the Digital Signature Algorithm [67] or its extension to elliptic curves ECDSA [59] are widely used protocols in public-key cryptography. All those cryptographic systems require two separate keys: a secret and a public one. The two keys are linked and they are constructed such that, by knowing the public key, it is extremely difficult to calculate the private. Public-key cryptography can be compared with a locking mechanism, where one of the keys of the key pair is used for locking (encrypting), the other for unlocking (decrypting). Clearly, we want those operations, as well as generation of key pairs, to be computationally easy.

From a mathematical point of view all those operations are calculations in Abelian groups: we have to build (large) multiples of an element of the group. A standard method to perform this scalar multiplication are double-and-add algorithms. Let n be a positive integer and P be a group element, then we write n in standard binary expansion, i.e., with base 2 and digits 0 and 1, and calculate nP by a Horner scheme. That is done digit-by-digit from the most significant digit of the binary expansion of n to least significant one. The number of doublings needed corresponds to the length of the expansion, the number of non-trivial additions to its weight (that is the number of non-zero digits). This means, in order to make the mentioned scalar multiplication fast, we want an expansion of the integer n which is short and has only a low number of non-zeros.

To achieve such expansions, we allow other bases than 2 and other digit sets. For example, if we use an additional digit -1 and keep base 2, then each integer has many representations. We choose one that fulfils the criteria above. A candidate for such a "good" expansion is the so called *non-adjacent form*, where from each two adjacent digits at most one is non-zero. We come back to those expansions later. In general, the strategy of using larger digit sets leads to representations with lower weights. The prize to pay are more precomputations (before the Horner scheme can be used, we have to calculate dP for each digit d). So one has to find a balance between the size of the

digit set (or re-formulated, the weight of an expansion) and the time needed for that precomputation.

1.1 Non-Adjacent Forms and Typical Questions that Arise

As mentioned above, we come back to non-adjacent forms. Let w be a positive integer, then an expansion is a *width-w non-adjacent form*, abbreviated by *w-NAF*, if in each block of w consecutive digits at most one is non-zero. This is an expansion with a low number of non-zero digits. Therefore, it is particularly suited for our scalar multiplication algorithm, since that leads to a low number of additions. It seems that the term non-adjacent form was first mentioned in Reitwiesner [86]. There it was used for faster arithmetic of binary expansions. The generalizations to w-NAFs are independently due to Blake, Seroussi and Smart [19], Cohen, Miyaji and Ono [77], Solinas [94, 95], and others.

Now, having the syntactic condition of the w-NAF, several questions arise and will be discussed in this thesis. A first natural question is, if it is always possible to write a number as the value of a w-NAF-expansion, i.e., if each element (e.g. each integer) admits such an expansion. Since this depends on the digit set, the question can be reformulated: Is there a digit set or how to choose the digit set such that each element has a w-NAF-expansion? Section 1.3 gives an introduction to such questions. See also Chapter 3.

It was mentioned above that the w-NAF is an expansion with a low weight. But is it an expansion with the minimal possible weight, or is there another expansion, which gives a "better" representation? This question, namely if the w-NAF is optimal/minimal, is discussed in Section 1.4 and in the Chapters 2 and 3.

A third question discussed in this thesis concerns the running time of the scalar multiplication algorithm. To answer that we need a precise analysis of the occurrence of a digit in all w-NAF-expansions (for example in some region around zero). Section 1.5 and Chapter 4 are devoted to that problem.

1.2 Frobenius-and-add Methods

We want to perform calculations in Abelian groups, but up to now we did not use any special structure or property of the group. When this group comes from elliptic curve cryptography or hyperelliptic curve cryptography, we can do better. To explain that more precisely, suppose we have an elliptic curve over a finite field with q elements and further that the Abelian group is the point group of the curve over a field extension of that finite field. Alternatively, and more generally, since the elliptic curve case is included, we can use hyperelliptic curves and the corresponding Jacobian variety. Then, since we are working over finite fields, we have a q-Frobenius endomorphism available, and that operation is "cheap" (meaning fast), especially using normal bases, see for example Ash, Blake and Vanstone [1]. So we want to replace the "expensive" doublings of the double-and-add algorithm by this Frobenius endomorphism. The result is called

Frobenius-and-add method, see Koblitz [63] and Solinas [94, 95], where this method is used in conjunction with Koblitz curves in characteristic two, or Smart [93] and Avanzi, Heuberger and Prodinger [4] for Koblitz curves in characteristic three.

Now let us look at those methods more closely. Suppose we can write n (for example a positive integer) as

$$n = \sum_{\ell=0}^{L-1} \xi_\ell \tau^\ell$$

with some digits ξ_ℓ out of a digit set and where the base τ is an algebraic integer. If τ is a zero of the characteristic polynomial of the q-Frobenius endomorphism on the Jacobian of the hyperelliptic curve, then for an element P of the Jacobian we can compute nP by

$$nP = \sum_{\ell=0}^{L-1} \xi_\ell \, \varphi^\ell(P) \,,$$

in which φ denotes the Frobenius endomorphism. Again, we would evaluate the previous calculation by a Horner scheme, where we use applications of the Frobenius endomorphism instead of doublings.

When we come from elliptic curves, then the Frobenius endomorphism fulfils a quadratic polynomial, see for example Koblitz [64] or Silverman [91], whereas in the general case, it is of degree $2g$, where g is the genus of the curve. Further, that characteristic polynomial is the reciprocal of the numerator of the zeta-function of the hyperelliptic curve, cf. Weil [107, 109]. Further, the Riemann Hypothesis of the Weil Conjectures, cf. Weil [108], Dwork [32] and Deligne [25], state that all conjugates of the zero τ have the same absolute value.

Therefore, numeral systems with general algebraic integer bases are of interest and worth discussing.

1.3 Results on the Existence Question

As discussed above, one major question is, which digit sets should we use in order to ensure that each element has an expansion as a w-NAF. If a digit set fulfils that property, then we call it a *w-non-adjacent digit set* (w-NADS). Note that for now we do not yet assume our expansions are minimal with respect to their Hamming-weight, cf. Section 1.4.

In the case that the base τ is an integer, results on that questions can be found in Reitwiesner [86], Solinas [94, 95], Muir and Stinson [79]. One can choose integers not divisible by the base τ with absolute value not larger than $\frac{1}{2}\,|\tau|^w$ as a digit set in those cases, see also Example 3.1.3. For some imaginary quadratic algebraic integers as a base, results are due to Solinas [94, 95], Koblitz [65], Lange [71], and Blake, Murty and Xu [16, 17, 18]. There, they choose a digit set consisting of representatives of minimal absolute value of residue classes modulo τ^w, which are not divisible by τ. Such a digit set is a generalization of the one used with rational integers above. For an arbitrary imaginary quadratic algebraic integer as base, this was generalized in Heuberger and

Krenn [49]. Other existence results, most of them for expansions without syntax, are due to Müller [80], Smart [93], Günther, Lange and Stein [46] in the quadratic case (coming from elliptic curves) and due to Ciet, Lange, Sica and Quisquater [23] and Lange [71] for higher degree cases (coming from Koblitz curve cryptosystems).

The aim of Chapter 3 is to answer the existence question for expansions with w-NAF-syntax for a general algebraic integer base. But the set-up in that chapter is even more general: In Section 3.1, which contains the definitions and some basic results, we work in an Abelian group and the base is represented by an injective endomorphism on that group. In the remaining article, starting with Section 3.2, the set-up is a lattice Λ in $\mathbb{R}^n$ and an injective endomorphism on Λ as base. The case of algebraic integer bases is a special case of this set-up, cf. Examples 3.1.2 and 3.1.7. One main step there was to use the Minkowski map to transform the τ-adic setting to a lattice.

In Section 3.2 we prove a necessary condition to be a w-NADS, namely that the endomorphism has to be expanding. Section 3.3 deals with the setting when the digit set comes from a tiling of $\mathbb{R}^n$. Theorem 3.3.3 states that we have a w-NADS if w is sufficiently large. The bound in that result is explicit. Another result of that kind is given in Section 3.4, generalising a result of Germán and Kovács [42] to $\mathcal{D}$-w-NAFs. There minimal norm digit sets are studied. Again we get a w-NADS if w is larger than a constant, which depends (only) on the eigenvalues of Φ, cf. Theorem 3.4.1. As an important example, we discuss the setting of bases τ coming from hyperelliptic curves, see above, in Example 3.4.3.

1.4 Results on the Optimality Question

A w-NAF-expansion of an element has low weight and therefore leads to quite efficient scalar multiplication in the double-and-add and Frobenius-and-add methods. This part of the introduction and also Chapter 2 is devoted to the following question: Does the w-NAF minimise the weight, i.e. the number of non-zero digits, among all possible representations (multi-expansions) with the same digit set? If the answer is affirmative, we call the w-NAF-expansion *optimal* or *minimal*.

To answer that question we will first work in general numeral systems: A *general numeral system* is an Abelian group $\mathcal{A}$ together with a group endomorphism Φ and a digit set $\mathcal{D}$, which is a finite subset of $\mathcal{A}$ including 0. The endomorphism acts as base in our numeral system. A common choice is multiplication by a fixed element. In this general setting we consider *multi-expansions*, which are simply finite sums with summands $\Phi^k(d)$, where $k \in \mathbb{N}_0$ and d is a non-zero digit in $\mathcal{D}$. The "multi" in the expression "multi-expansion" means that we allow several summands with the same k. If the k are pairwise distinct, we call the sum an *expansion*. Note that in the context of our scalar multiplication algorithms, multi-expansions are as good as expansions, as long as the weight is low.

In that general set-up, we give conditions equivalent to optimality in Section 2.2. We show that each group element has an optimal w-NAF-expansion if and only if the digit set is *w-subadditive*, which means that each multi-expansion with two summands has a

w-NAF-expansion with weight at most 2. This condition can be verified algorithmically, since there are only finitely many non-trivial cases to check. More precisely, one has to consider the w-NAFs corresponding to $w(\#\mathcal{D}-1)^2$ multi-expansions. Another way to verify w-subadditivity is to use the geometry of the digit set. This is done in the imaginary quadratic setting and then later in the general algebraic integer base setting, see below for more details.

Now consider some special cases of number systems, where optimality or non-optimality of the non-adjacent form is already known. Here, multiplication by a base element is chosen as endomorphism Φ. In the case of 2-NAFs with digit set $\{-1,0,1\}$ and base 2, optimality is known, cf. Reitwiesner [86]. This was reproved in Jedwab and Mitchell [56] and in Gordon [43]. That result was generalised in Avanzi [5], Muir and Stinson [79] and in Phillips and Burgess [85]. There, the optimality of the w-NAFs with base 2 was shown. As digit set, zero and all odd numbers with absolute value less than 2^{w-1} were used. In this setting, there is also another optimal expansion, cf. Muir and Stinson [78]. Using base 2 and a digit set $\{0,1,x\}$ with $x \in \mathbb{Z}$, optimality of the 2-NAFs is answered in Heuberger and Prodinger [52]. Some of these results will be reproved and extended to arbitrary rational integer bases with our tools in Section 2.3. That proof will show the main idea how to use the geometry of the digit set to show w-subadditivity and therefore optimality.

We come back to our imaginary quadratic setting, so suppose that the imaginary quadratic base τ is a solution of $\tau^2 - p\tau + q = 0$, where p and q are rational integers with $q > p^2/4$. Here, $\mathbb{Z}[\tau]$ plays the rôle of the group and multiplication by τ is taken as the endomorphism. We suppose that the digit set consists of 0 and one representative of minimal norm of every residue class modulo τ^w, which is not divisible by τ, and we call it a *minimal norm representatives digit set*, see Section 2.5 for a precise formulation.

First, consider the cases $|p| = 1$ and $q = 2$, which comes from a Koblitz curve in characteristic 2, cf. Koblitz [63], Meier and Staffelbach [76], and Solinas [94, 95]. There optimality of the w-NAFs can be shown for $w \in \{2,3\}$, cf. Avanzi, Heuberger and Prodinger [2, 3]. The case $w = 2$ can also be found in Gordon [43]. For the cases $w \in \{4,5,6\}$, non-optimality was shown, see Heuberger [47].

In Chapter 2 we give a general result on the optimality of the w-NAFs with imaginary quadratic bases, namely when $|p| \geq 3$, as well as some results for special cases. So let $|p| \geq 3$. If $w \geq 4$, then optimality of the w-NAFs could be shown in all cases. If we restrict the set-up to $|p| \geq 5$, then the w-NAFs are already optimal for $w \geq 3$. Further, we give a condition (p and q have to fulfil a special inequality), when 2-NAFs are optimal. All those results can be found in Section 2.6. There we show that the digit set in that cases is w-subadditive by using its geometry.

In the last four sections of Chapter 2 some special cases are examined. Important ones are the cases $|p| = 3$ and $q = 3$ coming from Koblitz curves in characteristic 3. In Kröll [70] optimality of the w-NAFs was shown for $w \in \{2,3,4,5,6,7\}$ by using a transducer and some heavy symbolic computations. Here, in Section 2.7, we prove that the w-NAF-expansions are optimal for all $w \geq 2$. In Section 2.8 we look at the cases $|p| = 2$ and $q = 2$. There the w-NAF-expansions are optimal if and only if w is odd. In the cases $p = 0$ and $q \geq 2$, see Section 2.9, non-optimality of the w-NAFs with odd w

could be shown.

The last section of Chapter 3 is devoted to an answer to the question of optimality in the case of general algebraic integer bases and in conjunction with lattices. We provide a positive answer for sufficiently large w and sufficiently large eigenvalues of the endomorphism in the lattice set-up in Theorem 3.5.4. This can then be used to answer optimality in the τ-adic setting.

1.5 Results on the Analysis

Now we talk about the third question discussed in Section 1.1. The work presented in Chapter 4 deals with analysing the number of occurrences of a digit in w-NAF-expansions with an algebraic integer base τ, where all conjugates have the same absolute value, cf. also the introduction in Section 1.2. This is needed for the analysis of the running time of the scalar multiplication algorithm (Frobenius-and-add) mentioned earlier in this introduction. As brought up in Section 1.3 and Chapter 3, we will do this analysis in the set-up of numeral systems in lattices, cf. Section 4.10. Our main result is the asymptotic formula

$$Z_\eta \sim N^n \lambda(U) E \log_{|\tau|} N.$$

for the number Z_η of occurrences of a fixed non-zero digit η in w-NAF-expansions in some region NU (e.g. a ball around 0). There, $\lambda(U)$ denotes the Lebesgue measure of U and E is a constant, see below. The main term of that formula coincides with the full block length analysis given in Heuberger and Krenn [49]. There an explicit expression for the expectation E and the variance of the occurrence of such a digit in all expansions of a fixed length is given. The result here is more precise: A periodic fluctuation in the second order term is also exhibited. Such structures—main term, oscillation term, smaller error term—are not uncommon in the context of digits counting, see for instance, Heuberger and Prodinger [52] or Grabner, Heuberger and Prodinger [44]. The result here is a generalisation of the one found in Heuberger and Krenn [49]. The proof, as the one in [49], follows Delange's method, cf. Delange [24], but several technical problems have to be taken into account.

The structure of Chapter 4 is as follows. We start with the formal definition of numeral systems and the non-adjacent form in Section 4.1. Sections 4.2 and 4.3 contain our primary set-up in a lattice. We will work in this set-up throughout the entire chapter. There also the used digit set, which comes from a tiling by the lattice, is defined. Additionally, some notations are fixed and some basic properties are given. The end of Section 4.2 is devoted to the full block length analysis theorem given in Heuberger and Krenn [49]. In Sections 4.4 to 4.8 a lot of properties of the investigated expansions, such as bounds of the value and the behaviour of the fundamental domain and the characteristic sets, are derived. Those are needed to prove our main result, the counting theorem in Section 4.9. The last section will forge a bridge to the τ-adic set-up. This is explained with details there and the counting theorem is restated in that set-up.

A last remark on the proofs given in Chapter 4. As that chapter is a generalisation of Heuberger and Krenn [49] several proofs of propositions and lemmata are skipped.

All those are straightforward generalisations of the ones for the quadratic case, which means, we have to do things like replacing $\mathbb{Z}[\tau]$ by the lattice, the multiplication by τ by a lattice endomorphism, the dimension 2 by n, using a norm instead of the absolute value, and so on. If the generalisation is not that obvious, the proofs are given.

Chapter 2

Optimality of the Width-w Non-adjacent Form

This chapter contains the article [48] with the title "Optimality of the Width-w Non-adjacent Form: General Characterisation and the Case of Imaginary Quadratic Bases". It is joint work with Clemens Heuberger. The article is submitted to *Journal de Théorie des Nombres de Bordeaux*. An introduction to this chapter can be found in Chapter 1, in particular Section 1.4.

Abstract

Efficient scalar multiplication in Abelian groups (which is an important operation in public key cryptography) can be performed using digit expansions. Apart from rational integer bases (double-and-add algorithm), imaginary quadratic integer bases are of interest for elliptic curve cryptography, because the Frobenius endomorphism fulfils a quadratic equation. One strategy for improving the efficiency is to increase the digit set (at the prize of additional precomputations). A common choice is the width-w non-adjacent form (w-NAF): each block of w consecutive digits contains at most one non-zero digit. Heuristically, this ensures a low weight, i.e. number of non-zero digits, which translates in few costly curve operations. This chapter investigates the following question: Is the w-NAF-expansion optimal, where optimality means minimising the weight over all possible expansions with the same digit set?

The main characterisation of optimality of w-NAFs can be formulated in the following more general setting: We consider an Abelian group together with an endomorphism (e.g., multiplication by a base element in a ring) and a finite digit set. We show that each group element has an optimal w-NAF-expansion if and only if this is the case for each sum of two expansions of weight 1. This leads both to an algorithmic criterion and to generic answers for various cases.

Imaginary quadratic integers of trace at least 3 (in absolute value) have optimal w-NAFs for $w \geq 4$. The same holds for the special case of base $(\pm 3 \pm \sqrt{-3})/2$ (four cases) and $w \geq 2$, which corresponds to Koblitz curves in characteristic three. In the case

of $\tau = \pm 1 \pm i$(again four cases), optimality depends on the parity of w. Computational results for small trace are given.

2.1 Expansions and Number Systems

This section contains the abstract definition of number systems and the definition of expansions. Further, we specify the width-w non-adjacent form and notions related to it.

Abstract number systems can be found in van de Woestijne [103], which are generalisations of the number systems used, for example, in Germán and Kovács [42]. We use that concept to define w-NAF-number systems.

Definition 2.1.1. A *pre-number system* is a triple $(\mathcal{A}, \Phi, \mathcal{D})$ where $\mathcal{A}$ is an Abelian group, Φ an endomorphism of $\mathcal{A}$ and the *digit set* $\mathcal{D}$ is a subset of $\mathcal{A}$ such that $0 \in \mathcal{D}$ and each non-zero digit is not in the image of Φ.

Note that we can assume Φ is not surjective, because otherwise the digit set would only consist of 0.

Before we define expansions and multi-expansions, we give a short introduction on multisets. We take the notation used, for example, in Knuth [62].

Notation 2.1.2. A *multiset* is like a set, but identical elements are allowed to appear more than once. For a multiset A, its cardinality $\#A$ is the number of elements in the multiset. For multisets A and B, we define new multisets $A \uplus B$ and $A \setminus B$ in the following way: If an element occurs exactly a times in A and b times in B, then it occurs exactly $a + b$ times in $A \uplus B$ and it occurs exactly $\max(a - b, 0)$ times in $A \setminus B$.

Now a pre-number system (and multisets) can be used to define what expansions and multi-expansions are.

Definition 2.1.3 (Expansion)**.** Let $(\mathcal{A}, \Phi, \mathcal{D})$ be a pre-number system, and let $\boldsymbol{\mu}$ be a multiset with elements $(d, n) \in (\mathcal{D} \setminus \{0\}) \times \mathbb{N}_0$. We define the following:

1. We set
$$\mathsf{weight}(\boldsymbol{\mu}) := \#\boldsymbol{\mu}$$
and call it the *Hamming-weight of* $\boldsymbol{\mu}$ or simply *weight of* $\boldsymbol{\mu}$. The multiset $\boldsymbol{\eta}$ is called *finite*, if its weight is finite.

2. We call an element $(d, n) \in \boldsymbol{\mu}$ an *atom* and $\Phi^n(d)$ the *value of the atom* (d, n).

3. Let $\boldsymbol{\mu}$ be finite. We call
$$\mathsf{value}(\boldsymbol{\mu}) := \sum_{(d,n) \in \boldsymbol{\mu}} \Phi^n(d)$$
the *value of* $\boldsymbol{\mu}$.

4. Let $z \in \mathcal{A}$. A *multi-expansion of* z is a finite $\boldsymbol{\mu}$ with $\mathsf{value}(\boldsymbol{\mu}) = z$.

5. Let $z \in \mathcal{A}$. An *expansion of* z is a multi-expansion $\boldsymbol{\mu}$ of z where all the n in $(d, n) \in \boldsymbol{\mu}$ are pairwise distinct.

We use the following conventions and notations. If necessary, we see an atom as a multi-expansion or an expansion of weight 1. We identify an expansion $\boldsymbol{\eta}$ with the sequence $(\eta_n)_{n\in\mathbb{N}_0} \in \mathcal{D}^{\mathbb{N}_0}$, where $\eta_n = d$ for $(d, n) \in \boldsymbol{\eta}$ and all other $\eta_n = 0$. For an expansion $\boldsymbol{\eta}$ (usually a bold, lower case Greek letter) we will use η_n (the same letter, but indexed and not bold) for the elements of the sequence. Further, we identify expansions (sequences) in $\mathcal{D}^{\mathbb{N}_0}$ with finite words over the alphabet $\mathcal{D}$ written from right (least significant digit) to left (most significant digit), except left-trailing zeros, which are usually skipped. Besides, we follow the terminology of Lothaire [73] for words.

Note, if $\boldsymbol{\eta}$ is an expansion, then the weight of $\boldsymbol{\eta}$ is

$$\mathsf{weight}(\boldsymbol{\eta}) = \#\{n \in \mathbb{N}_0 \,:\, \eta_n \neq 0\}$$

and the value of $\boldsymbol{\eta}$ is

$$\mathsf{value}(\boldsymbol{\eta}) = \sum_{n\in\mathbb{N}_0} \Phi^n(\eta_n)\,.$$

For the sake of completeness — although we do not need it in this paper — a pre-number system is called *number system* if each element of $\mathcal{A}$ has an expansion. We call the number system *non-redundant* if there is exactly one expansion for each element of $\mathcal{A}$, otherwise we call it *redundant*. We will modify this definition later for w-NAF number systems.

Before going any further, we want to see some simple examples for the given abstract definition of a number system. We use multiplication by an element τ as endomorphism Φ. This leads to values of the type

$$\mathsf{value}(\boldsymbol{\eta}) = \sum_{n\in\mathbb{N}_0} \eta_n \tau^n$$

for an expansion $\boldsymbol{\eta}$.

Example 2.1.4. The binary number system is the pre-number system

$$(\mathbb{N}_0, z \mapsto 2z, \{0, 1\}).$$

It is a non-redundant number system, since each integer admits exactly one binary expansion. We can extend the binary number system to the pre-number system

$$(\mathbb{Z}, z \mapsto 2z, \{-1, 0, 1\}),$$

which is a redundant number system.

In order to get a non-redundant number system out of a redundant one, one can restrict the language, i.e. we forbid some special configurations in an expansion. There is one special kind of expansion, namely the non-adjacent form, where no adjacent non-zeros are allowed. A generalisation of it is defined here.

Definition 2.1.5 (Width-w Non-Adjacent Form)**.** Let w be a positive integer and $\mathcal{D}$ be a digit set (coming from a pre-number system). Let $\boldsymbol{\eta} = (\eta_j)_{j\in\mathbb{N}_0} \in \mathcal{D}^{\mathbb{N}_0}$. The sequence $\boldsymbol{\eta}$ is called a *width-w non-adjacent form*, or *w-NAF* for short, if each factor[1] $\eta_{j+w-1}\dots\eta_j$, i.e. each block of length w, contains at most one non-zero digit.

A w-NAF-expansion is an expansion that is also a w-NAF.

Note that a w-NAF-expansion is finite. With the previous definition we can now define what a w-NAF number system is.

Definition 2.1.6. Let w be a positive integer. A pre-number system $(\mathcal{A}, \Phi, \mathcal{D})$ is called a *w-NAF number system* if each element of $\mathcal{A}$ admits a w-NAF-expansion, i.e. for each $z \in \mathcal{A}$ there is a w-NAF $\boldsymbol{\eta} \in \mathcal{D}^{\mathbb{N}_0}$ with $\mathsf{value}(\boldsymbol{\eta}) = z$. We call a w-NAF number system *non-redundant* if each element of $\mathcal{A}$ has a unique w-NAF-expansion, otherwise we call it *redundant*.

Now we continue the example started above.

Example 2.1.7. The redundant number system

$$(\mathbb{Z}, z \mapsto 2z, \{-1, 0, 1\})$$

is a non-redundant 2-NAF number system. This fact has been shown in Reitwiesner [86]. More generally, for an integer w at least 2, the number system

$$(\mathbb{Z}, z \mapsto 2z, \mathcal{D}),$$

where the digit set $\mathcal{D}$ consists of 0 and all odd integers with absolute value smaller than 2^{w-1}, is a non-redundant w-NAF number system, cf. Solinas [94, 95] or Muir and Stinson [79].

Finally, since this paper deals with the optimality of expansions, we have to define the term "optimal". This is done in the following definition.

Definition 2.1.8 (Optimal Expansion)**.** Let $(\mathcal{A}, \Phi, \mathcal{D})$ be a pre-number system, and let $z \in \mathcal{A}$. A multi-expansion or an expansion $\boldsymbol{\mu}$ of z is called *optimal* if for any multi-expansion $\boldsymbol{\nu}$ of z we have

$$\mathsf{weight}(\boldsymbol{\mu}) \le \mathsf{weight}(\boldsymbol{\nu}),$$

i.e. $\boldsymbol{\mu}$ minimises the Hamming-weight among all multi-expansions of z. Otherwise $\boldsymbol{\mu}$ is called *non-optimal*.

The "usual" definition of optimal, cf. [86, 56, 43, 5, 79, 85, 78, 52, 2, 3, 47], is more restrictive: An expansion of $z \in \mathcal{A}$ is optimal if it minimises the weight among all expansions of z. The difference is that in Definition 2.1.8 we minimise over all multi-expansions. The use of multi-expansions is motivated by applications: we want to do efficient operations. There it is no problem to take multi-expansions if they are "better", so it is more natural to minimise over all of them instead of just over all expansions.

[1] See Lothaire [73] for the used terminology on words.

2.2 The Optimality Result

This section contains our main theorem, the Optimality Theorem, Theorem 2.2.2. It contains four equivalences. One of them is a condition on the digit set and one is optimality of the w-NAF. We start with the definition of that condition on the digit set.

Definition 2.2.1. Let $(\mathcal{A}, \Phi, \mathcal{D})$ be a pre-number system, and let w be a positive integer. We say that the digit set $\mathcal{D}$ is *w-subadditive* if the sum of the values of two atoms has a w-NAF-expansion of weight at most 2.

In order to verify the w-subadditivity-condition it is enough to check atoms $(c, 0)$ and (d, n) with $n \in \{0, \ldots, w-1\}$ and non-zero digits c and d. Therefore, one has to consider $w\,(\#\mathcal{D}-1)^2$ multi-expansions.

Theorem 2.2.2 (Optimality Theorem)**.** *Let $(\mathcal{A}, \Phi, \mathcal{D})$ be a pre-number system with*

$$\bigcap_{m\in\mathbb{N}_0} \Phi^m(\mathcal{A}) = \{0\}\,,$$

and let w be a positive integer. Then the following statements are equivalent:

(1) The digit set $\mathcal{D}$ is w-subadditive.

(2) For all multi-expansions $\boldsymbol{\mu}$ there is a w-NAF-expansion $\boldsymbol{\xi}$ such that

$$\mathsf{value}(\boldsymbol{\xi}) = \mathsf{value}(\boldsymbol{\mu})$$

and

$$\mathsf{weight}(\boldsymbol{\xi}) \leq \mathsf{weight}(\boldsymbol{\mu})\,.$$

(3) For all w-NAF-expansions $\boldsymbol{\eta}$ and $\boldsymbol{\vartheta}$ there is a w-NAF-expansion $\boldsymbol{\xi}$ such that

$$\mathsf{value}(\boldsymbol{\xi}) = \mathsf{value}(\boldsymbol{\eta}) + \mathsf{value}(\boldsymbol{\vartheta})$$

and

$$\mathsf{weight}(\boldsymbol{\xi}) \leq \mathsf{weight}(\boldsymbol{\eta}) + \mathsf{weight}(\boldsymbol{\vartheta})\,.$$

(4) If $z \in \mathcal{A}$ admits a multi-expansion, then z also admits an optimal w-NAF-expansion.

Note that if we assume that each element $\mathcal{A}$ has at least one expansion (e.g. by assuming that we have a w-NAF number system), then we have the equivalence of w-subadditivity of the digit set and the existence of an optimal w-NAF-expansion for each group element.

We will use the term "addition" in the following way: The addition of two group elements x and y means finding a w-NAF-expansion of the sum $x + y$. Addition of two multi-expansions shall mean addition of their values.

Proof of Theorem 2.2.2. For a non-zero $z \in \mathcal{A}$, we define

$$L(z) := \max\{m \in \mathbb{N}_0 \,:\, z \in \Phi^m(\mathcal{A})\}\,.$$

The function L is well-defined, because

$$\bigcap_{m\in\mathbb{N}_0} \Phi^m(\mathcal{A}) = \{0\}\,.$$

We show that *(1)* implies *(2)* by induction on the pair $(\mathsf{weight}(\boldsymbol{\mu})\,, L(\mathsf{value}(\boldsymbol{\mu})))$ for the multi-expansion $\boldsymbol{\mu}$. The order on those pairs is lexicographic. In the case $\mathsf{value}(\boldsymbol{\mu}) = 0$, we choose $\boldsymbol{\xi} = 0$ and are finished. Further, if the multi-expansion $\boldsymbol{\mu}$ consists of less than two elements, then there is nothing to do, so we suppose $\mathsf{weight}(\boldsymbol{\eta}) \geq 2$.

We choose an atom $(d, n) \in \boldsymbol{\mu}$ (note that $d \in \mathcal{D} \setminus \{0\}$ and $n \in \mathbb{N}_0$) with minimal n. If $n > 0$, then we consider the multi-expansion $\boldsymbol{\mu}'$ arising from $\boldsymbol{\mu}$ by shifting all indices by n, use the induction hypothesis on $\boldsymbol{\mu}'$ and apply Φ^n. Note that $\boldsymbol{\mu}'$ and $\boldsymbol{\mu}$ have the same weight, but

$$L\big(\mathsf{value}\big(\boldsymbol{\mu}'\big)\big) = L(\mathsf{value}(\boldsymbol{\mu})) - n < L(\mathsf{value}(\boldsymbol{\mu}))\,.$$

So we can assume $n = 0$. Set $\boldsymbol{\mu}^\star := \boldsymbol{\mu} \setminus \{(d, 0)\}$. Using the induction hypothesis, there is a w-NAF-expansion $\boldsymbol{\eta}$ of $\mathsf{value}(\boldsymbol{\mu}^\star)$ with weight strictly smaller than $\mathsf{weight}(\boldsymbol{\mu}) = \mathsf{weight}(\boldsymbol{\mu}^\star) + 1$.

Consider the addition of $\boldsymbol{\eta}$ and the digit d. If the digits η_ℓ are zero for all $\ell \in \{0, \ldots, w-1\}$, then the result follows by setting $\boldsymbol{\xi} = \ldots \eta_{w+1}\eta_w 0^{w-1} d$. So we can assume

$$\boldsymbol{\eta} = \boldsymbol{\beta} 0^{w-k-1} b 0^k$$

with a w-NAF $\boldsymbol{\beta}$, a digit $b \neq 0$ and $k \in \{0, \ldots, w-1\}$. Note that there are at least $w-1$ zeros on the left hand-side of b in $\boldsymbol{\eta}$, but for our purposes, it is sufficient (and more convenient) to consider only $w-k-1$ zeros. Since the digit set $\mathcal{D}$ is w-subadditive, there is a w-NAF $\boldsymbol{\gamma}$ of $\Phi^k(b) + d$ with weight at most 2. If the weight is strictly smaller than 2, we use the induction hypothesis on the multi-expansion $\boldsymbol{\beta} 0^w \uplus \boldsymbol{\gamma}$ to get a w-NAF $\boldsymbol{\xi}$ with the desired properties and are done. Otherwise, denoting by J the smallest index with $\gamma_J \neq 0$, we distinguish between two cases: $J = 0$ and $J > 0$.

First let $J = 0$. The w-NAF $\boldsymbol{\beta}$ (seen as multi-expansion) has a weight less than $\mathsf{weight}(\boldsymbol{\eta})$, so, by induction hypothesis, there is a w-NAF $\boldsymbol{\xi}'$ with

$$\mathsf{value}\big(\boldsymbol{\xi}'\big) = \mathsf{value}(\boldsymbol{\beta}) + \mathsf{value}(\ldots \gamma_{w+1}\gamma_w)$$

and

$$\mathsf{weight}\big(\boldsymbol{\xi}'\big) \leq \mathsf{weight}(\boldsymbol{\beta}) + \mathsf{weight}(\ldots \gamma_{w+1}\gamma_w)\,.$$

We set $\boldsymbol{\xi} = \boldsymbol{\xi}'\gamma_{w-1}\ldots\gamma_0$. Since $\boldsymbol{\xi}$ is a w-NAF-expansion we are finished, because

$$\begin{aligned}\mathsf{value}(\boldsymbol{\xi}) &= \Phi^w(\mathsf{value}(\boldsymbol{\beta})) + \Phi^w(\mathsf{value}(\ldots\gamma_{w+1}\gamma_w)) + \mathsf{value}(\gamma_{w-1}\ldots\gamma_0)\\ &= \Phi^w(\mathsf{value}(\boldsymbol{\beta})) + \Phi^k(b) + d = \mathsf{value}(\boldsymbol{\eta}) + d = \mathsf{value}(\boldsymbol{\mu}^\star) + d = \mathsf{value}(\boldsymbol{\mu})\end{aligned}$$

and

$$\begin{aligned}\mathsf{weight}(\boldsymbol{\xi}) = \mathsf{weight}\big(\boldsymbol{\xi}'\big) + \mathsf{weight}(\gamma_{w-1}\dots\gamma_0) &\le \mathsf{weight}(\boldsymbol{\beta}) + \mathsf{weight}(\boldsymbol{\gamma}) \\ &\le \mathsf{weight}(\boldsymbol{\eta}) + 1 \le \mathsf{weight}(\boldsymbol{\mu}^\star) + 1 = \mathsf{weight}(\boldsymbol{\mu})\,.\end{aligned}$$

Now, in the case $J > 0$, we consider the multi-expansion $\boldsymbol{\nu} := \boldsymbol{\beta}0^w \uplus \boldsymbol{\gamma}$. We use the induction hypothesis for $\boldsymbol{\nu}$ shifted by J (same weight, L decreased by J) and apply Φ^J on the result.

The proofs of the other implications of the four equivalences are simple. To show that *(2)* implies *(3)*, take $\boldsymbol{\mu} := \boldsymbol{\eta} \uplus \boldsymbol{\vartheta}$, and *(3)* implies *(1)* is the special case when $\boldsymbol{\eta}$ and $\boldsymbol{\vartheta}$ are atoms.

Further, for *(2)* implies *(4)* take an optimal multi-expansion $\boldsymbol{\mu}$ (which exists, since z admits at least one multi-expansion). We get a w-NAF-expansion $\boldsymbol{\xi}$ with $\mathsf{weight}(\boldsymbol{\xi}) \le \mathsf{weight}(\boldsymbol{\mu})$. Since $\boldsymbol{\mu}$ was optimal, equality is obtained in the previous inequality, and therefore $\boldsymbol{\xi}$ is optimal, too. The converse, *(4)* implies *(2)*, follows using $z = \mathsf{value}(\boldsymbol{\mu})$ and the property that optimal expansions minimise the weight. □

Let X and Y be subsets in an additively written semigroup. Then we write

$$X + Y := \{x + y \,:\, x \in X, y \in Y\}\,,$$

see, for example, Hungerford [53]. We use that notion from now on.

Proposition 2.2.3. *Let $(\mathcal{A}, \Phi, \mathcal{D})$ be a pre-number system with*

$$\bigcap_{m\in\mathbb{N}_0} \Phi^m(\mathcal{A}) = \{0\}\,,$$

and let w be a positive integer. We have the following sufficient condition: Suppose we have sets U and S such that $\mathcal{D} \subseteq U$, $-\mathcal{D} \subseteq U$, $U \subseteq \Phi(U)$ and all elements in S are atoms. If $\mathcal{D}$ contains a representative for each residue class modulo $\Phi^w(\mathcal{A})$ which is not contained in $\Phi(\mathcal{A})$ and

$$\big(\Phi^{w-1}(U) + U + U\big) \cap \Phi^w(\mathcal{A}) \subseteq S \cup \{0\}\,, \tag{2.2.1}$$

then the digit set $\mathcal{D}$ is w-subadditive.

Sometimes it is more convenient to use (2.2.1) of this proposition instead of the definition of w-subadditive. For example, in Section 2.3 all digits lie in an interval U and all non-zero integers in that interval have a w-NAF expansion with weight 1. The same technique is used in the optimality result of Section 2.6.

Proof of Proposition 2.2.3. Let $(c, 0)$ and (d, n) be atoms with $n \in \{0, \dots, w-1\}$ and consider $y = \mathsf{value}((c, 0) \uplus (d, n))$. If $y = 0$, we have nothing to do, so we can assume $y \neq 0$. First suppose $y \notin \Phi(\mathcal{A})$. Because of our assumptions on $\mathcal{D}$ there is a digit a such that

$$z := \Phi^n(d) + c - a \in \Phi^w(\mathcal{A})\,.$$

If z is not zero, then, using our sufficient condition, there is an atom (b, m) with value z, and we have $m \geq w$. The w-NAF-expansion $b0^{m-1}a$ does what we want.

Now suppose $y \in \Phi^k(\mathcal{A})$ with a positive integer k, which is chosen maximally. That case can only happen when $n = 0$. Since $y \neq 0$ and our assumptions on $\mathcal{D}$ there is a w-NAF-expansion of y with an atom (a, k) as least significant digit. If $k \in \{0, \ldots, w-1\}$, then

$$z := d + c - \Phi^k(a) \in \Phi^{w+k}(\mathcal{A})\,.$$

If z is non-zero it is a value of an atom (b, m), $m \geq w + k$, because of (2.2.1), and we obtain a w-NAF-expansion of y with atoms (b, m) and (a, k). If $k \geq w$, then

$$z := d + c \in \Phi^{w+k}(\mathcal{A})$$

and z is the value of an atom (b, m) by (2.2.1). We get a w-NAF-expansion $b0^m$. □

Sometimes the w-subadditivity-condition is a bit too strong, so we do net get optimal w-NAFs. In that case one can check whether $(w-1)$-NAFs are optimal. This is stated in the following remark, where the w-subadditive-condition is weakened.

Remark 2.2.4. Suppose that we have the same setting as in Theorem 2.2.2. We call the digit set *w-weak-subadditive* if the sum of the values of two atoms (c, m) and (d, n) with $|m - n| \neq w - 1$ has a w-NAF-expansion with weight at most 2.

We get the following result: If the digit set $\mathcal{D}$ is w-weak-subadditive, then each element of $\mathcal{A}$, which has at least one multi-expansion, has an optimal $(w-1)$-NAF-expansion. The proof is similar to the proof of Theorem 2.2.2, except that a "rewriting" only happens when we have a $(w-1)$-NAF-violation.

2.3 Optimality for Integer Bases

In this section we give a first application of the abstract optimality theorem of the previous section. We reprove the optimality of the w-NAFs with a minimal norm digit set and base 2. But the result is more general: We prove optimality for all integer bases (with absolute value at least 2). This demonstrates one basic idea how to check whether a digit set is w-subadditive or not.

Let b be an integer with $|b| \geq 2$ and w be an integer with $w \geq 2$. Consider the non-redundant w-NAF number system

$$(\mathbb{Z}, z \mapsto bz, \mathcal{D})$$

where the digit set $\mathcal{D}$ consists of 0 and all integers with absolute value strictly smaller than $\frac{1}{2}|b|^w$ and not divisible by b. We mentioned the special case base 2 of that number system in Example 2.1.7. See also Reitwiesner [86] and Solinas [95].

The following optimality result can be shown. For proofs of the base 2 setting cf. Reitwiesner [86], Jedwab and Mitchell [56], Gordon [43], Avanzi [5], Muir and Stinson [79], and Phillips and Burgess [85].

Theorem 2.3.1. *With the setting above, the w-NAF-expansion for each integer is optimal.*

Proof. We show that the digit set $\mathcal{D}$ is w-subadditive by verifying the sufficient condition of Proposition 2.2.3. Then optimality follows from Theorem 2.2.2. First, note that the w-NAF-expansion of each integer with absolute value at most $\frac{1}{2}|b|^{w-1}$ has weight at most 1, because either the integer is already a digit, or one can divide by a power of b to get a digit. Further, we have $\mathcal{D} = -\mathcal{D}$. Set $U = \left[-\frac{1}{2}|b|^w, \frac{1}{2}|b|^w\right]$ and $S = b^w(U \cap \mathbb{Z} \setminus \{0\})$. We have to show

$$\left(b^{w-1}U + U + U\right) \cap b^w\mathbb{Z} \subseteq S \cup \{0\} = b^wU \cap b^w\mathbb{Z}.$$

If we can show

$$b^{-w}\left(b^{w-1}U + U + U\right) \subseteq \left[-\tfrac{1}{2}|b|^w, \tfrac{1}{2}|b|^w\right],$$

the inclusion above follows by multiplying with b^w and taking the intersection with $b^w\mathbb{Z}$.

So let

$$b^wz = b^{w-1}c + a + d$$

for some digits a, c and d. A digit has absolute value less than $\frac{1}{2}|b|^w$, so

$$|z| < |b|^{-w}\left(|b|^{w-1} + 2\right)\tfrac{1}{2}|b|^w = \left(|b|^{-1} + 2|b|^{-w}\right)\tfrac{1}{2}|b|^w \leq \tfrac{1}{2}|b|^w,$$

where we also used the assumptions $|b| \geq 2$ and $w \geq 2$. Thus, the desired inclusion is shown. □

2.4 Voronoi Cells

We first start to define Voronoi cells. Let $\tau \in \mathbb{C}$ be an algebraic integer that is imaginary quadratic, i.e. τ is solution of an equation $\tau^2 - p\tau + q = 0$ with $p, q \in \mathbb{Z}$ and such that $q - p^2/4 > 0$.

Definition 2.4.1 (Voronoi Cell)**.** We set

$$V := \{z \in \mathbb{C} : \forall y \in \mathbb{Z}[\tau]\colon |z| \leq |z - y|\}$$

and call it the *Voronoi cell for* 0 corresponding to the set $\mathbb{Z}[\tau]$. Let $u \in \mathbb{Z}[\tau]$. We define the *Voronoi cell for* u as

$$V_u := u + V = \{u + z : z \in V\} = \{z \in \mathbb{C} : \forall y \in \mathbb{Z}[\tau]\colon |z - u| \leq |z - y|\}.$$

The point u is called *centre of the Voronoi cell* or *lattice point corresponding to the Voronoi cell.*

An example of a Voronoi cell in a lattice $\mathbb{Z}[\tau]$ is shown in Figure 2.4.1. Two neighbouring Voronoi cells have at most a subset of their boundary in common. This can be a problem, when we tile the plane with Voronoi cells and want that each point is in exactly one cell. To fix this problem we define a restricted version of V. This is very similar to the construction used in Avanzi, Heuberger and Prodinger [4] and in Heuberger and Krenn [49].

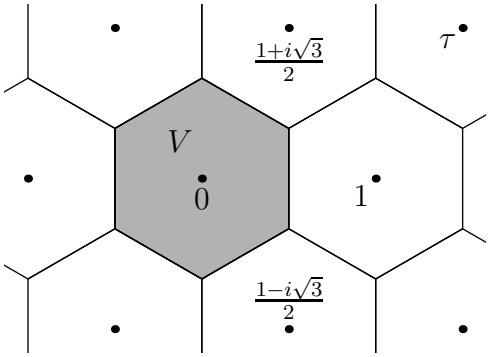


Figure 2.4.1: Voronoi cell V for 0 corresponding to the set $\mathbb{Z}[\tau]$ with $\tau = \frac{3}{2} + \frac{i}{2}\sqrt{3}$.

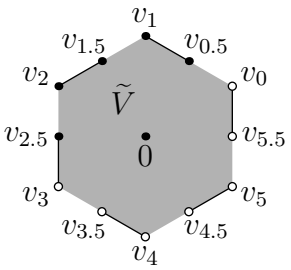


Figure 2.4.2: Restricted Voronoi cell $\widetilde{V}$ for 0 corresponding to the set $\mathbb{Z}[\tau]$ with $\tau = \frac{3}{2} + \frac{i}{2}\sqrt{3}$.

Definition 2.4.2 (Restricted Voronoi Cell)**.** Let V_u be a Voronoi cell with its centre u as above. Let $v_0, \ldots, v_{m-1}$ with appropriate $m \in \mathbb{N}$ be the vertices of V_u. We denote the midpoint of the line segment from v_k to v_{k+1} by $v_{k+1/2}$, and we use the convention that the indices are meant modulo m.

The *restricted Voronoi cell* $\widetilde{V}_u$ consists of

- the interior of V_u,
- the line segments from $v_{k+1/2}$ (excluded) to v_{k+1} (excluded) for all k,
- the points $v_{k+1/2}$ for $k \in \left\{0, \ldots, \left\lfloor \frac{m}{2} \right\rfloor - 1\right\}$, and
- the points v_k for $k \in \left\{1, \ldots, \left\lfloor \frac{m}{3} \right\rfloor\right\}$.

Again we set $\widetilde{V} := \widetilde{V}_0$.

In Figure 2.4.2 the restricted Voronoi cell of 0 is shown for $\tau = \frac{3}{2} + \frac{i}{2}\sqrt{3}$. The second condition in the definition is used because it benefits symmetries. The third condition is just to make the midpoints unique. Obviously, other rules[2] could have been used to define the restricted Voronoi cell.

[2]The rule has to make sure that the complex plane can be covered entirely and with no overlaps by restricted Voronoi cells, i.e. the condition $\mathbb{C} = \biguplus_{z \in \mathbb{Z}[\tau]} \widetilde{V}_z$ has to be fulfilled.

The statements (including proofs) of the following lemma can be found in Heuberger and Krenn [49]. We use the notation $\mathcal{B}(z,r)$ for an open ball with centre z and radius r and $\overline{\mathcal{B}}(z,r)$ for a closed ball.

Lemma 2.4.3 (Properties of Voronoi Cells)**.** *We have the following properties:*

(a) The vertices of V are given explicitly by

$$\begin{aligned}
v_0 &= 1/2 + \frac{i}{2\operatorname{Im}(\tau)}\left(\operatorname{Im}(\tau)^2 + \{\operatorname{Re}(\tau)\}^2 - \{\operatorname{Re}(\tau)\}\right),\\
v_1 &= \{\operatorname{Re}(\tau)\} - \frac{1}{2} + \frac{i}{2\operatorname{Im}(\tau)}\left(\operatorname{Im}(\tau)^2 - \{\operatorname{Re}(\tau)\}^2 + \{\operatorname{Re}(\tau)\}\right),\\
v_2 &= -1/2 + \frac{i}{2\operatorname{Im}(\tau)}\left(\operatorname{Im}(\tau)^2 + \{\operatorname{Re}(\tau)\}^2 - \{\operatorname{Re}(\tau)\}\right) = v_0 - 1,\\
v_3 &= -v_0,\\
v_4 &= -v_1
\end{aligned}$$

and

$$v_5 = -v_2.$$

All vertices have the same absolute value. If $\operatorname{Re}(\tau) \in \mathbb{Z}$, then $v_1 = v_2$ and $v_4 = v_5$, i.e. the hexagon degenerates to a rectangle.

(b) The Voronoi cell V is convex.

(c) We get $\overline{\mathcal{B}}\left(0, \frac{1}{2}\right) \subseteq V$.

(d) The inclusion $\tau^{-1}V \subseteq V$ holds.

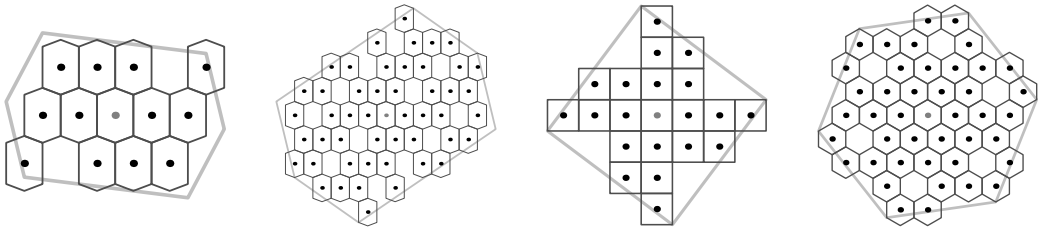
2.5 Digit Sets for Imaginary Quadratic Bases

In this section we assume that $\tau \in \mathbb{C}$ is an imaginary quadratic algebraic integer, i.e. τ is solution of an equation $\tau^2 - p\tau + q = 0$ with $p, q \in \mathbb{Z}$ and such that $q - p^2/4 > 0$. By V we denote the Voronoi cell of 0 of the lattice $\mathbb{Z}[\tau]$, by $\widetilde{V}$ the corresponding restricted Voronoi cell, cf. Section 2.4.

We consider w-NAF number systems

$$(\mathbb{Z}[\tau], z \mapsto \tau z, \mathcal{D}),$$

where the digit set $\mathcal{D}$ is the so called "minimal norm representatives digit set". The following definition specifies that digit set, cf. Solinas [94, 95], Blake, Murty and Xu [16] or Heuberger and Krenn [49]. It is used throughout this chapter, whenever we have the setting (imaginary quadratic base) mentioned above.



(a) Digit set for $\tau = \frac{3}{2} + \frac{i}{2}\sqrt{7}$ and $w = 2$.

(b) Digit set for $\tau = \frac{3}{2} + \frac{i}{2}\sqrt{7}$ and $w = 3$.

(c) Digit set for $\tau = 2 + i$ and $w = 2$.

(d) Digit set for $\tau = \frac{5}{2} + \frac{i}{2}\sqrt{3}$ and $w = 2$.

Figure 2.5.1: Minimal norm representatives digit sets modulo τ^w. For each digit η, the corresponding Voronoi cell V_η is drawn. The large scaled Voronoi cell is $\tau^w V$.

Definition 2.5.1 (Minimal Norm Representatives Digit Set)**.** Let w be an integer with $w \geq 2$ and $\mathcal{D} \subseteq \mathbb{Z}[\tau]$ consist of 0 and exactly one representative of each residue class of $\mathbb{Z}[\tau]$ modulo τ^w that is not divisible by τ. If all such representatives $\eta \in \mathcal{D}$ fulfil $\eta \in \tau^w \widetilde{V}$, then $\mathcal{D}$ is called the *minimal norm representatives digit set modulo* τ^w.

The previous definition uses the restricted Voronoi cell $\widetilde{V}$ for the point 0, see Definition 2.4.2, to choose a representative with minimal norm. Note that by construction of $\widetilde{V}$, there is only one such choice for the digit set. Some examples of such digit sets are shown in Figures 2.5.1, 2.7.1, 2.8.1 and 2.9.1.

Remark 2.5.2. The definition of a minimal norm representative digit set, Definition 2.5.1, depends on the definition of the restricted Voronoi cell $\widetilde{V}$, Definition 2.4.2. There we had some freedom in choosing which part of the boundary is included in $\widetilde{V}$, cf. the remarks after Definition 2.4.2. We point out that all results given here for imaginary quadratic bases are valid for any admissible configuration of the restricted Voronoi cell, although only the case corresponding to Definition 2.4.2 will be presented.

Using a minimal norm representatives digit set, each element of $\mathbb{Z}[\tau]$ corresponds to a unique w-NAF, i.e. the pre-number system given at the beginning of this section is indeed a w-NAF number system. This is stated in the following theorem, which can be found in Heuberger and Krenn [49].

Theorem 2.5.3 (Existence and Uniqueness Theorem)**.** *Let w be an integer with $w \geq 2$. Then the pre-number system*

$$(\mathbb{Z}[\tau], z \mapsto \tau z, \mathcal{D}),$$

where $\mathcal{D}$ is the minimal norm representatives digit set modulo τ^w, is a non-redundant w-NAF number system, i.e. each lattice point $z \in \mathbb{Z}[\tau]$ has a unique w-NAF-expansion $\boldsymbol{\eta} \in \mathcal{D}^{\mathbb{N}_0}$ with $z = \mathsf{value}(\boldsymbol{\eta})$.

2.6 Optimality for Imaginary Quadratic Bases

In this section we assume that $\tau \in \mathbb{C}$ is an imaginary quadratic algebraic integer, i.e. τ is solution of an equation $\tau^2 - p\tau + q = 0$ with $p, q \in \mathbb{Z}$ and such that $q - p^2/4 > 0$. Further let w be an integer with $w \geq 2$ and let

$$(\mathbb{Z}[\tau], z \mapsto \tau z, \mathcal{D})$$

be the non-redundant w-NAF number system with minimal norm representatives digit set modulo τ^w, cf. Section 2.5.

Our main question in this section, as well as for the remaining part of this article, is the following: For which bases and which w is the width-w non-adjacent form optimal? To answer this, we use the result from Section 2.2. If we can show that the digit set $\mathcal{D}$ is w-subadditive, then optimality follows. This is done in the lemma below. The result will then be formulated in Corollary 2.6.2, which, eventually, contains the optimality result for our mentioned configuration.

Lemma 2.6.1. *Suppose that one of the following conditions hold:*

(i) $w \geq 4$ *and* $|p| \geq 3$,

(ii) $w = 3$ *and* $|p| \geq 5$,

(iii) $w = 3$, $|p| = 4$ *and* $5 \leq q \leq 9$,

(iv) $w = 2$, p *even, and*

$$\left(\frac{1}{\sqrt{q}} + \frac{2}{q}\right)^2 \left(q - \frac{p^2}{4} + 1\right) < 1$$

or equivalently

$$|p| > 2\sqrt{q + 1 - \frac{q^2}{\left(2 + \sqrt{q}\right)^2}},$$

(v) $w = 2$, p *odd and*

$$\left(\frac{1}{\sqrt{q}} + \frac{2}{q}\right)^2 \left(q - \frac{p^2}{4} + \frac{1}{4}\right)^2 \left(q - \frac{p^2}{4}\right)^{-1} < 1.$$

Then the digit set $\mathcal{D}$ is w-subadditive.

The conditions (iv) and (v) of Lemma 2.6.1, i.e. the case $w = 2$, are illustrated graphically in Figure 2.6.1.

Proof. We denote the interior of V by $\operatorname{int}(V)$. If

$$\tau^{w-1}V + V + V \subseteq \tau^w \operatorname{int}(V)$$

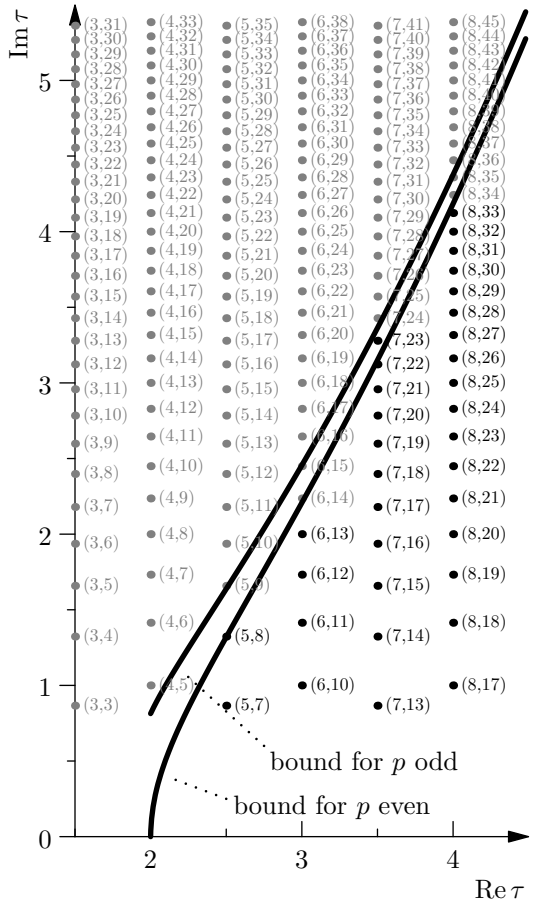


Figure 2.6.1: Bounds for the optimality of 2-NAFs. The two curves correspond to the conditions (iv) and (v) of Lemma 2.6.1. A dot corresponds to a valid τ. If the dot is black, then the 2-NAFs of that τ are optimal, gray means not decidable with this method. Each dot is labelled with $(|p|, q)$.

holds, then the digit set $\mathcal{D}$ is w-subadditive since $\mathcal{D} \subseteq \tau^w V$, $-\mathcal{D} \subseteq \tau^w V$, $V \subseteq \tau V$ and $z \in \tau^w \operatorname{int}(V) \cap \mathbb{Z}[\tau]$ implies that there is an integer $\ell \geq 0$ with $z \in \tau^\ell \mathcal{D}$. The sufficient condition of Proposition 2.2.3 was used with $U = \tau^w V$ and $S = \tau^w \operatorname{int}(V) \setminus \{0\}$.

Since V is convex, it is sufficient to show that

$$\tau^{w-1} V + 2V \subseteq \tau^w \operatorname{int}(V)\,.$$

This will be done by showing

$$\left(|\tau|^{-1} + 2\,|\tau|^{-w}\right) |V| < \tfrac{1}{2},$$

where $|V|$ denotes the radius of the smallest closed disc with centre 0 containing V. By setting

$$T(p,q,w) := 2\left(|\tau|^{-1} + 2\,|\tau|^{-w}\right) |V|\,,$$

we have to show that

$$T(p,q,w) < 1.$$

Note that $T(p,q,w) > 0$, so it is sufficient to show

$$T^2(p,q,w) < 1.$$

For each of the different conditions given, we will check that the inequality holds for special values of p, q and w and then use a monotonicity argument to get the result for other values of p, q and w. In the following we distinguish between even and odd p.

Let first p be even, first. Then $\frac{1}{2} + \frac{i}{2}\operatorname{Im}(\tau)$ is a vertex of the Voronoi cell V. This means $|V| = \frac{1}{2}\sqrt{1 + q - p^2/4}$. Inserting that and $|\tau| = \sqrt{q}$ in the asserted inequality yields

$$T^2(p,q,w) = \left(\frac{1}{\sqrt{q}} + 2q^{-w/2}\right)^2 \left(1 + q - \frac{p^2}{4}\right) < 1.$$

It is easy to see that the left hand side of this inequality is monotonically decreasing in $|p|$ (as long as the condition $q > p^2/4$ is fulfilled) and monotonically decreasing in w. We assume $p \geq 0$.

If we set $p = 4$ and $w = 4$, we get

$$T^2(4,q,4) = -\frac{12}{q^4} + \frac{4}{q^3} - \frac{12}{q^{5/2}} + \frac{4}{q^{3/2}} - \frac{3}{q} + 1,$$

which is strictly monotonically increasing for $q \geq 5$. Further we get

$$\lim_{q\to\infty} T^2(4,q,4) = 1.$$

This means $T^2(4,q,4) < 1$ for all $q \geq 5$. Since $p \geq 4$ implies $q \geq 5$ and because of the monotonicity mentioned before, the case (i) for the even p is completed.

If we set $p = 6$ and $w = 3$, we get

$$T^2(6,q,3) = -\frac{32}{q^3} - \frac{28}{q^2} - \frac{4}{q} + 1,$$

which is obviously less than 1. Therefore, again by monotonicity, the case (ii) is done for the even p.

If we set $p = 4$ and $w = 3$, we obtain

$$T^2(4, q, 3) = -\frac{12}{q^3} - \frac{8}{q^2} + \frac{1}{q} + 1,$$

which is monotonically increasing for $5 \geq q \geq 18$. Further we get

$$T^2(4, 9, 3) = \frac{242}{243} < 1$$

and $T^2(4, 10, 3) > 1$. This means $T^2(4, q, 3) < 1$ for all q with $5 \leq q \leq 9$. So case (iii) is completed.

The condition given in (iv) is exactly

$$T^2(p, q, 2) < 1$$

for even p, so the result follows immediately.

Now, let p be odd. Then $\frac{i}{2\,\mathrm{Im}(\tau)}\left(\mathrm{Im}(\tau)^2 + \frac{1}{4}\right)$ is a vertex of the Voronoi cell V. This means

$$|V| = \frac{1}{2}\left(q - \frac{p^2}{4}\right)^{-1/2}\left(q - \frac{p^2}{4} + \frac{1}{4}\right).$$

Inserting that in the asserted inequality yields

$$T^2(p, q, w) = \left(q - \frac{p^2}{4}\right)^{-1}\left(q - \frac{p^2}{4} + \frac{1}{4}\right)^2\left(2q^{-w/2} + q^{-1/2}\right)^2 < 1.$$

Again, it is easy to verify that the left hand side of this inequality is monotonically decreasing in p (as long as the condition $q \geq p^2/4 + 1/4$ is fulfilled) and monotonically decreasing in w. We assume $p \geq 0$.

If we set $p = 3$ and $w = 4$, we get

$$T^2(3, q, 4) = \frac{4(q-2)^2\left(q^{3/2} + 2\right)^2}{q^4(4q - 9)}$$

which is strictly monotonically increasing for $q \geq 3$. Further we get

$$\lim_{q\to\infty} T^2(3, q, 4) = 1.$$

This means $T^2(3, q, 4) < 1$ for all $q \geq 3$. Since $p \geq 3$ implies $q \geq 3$ and because of the monotonicity mentioned before, the case (i) for the odd p is finished.

If we set $p = 5$ and $w = 3$, we get

$$T^2(5, q, 3) = \frac{4(q-6)^2(q+2)^2}{q^3(4q - 25)}$$

which is strictly monotonically increasing for $q \geq 7$. Further we get

$$\lim_{q\to\infty} T^2(5, q, 3) = 1.$$

This means $0 < T^2(5, q, 3) < 1$ for all $q \geq 7$. As $p \geq 5$ implies $q \geq 7$, using monotonicity again, the case (ii) is done for the odd p.

The condition given in (v) is exactly

$$T^2(p, q, 2) < 1$$

for odd p, so the result follows immediately.

Since we have now analysed all the conditions, the proof is finished. □

Now we can prove the following optimality corollary, which is a consequence of Theorem 2.2.2.

Corollary 2.6.2. *Suppose that one of the conditions (i) to (v) of Lemma 2.6.1 holds. Then the width-w non-adjacent form expansion for each element of $\mathbb{Z}[\tau]$ is optimal.*

Proof. Lemma 2.6.1 implies that the digit set $\mathcal{D}$ is w-subadditive, therefore Theorem 2.2.2 can be used directly to get the desired result. □

Remark 2.6.3. We have the following weaker optimality result. Let p, q and w be integers with $|p| \geq p_0$, $q \geq q_0$ and $w \geq w_0$ for a $(p_0, q_o, w_0) \in Y$, where

$$Y = \{(0,10,2),(0,5,3),(0,4,4),(0,3,5),(0,2,10),\\ (1,2,8),(2,3,4),(2,2,7),(3,7,2),(3,3,3),(4,5,2)\}.$$

Then we can show that the minimal norm representatives digit set modulo τ^w coming from a τ with (p, q) is w-weak-subadditive, and therefore, by Remark 2.2.4, we obtain optimality of a $(w-1)$-NAF of each element of $\mathbb{Z}[\tau]$. The results are visualised graphically in Figure 2.6.2.

To show that the digit set is w-weak-subadditive we proceed in the same way as in the proof of Lemma 2.6.1. We have to show the condition

$$T'(p, q, w) < 1$$

where

$$T'(p, q, w) = 2\left(|\tau|^{-2} + 2\,|\tau|^{-w}\right)|V|$$

with $|\tau| = \sqrt{q}$. When p is even, we have

$$|V| = \frac{1}{2}\sqrt{1 + q - \frac{p^2}{4}},$$

and when p is odd, we have

$$|V| = \frac{1}{2}\left(q - \frac{p^2}{4}\right)^{-1/2}\left(q - \frac{p^2}{4} + \frac{1}{4}\right).$$

Using monotonicity arguments as in the proof of Lemma 2.6.1 yields the list Y of "critical points".

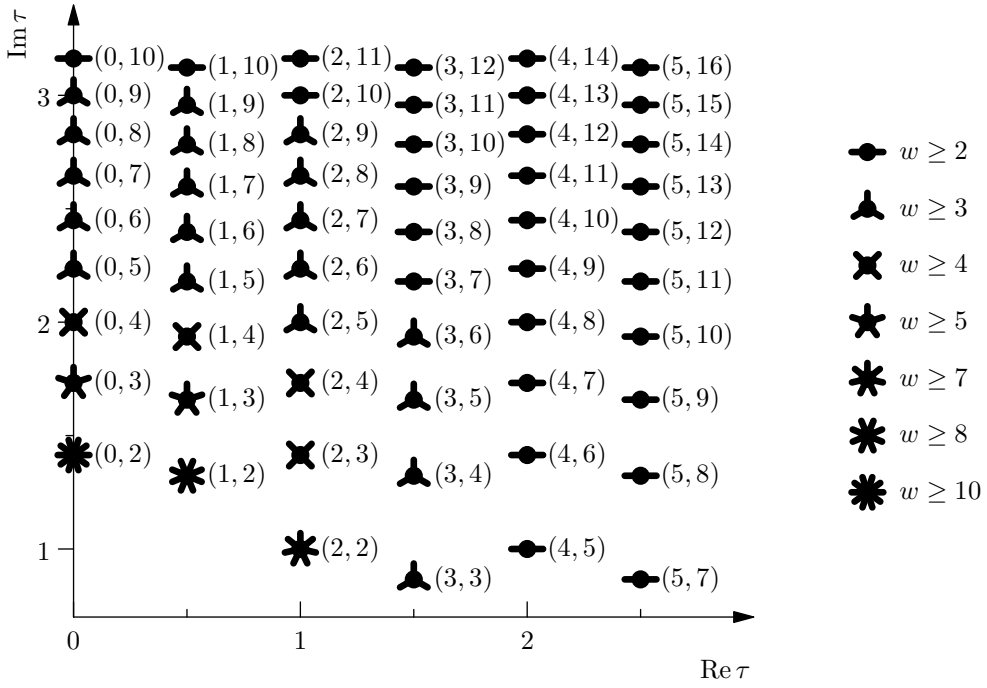
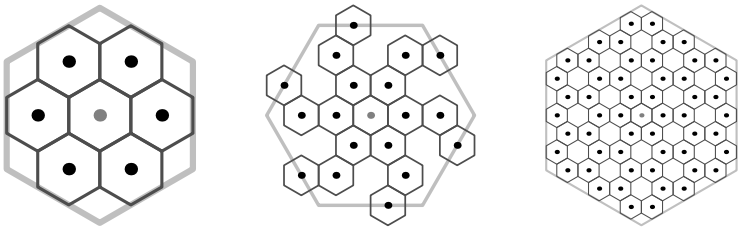


Figure 2.6.2: Optimality of a $(w-1)$-NAF-expansion with a digit set used for w-NAFs. Each symbol is labelled with $(|p|, q)$ and represents the minimal w for which there is an optimal $(w-1)$-NAF-expansion of each element of $\mathbb{Z}[\tau]$.



(a) Digit set for $\tau = \frac{3}{2} + \frac{i}{2}\sqrt{3}$ and $w = 2$.

(b) Digit set for $\tau = \frac{3}{2} + \frac{i}{2}\sqrt{3}$ and $w = 3$.

(c) Digit set for $\tau = \frac{3}{2} + \frac{i}{2}\sqrt{3}$ and $w = 4$.

Figure 2.7.1: Minimal norm representatives digit sets modulo τ^w. For each digit η, the corresponding Voronoi cell V_η is drawn. The large scaled Voronoi cell is $\tau^w V$.

2.7 The p-is-3-q-is-3-Case

One important case can be proved by using the Optimality Theorem of Section 2.2, too, namely when τ comes from a Koblitz curve in characteristic 3. We specialise the setting of Section 2.6 to $p = 3\mu$ with $\mu \in \{-1, 1\}$ and $q = 3$. We continue looking at w-NAF-number systems with minimal norm representative digit set modulo τ^w with $w \geq 2$. Some examples of those digit sets are shown in Figure 2.7.1. We have the following optimality result.

Corollary 2.7.1. *With the setting above, the width-w non-adjacent form expansion for each element of $\mathbb{Z}[\tau]$ is optimal.*

Proof. Using the statement of Lemma 2.6.1 and Theorem 2.2.2 yields the optimality for all $w \geq 4$.

Let $w = 2$. Then our minimal norm representatives digit set is

$$\mathcal{D} = \{0\} \cup \bigcup_{0 \leq k < 6} \zeta^k \{1\},$$

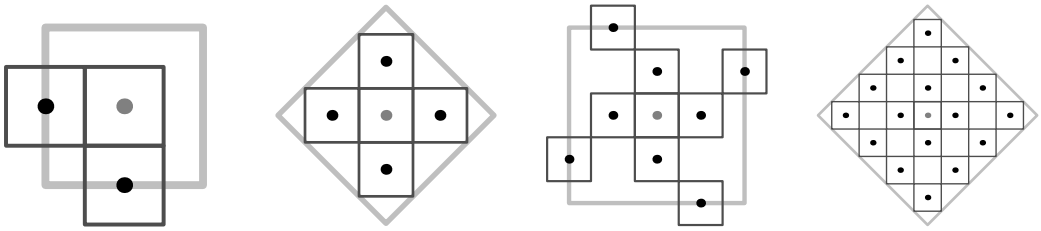
where ζ is a primitive sixth root of unity, see Avanzi, Heuberger and Prodinger [4]. Therefore we obtain $|\mathcal{D}| = 1$ and $\mathcal{D} = -\mathcal{D}$. For $k \in \{0, 1\}$ we get

$$\tau^k \mathcal{D} + \mathcal{D} + \mathcal{D} \subseteq \overline{\mathcal{B}}\left(0, \sqrt{3} + 2\right) \subseteq \sqrt{3}^4 \, \mathcal{B}\left(0, \frac{1}{2}\right) \subseteq \tau^{2w} \operatorname{int}(V),$$

so the digit set $\mathcal{D}$ is w-subadditive by the same arguments as in the beginning of the proof of Lemma 2.6.1, and we can apply the Optimality Theorem to get the desired result.

Let $w = 3$. Then our minimal norm representatives digit set is

$$\mathcal{D} = \{0\} \cup \bigcup_{0 \leq k < 6} \zeta^k \{1, 2, 4 - \mu\tau\},$$



(a) Digit set for $\tau = 1+i$ and $w = 2$. (b) Digit set for $\tau = 1+i$ and $w = 3$. (c) Digit set for $\tau = 1+i$ and $w = 4$. (d) Digit set for $\tau = 1+i$ and $w = 5$.

Figure 2.8.1: Minimal norm representatives digit sets modulo τ^w. For each digit η, the corresponding Voronoi cell V_η is drawn. The large scaled Voronoi cell is $\tau^w V$.

where ζ is again a primitive sixth root of unity, again [4]. Therefore we obtain $|\mathcal{D}| = |4 - \mu\tau| = \sqrt{7}$ and again $\mathcal{D} = -\mathcal{D}$. For $k \in \{0, 1, 2\}$, we get $|\tau|^k |\mathcal{D}| \leq 3\sqrt{7}$, because $|\tau| = \sqrt{3}$. Therefore

$$\tau^k \mathcal{D} + \mathcal{D} + \mathcal{D} \subseteq \overline{\mathcal{B}}\Big(0, 5\sqrt{7}\Big) \subseteq \sqrt{3}^6 \, \mathcal{B}\left(0, \frac{1}{2}\right) \subseteq \tau^{2w} \operatorname{int}(V)\,,$$

so we can use Theorem 2.2.2 again to get the optimality. □

2.8 The p-is-2-q-is-2-Case

In this section we look at another special base τ. We assume that $p \in \{-2, 2\}$ and $q = 2$. Again, we continue looking at w-NAF-number systems with minimal norm representative digit set modulo τ^w with $w \geq 2$. Some examples of those digit sets are shown in Figure 2.8.1.

For all possible τ of this section, the corresponding Voronoi cell can be written explicitly as

$$V = \mathsf{polygon}\big(\big\{\tfrac{1}{2}(1+i), \tfrac{1}{2}(-1+i), \tfrac{1}{2}(-1-i), \tfrac{1}{2}(1-i)\big\}\big)\,.$$

Remark that V is an axis-parallel square and that we have

$$\tau V = \mathsf{polygon}\big(\big\{i^j \,:\, j \in \{0, 1, 2, 3\}\big\}\big)\,.$$

In this section we will prove that the w-NAFs are optimal if and only if w is odd. The first part, optimality for odd w, is written down as the theorem below. The non-optimality part for even w can be found as Proposition 2.8.3.

Theorem 2.8.1. *Let w be an odd integer with $w \geq 3$, and let $z \in \mathbb{Z}[\tau]$. Then the width-w non-adjacent form expansion of z is optimal.*

Remark 2.8.2. Let w be an odd integer with $w \geq 3$. Let $z \in \tau^w V \cap \mathbb{Z}[\tau]$, then z can be represented as a w-NAF expansion with weight at most 1. To see this, consider

the boundary of $\tau^w V$. Its vertices are $2^{(w-1)/2} i^m$ for $m \in \{0,1,2,3\}$. All elements of $\partial(\tau^w V) \cap \mathbb{Z}[\tau]$ can be written as $2^{(w-1)/2} i^m + k(1+i)i^n$ for some integers k, m and n. Further, all those elements are divisible by τ. Therefore each digit lies in the interior of $\tau^w V$, and for each $z \in \tau^w V \cap \mathbb{Z}[\tau]$ there is an integer $\ell \geq 0$ such that $\tau^{-\ell} z \in \mathcal{D}$, because $\tau^{-1} V \subseteq V$ and $|\tau| > 1$.

Proof of Theorem 2.8.1. We prove that the digit set $\mathcal{D}$ is w-subadditive. Hence, optimality follows using Theorem 2.2.2. Using the remark above, $\mathcal{D} = -\mathcal{D}$ and the ideas of Proposition 2.2.3, it is sufficient to show

$$\tau^{-w} \left(\tau^k \mathcal{D} + \mathcal{D} + \mathcal{D} \right) \cap \mathbb{Z}[\tau] \subseteq \tau^w V$$

for $k \in \{0, \ldots, w-1\}$.

Let $k = w - 1$. We show that

$$\left(\mathcal{D} + \tau^{-(w-1)} (\mathcal{D} + \mathcal{D}) \right) \cap \tau \mathbb{Z}[\tau] \subseteq \tau^{w+1} V. \tag{2.8.1}$$

So let $y = b + a$ be an element of the left hand side of (2.8.1) with $b \in \mathcal{D}$ and $a \in \tau^{-(w-1)} (\mathcal{D} + \mathcal{D})$. We can assume $y \neq 0$. Since $y \in \mathbb{Z}[\tau]$ and $\mathcal{D} \subseteq \mathbb{Z}[\tau]$, we have $a \in \mathbb{Z}[\tau]$. Since $\mathcal{D} \subseteq \tau^w V$, we obtain

$$\tau^{-(w-1)} (\mathcal{D} + \mathcal{D}) \subseteq 2\tau V.$$

The case $b = 0$ is easy, because $2\tau V = \tau^3 V \subseteq \tau^w V$. So we can assume $b \neq 0$. This means $\tau \nmid b$. Since $\tau \mid y$, we have $\tau \nmid a$. The set $2\tau V \cap \mathbb{Z}[\tau]$ consists exactly of 0, i^m, $2i^m$ and τi^m for $m \in \{0,1,2,3\}$. The only elements in that set not divisible by τ are the i^m. Therefore $a = i^m$ for some m. The digit b is in the interior of $\tau^w V$, thus $y = b + a$ is in $\tau^w V \subseteq \tau^{w+1} V$.

Now let $k \in \{0, \ldots, w-2\}$. If $w \geq 5$, then

$$\tau^{-w} \left(\tau^k \mathcal{D} + \mathcal{D} + \mathcal{D} \right) \subseteq \tau^{w-2} V + 2V,$$

using $\mathcal{D} \subseteq \tau^w V$ and properties of the Voronoi cell V. Consider the two squares $\tau^{w-2} V$ and $\tau^w V = 2\tau^{w-2} V$. The distance between the boundaries of them is at least $\frac{1}{2} |\tau|^{w-2}$, which is at least $\sqrt{2}$. Since $2V$ is contained in a disc with radius $\sqrt{2}$, we obtain $\tau^{w-2} V + 2V \subseteq \tau^w V$.

We are left with the case $w = 3$ and $k \in \{0,1\}$. There the digit set $\mathcal{D}$ consists of 0 and i^m for $m \in \{0,1,2,3\}$. Therefore we have $\mathcal{D} \subseteq \tau V$ (instead of $\mathcal{D} \subseteq \tau^3 V$). By the same arguments as in the previous paragraph we get

$$\tau^{-3} \left(\tau^k \mathcal{D} + \mathcal{D} + \mathcal{D} \right) \subseteq \tfrac{1}{2} (\tau V + 2V) \subseteq \tau^3 V,$$

so the proof is complete. □

The next result is the non-optimality result for even w.

Proposition 2.8.3. *Let w be an even integer with $w \geq 2$. Then there is an element of $\mathbb{Z}[\tau]$ whose w-NAF-expansion is non-optimal.*

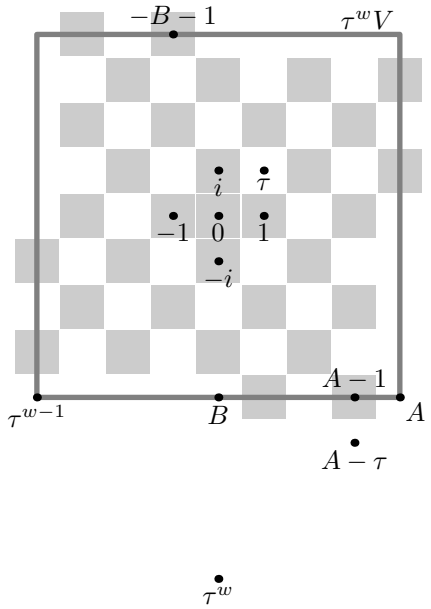


Figure 2.8.2: The w-is-even situation. The figure shows the configuration $p = 2$, $q = 2$, $w = 6$, $s = 1$. A polygon filled grey represents a digit, a dot represents a point of interest in Lemma 2.8.4.

Again, some examples of the digit sets used are shown in Figure 2.8.1. The proof of the proposition is split up: Lemma 2.8.4 handles the general case for even $w \geq 4$ and Lemma 2.8.5 gives a counter-example (to optimality) for $w = 2$.

For the remaining section—it contains the proof of Proposition 2.8.3—we will assume $\tau = 1 + i$. All other cases are analogous.

Lemma 2.8.4. *Let the assumptions of Proposition 2.8.3 hold and suppose $w \geq 4$. Define $A := |\tau|^w \frac{1}{2}(1-i)$ and $B := \frac{1}{\tau}A$ and set $s = -i^{1-w/2}$. Then*

(a) 1, i, -1 and $-i$ are digits,

(b) $A - 1$ is a digit,

(c) $-B - 1$ is a digit,

(d) $i\tau^{w-1} - s^{-1}$ is a digit, and

(e) we have

$$(A-1)\tau^{w-1} + (-s^{-1}) = s\tau^{2w} + (-B-1)\tau^w + (i\tau^{w-1} - s^{-1}).$$

Figure 2.8.2 shows the digits used in Lemma 2.8.4 for a special configuration.

Proof. (a) A direct calculation shows that the lattice elements 1, i, -1 and $-i$ are in the interior of

$$\tau^w V = \left[-2^{w/2-1}, 2^{w/2-1}\right] + \left[-2^{w/2-1}, 2^{w/2-1}\right] i$$

and are not divisible by τ. So all of them are digits.

(b) We can rewrite A as

$$A = 2^{w/2-1}(1-i) = -2^{w/2-1} i\tau,$$

therefore $\tau^2 \mid A$. We remark that A is a vertex (the lower-right vertex) of the scaled Voronoi cell $\tau^w V$ and that the edges of $\tau^w V$ are parallel to the real and imaginary axes. This means that $A-1$ is on the boundary, too, and its real part is larger than 0. By using the construction of the restricted Voronoi cell, cf. Definition 2.4.2, we know that $A-1$ is in $\tau^w \widetilde{V}$. Since it is clearly not divisible by τ, it is a digit.

(c) We have

$$B = \tfrac{1}{\tau} A = -2^{w/2-1} i.$$

Therefore $\tau \mid B$, and we know that B halves the edge at the bottom of the Voronoi cell $\tau^w V$. By construction of the scaled restricted Voronoi cell $\tau^w \widetilde{V}$, cf. Definition 2.4.2, we obtain that $B+1$ is a digit, and therefore, by symmetry, $-B-1$ is a digit, too.

(d) Rewriting yields

$$i\tau^{w-1} - s^{-1} = s^{-1}(is\tau^{w-1} - 1),$$

and we obtain

$$s\tau^w = -i^{1-w/2}(1+i)^w = -2^{w/2} i,$$

since $(1+i)^2 = 2i$. Further we can check that the vertices of $\tau^w V$ are $i^k \tau^{w-1}$ for an appropriate $k \in \mathbb{Z}$.

Now consider $is\tau^{w-1}$. This is exactly the lower-right vertex A of $\tau^w V$. Therefore, we have

$$i\tau^{w-1} - s^{-1} = s^{-1}(A-1).$$

Using that $A-1$ is a digit and the rotational symmetry of the restricted Voronoi cell, $i\tau^{w-1} - s^{-1}$ is a digit.

(e) As before, we remark that $s\tau^w = -2^{w/2} i$. Therefore we obtain

$$B - 1 - s\tau^w = -B - 1.$$

Now, by rewriting, we get

$$\begin{aligned}
(A-1)\tau^{w-1} + (-s^{-1}) &= (A-\tau)\tau^{w-1} + (i\tau^{w-1} - s^{-1}) \\
&= (B-1)\tau^w + (i\tau^{w-1} - s^{-1}) \\
&= s\tau^{2w} + (B - 1 - s\tau^w)\tau^w + (i\tau^{w-1} - s^{-1}) \\
&= s\tau^{2w} + (-B-1)\tau^w + (i\tau^{w-1} - s^{-1}),
\end{aligned}$$

which was to prove. □

Lemma 2.8.5. *Let the assumptions of Proposition 2.8.3 hold and suppose $w = 2$. Then*

(a) -1 and $-i$ are digits and

(b) we have

$$-\tau - 1 = -i\tau^6 - \tau^4 - i\tau^2 - i.$$

Proof. (a) The elements -1 and $-i$ are on the boundary of the Voronoi cell $\tau^2 V$, cf. Figure 2.8.1(a). More precisely, each is halving an edge of the Voronoi cell mentioned. The construction of the restricted Voronoi cell, together with the rotation and scaling of $\tau^2 = 2i$, implies that -1 and $-i$ are in $\tau^2\widetilde{V}$. Since none of them is divisible by τ, both are digits.

(b) The element i has the 2-NAF-representation

$$i = -i\tau^4 - \tau^2 - i.$$

Therefore we obtain

$$-\tau - 1 = (-1 + i)\tau + (i\tau - 1) = i\tau^2 + (-i) = -i\tau^6 - \tau^4 - i\tau^2 - i$$

as required. □

Finally, we are able to prove the non-optimality result.

Proof of Proposition 2.8.3. Let $w \geq 4$. Everything needed can be found in Lemma 2.8.4: We have the equation

$$(A - 1)\tau^{w-1} + (-s^{-1}) = s\tau^{2w} + (-B - 1)\tau^w + (i\tau^{w-1} - s^{-1}),$$

in which the left and the right hand side are both valid expansion (the coefficients are digits). The left hand side has weight 2 and is not a w-NAF, whereas the right hand side has weight 3 and is a w-NAF.

Similarly the case $w = 2$ is shown in Lemma 2.8.5: We have the equation

$$-\tau - 1 = -i\tau^6 - \tau^4 - i\tau^2 - i,$$

which again is a counter-example to the optimality of the 2-NAFs. □

2.9 The p-is-0-Case

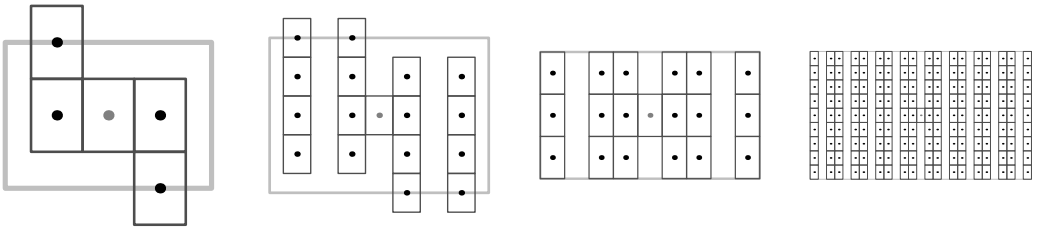
This section contains another special base τ. We assume that $p = 0$ and that we have an integer $q \geq 2$. Again, we continue looking at w-NAF-number systems with minimal norm representative digit set modulo τ^w with $w \geq 2$. Some examples of the digit sets used are shown in Figure 2.9.1.

For all possible τ of this section, the corresponding Voronoi cell can be written explicitly as

$$V = \mathsf{polygon}\left(\left\{\tfrac{1}{2}(\tau + 1), \tfrac{1}{2}(\tau - 1), \tfrac{1}{2}(-\tau - 1), \tfrac{1}{2}(-\tau + 1)\right\}\right).$$

Remark that V is an axis-parallel rectangle.

In this section we prove the following non-optimality result.



(a) Digit set for $\tau = i\sqrt{2}$ and $w = 3$. (b) Digit set for $\tau = i\sqrt{2}$ and $w = 5$. (c) Digit set for $\tau = i\sqrt{3}$ and $w = 3$. (d) Digit set for $\tau = i\sqrt{3}$ and $w = 5$.

Figure 2.9.1: Minimal norm representatives digit sets modulo τ^w. For each digit η, the corresponding Voronoi cell V_η is drawn. The large scaled Voronoi cell is $\tau^w V$.

Proposition 2.9.1. *Let w be an odd integer with $w \geq 3$ and the setting as above. Then there is an element of $\mathbb{Z}[\tau]$ whose w-NAF-expansion is non-optimal.*

For the remaining section—it contains the proof of the proposition above—we will assume $\tau = i\sqrt{q}$. The case $\tau = -i\sqrt{q}$ is analogous. Before we start with the proof of Proposition 2.9.1, we need the following two lemmata.

Lemma 2.9.2. *Let the assumptions of Proposition 2.9.1 hold, and suppose that q is even. Define $A := \frac{1}{2}|\tau|^{w+1}$ and $B := \frac{1}{\tau}A$, and set $s = (-1)^{\frac{1}{2}(w+1)}$. Then*

(a) 1 and -1 are digits,

(b) $A - 1 - \tau$ is a digit,

(c) $-B - 1$ is a digit,

(d) $-s - \tau^{w-1}$ is a digit, and

(e) we have

$$(A - 1 - \tau)\tau^{w-1} - s = s\tau^{2w} + (-B - 1)\tau^w + (-s - \tau^{w-1}).$$

Figure 2.9.2 shows the digits used in Lemma 2.9.2 for a special configuration.

Proof. (a) A direct calculation shows that -1 and 1 are in an open disc with radius $\frac{1}{2}|\tau|^w$, which itself is contained in $\tau^w V$. Both are not divisible by τ, so both are digits.

(b) Because w is odd, q is even and $\tau = i\sqrt{q}$, we can rewrite the point A as

$$A = \frac{1}{2}|\tau|^{w+1} = \frac{q}{2}q^{\frac{1}{2}(w-1)}$$

and see that A is a (positive) rational integer and that $\tau^{w-1} \mid A$. Furthermore, A halves an edge of $\tau^w V$. Therefore, $A - 1$ is inside $\tau^w V$. If $q \geq 4$ or $w \geq 5$, the

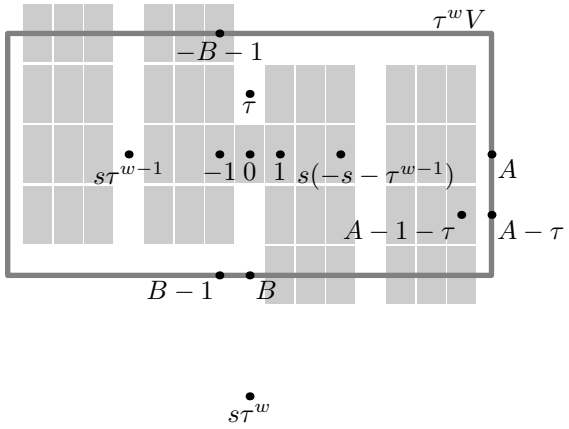


Figure 2.9.2: The q-is-even situation. The figure shows the configuration $p = 0$, $q = 4$, $\tau = 2i$, $w = 3$, $s = 1$. A polygon filled grey represents a digit, a dot represents a point of interest in Lemma 2.9.2.

point $A-1-\tau$ is inside $\tau^w V$, too, since the vertical (parallel to the imaginary axis) side-length of $\tau^w V$ is $|\tau|^w$ and $|\tau| < \frac{1}{2}|\tau|^w$. Since $\tau^2 \mid A$, we obtain $\tau \nmid A-1-\tau$, so $A-1-\tau$ is a digit. If $q = 2$ and $w = 3$, we have $A-1-\tau = 1-\tau$. Due to the definition of the restricted Voronoi cell $\widetilde{V}$, cf. Definition 2.4.2, we obtain that $1-\tau$ is a digit.

(c) Previously we saw $\tau^{w-1} \mid A$. Using the definition of B and $w \geq 3$ yields $\tau \mid B$. It is easy to check that $B = \frac{1}{2}s\tau^w$. Furthermore, we see that B is on the boundary of the Voronoi cell $\tau^w V$. By a symmetry argument we get the same results for $-B$. By the construction of the restricted Voronoi cell $\widetilde{V}$, cf. Definition 2.4.2, we obtain that $-B-1$ is in $\tau^w\widetilde{V}$ and since clearly $\tau \nmid (-B-1)$, we get that $-B-1$ is a digit.

(d) We first remark that $\tau^{w-1} \in \mathbb{Z}$ and that $\left|\tau^{w-1}\right| \leq A$. Even more, we get $0 < -s\tau^{w-1} \leq A$. Since A is on the boundary of $\tau^w V$, we obtain $-1-s\tau^{w-1} \in \tau^w \operatorname{int}(V)$. By symmetry the result is true for $-s-\tau^{w-1}$ and clearly $\tau \nmid (-s-\tau^{w-1})$, so $-s-\tau^{w-1}$ is a digit.

(e) We get

$$\begin{aligned}(A-1-\tau)\tau^{w-1} + (-s) &= (A-\tau)\tau^{w-1} + (-s-\tau^{w-1}) \\ &= (B-1)\tau^w + (-s-\tau^{w-1}) \\ &= s\tau^{2w} + (B-1-s\tau^w)\tau^w + (-s-\tau^{w-1}) \\ &= s\tau^{2w} + (-B-1)\tau^w + (-s-\tau^{w-1}),\end{aligned}$$

which can easily be verified. We used $B = \frac{1}{\tau}A$. □

Lemma 2.9.3. *Let the assumptions of Proposition 2.9.1 hold, and suppose that q is odd. Define $A' := \frac{1}{2}|\tau|^{w+1}$, $B' := \frac{1}{\tau}A$, $A := A' - \frac{1}{2}$ and $B := B' + \frac{\tau}{2}$, and set $C = -A$, $t = (q+1)/2$ and $s = (-1)^{\frac{1}{2}(w+1)} \in \{-1, 1\}$. Then*

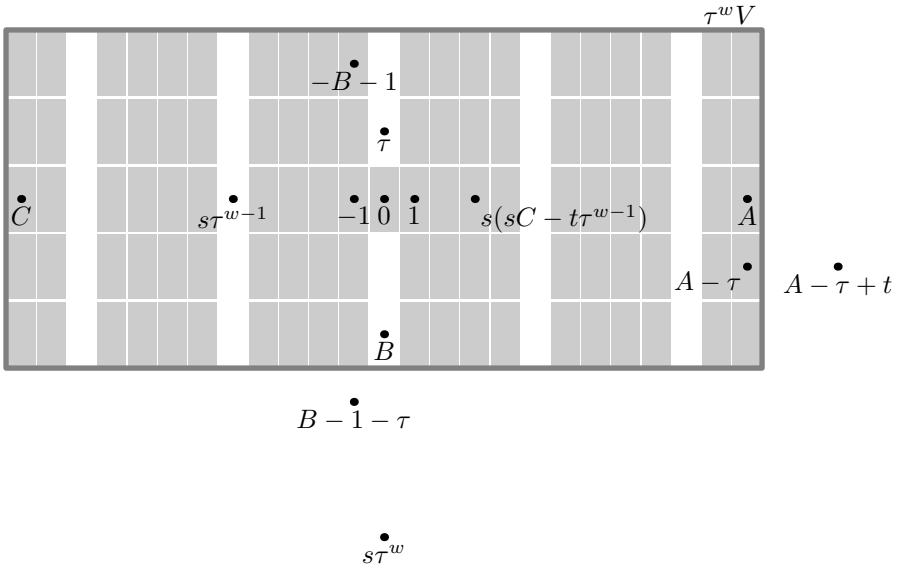


Figure 2.9.3: The q-is-odd situation, The figure shows the configuration $p = 0$, $q = 5$, $\tau = i\sqrt{5}$, $w = 3$, $s = 1$. A polygon filled grey represents a digit, a dot represents a points of interest in Lemma 2.9.3.

(a) 1 *and* -1 *are digits,*

(b) $A - \tau$ *is a digit,*

(c) sC *is a digit,*

(d) $-B - 1$ *is a digit,*

(e) $sC - t\tau^{w-1}$ *is a digit, and*

(f) we have

$$(A-\tau)\tau^{w-1} + (sC) = s\tau^{2w} + (-B-1)\tau^w + (sC - t\tau^{w-1}).$$

Figure 2.9.3 shows the digits used in Lemma 2.9.3 for a special configuration.

Proof. (a) See the proof of Lemma 2.9.2.

(b) We can rewrite the point A as

$$A = \frac{1}{2}|\tau|^{w+1} - \frac{1}{2} = \frac{1}{2}\left(q^{\frac{1}{2}(w+1)} - 1\right).$$

Since q is odd with $q \geq 3$ and w is odd with $w \geq 3$, we obtain $A \in \mathbb{Z}$ with $0 < A < \frac{1}{2}|\tau|^{w+1}$ and $q \nmid A$. Therefore $\tau \nmid A$ and A is in the interior of the Voronoi cell $\tau^w V$. The vertical (parallel to the imaginary axis) side-length of $\tau^w V$ is $|\tau|^w$ and $|\tau| < \frac{1}{2}|\tau|^w$, so $A - \tau$ is in the interior of $\tau^w V$, too. Since $\tau \nmid A - \tau$, the element $A - \tau$ is a digit.

(c) We got $\tau \nmid A$ and A is in the interior of the Voronoi cell $\tau^w V$. Therefore A is a digit, and—by symmetry—sC is a digit, too.

(d) We obtain

$$B = -\frac{1}{2} i \sqrt{q} \left|\tau\right|^{w-1} + i \frac{1}{2} \sqrt{q} = \frac{1}{2} \tau \left(- \left|\tau\right|^{w-1} + 1 \right),$$

which is inside $\tau^w V$. Therefore the same is true for $-B$. The horizontal (parallel to the real axis) side-length of $\tau^w V$ is larger than 2, therefore $-B - 1$ is inside $\tau^w V$, too. Since $\tau \mid B$ we get $\tau \nmid (-B - 1)$, so $-B - 1$ is a digit.

(e) We obtain

$$\begin{aligned} 0 < s \left(sC - t\tau^{w-1} \right) &= \frac{1}{2} \left((q+1) \left|\tau\right|^{w-1} - \left|\tau\right|^{w+1} + 1 \right) \\ &= \frac{1}{2} \left(\left|\tau\right|^{w-1} + 1 \right) < \frac{1}{2} \left|\tau\right|^{w+1}. \end{aligned}$$

This means that $sC - t\tau^{w-1}$ is in the interior of the Voronoi cell $\tau^w V$. Since $\tau \nmid (-A) = C$, the same is true for $sC - t\tau^{w-1}$, i.e. it is a digit.

(f) We get

$$\begin{aligned} (A - \tau)\tau^{w-1} + (sC) &= (A - \tau + t)\tau^{w-1} + (sC - t\tau^{w-1}) \\ &= (B - 1 - \tau)\tau^w + (sC - t\tau^{w-1}) \\ &= s\tau^{2w} + (B - 1 - \tau - s\tau^w)\tau^w + (sC - t\tau^{w-1}) \\ &= s\tau^{2w} + (-B - 1)\tau^w + (sC - t\tau^{w-1}), \end{aligned}$$

which can be checked easily. □

The two lemmata above now allow us to prove the non-optimality result of this section.

Proof of Proposition 2.9.1. Let q be even. In Lemma 2.9.2 we got

$$(A - 1 - \tau)\tau^{w-1} - s = s\tau^{2w} + (-B - 1)\tau^w + (-s - \tau^{w-1})$$

and that all the coefficients there were digits, i.e. we have valid expansions on the left and right hand side. The left hand side has weight 2 and is not a w-NAF, whereas the right hand side has weight 3 and is a w-NAF. Therefore a counter-example to the optimality was found.

The case q is odd works analogously. We got the counter-example

$$(A - \tau)\tau^{w-1} + (sC) = s\tau^{2w} + (-B - 1)\tau^w + (sC - t\tau^{w-1})$$

in Lemma 2.9.3. □

Figure 2.10.1: The optimality map including computational results. Below each block the parameters $|p|$ and q (fulfilling $\tau^2 - p\tau + q = 0$) are printed. A block is positioned according to τ in the complex plane. Above each block are the w. The symbol O means that the w-NAF-expansions are optimal, N means there are non-optimal w-NAF-expansions. If a result is surrounded by a circle, then it is a computational result. Otherwise, if there is no circle, then the result comes from a theorem given here or was already known. A dot means that there is no result available.

2.10 Computational Results

This section contains computational results on the optimality of w-NAFs for some special imaginary quadratic bases τ and integers w. We assume that we have a τ coming from integers p and q with $q > p^2/4$. Again, we continue looking at w-NAF-number systems with minimal norm representative digit set modulo τ^w with $w \geq 2$.

As mentioned in Section 2.2, the condition w-subadditivity-condition—and therefore optimality—can be verified by finding a w-NAF-expansion with weight at most 2 in $w\,(\#\mathcal{D} - 1)$ cases. The computational results can be found in Figure 2.10.1. The calculations were performed in Sage [96].

Chapter 3

Existence and Optimality of w-Non-adjacent Forms with an Algebraic Integer Base

This chapter contains the article [50] with the title "Existence and Optimality of w-Non-adjacent Forms with an Algebraic Integer Base". It is joint work with Clemens Heuberger. The article was accepted for publication by *Acta Mathematica Hungarica* on October 31, 2012. An introduction to this chapter can be found in Chapter 1, in particular Sections 1.3 and 1.4.

Abstract

We consider digital expansions in lattices with endomorphisms acting as base. We focus on the w-non-adjacent form (w-NAF), where each block of w consecutive digits contains at most one non-zero digit. We prove that for sufficiently large w and an expanding endomorphism, there is a suitable digit set such that each lattice element has an expansion as a w-NAF.

If the eigenvalues of the endomorphism are large enough and w is sufficiently large, then the w-NAF is shown to minimise the weight among all possible expansions of the same lattice element using the same digit system.

3.1 w-Non-Adjacent Forms and Digit Sets

In this section, we recall the notion of w-non-adjacent forms and formally introduce w-non-adjacent digits sets.

We use the abstract setting mentioned in the introduction (and already used in the previous chapter): We consider an Abelian group $\mathcal{A}$, an injective endomorphism Φ of $\mathcal{A}$ and an integer $w \geq 1$. Let $\mathcal{D}^\bullet$ be a system of representatives of those residue classes of $\mathcal{A}$ modulo $\Phi^w(\mathcal{A})$ which are not contained in $\Phi(\mathcal{A})$. We set $\mathcal{D} = \mathcal{D}^\bullet \cup \{0\}$.

We call the triple $(\mathcal{A}, \Phi, \mathcal{D})$ a *pre-w-non-adjacent digit set* (*pre-w-NADS*).

Definition 3.1.1. 1. A word $\boldsymbol{\eta} = \eta_{\ell-1} \dots \eta_0$ over the alphabet $\mathcal{D}$ is said to be a *$\mathcal{D$-w-non-adjacent form* (*$\mathcal{D$-w-NAF*), if every factor $\eta_{j+w-1} \dots \eta_j$, $0 \le j \le \ell - w$, contains at most one non-zero letter η_k. Its *value* is defined to be

$$\mathsf{value}(\eta_{\ell-1} \dots \eta_0) = \sum_{j=0}^{\ell-1} \Phi^j(\eta_j).$$

We say that $\boldsymbol{\eta}$ is a *$\mathcal{D$-w-NAF of* $\alpha \in \mathcal{A}$ if $\mathsf{value}(\boldsymbol{\eta}) = \alpha$.

2. We say that $\mathcal{D}$ is a *w-non-adjacent digit set* (*w-NADS*), if every $\alpha \in \mathcal{A}$ admits a $\mathcal{D}$-w-NAF.

Example 3.1.2. Let K be a number field of degree n, $\mathfrak{O}$ be an order in K and $\tau \in \mathfrak{O}$. We consider the endomorphism $\Phi_\tau \colon \mathfrak{O} \to \mathfrak{O}$ with $\alpha \mapsto \tau\alpha$, i.e., multiplication by τ. Then let $\mathcal{D}^\bullet$ be a system of representatives of those residue classes of $\mathfrak{O}$ modulo τ^w which are not divisible by τ and $\mathcal{D} = \mathcal{D}^\bullet \cup \{0\}$. Then $(\mathfrak{O}, \Phi_\tau, \mathcal{D})$ is a pre-w-NADS. Note that

$$\mathsf{value}(\eta_{\ell-1} \dots \eta_0) = \sum_{j=0}^{\ell-1} \eta_j \tau^j$$

for a word $\eta_{\ell-1} \dots \eta_0$ over the alphabet $\mathcal{D}$.

We state a few special cases.

Example 3.1.3. Let $\tau \in \mathbb{Z}$, $|\tau| \ge 2$ and $w \ge 1$ be an integer. Consider

$$\mathcal{D}^\bullet = \left\{ d \in \mathbb{Z} : -\frac{|\tau|^w}{2} < d \le \frac{|\tau|^w}{2}, \tau \nmid d \right\}$$

and $\mathcal{D} = \mathcal{D}^\bullet \cup \{0\}$. Then $(\mathbb{Z}, \Phi_\tau, \mathcal{D})$ is a pre-w-NADS, where Φ_τ still denotes multiplication by τ. It can be shown that $(\mathbb{Z}, \Phi_\tau, \mathcal{D})$ is a w-NADS. This will also be a consequence of Theorem 3.4.1.

Example 3.1.4. Let τ be an imaginary quadratic integer and $\mathcal{D}^\bullet$ a system of representatives of those residue classes of $\mathbb{Z}[\tau]$ modulo τ^w which are not divisible by τ with the property that

$$\text{if } \alpha \equiv \beta \pmod{\tau^w} \text{ and } \alpha \in \mathcal{D}^\bullet, \text{ then } |\alpha| \le |\beta|$$

holds for α, $\beta \in \mathbb{Z}[\tau]$ which are not divisible by τ. This means that $\mathcal{D}$ contains a representative of minimal absolute value of each residue class not divisible by τ. As always, we set $\mathcal{D} = \mathcal{D}^\bullet \cup \{0\}$.

Then, for $w \ge 2$, $(\mathbb{Z}[\tau], \Phi_\tau, \mathcal{D})$ is a w-NADS (cf. Heuberger and Krenn [49]), where Φ_τ still denotes multiplication by τ.

For $\tau \in \{(\pm 1 \pm \sqrt{-7})/2, (\pm 3 \pm \sqrt{-3})/2, 1 + \sqrt{-1}, \sqrt{-2}, (1 + \sqrt{-11})/2\}$, this has been shown by Solinas [94, 95] and Blake, Murty and Xu [16, 18], cf. also Blake, Murty and Xu [17] for other digit sets to the bases $(\pm 1 \pm \sqrt{-7})/2$.

At several occurrences, it is useful to consider *equivalent* pre-w-NADS.

Definition 3.1.5. The pre-w-NADS $(\mathcal{A}, \Phi, \mathcal{D})$ and $(\mathcal{A}', \Phi', \mathcal{D}')$ are said to be *equivalent*, if there is a group isomorphism $Q\colon \mathcal{A} \to \mathcal{A}'$ such that the diagram

$$\begin{array}{ccc} \mathcal{A} & \xrightarrow[\Phi]{} & \mathcal{A} \\ {\scriptstyle Q}\downarrow & & \downarrow{\scriptstyle Q} \\ \mathcal{A}' & \xrightarrow[\Phi']{} & \mathcal{A}' \end{array}$$

commutes and such that $\mathcal{D}' = Q(\mathcal{D})$.

It is then clear that the following proposition holds.

Proposition 3.1.6. *Let $(\mathcal{A}, \Phi, \mathcal{D})$ and $(\mathcal{A}', \Phi', \mathcal{D}')$ two equivalent pre-w-NADS. Then $\mathcal{D}$ is a w-NADS if and only if $\mathcal{D}'$ is a w-NADS.*

Proof. Straightforward. □

Example 3.1.7. We continue Example 3.1.2, i.e., K is a number field, $\mathfrak{O}$ an order in K, $\tau \in \mathfrak{O}$, the endomorphism considered is Φ_τ, the multiplication by τ, and the digit set $\mathcal{D}$ is as in Example 3.1.2.

The real embeddings of K are denoted by $\sigma_1, \ldots, \sigma_s$; the non-real complex embeddings of K are denoted by σ_{s+1}, $\overline{\sigma_{s+1}}$, $\ldots$, σ_{s+t}, $\overline{\sigma_{s+t}}$, where $\overline{\cdot}$ denotes complex conjugation and $n = s + 2t$. The *Minkowski map* $\Sigma\colon K \to \mathbb{R}^n$ maps $\alpha \in K$ to

$$(\sigma_1(\alpha), \ldots, \sigma_s(\alpha), \Re\sigma_{s+1}(\alpha), \Im\sigma_{s+1}(\alpha), \ldots, \Re\sigma_{s+t}(\alpha), \Im\sigma_{s+t}(\alpha)) \in \mathbb{R}^n.$$

We write $\Lambda = \Sigma(\mathfrak{O})$ for the image of $\mathfrak{O}$ under Σ. Note that Λ is a lattice in $\mathbb{R}^n$. We consider the $n \times n$ block diagonal matrix

$$A_\tau := \operatorname{diag}\left(\sigma_1(\tau), \ldots, \sigma_s(\tau), \begin{pmatrix} \Re\sigma_{s+1}(\tau) & -\Im\sigma_{s+1}(\tau) \\ \Im\sigma_{s+1}(\tau) & \Re\sigma_{s+1}(\tau) \end{pmatrix}, \ldots, \begin{pmatrix} \Re\sigma_{s+t}(\tau) & -\Im\sigma_{s+t}(\tau) \\ \Im\sigma_{s+t}(\tau) & \Re\sigma_{s+t}(\tau) \end{pmatrix}\right)$$

and set $\mathcal{D}' := \Sigma(\mathcal{D})$. Then the pre-$w$-NADS $(\mathfrak{O}, \Phi_\tau, \mathcal{D})$ and $(\Lambda, \Phi'_\tau, \mathcal{D}')$ are easily seen to be equivalent, where $\Phi'_\tau(x) := A_\tau \cdot x$ for $x \in \mathbb{R}^n$.

Note that if K is an imaginary quadratic number field (cf. Example 3.1.4), this construction merely corresponds to a straight-forward identification of $\mathbb{C}$ with $\mathbb{R}^2$.

In order to investigate the w-NADS property further, it is convenient to consider the following two maps.

Definition 3.1.8. Let $(\mathcal{A}, \Phi, \mathcal{D})$ be a pre-w-NADS. We define

1. $d\colon \mathcal{A} \to \mathcal{D}$ with $d(\alpha) = 0$ for $\alpha \in \Phi(\mathcal{A})$ and $d(\alpha) \equiv \alpha \pmod{\Phi^w(\mathcal{A})}$ for all other $\alpha \in \mathcal{A}$,
2. $T\colon \mathcal{A} \to \mathcal{A}$ with $\alpha \mapsto \Phi^{-1}(\alpha - d(\alpha))$.

Note that the map d is well-defined as $\mathcal{D}^\bullet$ contains exactly one representative of every residue class of $\mathcal{A}$ modulo $\Phi^w(\mathcal{A})$ which is not contained in $\Phi(\mathcal{A})$. Furthermore, we have $\alpha \equiv d(\alpha) \pmod{\Phi(\mathcal{A})}$ for all $\alpha \in \mathcal{A}$. Therefore and by the injectivity of Φ, the map T is well-defined. We remark that by definition, we have $T(0) = 0$.

We get the following characterisation, which corresponds to the backwards division algorithm for computing digital expansions from right (least significant digit) to left (most significant digit).

Lemma 3.1.9. *Let $\alpha \in \mathcal{A}$. Then α has a $\mathcal{D}$-w-NAF $\eta_{\ell-1} \dots \eta_0$ if and only if $T^\ell(\alpha) = 0$. In this case, we have $\eta_k = d(T^k(\alpha))$ for $0 \le k < \ell$. In particular, the $\mathcal{D}$-w-NAF of an $\alpha \in \mathcal{A}$, if it exists, is unique up to leading zeros.*

Proof. Assume that $\eta_{\ell-1} \dots \eta_0$ is a $\mathcal{D}$-w-NAF of α. We clearly have $\alpha \equiv \eta_0 \pmod{\Phi(\Lambda)}$, so that α is an element of $\Phi(\Lambda)$ if and only if $\eta_0 = 0$. Otherwise, the w-NAF-condition ensures that $\alpha \equiv \eta_0 \pmod{\Phi^w(\Lambda)}$. In both cases, we get $d(\alpha) = \eta_0$ and therefore

$$T(\alpha) = \mathsf{value}(\eta_{\ell-1} \dots \eta_1).$$

Iterating this process yields $T^k(\alpha) = \mathsf{value}(\eta_{\ell-1} \dots \eta_k)$ for $0 \le k \le \ell$, where $\eta_{\ell-1} \dots \eta_k$ is a $\mathcal{D}$-w-NAF. For $k = \ell$, we see that $T^\ell(\alpha)$ is the value of the empty word, which is zero by the definition of the empty sum.

Conversely, we assume that $T^\ell(\alpha) = 0$. We note that if $d(\beta) \neq 0$ for some $\beta \in \mathcal{A}$, we have $\beta - d(\beta) \equiv 0 \pmod{\Phi^w(\Lambda)}$, which results in $T^j(\beta) = \Phi^{-j}(\beta - d(\beta)) \equiv 0 \pmod{\Phi(\Lambda)}$ and $d(T^j(\beta)) = 0$ for $1 \le j \le w-1$. Therefore, the word $\boldsymbol{\eta} = d(T^{\ell-1}(\alpha)) \dots d(T(\alpha))d(\alpha)$ is a $\mathcal{D}$-w-NAF. Iterating the relation $\beta = \Phi(T(\beta)) + d(\beta)$ valid for all $\beta \in \mathcal{A}$, we conclude that $\alpha = \Phi^\ell(T^\ell(\alpha)) + \mathsf{value}(\boldsymbol{\eta}) = \mathsf{value}(\boldsymbol{\eta})$. □

3.2 Lattices and $\mathcal{D}$-w-NAFs

We now specialise our investigations to the case that the abstract Abelian group $\mathcal{A}$ is replaced by a lattice in $\mathbb{R}^n$, i.e., $\mathcal{A} = \Lambda = w_1\mathbb{Z} \oplus \dots \oplus w_n\mathbb{Z}$ for linearly independent $w_1, \dots, w_n \in \mathbb{R}^n$. Further let Φ be an injective endomorphism of $\mathbb{R}^n$ with $\Phi(\Lambda) \subseteq \Lambda$, $w \ge 1$ be an integer, and $\mathcal{D}^\bullet$ a system of representatives of those residue classes of Λ modulo $\Phi^w(\Lambda)$ which are not contained in $\Phi(\Lambda)$, and set $\mathcal{D} = \mathcal{D}^\bullet \cup \{0\}$.

The results are still applicable to the case of multiplication by τ in the order of a number field, as the purpose of Example 3.1.7 was to describe it as equivalent to a lattice $\Lambda \subseteq \mathbb{R}^n$ via the isomorphism Σ.

The aim of this section is to prove a necessary criterion for a pre-w-NADS to be a w-NADS.

Proposition 3.2.1. *Let $\mathcal{D}$ be a w-NADS. Then Φ is expanding, i.e., $|\lambda| > 1$ holds for all eigenvalues λ of Φ.*

Proof. 1. We first consider the case that there is an eigenvalue λ of Φ with $|\lambda| < 1$.

In a somewhat different wording, this has been led to a contradiction by Vince [104]. The idea is the following: After a suitable change of variables, the endomorphism

Φ can be represented by a Jordan matrix such that the first k coordinates, say, correspond to the eigenvalue λ. Thus the first k coefficients of $\mathsf{value}(\boldsymbol{\eta})$ are bounded independently of the word $\boldsymbol{\eta}$ over the alphabet $\mathcal{D}$. Thus it is impossible to have a representation of all elements of Λ. This is completely independent of the w-NAF-condition (and gives, in fact, a stronger result, as representability by any word over the digit set is impossible).

2. We next consider the case that $|\lambda| \geq 1$ for all eigenvalues λ of Φ with equality $|\lambda_0| = 1$ for at least one eigenvalue λ_0.

   We again follow Vince [104], see also Kovács and Pethő [66], to see that λ_0 must be a root of unity. The idea is that λ_0 is a unit in $\mathbb{Z}[\lambda_0, \overline{\lambda_0}]$, as $\overline{\lambda_0}$ is its inverse. Therefore, λ has absolute norm ± 1. As we already assumed that all its absolute conjugates are at least 1 in absolute value, this implies that all absolute conjugates of λ_0 lie on the unit circle. Thus λ_0 is a root of unity.

   As a consequence, there is some ℓ such that $\lambda_0^\ell = 1$. In other words, 1 is an eigenvalue of Φ^ℓ. After a suitable change of coordinates, Λ can be assumed to be $\mathbb{Z}^n$ and Φ can be represented by a matrix with integer entries. Let α be an eigenvector of Φ^ℓ with eigenvalue 1. Multiplying α by a suitable integer if necessary, we can assume that $\alpha \in \mathbb{Z}^n = \Lambda$. As $\alpha = \Phi^\ell(\alpha)$, we get $\alpha \in \Phi^k(\Lambda)$ for all integers $k \geq 0$, which implies that $d(T^k(\alpha)) = 0$ holds for all k. Furthermore, we cannot have $T^k(\alpha) = 0$ for any $k \geq 0$. Thus, α cannot be represented. □

3.3 Tiling Based Digit Sets

In this section, we consider a fixed lattice $\Lambda \subseteq \mathbb{R}^n$ and an expanding endomorphism Φ of $\mathbb{R}^n$ with $\Phi(\Lambda) \subseteq \Lambda$. We will discuss digit sets constructed from tilings.

Definition 3.3.1. Let V be a subset of $\mathbb{R}^n$. We say that V tiles $\mathbb{R}^n$ by the lattice Λ, if the following two properties hold:

1. $\bigcup_{z\in\Lambda}(z + V) = \mathbb{R}^n$,
2. $V \cap (z + V) \subseteq \partial V$ holds for all $z \in \Lambda$ with $z \neq 0$.

We now assume that V be a subset of $\mathbb{R}^n$ tiling $\mathbb{R}^n$ by Λ.

Lemma 3.3.2. *Let $w \geq 1$ and*

$$\widetilde{\mathcal{D}} := \{\alpha \in \Lambda : \Phi^{-w}(\alpha) \in V\}.$$

Then $\widetilde{\mathcal{D}}$ contains a complete residue system of Λ modulo $\Phi^w(\Lambda)$.

Furthermore, if α, $\alpha' \in \widetilde{\mathcal{D}}$ with $\alpha \neq \alpha'$ and $\alpha \equiv \alpha' \pmod{\Phi^w(\Lambda)}$, then $\Phi^{-w}(\alpha)$, $\Phi^{-w}(\alpha') \in \partial V$.

Proof. Let $\beta \in \Lambda$. Then there is a $\gamma \in \Lambda$ and a $v \in V$ such that $\Phi^{-w}(\beta) = \gamma + v$. Setting $\alpha := \beta - \Phi^w(\gamma)$, this implies that

$$\Phi^{-w}(\alpha) = \Phi^{-w}(\beta) - \gamma = v \in V,$$

i.e., $\alpha \in \widetilde{\mathcal{D}}$ and $\beta \equiv \alpha \pmod{\Phi^w(\Lambda)}$.

Assume now α, $\alpha' \in \widetilde{\mathcal{D}}$ with $\alpha \neq \alpha'$ and $\alpha \equiv \alpha' \pmod{\Phi^w(\Lambda)}$. We write $\alpha' = \alpha + \Phi^w(\gamma)$ for a suitable $\gamma \in \Lambda$. We obtain

$$\Phi^{-w}(\alpha') = \Phi^{-w}(\alpha) + \gamma,$$

which implies that $\Phi^{-w}(\alpha') \in \partial V$. Analogously, we get $\Phi^{-w}(\alpha) \in \partial V$. □

For an integer $w \geq 1$, we choose a subset $\mathcal{D}^\bullet$ of $\widetilde{\mathcal{D}}$ in such a way that $\mathcal{D}^\bullet$ contains exactly one representative of every residue class modulo $\Phi^w(\Lambda)$ which is not contained in $\Phi(\Lambda)$. We also set $\mathcal{D} := \mathcal{D}^\bullet \cup \{0\}$.

Theorem 3.3.3. *Let $\| \cdot \|$ be a vector norm on $\mathbb{R}^n$ such that for the corresponding induced operator norm, also denoted by $\| \cdot \|$, the inequality $\|\Phi^{-1}\| < 1$ holds. Let r and R be positive reals with*

$$\{x \in \mathbb{R}^n : \|x\| \leq r\} \subseteq V \subseteq \{x \in \mathbb{R}^n : \|x\| \leq R\}. \tag{3.3.1}$$

If w is a positive integer such that

$$\|\Phi^{-1}\|^w < \frac{1}{1 + R/r}, \tag{3.3.2}$$

then $\mathcal{D}$ is a w-NADS.

Remark 3.3.4. In the case of expansions in an order of a number field (Example 3.1.7), we may take $\| \cdot \|$ to be the Euclidean norm $\| \cdot \|_2$, as the corresponding operator norm fulfils $\|A_\tau^{-1}\|_2 = \max\{1/|\sigma_j(\tau)| : 1 \leq j \leq s + t\}$. In this case, (3.3.2) is equivalent to $|\sigma_j(\tau)|^w > 1 + R/r$ for all $1 \leq j \leq s + t$.

Proof of Theorem 3.3.3. Let $\alpha \in \Lambda$. We claim that

$$\|T^k(\alpha)\| \leq \frac{R}{1 - \|\Phi^{-1}\|^w} + \|\Phi^{-1}\|^k \cdot \|\alpha\| \tag{3.3.3}$$

holds for all k with the property that $d(T^{k'}(\alpha)) = 0$ holds for all non-negative k' with $k - w < k' \leq k$.

For $k = 0$, (3.3.3) is obviously true. We assume that (3.3.3) holds for some k. As an abbreviation, we write $\beta = T^k(\alpha)$ and $\eta = d(\beta)$. If $\eta = 0$, then we have

$$\|T^{k+1}(\alpha)\| = \|T(\beta)\| = \|\Phi^{-1}(\beta)\| \leq \|\Phi^{-1}\| \cdot \|\beta\| \leq \frac{\|\Phi^{-1}\| \cdot R}{1 - \|\Phi^{-1}\|^w} + \|\Phi^{-1}\|^{k+1} \cdot \|\alpha\|,$$

which proves (3.3.3) for $k + 1$.

In the case $\eta \neq 0$, we get

$$\begin{aligned}\|T^{k+w}(\alpha)\| &= \|\Phi^{-w}(\beta-\eta)\| \leq \|\Phi^{-1}\|^w \cdot \|\beta\| + \|\Phi^{-1}\|^w \cdot \|\eta\| \leq \|\Phi^{-1}\|^w \cdot \|\beta\| + 1^w R \\ &\leq \|\Phi^{-1}\|^w \left(\frac{R}{1-\|\Phi^{-1}\|^w} + \|\Phi^{-1}\|^k \cdot \|\alpha\|\right) + R \\ &= \frac{R}{1-\|\Phi^{-1}\|^w} + \|\Phi^{-1}\|^{k+w} \cdot \|\alpha\|,\end{aligned}$$

which is (3.3.3) for $k+w$.

By (3.3.2) and (3.3.3), we can choose a k_0 such that

$$\|\Phi^{-w}(T^k(\alpha))\| \leq \frac{\|\Phi^{-1}\|^w}{1-\|\Phi^{-1}\|^w} R + \|\Phi^{-1}\|^{k+w} \cdot \|\alpha\| < r \tag{3.3.4}$$

holds for all $k \geq k_0$.

If $T^{k_0}(\alpha) = 0$, then α admits a $\mathcal{D}$-w-NAF by Lemma 3.1.9. Otherwise, choose $k \geq k_0$ maximally such that $T^{k_0}(\alpha) \in \Phi^{k-k_0}(\Lambda)$. This is possible because Φ is expanding. This results in $T^k(\alpha) \notin \Phi(\Lambda)$. Then (3.3.4) implies that

$$\|\Phi^{-w}(T^k(\alpha))\| < r.$$

By (3.3.1), we conclude that $\Phi^{-w}(T^k(\alpha))$ is an element of the interior of V.

By Lemma 3.3.2, we obtain $\Phi^{-w}(T^k(\alpha)) \in \mathcal{D}^\bullet$, hence $d(T^k(\alpha)) = T^k(\alpha)$ and $T^{k+1}(\alpha) = 0$. Thus α admits a $\mathcal{D}$-w-NAF by Lemma 3.1.9. □

3.4 Minimal Norm Digit Set

In this section, we study a special digit set, the minimal norm digit set. In the case of an imaginary quadratic integer τ, this notion coincides with the minimal norm representative digit sets introduced by Solinas [94, 95].

Let again Λ be a lattice in $\mathbb{R}^n$ and Φ an expansive endomorphism of $\mathbb{R}^n$ with $\Phi(\Lambda) \subseteq \Lambda$. Choose a positive integer w_0 such that $|\lambda| > 2^{1/w_0}$ holds for all eigenvalues λ of Φ. Thus the spectral radius of Φ^{-1} is less than $1/2^{1/w_0}$. We choose a vector norm $\|\cdot\|$ on $\mathbb{R}^n$ such that the induced operator norm (also denoted by $\|\cdot\|$) fulfils $\|\Phi^{-1}\| < 1/2^{1/w_0}$. As a consequence, we have $\|\Phi^{-1}\|^w < 1/2$ for all $w \geq w_0$.

Again, in the case of expansions in an order of a number field (Example 3.1.7), we may take $\|\cdot\|$ to be the Euclidean norm $\|\cdot\|_2$, cf. Remark 3.3.4.

Let V be the Voronoi cell of the origin with respect to the point set Λ and the vector norm $\|\cdot\|$, i.e.,

$$V = \{z \in \mathbb{R}^n : \|z\| \leq \|z+\alpha\| \text{ holds for all } \alpha \in \Lambda\}.$$

While V does not necessarily tile $\mathbb{R}^n$ by Λ (consider the norm $\|\cdot\|_\infty$ and the lattice generated by $(1,0)$ and $(0,10)$ in $\mathbb{R}^2$), for a given integer $w \geq 1$, we can still select a set

$\mathcal{D}^\bullet$ of representatives of those residue classes of Λ modulo $\Phi^w(\Lambda)$ which are not contained in $\Phi^w(\Lambda)$ such that

$$\mathcal{D}^\bullet \subseteq \{\alpha \in \Lambda : \Phi^{-w}(\alpha) \in V\}.$$

As usual, we also set $\mathcal{D} := \mathcal{D}^\bullet \cup \{0\}$ and call it a *minimal norm digit set modulo* Φ^w.

Adapting ideas of Germán and Kovács [42] to our setting, we prove the following theorem.

Theorem 3.4.1. *If $w \geq w_0$, then $\mathcal{D}$ is a w-NADS.*

Proof. We set $\widetilde{M} := \max\{\|\eta\| : \eta \in \mathcal{D}\}$. For $\beta \in \Lambda$, we have

$$\|T(\beta)\| = \|\Phi^{-1}(\beta - d(\beta))\| \leq \|\Phi^{-1}\|(\|\beta\| + \widetilde{M}).$$

Setting

$$M := \frac{\|\Phi^{-1}\|}{1 - \|\Phi^{-1}\|}\widetilde{M},$$

we see that

$$\begin{aligned} \|T(\beta)\| &< \|\beta\| \quad \text{if } \|\beta\| > M, \\ \|T(\beta)\| &\leq M \quad \text{if } \|\beta\| \leq M. \end{aligned}$$

As Λ is a discrete subset of $\mathbb{R}^n$, we conclude that the sequence $(T^k(\alpha))_{k\geq 0}$ is eventually periodic for all $\alpha \in \Lambda$.

For $\beta \in \Phi(\Lambda)$ with $\beta \neq 0$, we have

$$\|T(\beta)\| = \|\Phi^{-1}(\beta)\| \leq \|\Phi^{-1}\| \cdot \|\beta\| < \|\beta\|.$$

Consider the set

$$P := \{\beta \in \Lambda : \beta \notin \Phi(\Lambda) \text{ and } (T^k(\beta))_{k\geq 0} \text{ is purely periodic}\}.$$

The set P is empty if and only if for each $\alpha \in \Lambda$, there is an ℓ with $T^\ell(\alpha) = 0$, i.e., α admits a $\mathcal{D}$-w-NAF. Therefore, by Lemma 3.1.9, P is empty if and only if $\mathcal{D}$ is a w-NADS.

We therefore assume that P is nonempty. We choose an $\alpha \in P$ such that $\|\Phi^{-w}(\alpha)\| \geq \|\Phi^{-w}(\beta)\|$ holds for all $\beta \in P$. This is possible, since all elements β of P fulfil $\|\beta\| \leq M$, which implies that P is a finite set.

Next, we choose $\ell > 0$ with $T^\ell(\alpha) = \alpha$ and set $\eta_k = d(T^k(\alpha))$ for $0 \leq k \leq \ell$. We set

$$N := \{0 \leq k \leq \ell : \eta_k \neq 0\}.$$

By the w-NAF-condition, we have $|k - k'| \geq w$ for distinct elements k and k' of N.

By definition of T, we have

$$\alpha = T^\ell(\alpha) = \Phi^{-\ell}\Big(\alpha - \sum_{k=0}^{\ell-1} \Phi^k(\eta_k)\Big) = \Phi^{-\ell}(\alpha) - \sum_{k=0}^{\ell-1} \Phi^{k-\ell}(\eta_k).$$

Applying Φ^{-w} once more and rearranging yields

$$\Phi^{-w}(\alpha) = (\mathrm{id} - \Phi^{-\ell})^{-1}\Big(-\sum_{\substack{k=0\\k\in N}}^{\ell-1} \Phi^{k-\ell}(\Phi^{-w}(\eta_k))\Big). \tag{3.4.1}$$

Note that we restricted the sum to those k corresponding to non-zero digits.

We claim that

$$\|\Phi^{-w}(\eta_k)\| \le \|\Phi^{-w}(T^k(\alpha))\| \le \|\Phi^{-w}(\alpha)\| \tag{3.4.2}$$

holds for $k \in N$. The first inequality is an immediate consequence of the definition of $\mathcal{D}^\bullet$, as $\Phi^{-w}(T^k(\alpha)) = \Phi^{-w}(\eta_k) + \gamma$ for a suitable $\gamma \in \Lambda$. Here, we used that $\eta_k \neq 0$ implies that $T^k(\alpha) \notin \Phi(\Lambda)$. Therefore and as $T^{k+\ell}(\alpha) = T^k(T^\ell(\alpha)) = T^k(\alpha)$, we also get $T^k(\alpha) \in P$. By the choice of α, we conclude the second inequality in (3.4.2).

Taking norms in (3.4.1) yields

$$\|\Phi^{-w}(\alpha)\| \le \frac{\|\Phi^{-w}(\alpha)\|}{1 - \|\Phi^{-1}\|^\ell} \sum_{\substack{k=0\\k\in N}}^{\ell-1} \|\Phi^{-1}\|^{\ell-k}. \tag{3.4.3}$$

As $\ell \in N$, we have

$$\sum_{\substack{k=0\\k\in N}}^{\ell-1} \|\Phi^{-1}\|^{\ell-k} \le \|\Phi^{-1}\|^w + \|\Phi^{-1}\|^{2w} + \cdots + \|\Phi^{-1}\|^{mw} = \|\Phi^{-1}\|^w \frac{1 - \|\Phi^{-1}\|^{mw}}{1 - \|\Phi^{-1}\|^w}, \tag{3.4.4}$$

where $m = \lfloor \ell/w \rfloor$. Combining (3.4.3) and (3.4.4) yields

$$\|\Phi^{-w}(\alpha)\| \le \frac{\|\Phi^{-1}\|^w}{1 - \|\Phi^{-1}\|^w} \frac{1 - \|\Phi^{-1}\|^{mw}}{1 - \|\Phi^{-1}\|^\ell} \|\Phi^{-w}(\alpha)\| < \|\Phi^{-w}(\alpha)\|,$$

as $\|\Phi^{-1}\|^w < 1/2$, contradiction. □

We restate this result explicitly for expansion in orders of algebraic number fields.

Corollary 3.4.2. *Let K be an algebraic number field of degree n, $\sigma_1, \ldots, \sigma_s$ the real embeddings and $\sigma_{s+1}, \overline{\sigma_{s+1}}, \ldots, \sigma_{s+t}, \overline{\sigma_{s+t}}$ be the non-real complex embeddings of K.*

Let $\mathfrak{O}$ be an order of K and $\tau \in \mathfrak{O}$ such that $|\sigma_j(\tau)| > 1$ holds for all j. Let w be an integer with

$$w > \max\left\{ \frac{\log 2}{\log |\sigma_j(\tau)|} : 1 \le j \le s+t \right\}.$$

Let $\mathcal{D}^\bullet$ be a system of representatives of those residue classes of $\mathfrak{O}$ modulo τ^w which are not divisible by τ such that

$$\text{if } \alpha \equiv \beta \pmod{\tau^w} \text{ with } \tau \nmid \alpha \text{ and } \alpha \in \mathcal{D}, \text{ then } \sum_{j=1}^{s+t} a_j \Big|\sigma_j\Big(\frac{\alpha}{\tau^w}\Big)\Big|^2 \le \sum_{j=1}^{s+t} a_j \Big|\sigma_j\Big(\frac{\beta}{\tau^w}\Big)\Big|^2,$$

where $a_j = 1$ for $j \in \{1, \ldots, s\}$ and $a_j = 2$ for $j \in \{s+1, \ldots, s+t\}$. Then $\mathcal{D} := \mathcal{D}^\bullet \cup \{0\}$ is a w-NADS.

Example 3.4.3. Let C be an algebraic curve of genus g defined over $\mathbb{F}_q$ (a field with q elements). The Frobenius endomorphism operates on the Jacobian variety of C and satisfies a characteristic polynomial $\chi \in \mathbb{Z}[T]$ of degree $2g$. Let τ be a root of χ. Set $K = \mathbb{Q}[\tau]$ and $\mathfrak{O} = \mathbb{Z}[\tau]$, and denote the embeddings of K by σ_j. Using Corollary 3.4.2, a minimal norm digit set modulo τ^w is a w-NADS if

$$w > \frac{\log 4}{\log q}.$$

This is true because of the following reasons: The polynomial χ fulfils the equation

$$\chi(T) = T^{2g} Y(1/T),$$

where $Y(T)$ denotes the numerator of the zeta-function of C over $\mathbb{F}_q$, cf. Weil [107, 109]. The Riemann Hypothesis of the Weil Conjectures, cf. Weil [108], Dwork [32] and Deligne [25], state that all zeros of Y have absolute value $1/\sqrt{q}$. Therefore $|\sigma_j(\tau)| = \sqrt{q}$, which was to show.

3.5 Optimality of $\mathcal{D}$-w-NAFs

In this section, we consider a lattice $\Lambda \subseteq \mathbb{R}^n$ and an expanding endomorphism Φ of $\mathbb{R}^n$ with $\Phi(\Lambda) \subseteq \Lambda$. First, we recall the definition of weight as in the previous chapter.

Definition 3.5.1. Let $\boldsymbol{\eta} = \eta_{\ell-1} \dots \eta_0$ be a word over the alphabet $\mathcal{D}$. Its *(Hamming) weight* is the cardinality of $\{j : \eta_j \neq 0\}$, i.e., the number of non-zero digits in $\boldsymbol{\eta}$.

Let $z = \mathsf{value}(\boldsymbol{\eta})$. The expansion $\boldsymbol{\eta}$ is said to be *optimal* if it minimises the weight among all possible expansions of z, i.e., if the weight of $\boldsymbol{\eta}$ is at most the weight of $\boldsymbol{\xi}$ for all words $\boldsymbol{\xi}$ over $\mathcal{D}$ with $\mathsf{value}(\boldsymbol{\xi}) = z$.

We will show an optimality result for $\mathcal{D}$-w-NAFs in Theorem 3.5.4, where the digit set comes from a tiling as in Section 3.3.

Lemma 3.5.2. *We have*

$$\lim_{m\to\infty} \Phi^m(\Lambda) := \bigcap_{m\in\mathbb{N}_0} \Phi^m(\Lambda) = \{0\}.$$

Proof. Let $\alpha \in \lim_{m\to\infty} \Phi^m(\Lambda) = \bigcap_{m\in\mathbb{N}_0} \Phi^m(\Lambda)$. Then there is a sequence $(\beta_m)_{m\in\mathbb{N}_0}$, all $\beta_m \in \Lambda$ and with $\beta_m = \Phi^{-m}(\alpha)$. As Φ is expanding, we obtain $\beta_m \to 0$ as m tends to infinity. The lattice Λ is discrete, so $\beta_m = 0$ for sufficiently large m. We conclude that $\alpha = 0$. □

Now we define the digit set: We start with a subset V of $\mathbb{R}^n$ tiling $\mathbb{R}^n$ by Λ. For a positive integer w let

$$\widetilde{\mathcal{D}} := \{\alpha \in \Lambda : \Phi^{-w}(\alpha) \in V\}$$

and

$$\widetilde{\mathcal{D}}_{\text{int}} := \{\alpha \in \Lambda : \Phi^{-w}(\alpha) \in \text{int}(V)\},$$

where $\mathrm{int}(V)$ denotes the interior of V. We choose a subset $\mathcal{D}^\bullet$ of $\widetilde{\mathcal{D}}$ in such a way that $\mathcal{D}^\bullet$ contains exactly one representative of every residue class modulo $\Phi^w(\Lambda)$ which is not contained in $\Phi(\Lambda)$. We also set $\mathcal{D} := \mathcal{D}^\bullet \cup \{0\}$. This is the same construction as in Section 3.3.

Lemma 3.5.3. *Assume that $V \subseteq \Phi(V)$. Then each element of $\widetilde{\mathcal{D}}_{\mathrm{int}} \setminus \{0\}$ has an expansion of weight 1.*

Proof. Let $\alpha \in \widetilde{\mathcal{D}}_{\mathrm{int}} \setminus \{0\}$, and let $\beta = \Phi^{-\ell}(\alpha) \in \Lambda$ such that the non-negative integer ℓ is maximal. Therefore $\beta \notin \Phi(\Lambda)$. We have that $\Phi^{-w}(\beta) = \Phi^{-w-\ell}(\alpha)$ is in the interior of $\Phi^{-\ell}(V)$. Using $V \subseteq \Phi(V)$ yields $\Phi^{-w}(\beta) \in \mathrm{int}(V)$, and therefore, by Lemma 3.3.2, $\beta \in \mathcal{D}^\bullet$. Thus $\alpha = \Phi^\ell(\beta)$ has an expansion of weight 1. □

Theorem 3.5.4. *Assume that $V \subseteq \Phi(V)$, $V = -V$ and that there are a vector norm $\|\cdot\|$ on $\mathbb{R}^n$ and positive reals r and R such that*

$$\{x \in \mathbb{R}^n : \|x\| \le r\} \subseteq V \subseteq \{x \in \mathbb{R}^n : \|x\| \le R\} \tag{3.5.1}$$

and such that the induced operator norm (also denoted by $\|\cdot\|$) fulfils $\|\Phi^{-1}\| < \frac{r}{R}$.

If w is a positive integer such that

$$\|\Phi^{-1}\|^w < \frac{1}{2}\left(\frac{r}{R} - \|\Phi^{-1}\|\right) \tag{3.5.2}$$

and $\mathcal{D}$ is a w-NADS, then the $\mathcal{D}$-w-NAF-expansion of each element of Λ is optimal.

The proof relies on the following optimality result.

Theorem (Heuberger and Krenn [48])**.** *If*

$$\lim_{m\to\infty} \Phi^m(\Lambda) = \{0\}, \tag{3.5.3}$$

and if there are sets U and S such that $\mathcal{D} \subseteq U$, $-\mathcal{D} \subseteq U$, $U \subseteq \Phi(U)$, all elements in $S \cap \Lambda$ are singletons (have expansions of weight 1) and if

$$\left(\Phi^{-1}(U) + \Phi^{-w}(U) + \Phi^{-w}(U)\right) \cap \Lambda \subseteq S \cup \{0\},$$

then every $\mathcal{D}$-w-NAF is optimal.

Proof of Theorem 3.5.4. Condition (3.5.3) is shown in Lemma 3.5.2. For the second condition, we choose $U = \Phi^w(V)$ and $S = \Phi^w(\mathrm{int}(V)) \setminus \{0\}$, and we show

$$\left(\Phi^{-1}(V) + \Phi^{-w}(V) + \Phi^{-w}(V)\right) \subseteq \mathrm{int}(V)\,.$$

Optimality then follows, since each element in $S \cap \Lambda$ has a weight 1 expansion by Lemma 3.5.3. So let z be an element of the left hand side of the inclusion above. Using (3.5.1) and (3.5.2) yields

$$\|z\| \le \|\Phi^{-1}\|R + 2\|\Phi^{-1}\|^w R < r,$$

therefore z is in the interior of V. □

Chapter 4

Analysis of the Width-w Non-Adjacent Form

This chapter contains the article [68] with the title "Analysis of the Width-w Non-Adjacent Form in Conjunction with Hyperelliptic Curve Cryptography and with Lattices". The article is submitted to *Theoretical Computer Science*. An introduction to this chapter can be found in Chapter 1, in particular Section 1.5.

Abstract

We analyse the number of occurrences of a fixed non-zero digit in the width-w non-adjacent forms of all elements of a lattice in some region (e.g. a ball). Our result is an asymptotic formula, where its main term coincides with the full block length analysis. In its second order term a periodic fluctuation is exhibited. The proof follows Delange's method. This results in a general lattice set-up, which is then used for numeral systems with an algebraic integer as base. Those come from efficient scalar multiplication methods (Frobenius-and-add methods) in hyperelliptic curves cryptography, and our result is needed for analysing the running time of such algorithms.

4.1 Non-Adjacent Forms

This section is devoted to the formal introduction of width-w non-adjacent forms. Let Λ be an Abelian group, Φ an injective endomorphism of Λ and w a positive integer. Later, starting with the next section, the group Λ will be a lattice with the usual addition of lattice points.

We start with the definition of the digit set used throughout this article.

Definition 4.1.1 (Reduced Residue Digit Set)**.** Let $\mathcal{D} \subseteq \Lambda$. The set $\mathcal{D}$ is called a *reduced residue digit set modulo* Φ^w, if it is consists of 0 and exactly one representative for each residue class of Λ modulo $\Phi^w\Lambda$ that is not contained in $\Phi\Lambda$.

Next we define the syntactic condition of our expansions. This syntax is used to get unique expansions, because our numeral systems are redundant.

Definition 4.1.2 (Width-w Non-Adjacent Forms)**.** Let $\boldsymbol{\eta} = (\eta_j)_{j\in\mathbb{Z}} \in \mathcal{D}^{\mathbb{Z}}$. The sequence $\boldsymbol{\eta}$ is called a *width-w non-adjacent form*, or *w-NAF* for short, if each factor $\eta_{j+w-1}\dots\eta_j$, i.e., each block of width w, contains at most one non-zero digit.

Let $J := \{j \in \mathbb{Z} : \eta_j \neq 0\}$. We call $\sup(\{0\} \cup (J+1))$, where $J+1 = \{j+1 : j \in J\}$, the *left-length of the w-NAF $\boldsymbol{\eta}$* and $-\inf(\{0\} \cup J)$ the *right-length of the w-NAF $\boldsymbol{\eta}$*. Let ℓ and r be elements of $\mathbb{N}_0 \cup \{\mathsf{fin}, \infty\}$, where fin means finite. We denote the *set of all w-NAFs of left-length at most ℓ and right-length at most r* by $\mathbf{NAF}_w^{\ell.r}$. The elements of the set $\mathbf{NAF}_w^{\mathsf{fin}.0}$ will be called *integer w-NAFs*. The *most-significant digit* of a $\boldsymbol{\eta} \in \mathbf{NAF}_w^{\mathsf{fin}.\infty}$ is the digit $\eta_j \neq 0$, where j is chosen maximally with that property.

For $\boldsymbol{\eta} \in \mathbf{NAF}_w^{\mathsf{fin}.\infty}$ we call

$$\mathsf{value}(\boldsymbol{\eta}) := \sum_{j\in\mathbb{Z}} \Phi^j \eta_j$$

the *value of the w-NAF $\boldsymbol{\eta}$*.

The following notations and conventions are used. A block of digits zero is denoted by $\mathbf{0}$. For a digit η and $k \in \mathbb{N}_0$ we will use

$$\eta^k := \underbrace{\eta\dots\eta}_{k},$$

with the convention $\eta^0 := \varepsilon$, where ε denotes the empty word. A w-NAF $\boldsymbol{\eta} = (\eta_j)_{j\in\mathbb{Z}}$ will be written as $\boldsymbol{\eta}_I.\boldsymbol{\eta}_F$, where $\boldsymbol{\eta}_I$ contains the η_j with $j \geq 0$ and $\boldsymbol{\eta}_F$ contains the η_j with $j < 0$. $\boldsymbol{\eta}_I$ is called *integer part*, $\boldsymbol{\eta}_F$ *fractional part*, and the dot is called *Φ-point*. Left-leading zeros in $\boldsymbol{\eta}_I$ can be skipped, except η_0, and right-trailing zeros in $\boldsymbol{\eta}_F$ can be skipped as well. If $\boldsymbol{\eta}_F$ is a sequence containing only zeros, the Φ-point and this sequence are not drawn.

Further, for a w-NAF $\boldsymbol{\eta}$ (a bold, usually small Greek letter) we will always use η_j (the same letter, but indexed and not bold) for the elements of the sequence.

The set $\mathbf{NAF}_w^{\mathsf{fin}.\infty}$ can be equipped with a metric. It is defined in the following way. Let $\rho > 1$. For $\boldsymbol{\eta} \in \mathbf{NAF}_w^{\mathsf{fin}.\infty}$ and $\boldsymbol{\xi} \in \mathbf{NAF}_w^{\mathsf{fin}.\infty}$ define

$$\mathrm{d}_{\mathsf{NAF}}(\boldsymbol{\eta}, \boldsymbol{\xi}) := \begin{cases} \rho^{\max\{j\in\mathbb{Z}\,:\,\eta_j\neq\xi_j\}} & \text{if } \boldsymbol{\eta} \neq \boldsymbol{\xi}, \\ 0 & \text{if } \boldsymbol{\eta} = \boldsymbol{\xi}. \end{cases}$$

So the largest index, where the two w-NAFs differ, decides their distance. See for example Edgar [34] for details on such metrics.

We get a compactness result on the metric space $\mathbf{NAF}_w^{\ell.\infty} \subseteq \mathbf{NAF}_w^{\mathsf{fin}.\infty}$, $\ell \in \mathbb{N}_0$, see the proposition below. The metric space $\mathbf{NAF}_w^{\mathsf{fin}.\infty}$ is not compact, because if we fix a non-zero digit η, then the sequence $\left(\eta 0^j\right)_{j\in\mathbb{N}_0}$ has no convergent subsequence, but all $\eta 0^j$ are in the set $\mathbf{NAF}_w^{\mathsf{fin}.\infty}$.

Proposition 4.1.3. *For every $\ell \geq 0$ the metric space $\left(\mathbf{NAF}_w^{\ell.\infty}, \mathrm{d}_{\mathsf{NAF}}\right)$ is compact.*

This is a consequence of Tychonoff's Theorem, see [49] for details.

4.2 The Set-Up and Notations

In this section we describe the set-up, which we use throughout this article.

1. Let Λ be a lattice in $\mathbb{R}^n$ with full rank, i.e., $\Lambda = w_1\mathbb{Z} \oplus \cdots \oplus w_n\mathbb{Z}$ for linearly independent $w_1, \ldots, w_n \in \mathbb{R}^n$.

2. Let $n \in \mathbb{N}$ and Φ be an endomorphism of $\mathbb{R}^n$ with $\Phi(\Lambda) \subseteq \Lambda$. We assume that each eigenvalue of Φ has the same absolute value ρ, where ρ is a fixed real constant with $\rho > 1$. Further we assume that $\rho^n \in \mathbb{N}$. Additionally, we take this ρ as parameter in the definition of the metric $\mathrm{d}_{\mathsf{NAF}}$.

3. Suppose that the set $T \subseteq \mathbb{R}^n$ tiles the space $\mathbb{R}^n$ by the lattice Λ, i.e., the following two properties hold:

   a) $\bigcup_{z\in\Lambda}(z + T) = \mathbb{R}^n$,

   b) $T \cap (z + T) \subseteq \partial T$ holds for all $z \in \Lambda$ with $z \neq 0$.

   Further, we assume that T is closed and that $\lambda(\partial T) = 0$, where λ denotes the n-dimensional Lebesgue measure. We set $d_\Lambda := \lambda(T)$.

4. Let $\|\cdot\|$ be a vector norm on $\mathbb{R}^n$ such that for the corresponding induced operator norm, also denoted by $\|\cdot\|$, the equalities $\|\Phi\| = \rho$ and $\left\|\Phi^{-1}\right\| = \rho^{-1}$ hold.

   For a $z \in \Lambda$ and non-negative $r \in \mathbb{R}$ the *open ball with centre z and radius r* is denoted by

   $$\mathcal{B}(z, r) := \{y \in \Lambda \,:\, \|z - y\| < r\}$$

   and the *closed ball with centre z and radius r* by

   $$\overline{\mathcal{B}}(z, r) := \{y \in \Lambda \,:\, \|z - y\| \leq r\}\,.$$

5. Let r and R be positive reals with

   $$\overline{\mathcal{B}}(0, r) \subseteq T \subseteq \overline{\mathcal{B}}(0, R)\,. \tag{4.2.1}$$

6. Let w be a positive integer such that

   $$\frac{R}{r} < \rho^w - 1. \tag{4.2.2}$$

7. Let $\mathcal{D}$ be a reduced residue digit set modulo Φ^w, cf. Definition 4.1.1, corresponding to the tiling T, i.e. the digit set $\mathcal{D}$ fulfils $\mathcal{D} \subseteq \Phi^w T$.

   Further, suppose that the cardinality of the digit set $\mathcal{D}$ is

   $$\rho^{n(w-1)}\,(\rho^n - 1) + 1.$$

We use the following notation concerning our tiling: for a lattice element $z \in \Lambda$ we set $T_z := z + T$. Therefore $\bigcup_{z\in\Lambda} T_z = \mathbb{R}^n$ and $T_y \cap T_z \subseteq \partial T_z$ for all distinct $y, z \in \Lambda$.

Next we define a fractional part function in $\mathbb{R}^n$ with respect to the lattice Λ, which should be a generalisation of the usual fractional part of elements in $\mathbb{R}$ with respect to the rational integers $\mathbb{Z}$. Our tiling T induces such a fractional part.

Definition 4.2.1 (Fractional Part)**.** Let $\widetilde{T}$ be a tiling arising from T in the following way: Restrict the set $\widetilde{T} \subseteq T$ such that it fulfils $\biguplus_{z\in\Lambda}(z + \widetilde{T}) = \mathbb{R}^n$.

For $z \in \mathbb{R}^n$ with $z = u + v$, where $u \in \Lambda$ and $v \in \widetilde{T}$ define the *fractional part corresponding to the lattice* Λ by $\{z\}_\Lambda := v$.

Note that this fractional part depends on the tiling T (or more precisely, on the tiling $\widetilde{T}$). We omit this dependency, since we assume that our tiling is fixed.

4.3 Some Basic Properties and some Remarks

The previous section contained our set-up. Some basic implications of that set-up are now given in this section. Further we give remarks on the tilings and on the digit sets used, and there are also comments on the existence of w-NAF-expansions in the lattice.

We start with two remarks on our mapping Φ.

Remark 4.3.1. Since all eigenvalues of Φ have an absolute value larger than 1, the function Φ is injective. Note that we already assumed injectivity of the endomorphism Φ in the basic definitions given in Section 4.1.

Remark 4.3.2. We have assumed $\|\Phi\| = \rho$ and $\left\|\Phi^{-1}\right\| = \rho^{-1}$. Therefore, for all $J \in \mathbb{Z}$ the equality $\left\|\Phi^J\right\| = \rho^J$ follows.

Remark 4.3.3. The endomorphism Φ is diagonalisable. This follows from the assumptions that all eigenvalues have the same absolute value ρ and the existence of a norm with $\|\Phi\| = \rho$.

One special tiling comes from the Voronoi diagram of the lattice. This is stated in the remark below.

Remark 4.3.4. Let

$$V := \{z \in \mathbb{R}^n \,:\, \forall y \in \Lambda : \|z\| \leq \|z - y\|\}\,.$$

We call V the *Voronoi cell for* 0 corresponding to the lattice Λ. Let $u \in \Lambda$. We define the *Voronoi cell for* u as $V_u := u + V$.

Now choosing $T = V$ results in a tiling of the $\mathbb{R}^n$ by the lattice Λ.

In our set-up the digit set corresponds to the tiling. In Remark 4.3.5 this is explained in more detail. The Voronoi tiling mentioned above gives rise to a special digit set, namely the minimal norm digit set. There, for each digit a representative of minimal norm is chosen.

Remark 4.3.5. The condition $\frac{R}{r} < \rho^w - 1$ in the our set-up implies the existence of w-NAFs: each element of Λ has a unique w-NAF-expansion with the digit set $\mathcal{D}$. See Heuberger and Krenn [50] for details. There numeral systems in lattices with w-NAF-condition and digit sets coming from tilings are explained in detail. Further it is shown that each tiling and positive integer w give rise to a digit set $\mathcal{D}$.

Because $\mathcal{D} \subseteq \Phi^w T$, we have

$$\rho^w r \le \|d\| \le \rho^w R$$

for each non-zero digit $d \in \mathcal{D}$.

Further, we get the following continuity result.

Proposition 4.3.6. *The value function* value *is Lipschitz continuous on* $\mathbf{NAF}_w^{\mathsf{fin},\infty}$.

This result is a consequence of the boundedness of the digit set, see [49] for a formal proof.

We need the full block length distribution theorem from Heuberger and Krenn [49]. This was proved for numeral systems with algebraic integer τ as base. But the result does not depend on τ directly, only on the size of the digit set, which is dependent on the norm of τ. In our case this norm equals ρ^n. That replacement is already done in the theorem written down below.

Theorem 4.3.7 (Full Block Length Distribution Theorem)**.** *Denote the number of w-NAFs of length $m \in \mathbb{N}_0$ by C_m. We get*

$$C_m = \frac{1}{(\rho^n - 1)w + 1} \rho^{n(m+w)} + \mathcal{O}((\mu\rho^n)^m)\,,$$

where $\mu = (1 + \frac{1}{\rho^n w^3})^{-1} < 1$.

Further let $0 \neq \eta \in \mathcal{D}$ be a fixed digit and define the random variable $X_{m,\eta}$ to be the number of occurrences of the digit η in a random w-NAF of length m, where every w-NAF of length m is assumed to be equally likely. Then we get

$$\mathbb{E}(X_{m,\eta}) = Em + \mathcal{O}(1)$$

for the expectation, where

$$E = \frac{1}{\rho^{n(w-1)}((\rho^n - 1)w + 1)}.$$

The theorem in [49] gives more details, which we do not need for the results in this article: We have

$$\mathbb{E}(X_{m,\eta}) = Em + E_0 + \mathcal{O}(m\mu^m)$$

with an explicit constant term E_0. Further the variance

$$\mathbb{V}(X_{n,w,\eta}) = Vm + V_0 + \mathcal{O}\big(m^2\mu^m\big)$$

with explicit constants V and V_0 is calculated, and a central limit theorem is proved.

4.4 Bounds for the Value of Non-Adjacent Forms

In this section we have a closer look at the value of a w-NAF. We want to find upper bounds, as well as, a lower bound for it. In the proofs of all those bounds we use bounds for the norm $\|\cdot\|$. More precisely, geometric parameters of the tiling T, i.e., the already defined reals r and R, are used.

The following proposition deals with three upper bounds, one for the norm of the value of a w-NAF-expansion and two give us bounds in conjunction with the tiling.

Proposition 4.4.1 (Upper Bounds)**.** *Let $\boldsymbol{\eta} \in \mathbf{NAF}_w^{\mathsf{fin}.\infty}$, and denote the position of the most significant digit of $\boldsymbol{\eta}$ by J. Let*

$$B_U = \frac{\rho^w R}{1-\rho^{-w}}.$$

Then the following statements are true:

(a) We get

$$\|\mathsf{value}(\boldsymbol{\eta})\| \leq \rho^J B_U$$

(b) We have

$$\mathsf{value}(\boldsymbol{\eta}) \in \bigcup_{z \in \Phi^{w+J}T} \overline{\mathcal{B}}\big(z, \rho^{-w+J} B_U\big)\,.$$

(c) We get

$$\mathsf{value}(\boldsymbol{\eta}) \in \Phi^{2w+J}T.$$

(d) For each $\ell \in \mathbb{N}_0$ we have

$$\mathsf{value}(0.\eta_{-1}\dots\eta_{-\ell}) + \Phi^{-\ell}T \subseteq \Phi^{2w-1}T.$$

Note that $\rho^J = \mathrm{d}_{\mathsf{NAF}}(\boldsymbol{\eta}, \mathbf{0})$, so we can rewrite the statements of the proposition above in terms of that metric, see also Corollary 4.4.3.

Proof. (a) In the calculations below, we use the Iversonian notation $[\text{expr}] = 1$ if expr is true and $[\text{expr}] = 0$ otherwise, cf. Graham, Knuth and Patashnik [45].

The result follows trivially for $\boldsymbol{\eta} = \mathbf{0}$. First assume that the most significant digit of $\boldsymbol{\eta}$ is at position 0. Since $\|\eta_{-j}\| \leq \rho^w R$, see Remark 4.3.5 on the preceding page, $\rho > 1$ and $\boldsymbol{\eta}$ is fulfilling the w-NAF-condition we obtain

$$\begin{aligned}
\|\mathsf{value}(\boldsymbol{\eta})\| = \left\|\sum_{j=0}^{\infty} \Phi^{-j}\eta_{-j}\right\| &\leq \sum_{j=0}^{\infty} \left\|\Phi^{-1}\right\|^j \|\eta_{-j}\| = \sum_{j=0}^{\infty} \rho^{-j}\,\|\eta_{-j}\| \\
&\leq \rho^w R \sum_{j=0}^{\infty} \rho^{-j}\,[\eta_{-j} \neq 0] \leq \rho^w R \sum_{j=0}^{\infty} \rho^{-j}\,[-j \equiv 0 \pmod w] \\
&= \rho^w R \sum_{k=0}^{\infty} \rho^{-wk} = \frac{\rho^w R}{1-\rho^{-w}} = B_U.
\end{aligned}$$

In the general case we have the most significant digit of $\boldsymbol{\eta}$ at a position J. Then $\mathsf{value}(\boldsymbol{\eta}) = \Phi^J \mathsf{value}(\boldsymbol{\eta}')$ for a w-NAF $\boldsymbol{\eta}'$ with most significant digit at position 0. Therefore

$$\|\mathsf{value}(\boldsymbol{\eta})\| = \left\|\Phi^J \mathsf{value}\big(\boldsymbol{\eta}'\big)\right\| \leq \|\Phi\|^J \left\|\mathsf{value}\big(\boldsymbol{\eta}'\big)\right\| \leq \rho^J B_U,$$

which was to prove.

(b) There is nothing to show if the w-NAF $\boldsymbol{\eta}$ is zero. First suppose that the most significant digit is at position w. Then, using (a), we have

$$\|\mathsf{value}(\boldsymbol{\eta}) - \Phi^w \eta_w\| \leq B_U,$$

therefore

$$\mathsf{value}(\boldsymbol{\eta}) \in \overline{\mathcal{B}}(\Phi^w \eta_w, B_U)\,.$$

Since $\eta_w \in \Phi^w T$, the statement follows for the special case. The general case is again obtained by shifting.

(c) Using the upper bound found in (a) and the assumption (4.2.2) yields

$$\|\mathsf{value}(\boldsymbol{\eta})\| \leq \rho^J B_U = \rho^J \frac{\rho^w R}{1-\rho^{-w}} \leq r\rho^{2w+J}.$$

Since $\overline{\mathcal{B}}\big(0, r\rho^{2w+J}\big) \subseteq \Phi^{2w+J}T$, the statement follows.

(d) Analogously to the proof of (a), except that we use ℓ for the upper bound of the sum, we obtain for $v \in T$

$$\begin{aligned}
\left\|\mathsf{value}(0.\eta_{-1}\dots\eta_{-\ell}) + \Phi^{-\ell}v\right\| &\leq \|\mathsf{value}(0.\eta_{-1}\dots\eta_{-\ell})\| + \rho^{-\ell}R \\
&\leq \rho^{-1}\frac{\rho^w R}{1-\rho^{-w}}\left(1-\rho^{-w\left\lfloor\frac{\ell-1+w}{w}\right\rfloor}\right) + \rho^{-\ell}R \\
&\leq \frac{\rho^{w-1}R}{1-\rho^{-w}}\left(1-\rho^{-\ell+1-w}+\rho^{-\ell+1-w}\left(1-\rho^{-w}\right)\right) \\
&= \frac{\rho^{w-1}R}{1-\rho^{-w}}\left(1-\rho^{-\ell+1-2w}\right).
\end{aligned}$$

Since $1-\rho^{-\ell+1-2w} < 1$ we get

$$\left\|\mathsf{value}(0.\eta_{-1}\dots\eta_{-\ell}) + \Phi^{-\ell}T\right\| \leq \rho^{-1}\frac{\rho^w R}{1-\rho^{-w}} = \rho^{-1}B_U$$

for all $\ell \in \mathbb{N}_0$. By the same argumentation as in the proof of (c), the statement follows. □

Next we want to find a lower bound for the value of a w-NAF. Clearly the w-NAF $\mathbf{0}$ has value 0, so we are interested in cases where we have a non-zero digit somewhere.

Proposition 4.4.2 (Lower Bound)**.** *Let $\boldsymbol{\eta} \in \mathbf{NAF}_w^{\mathsf{fin},\infty}$ be non-zero, and denote the position of the most significant digit of $\boldsymbol{\eta}$ by J. Then we have*

$$\|\mathsf{value}(\boldsymbol{\eta})\| \geq \rho^J B_L,$$

where

$$B_L = r - \rho^{-2w} B_U = r - \frac{R}{\rho^w - 1}.$$

Note that $B_L > 0$ is equivalent to $\frac{R}{r} < \rho^w - 1$, i.e. the assumption (4.2.2). Moreover, we have

$$\frac{R}{r - B_L} = \rho^w - 1.$$

Proof of Proposition 4.4.2. First suppose the most significant digit of the w-NAF $\boldsymbol{\eta}$ is at position 0 and the second non-zero digit (read from left to right) at position J. Then

$$\mathsf{value}(\boldsymbol{\eta}) - \eta_0 = \sum_{k=w}^{\infty} \Phi^{-k} \eta_{-k} \in \bigcup_{z \in T} \overline{\mathcal{B}}\big(z, \rho^{-w+J} B_U\big) \subseteq \bigcup_{z \in T} \overline{\mathcal{B}}\big(z, \rho^{-2w} B_U\big)$$

according to (b) of Proposition 4.4.1 on page 56. Therefore

$$\mathsf{value}(\boldsymbol{\eta}) \in \bigcup_{z \in T_{\eta_0}} \overline{\mathcal{B}}\big(z, \rho^{-2w} B_U\big)\,.$$

This means that $\mathsf{value}(\boldsymbol{\eta})$ is in T_{η_0} or in a $\rho^{-2w} B_U$-strip around this cell. The two tiling cells T_{η_0} for η_0 and $T_0 = T$ for 0 are disjoint, except for parts of the boundary, if they are adjacent. Since a ball with radius r is contained in each tiling cell, we deduce that

$$\|\mathsf{value}(\boldsymbol{\eta})\| \geq r - \rho^{-2w} B_U = r - \frac{R}{\rho^w - 1} = B_L,$$

which was to show. The case of a general J is again, as in the proof of Proposition 4.4.1 obtained by shifting. □

Combining the previous two propositions leads to the following corollary, which gives an upper and a lower bound for the norm of the value of a w-NAF by looking at the largest non-zero index.

Corollary 4.4.3 (Bounds for the Value)**.** *Let $\boldsymbol{\eta} \in \mathbf{NAF}_w^{\mathsf{fin},\infty}$, then we get*

$$\mathrm{d}_{\mathsf{NAF}}(\boldsymbol{\eta}, \mathbf{0})\, B_L \leq \|\mathsf{value}(\boldsymbol{\eta})\| \leq \mathrm{d}_{\mathsf{NAF}}(\boldsymbol{\eta}, \mathbf{0})\, B_U.$$

Proof. Follows directly from Proposition 4.4.1 on page 56 and Proposition 4.4.2, since the term ρ^J is equal to $\mathrm{d}_{\mathsf{NAF}}(\boldsymbol{\eta}, \mathbf{0})$. □

Last in this section, we want to find out if there are special w-NAFs for which we know for sure that all their expansions start with a certain finite w-NAF. This is formulated in the following lemma.

Lemma 4.4.4. *There is a $k_0 \in \mathbb{N}_0$ such that for all $k \geq k_0$ the following holds: If $\boldsymbol{\eta} \in \mathbf{NAF}_w^{0.\infty}$ starts with the word 0^k, i.e., $\eta_{-1} = 0, \ldots, \eta_{-k} = 0$, then we get for all $\boldsymbol{\xi} \in \mathbf{NAF}_w^{\mathsf{fin}.\infty}$ that $\mathsf{value}(\boldsymbol{\xi}) = \mathsf{value}(\boldsymbol{\eta})$ implies $\boldsymbol{\xi} \in \mathbf{NAF}_w^{0.\infty}$.*

Proof. Let $\boldsymbol{\xi} = \boldsymbol{\xi}_I.\boldsymbol{\xi}_F$. Then $\|\mathsf{value}(\boldsymbol{\xi}_I.\boldsymbol{\xi}_F)\| < B_L$ implies $\boldsymbol{\xi}_I = \mathbf{0}$, cf. Corollary 4.4.3. Further, for our $\boldsymbol{\eta}$ we obtain $z = \|\mathsf{value}(\boldsymbol{\eta})\| \leq \rho^{-k} B_U$. So it is sufficient to show that

$$\rho^{-k} B_U < B_L,$$

which is equivalent to

$$k > \log_\rho \frac{B_U}{B_L}.$$

We obtain

$$k > 2w - \log_\rho \left(\frac{r}{R} \left(\rho^w - 1 \right) - 1 \right),$$

where we just inserted the formulas for B_U and B_L. Choosing an appropriate k_0 is now easily possible. □

Note that we can find a constant k_1 independent from w such that for all $k \geq 2w + k_1$ the assertion of Lemma 4.4.4 holds. This can be seen in the proof, since $\frac{r}{R} \left(\rho^w - 1 \right) - 1$ is monotonically increasing in w.

4.5 Right-infinite Expansions

We have the existence of a (finite integer) w-NAF-expansion for each element of the lattice $\Lambda \subseteq \mathbb{R}^n$, cf. Remark 4.3.5. But that existence condition is also sufficient to get w-NAF-expansions for all elements in $\mathbb{R}^n$. Those expansions possibly have an infinite right-length. The aim of this section is to show that result. The proofs themselves are a minor generalisation of the ones given in [49] for the quadratic case.

We will use the following abbreviation in this section. We define

$$[\Phi^{-1}]\Lambda := \bigcup_{j \in \mathbb{N}_0} \Phi^{-j} \Lambda.$$

Note that $\Lambda \subseteq \Phi^{-1}\Lambda$.

To prove the existence theorem of this section, we need the following three lemmata.

Lemma 4.5.1. *The function $\mathsf{value}|_{\mathbf{NAF}_w^{\mathsf{fin}.\mathsf{fin}}}$ is injective.*

Proof. Let $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ be elements of $\mathbf{NAF}_w^{\mathsf{fin}.\mathsf{fin}}$ with $\mathsf{value}(\boldsymbol{\eta}) = \mathsf{value}(\boldsymbol{\xi})$. This implies that $\Phi^J \mathsf{value}(\boldsymbol{\eta}) = \Phi^J \mathsf{value}(\boldsymbol{\xi}) \in \Lambda$ for some $J \in \mathbb{Z}$. By uniqueness of the integer w-NAFs we conclude that $\boldsymbol{\eta} = \boldsymbol{\xi}$. □

Lemma 4.5.2. *We have $\mathsf{value}\left(\mathbf{NAF}_w^{\mathsf{fin}.\mathsf{fin}}\right) = [\Phi^{-1}]\Lambda$.*

Proof. Let $\boldsymbol{\eta} \in \mathbf{NAF}_w^{\mathsf{fin}.\mathsf{fin}}$. There are only finitely many $\eta_j \neq 0$, so there is a $J \in \mathbb{N}_0$ such that $\mathsf{value}(\boldsymbol{\eta}) \in \Phi^{-J}\Lambda$. Conversely, if $z \in \Phi^{-J}\Lambda$, then there is an integer w-NAF of $\Phi^J z$, and therefore, there is a $\boldsymbol{\xi} \in \mathbf{NAF}_w^{\mathsf{fin}.\mathsf{fin}}$ with $\mathsf{value}(\boldsymbol{\xi}) = z$. □

Lemma 4.5.3. *$[\Phi^{-1}]\Lambda$ is dense in $\mathbb{R}^n$.*

Proof. Let $\Lambda = w_1\mathbb{Z} \oplus \cdots \oplus w_n\mathbb{Z}$ for linearly independent $w_1, \ldots, w_n \in \mathbb{R}^n$. Let $z \in \mathbb{R}^n$ and $K \in \mathbb{N}_0$. Then $\Phi^K z = z_1 w_1 + \cdots + z_n w_n$ for some reals $z_1, \ldots, z_n$. We have

$$\left\| z - \left(\lfloor z_1 \rfloor \, \Phi^{-K} w_1 + \cdots + \lfloor z_n \rfloor \, \Phi^{-K} w_n \right) \right\| < \rho^{-K} \left(\|w_1\| + \cdots + \|w_n\| \right),$$

which proves the lemma. □

Now we can prove the following theorem.

Theorem 4.5.4 (Existence Theorem concerning $\mathbb{R}^n$)**.** *Let $z \in \mathbb{R}^n$. Then there is an $\boldsymbol{\eta} \in \mathbf{NAF}_w^{\mathsf{fin},\infty}$ such that $z = \mathsf{value}(\boldsymbol{\eta})$, i.e., each element in $\mathbb{R}^n$ has a w-NAF-expansion.*

Proof. By Lemma 4.5.3, there is a sequence $z_n \in [\Phi^{-1}]\Lambda$ converging to z. By Lemma 4.5.2 on the previous page, there is a sequence $\boldsymbol{\eta}_n \in \mathbf{NAF}_w^{\mathsf{fin},\mathsf{fin}}$ with $\mathsf{value}(\boldsymbol{\eta}_n) = z_n$ for all n. By Corollary 4.4.3 on page 58 the sequence $\mathrm{d}_{\mathsf{NAF}}(\boldsymbol{\eta}_n, 0)$ is bounded from above, so there is an ℓ such that $\boldsymbol{\eta}_n \in \mathbf{NAF}_w^{\ell,\mathsf{fin}} \subseteq \mathbf{NAF}_w^{\ell,\infty}$. By Proposition 4.1.3 on page 52, we conclude that there is a convergent subsequence $\boldsymbol{\eta}'_n$ of $\boldsymbol{\eta}_n$. Set $\boldsymbol{\eta} := \lim_{n\to\infty} \boldsymbol{\eta}'_n$. By continuity of value, see Proposition 4.3.6 on page 55, we conclude that $\mathsf{value}(\boldsymbol{\eta}) = z$. □

4.6 The Fundamental Domain

We now derive properties of the *Fundamental Domain*, i.e., the subset of $\mathbb{R}^n$ representable by w-NAFs which vanish left of the Φ-point. The boundary of the fundamental domain is shown to correspond to elements which admit more than one w-NAFs differing left of the Φ-point. Finally, an upper bound for the Hausdorff dimension of the boundary is derived.

All the results in this section are generalisations of the propositions and remarks found in [49]. For some of those results given here, the proof is the same as in the quadratic case or a straightforward generalisation of it. In those cases the proofs will be skipped.

We start with the formal definition of the fundamental domain.

Definition 4.6.1 (Fundamental Domain)**.** The set

$$\mathcal{F} := \mathsf{value}\big(\mathbf{NAF}_w^{0,\infty}\big) = \left\{ \mathsf{value}(\boldsymbol{\xi}) \,:\, \boldsymbol{\xi} \in \mathbf{NAF}_w^{0,\infty} \right\}.$$

is called *fundamental domain*.

The pictures in Figure 4.8.1 on page 68 show some fundamental domains for lattices coming from imaginary-quadratic algebraic integers τ. We continue with some properties of fundamental domains. We have the following compactness result.

Proposition 4.6.2. *The fundamental domain $\mathcal{F}$ is compact.*

Proof. The proof is a straightforward generalisation of the proof of the quadratic case in [49]. □

We can also compute the Lebesgue measure of the fundamental domain. This result can be found in Remark 4.8.3 on page 69. To calculate $\lambda(\mathcal{F})$, we will need the results of Sections 4.7 and 4.8.

The space $\mathbb{R}^n$ has a tiling property with respect to the fundamental domain. This fact is stated in the following proposition.

Proposition 4.6.3 (Tiling Property)**.** *The space $\mathbb{R}^n$ can be tiled with scaled versions of the fundamental domain $\mathcal{F}$. Only finitely many different sizes are needed. More precisely: Let $K \in \mathbb{Z}$, then*

$$\mathbb{R}^n = \bigcup_{\substack{k\in\{K,K+1,\dots,K+w-1\} \\ \boldsymbol{\xi}\in\mathbf{NAF}_w^{\text{fin}.0} \\ k \neq K+w-1 \text{ implies } \boldsymbol{\xi}_0 \neq 0}} \left(\Phi^k \,\mathsf{value}(\boldsymbol{\xi}) + \Phi^{k-w+1}\mathcal{F}\right),$$

and the intersection of two different $\Phi^k \,\mathsf{value}(\boldsymbol{\xi}) + \Phi^{k-w+1}\mathcal{F}$ in this union is a subset of the intersection of their boundaries.

Proof. The proof is a straightforward generalisation of the proof of the quadratic case in [49]. □

Note that the intersection of the two different sets of the tiling in the previous corollary has Lebesgue measure 0. This will be a consequence of Proposition 4.6.6 on the next page.

Remark 4.6.4 (Iterated Function System). Define $f_0(z) = \Phi^{-1}z$ and for a non-zero digit $\vartheta \in \mathcal{D}^\bullet$ define $f_\vartheta(z) = \Phi^{-1}\vartheta + \Phi^{-w}z$. Then the *(affine) iterated function system* $(f_\vartheta)_{\vartheta\in\mathcal{D}}$, cf. Edgar [34] or Barnsley [6], has the fundamental domain $\mathcal{F}$ as an invariant set, i.e.,

$$\mathcal{F} = \bigcup_{\vartheta\in\mathcal{D}} f_\vartheta(\mathcal{F}) = \Phi^{-1}\mathcal{F} \cup \bigcup_{\vartheta\in\mathcal{D}^\bullet} \left(\Phi^{-1}\vartheta + \Phi^{-w}\mathcal{F}\right).$$

That formula also reflects the fact that we have two possibilities building the elements $\boldsymbol{\xi} \in \mathbf{NAF}_w^{0.\infty}$ from left to right: We can either append 0, what corresponds to an application of Φ^{-1}, or we can append a non-zero digit $\vartheta \in \mathcal{D}^\bullet$ and then add $w-1$ zeros.

Furthermore, the iterated function system $(f_\vartheta)_{\vartheta\in\mathcal{D}}$ fulfils *Moran's open set condition*[1], cf. Edgar [34] or Barnsley [6]. The *Moran open set* used is $\operatorname{int}\mathcal{F}$. This set satisfies

$$f_\vartheta(\operatorname{int}\mathcal{F}) \cap f_{\vartheta'}(\operatorname{int}\mathcal{F}) = \emptyset$$

for $\vartheta \neq \vartheta' \in \mathcal{D}$ and

$$\operatorname{int}\mathcal{F} \supseteq f_\vartheta(\operatorname{int}\mathcal{F})$$

for all $\vartheta \in \mathcal{D}$. We remark that the first condition follows directly from the tiling property in Corollary 4.6.3 with $K = -1$. The second condition follows from the fact that f_ϑ is an open mapping.

[1] "Moran's open set condition" is sometimes just called "open set condition"

Next we want to have a look at the Hausdorff dimension of the boundary of $\mathcal{F}$. We will need the following characterisation of the boundary.

Proposition 4.6.5 (Characterisation of the Boundary)**.** *Let $z \in \mathcal{F}$. Then $z \in \partial\mathcal{F}$ if and only if there exists a w-NAF $\boldsymbol{\xi}_I.\boldsymbol{\xi}_F \in \mathbf{NAF}_w^{\mathsf{fin}.\infty}$ with $\boldsymbol{\xi}_I \neq \mathbf{0}$ such that $z = \mathsf{value}(\boldsymbol{\xi}_I.\boldsymbol{\xi}_F)$.*

Proof. The proof is a straightforward generalisation of the proof of the quadratic case in [49]. □

The following proposition deals with the Hausdorff dimension of the boundary of $\mathcal{F}$.

Proposition 4.6.6. *For the Hausdorff dimension of the boundary of the fundamental domain we get* $\dim_H \partial\mathcal{F} < n$.

The idea of this proof is similar to a proof in Heuberger and Prodinger [52], and it is a generalisation of the one given in [49].

Proof. Set $k := k_0 + w - 1$ with k_0 from Lemma 4.4.4 on page 59. For $j \in \mathbb{N}$ define

$$U_j := \left\{ \boldsymbol{\xi} \in \mathbf{NAF}_w^{0.j} \,:\, \xi_{-\ell}\xi_{-(\ell+1)}\dots\xi_{-(\ell+k-1)} \neq 0^k \text{ for all } \ell \in \{1,\dots,j-k+1\} \right\}.$$

The elements of U_j, or more precisely the digits from index -1 to $-j$, can be described by the regular expression

$$\left(\varepsilon + \sum_{d\in\mathcal{D}^\bullet}\sum_{\ell=0}^{w-2} 0^\ell d\right)\left(\sum_{d\in\mathcal{D}^\bullet}\sum_{\ell=w-1}^{k-1} 0^\ell d\right)^*\left(\sum_{\ell=0}^{k-1} 0^\ell\right).$$

This can be translated to the generating function

$$G(Z) = \sum_{j\in\mathbb{N}} \#U_j Z^j = \left(1 + \#\mathcal{D}^\bullet \sum_{\ell=0}^{w-2} Z^{\ell+1}\right) \frac{1}{1 - \#\mathcal{D}^\bullet \sum_{\ell=w-1}^{k-1} Z^{\ell+1}} \left(\sum_{\ell=0}^{k-1} Z^\ell\right)$$

used for counting the number of elements in U_j. Rewriting yields

$$G(Z) = \frac{1-Z^k}{1-Z}\,\frac{1 + (\#\mathcal{D}^\bullet - 1)Z - \#\mathcal{D}^\bullet Z^w}{1 - Z - \#\mathcal{D}^\bullet Z^w + \#\mathcal{D}^\bullet Z^{k+1}},$$

and we set

$$q(Z) := 1 - Z - \#\mathcal{D}^\bullet Z^w + \#\mathcal{D}^\bullet Z^{k+1}.$$

Now we define

$$\widetilde{U}_j := \{\boldsymbol{\xi} \in U_j \,:\, \xi_{-j} \neq 0\}$$

and consider $\widetilde{U} := \bigcup_{j\in\mathbb{N}} \widetilde{U}_j$. Suppose $w \geq 2$. The w-NAFs in that set, or more precisely the finite strings from index -1 to the smallest index of a non-zero digit, will be recognised by the automaton $\mathcal{A}$ which is shown in Figure 4.6.1 on the facing page and reads its input from right to left. It is easy to see that the underlying directed graph $G_{\mathcal{A}}$ of the automaton $\mathcal{A}$ is strongly connected, therefore its adjacency matrix $M_{\mathcal{A}}$ is irreducible.

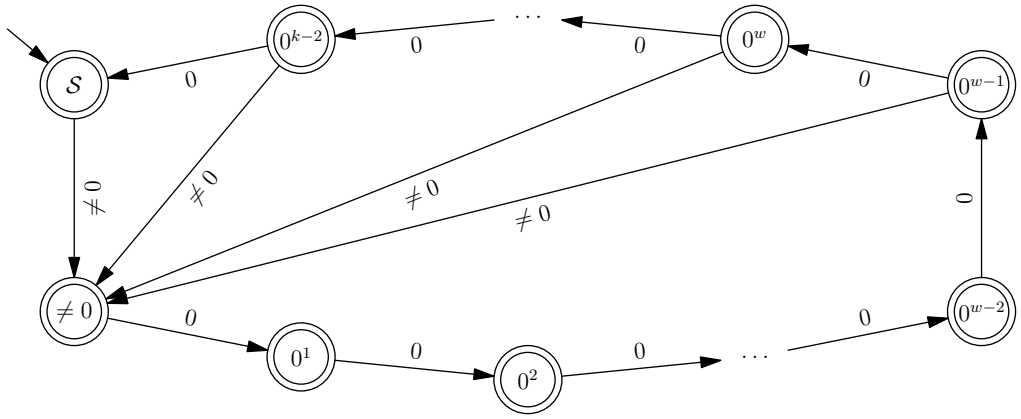


Figure 4.6.1: Automaton $\mathcal{A}$ recognising $\bigcup_{j\in\mathbb{N}} \widetilde{U}_j$ from right to left, see proof of Proposition 4.6.6. The state $\mathcal{S}$ is the starting state, all states are valid end states. An edge marked with $\neq 0$ means one edge for each non-zero digit in the digit set $\mathcal{D}$. The state $\neq 0$ means that there was an non-zero digit read, a state 0^ℓ means that ℓ zeros have been read.

Since there are cycles of length w and $w+1$ in the graph and $\gcd(w, w+1) = 1$, the adjacency matrix is primitive. Thus, using the Perron-Frobenius theorem we obtain

$$\begin{aligned}\#\widetilde{U}_j &= \#(\text{walks in } G_{\mathcal{A}} \text{ of length } j \text{ from starting state } \mathcal{S} \text{ to some other state})\\ &= \begin{pmatrix}1 & 0 & \dots & 0\end{pmatrix} M_{\mathcal{A}}^j \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = \widetilde{c}\,(\sigma\rho^n)^j \left(1 + \mathcal{O}\big(s^j\big)\right)\end{aligned}$$

for a $\widetilde{c} > 0$, a $\sigma > 0$, and an s with $0 \le s < 1$. Since the number of w-NAFs of length j is $\mathcal{O}\big(\rho^{nj}\big)$, see Theorem 4.3.7 on page 55, we get $\sigma \le 1$.

We clearly have

$$U_j = \biguplus_{\ell=j-k+1}^{j} \widetilde{U}_\ell,$$

so we get

$$\#U_j = \big[Z^j\big]\, G(Z) = c\,(\sigma\rho^n)^j \left(1 + \mathcal{O}\big(s^j\big)\right)$$

for some constant $c > 0$.

To rule out $\sigma = 1$, we insert the "zero" ρ^{-n} in $q(Z)$. We obtain

$$\begin{aligned}q\big(\rho^{-n}\big) &= 1 - \rho^{-n} - \#\mathcal{D}^\bullet \rho^{-nw} + \#\mathcal{D}^\bullet \rho^{-n(k+1)}\\ &= 1 - \rho^{-n} - \rho^{n(w-1)}\,(\rho^n - 1)\,\rho^{-nw} + \rho^{n(w-1)}\,(\rho^n - 1)\,\rho^{-n(k+1)}\\ &= (\rho^n - 1)\,\rho^{n(w-k-2)} > 0,\end{aligned}$$

where we used the cardinality of $\mathcal{D}^\bullet$ from our set-up in Section 4.2 and $\rho > 1$. Therefore we get $\sigma < 1$. It is easy to check, that the result for $\#U_j$ holds in the case $w = 1$, too.

Define

$$U := \left\{ \mathsf{value}(\boldsymbol{\xi}) \,:\, \boldsymbol{\xi} \in \mathbf{NAF}_w^{0.\infty} \text{ with } \xi_{-\ell}\xi_{-(\ell+1)}\ldots\xi_{-(\ell+k-1)} \neq 0^k \text{ for all } \ell \geq 1 \right\}.$$

We want to cover U with hypercubes. Let $C \subseteq \mathbb{R}^n$ be the closed paraxial hypercube with centre 0 and width 2. Using Proposition 4.4.1 on page 56 yields

$$U \subseteq \bigcup_{z \in \mathsf{value}(U_j)} \left(z + B_U \rho^{-j} C\right)$$

for all $j \in \mathbb{N}$, i.e., U can be covered with $\#U_j$ boxes of size $2B_U\rho^{-j}$. Thus we get for the upper box dimension, cf. Edgar [34],

$$\overline{\dim}_B U \leq \lim_{j\to\infty} \frac{\log \#U_j}{-\log(2B_U\rho^{-j})}.$$

Inserting the cardinality $\#U_j$ from above, using the logarithm to base ρ and $0 \leq s < 1$ yields

$$\overline{\dim}_B U \leq \lim_{j\to\infty} \frac{\log_\rho c + j\log_\rho(\sigma\rho^n) + \log_\rho(1+\mathcal{O}(s^j))}{j+\mathcal{O}(1)} = n + \log_\rho \sigma.$$

Since $\sigma < 1$, we get $\overline{\dim}_B U < 2$.

Now we will show that $\partial\mathcal{F} \subseteq U$. Clearly $U \subseteq \mathcal{F}$, so the previous inclusion is equivalent to $\mathcal{F}\setminus U \subseteq \mathrm{int}(\mathcal{F})$. So let $z \in \mathcal{F}\setminus U$. Then there is a $\boldsymbol{\xi} \in \mathbf{NAF}_w^{0.\infty}$ such that $z = \mathsf{value}(\boldsymbol{\xi})$ and $\boldsymbol{\xi}$ has a block of at least k zeros somewhere on the right hand side of the Φ-point. Let ℓ denote the starting index of this block, i.e.,

$$\boldsymbol{\xi} = 0.\underbrace{\xi_{-1}\ldots\xi_{-(\ell-1)}}_{=:\boldsymbol{\xi}_A} 0^k \xi_{-(\ell+k)}\xi_{-(\ell+k+1)}\cdots.$$

Let $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_I.\boldsymbol{\vartheta}_A\vartheta_{-\ell}\vartheta_{-(\ell+1)}\ldots \in \mathbf{NAF}_w^{\mathsf{fin}.\infty}$ with $\mathsf{value}(\boldsymbol{\vartheta}) = z$. We have

$$z = \mathsf{value}(0.\boldsymbol{\xi}_A) + \Phi^{-\ell-w} z_\xi = \mathsf{value}(\boldsymbol{\vartheta}_I.\boldsymbol{\vartheta}_A) + \Phi^{-\ell-w} z_\vartheta$$

for appropriate z_ξ and z_ϑ. By Lemma 4.4.4 on page 59, all expansions of z_ξ are in $\mathbf{NAF}_w^{0.\infty}$. Thus all expansions of

$$\mathsf{value}(\boldsymbol{\vartheta}_I\boldsymbol{\vartheta}_A) + \Phi^{-(w-1)} z_\vartheta - \mathsf{value}(\boldsymbol{\xi}_A) = \Phi^{\ell-1} z - \mathsf{value}(\boldsymbol{\xi}_A) = \Phi^{-(w-1)} z_\xi$$

start with 0.0^{w-1}, since our choice of k is $k_0 + w - 1$. As the unique w-NAF of $\mathsf{value}(\boldsymbol{\vartheta}_I\boldsymbol{\vartheta}_A) - \mathsf{value}(\boldsymbol{\xi}_A)$ concatenated with any w-NAF of $\Phi^{-(w-1)} z_\vartheta$ gives rise to such an expansion, we conclude that $\mathsf{value}(\boldsymbol{\vartheta}_I\boldsymbol{\vartheta}_A) - \mathsf{value}(\boldsymbol{\xi}_A) = 0$ and therefore $\boldsymbol{\vartheta}_I = \mathbf{0}$ and $\boldsymbol{\vartheta}_A = \boldsymbol{\xi}_A$. So we conclude that all representations of z as a w-NAF have to be of the form $0.\boldsymbol{\xi}_A 0^{w-1}\boldsymbol{\eta}$ for some w-NAF $\boldsymbol{\eta}$. Thus, by using Proposition 4.6.5 on page 62, we get $z \notin \partial\mathcal{F}$ and therefore $z \in \mathrm{int}(\mathcal{F})$.

Until now we have proved

$$\overline{\dim}_B \partial\mathcal{F} \leq \overline{\dim}_B U < n.$$

Because the Hausdorff dimension of a set is at most its upper box dimension, cf. Edgar [34] again, the desired result follows. □

4.7 Cell Rounding Operations

In this section we define operators working on subsets of the space $\mathbb{R}^n$. These will use the lattice Λ and the tiling T. They will be a very useful concept to prove Theorem 4.9.1 on page 69.

Definition 4.7.1 (Cell Rounding Operations)**.** Let $B \subseteq \mathbb{R}^n$ and $j \in \mathbb{Z}$. We define the *cell packing of* B ("floor B")

$$\lfloor B \rfloor_T := \bigcup_{\substack{z \in \Lambda \\ T_z \subseteq B}} T_z \qquad \text{and} \qquad \lfloor B \rfloor_{T,j} := \Phi^{-j}\big(\lfloor \Phi^j B \rfloor_T\big),$$

the *cell covering of* B ("ceil B")

$$\lceil B \rceil_T := \overline{\lfloor B^C \rfloor_T^C} \qquad \text{and} \qquad \lceil B \rceil_{T,j} := \Phi^{-j}\big(\lceil \Phi^j B \rceil_T\big),$$

the *fractional cells of* B

$$\{B\}_T := B \setminus \lfloor B \rfloor_T \qquad \text{and} \qquad \{B\}_{T,j} := \Phi^{-j}\big(\{\Phi^j B\}_T\big),$$

the *cell covering of the boundary of* B

$$\partial(B)_T := \overline{\lceil B \rceil_T \setminus \lfloor B \rfloor_T} \qquad \text{and} \qquad \partial(B)_{T,j} := \Phi^{-j}\big(\partial\big(\Phi^j B\big)_T\big),$$

the *cell covering of the lattice points inside* B

$$\lfloor B \rceil_T := \bigcup_{z \in B \cap \Lambda} T_z \qquad \text{and} \qquad \lfloor B \rceil_{T,j} := \Phi^{-j}\big(\lfloor \Phi^j B \rceil_T\big),$$

and the *number of lattice points inside* B as

$$\#(B)_T := \#(B \cap \Lambda) \qquad \text{and} \qquad \#(B)_{T,j} := \#\big(\Phi^j B\big)_T.$$

For the cell covering of a set B an alternative, perhaps more intuitive description can be given by

$$\lceil B \rceil_T := \bigcup_{\substack{z \in \Lambda \\ T_z \cap B \neq \emptyset}} T_z.$$

The following proposition deals with some basic properties that will be helpful when working with those operators.

Proposition 4.7.2 (Basic Properties of Cell Rounding Operations)**.** *Let* $B \subseteq \mathbb{R}^n$ *and* $j \in \mathbb{Z}$.

(a) *We have the inclusions*

$$\lfloor B \rfloor_{T,j} \subseteq B \subseteq \overline{B} \subseteq \lceil B \rceil_{T,j}$$

and

$$\lfloor B \rfloor_{T,j} \subseteq \lfloor B \rceil_{T,j} \subseteq \lceil B \rceil_{T,j}\,.$$

For $B' \subseteq \mathbb{R}^n$ with $B \subseteq B'$ we get $\lfloor B \rfloor_{T,j} \subseteq \lfloor B' \rfloor_{T,j}$, $\lfloor B \rceil_{T,j} \subseteq \lfloor B' \rceil_{T,j}$ and $\lceil B \rceil_{T,j} \subseteq \lceil B' \rceil_{T,j}$, i.e., monotonicity with respect to inclusion.

(b) *The inclusion*

$$\{B\}_{T,j} \subseteq \partial(B)_{T,j}$$

holds.

(c) *We have $\partial B \subseteq \partial(B)_{T,j}$ and for each cell T' in $\partial(B)_{T,j}$ we have $T' \cap \partial B \neq \emptyset$.*

(d) *For $B' \subseteq \mathbb{R}^n$ with B' disjoint from B, we get*

$$\#\big(B \cup B'\big)_{T,j} = \#(B)_{T,j} + \#\big(B'\big)_{T,j}\,,$$

and therefore the number of lattice points operation is monotonic with respect to inclusion, i.e., for $B'' \subseteq \mathbb{R}^n$ with $B'' \subseteq B$ we have $\#(B'')_{T,j} \leq \#(B)_{T,j}$. Further we get

$$\#(B)_{T,j} = \#\Big(\lfloor B \rceil_{T,j}\Big)_{T,j} = |\det \Phi|^j \, \frac{\lambda\Big(\lfloor B \rceil_{T,j}\Big)}{d_\Lambda}$$

Proof. The proof is a straightforward generalisation of the proof for Voronoi-tilings in the quadratic case in [49]. □

We will need some more properties concerning cardinality. We want to know the number of points inside a region after using one of the operators. Especially we are interested in the asymptotic behaviour, i.e., if our region becomes scaled very large. The following proposition provides information about that.

Proposition 4.7.3. *Let $U \subseteq \mathbb{R}^n$ bounded, measurable, and such that*

$$\#(\partial(\Psi U)_T)_T = \mathcal{O}\Big(|\det \Psi|^{\delta/n}\Big)$$

for $|\det \Psi| \to \infty$ with maps $\Psi\colon \mathbb{R}^n \to \mathbb{R}^n$ and a fixed $\delta \in \mathbb{R}$ with $\delta > 0$.

(a) *We get that each of $\#(\lfloor \Psi U \rfloor_T)_T$, $\#(\lceil \Psi U \rceil_T)_T$, $\#(\lfloor \Psi U \rceil_T)_T$ and $\#(\Psi U)_T$ equals*

$$|\det \Psi| \, \frac{\lambda(U)}{d_\Lambda} + \mathcal{O}\Big(|\det \Psi|^{\delta/n}\Big)\,.$$

In particular, let $N \in \mathbb{R}$, $N > 0$, and set $\Psi = \operatorname{diag}(N, \ldots, N)$, which we identify with N. Then we get that each one of $\#(\lfloor NU \rfloor_T)_T$, $\#(\lceil NU \rceil_T)_T$, $\#(\lfloor NU \rceil_T)_T$ and $\#(NU)_T$ equals

$$N^n \frac{\lambda(U)}{d_\Lambda} + \mathcal{O}\Big(N^\delta\Big)\,.$$

(b) Let $N \in \mathbb{R}$, $N > 0$, and set $\Psi = \operatorname{diag}(N, \dots, N)$, which we identify with N. Then we get

$$\#((N+1)U \setminus NU)_T = \mathcal{O}\Big(N^{\delta}\Big).$$

Proof. Again, the proof is a straightforward generalisation of the proof for Voronoi-tilings in the quadratic case in [49]. □

Note that $\delta = n - 1$ if U is, for example, a ball or a polyhedron.

4.8 The Characteristic Sets

In this section we define characteristic sets for a digit at a specified position in the w-NAF expansion and prove some basic properties of them. Those will be used in the proof of Theorem 4.9.1.

Definition 4.8.1 (Characteristic Sets)**.** Let $\eta \in \mathcal{D}^{\bullet}$. For $j \in \mathbb{N}_0$ define

$$\mathcal{W}_{\eta,j} := \big\{\mathsf{value}(\boldsymbol{\xi}) \,:\, \boldsymbol{\xi} \in \mathbf{NAF}_w^{0.j+w} \text{ with } \xi_{-w} = \eta\big\}.$$

We call $\lfloor \mathcal{W}_{\eta,j} \rceil_{T,j+w}$ the *jth approximation of the characteristic set for η*, and we define

$$W_{\eta,j} := \Big\{ \lfloor \mathcal{W}_{\eta,j} \rceil_{T,j+w} \Big\}_{\Lambda}.$$

Further we define the *characteristic set for η*

$$\mathcal{W}_{\eta} := \big\{\mathsf{value}(\boldsymbol{\xi}) \,:\, \boldsymbol{\xi} \in \mathbf{NAF}_w^{0.\infty} \text{ with } \xi_{-w} = \eta\big\}$$

and

$$W_{\eta} := \{\mathcal{W}_{\eta}\}_{\Lambda}.$$

For $j \in \mathbb{N}_0$ we set

$$\beta_{\eta,j} := \lambda\Big(\lfloor \mathcal{W}_{\eta,j} \rceil_{T,j+w}\Big) - \lambda(\mathcal{W}_{\eta}).$$

Note that sometimes the set W_{η} will also be called *characteristic set for η*, and analogously for the set $W_{\eta,j}$. In Figure 4.8.1 on the following page some of these characteristic sets, more precisely some approximations of the characteristic sets, are shown.

The following proposition deals with some properties of those defined sets.

Proposition 4.8.2 (Properties of the Characteristic Sets)**.** *Let $\eta \in \mathcal{D}^{\bullet}$.*

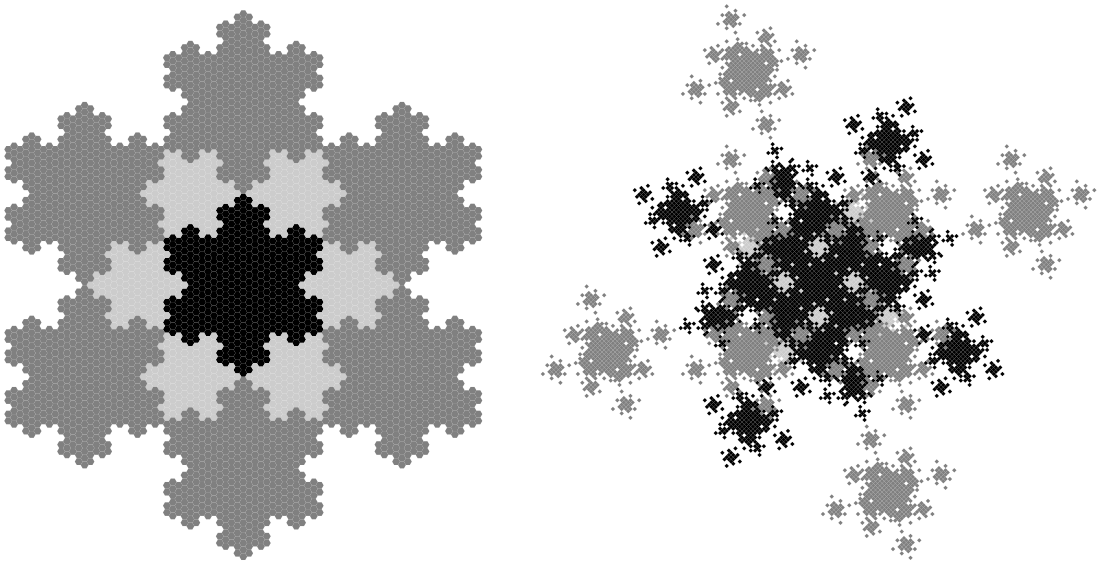
(a) We have

$$\mathcal{W}_{\eta} = \eta\tau^{-w} + \Phi^{-2w+1}\mathcal{F}.$$

(b) The set $\mathcal{W}_{\eta}$ is compact.

(c) We get

$$\mathcal{W}_{\eta} = \overline{\bigcup_{j \in \mathbb{N}_0} \mathcal{W}_{\eta,j}} = \overline{\lim_{j \to \infty} \mathcal{W}_{\eta,j}}.$$



(a) $\mathcal{W}_{\eta,j}$ for a lattice coming from τ with $\tau^2 - 3\tau + 3 = 0$, $w = 2$ and $j = 7$

(b) $\mathcal{W}_{\eta,j}$ for a lattice coming from τ with $\tau^2 - 2\tau + 2 = 0$, $w = 4$ and $j = 11$

Figure 4.8.1: Fundamental domains and characteristic sets $\mathcal{W}_\eta$. Each figure shows a fundamental domain. The light-gray coloured parts represent the approximations $\mathcal{W}_{\eta,j}$ of the characteristic sets $\mathcal{W}_\eta$.

(d) The set $\lfloor \mathcal{W}_{\eta,j} \rceil_{T,j+w}$ is indeed an approximation of $\mathcal{W}_\eta$, i.e., we have

$$\mathcal{W}_\eta = \overline{\liminf_{j \in \mathbb{N}_0} \lfloor \mathcal{W}_{\eta,j} \rceil_{T,j+w}} = \overline{\limsup_{j \in \mathbb{N}_0} \lfloor \mathcal{W}_{\eta,j} \rceil_{T,j+w}}.$$

(e) We have $\operatorname{int} \mathcal{W}_\eta \subseteq \liminf_{j \in \mathbb{N}_0} \lfloor \mathcal{W}_{\eta,j} \rceil_{T,j+w}$.

(f) We get $\mathcal{W}_\eta - \Phi^{-w}\eta \subseteq T$, and for $j \in \mathbb{N}_0$ we obtain $\lfloor \mathcal{W}_{\eta,j} \rceil_{T,j+w} - \Phi^{-w}\eta \subseteq T$.

(g) For the Lebesgue measure of the characteristic set we obtain $\lambda(\mathcal{W}_\eta) = \lambda(W_\eta)$ and for its approximation $\lambda\Big(\lfloor \mathcal{W}_{\eta,j} \rceil_{T,j+w}\Big) = \lambda(W_{\eta,j})$.

(h) Let $j \in \mathbb{N}_0$, then

$$\lambda\Big(\lfloor \mathcal{W}_{\eta,j} \rceil_{T,j+w}\Big) = d_\Lambda E + \mathcal{O}\big(\mu^j\big)$$

with E and $\mu < 1$ from Theorem 4.3.7 on page 55.

(i) The Lebesgue measure of W_η is

$$\lambda(W_\eta) = d_\Lambda E,$$

again with E from Theorem 4.3.7 on page 55.

(j) Let $j \in \mathbb{N}_0$. We get

$$\beta_{\eta,j} = \int_{x \in T} \left(\mathbb{1}_{W_{\eta,j}} - \mathbb{1}_{W_\eta}\right)(x)\,\mathrm{d}x = \mathcal{O}\left(\mu^j\right).$$

Again $\mu < 1$ can be found in Theorem 4.3.7 on page 55.

Proof. The proof is a straightforward generalisation of the proof in [49]. □

We can also determine the Lebesgue measure of the fundamental domain $\mathcal{F}$ defined in Section 4.6.

Remark 4.8.3 (Lebesgue Measure of the Fundamental Domain). We get

$$\lambda(\mathcal{F}) = \rho^{n(2w-1)} E d_\Lambda = \frac{\rho^{nw} d_\Lambda}{(\rho^n - 1)w + 1},$$

using (a) and (i) from Proposition 4.8.2 on page 67 and E from Theorem 4.3.7 on page 55.

The next lemma makes the connection between the w-NAFs of elements of the lattice Λ and the characteristic sets $W_{\eta,j}$.

Lemma 4.8.4. *Let $\eta \in \mathcal{D}^\bullet$, $j \geq 0$. Let $z \in \Lambda$ and let $\boldsymbol{\xi} \in \mathbf{NAF}_w^{\mathsf{fin},0}$ be its w-NAF. Then the following statements are equivalent:*

(1) The jth digit of $\boldsymbol{\xi}$ equals η.

(2) The condition $\left\{\Phi^{-(j+w)} z\right\}_\Lambda \in W_{\eta,j}$ holds.

(3) The inclusion $\left\{\Phi^{-(j+w)} T_z\right\}_\Lambda \subseteq W_{\eta,j}$ holds.

Proof. The proof is a straightforward generalisation of the proof of the quadratic case in [49]. □

4.9 Counting the Occurrences of a non-zero Digit in a Region

In this section we will prove our main result on the asymptototic number of occurrences of a digit in a given region.

Note that Iverson's notation [expr] = 1 if expr is true and [expr] = 0 otherwise, cf. Graham, Knuth and Patashnik [45], will be used.

Theorem 4.9.1 (Counting Theorem)**.** *Let $0 \neq \eta \in \mathcal{D}$ and $N \in \mathbb{R}$ with $N > 0$. Further let $U \subseteq \mathbb{R}^n$ be measurable with respect to the Lebesgue measure and bounded with $U \subseteq \mathcal{B}(0,d)$ for a finite d, and set δ such that $\#(\partial(NU)_T)_T = \mathcal{O}\left(N^\delta\right)$ with $1 \leq \delta < n$. We denote the number of occurrences of the digit η in all integer width-w non-adjacent forms with value in the region NU by*

$$Z_\eta(N) = \sum_{z \in NU \cap \Lambda} \sum_{j \in \mathbb{N}_0} [j\text{th digit of } z \text{ in its } w\text{-NAF-expansion equals } \eta].$$

Then we get

$$Z_\eta(N) = N^n\,\lambda(U)\,E\log_\rho N + N^n\,\psi_\eta\big(\log_\rho N\big) + \mathcal{O}\big(N^\alpha \log_\rho N\big) + \mathcal{O}\Big(N^\delta \log_\rho N\Big)\,,$$

in which the expressions described below are used. The Lebesgue measure on $\mathbb{R}^n$ *is denoted by* λ*. We have the constant of the expectation*

$$E = \frac{1}{\rho^{n(w-1)}((\rho^n-1)w+1)},$$

cf. Theorem 4.3.7 on page 55. Then there is the function

$$\psi_\eta(x) = \psi_{\eta,\mathcal{M}}(x) + \psi_{\eta,\mathcal{P}}(x) + \psi_{\eta,\mathcal{Q}}(x)\,,$$

where

$$\psi_{\eta,\mathcal{M}}(x) = \lambda(U)\,(J_0 + 1 - \{x\})\,E,$$

$$\psi_{\eta,\mathcal{P}}(x) = \frac{\rho^{n(J_0-\{x\})}}{d_\Lambda} \sum_{j=0}^{\infty} \int_{y\in\left\{\Phi^{-\lfloor x\rfloor - J_0}\rho^x U\right\}_{T,j-w}} \big(\mathbb{1}_W\big(\big\{\Phi^{j-w}y\big\}_\Lambda\big) - \lambda(W)\big)\,\mathrm{d}y,$$

and

$$\psi_{\eta,\mathcal{Q}} = \frac{\lambda(U)}{d_\Lambda^2} \sum_{j=0}^{\infty} \beta_j.$$

We have $\alpha = n + \log_\rho \mu < n$*, with* $\mu = \left(1 + \frac{1}{\rho^n w^3}\right)^{-1} < 1$*, and*

$$J_0 = \big\lfloor \log_\rho d - \log_\rho B_L \big\rfloor + 1$$

with the constant B_L *of Proposition 4.4.2 on page 58.*

Further, let

$$\Phi = Q\,\mathrm{diag}\,\rho e^{i\theta_1}, \ldots, \rho e^{i\theta_n} Q^{-1},$$

where Q *is a regular matrix. If there is a* $p \in \mathbb{N}$ *such that*

$$Q\,\mathrm{diag}\,e^{i\theta_1 p}, \ldots, e^{i\theta_n p} Q^{-1} U = U,$$

then ψ_η *is p-periodic. Moreover, if* ψ_η *is p-periodic for some* $p \in \mathbb{N}$*, then it is also continuous.*

Remark 4.9.2. Consider the main term of our result. When N tends to infinity, we get the asymptotic formula

$$Z_\eta \sim N^n\,\lambda(U)\,E\log_\rho N.$$

This result is not surprising, since intuitively the number of lattice points in the region NU corresponds to the Lebesgue measure $N^n\,\lambda(U)$ of this region, and each of that elements can be represented as an integer w-NAF with length about $\log_\rho N$. Therefore, using the expectation of Theorem 4.3.7 on page 55, we get an explanation for this term.

Remark 4.9.3. If $\delta = n$ in the theorem, then the statement stays true, but degenerates to

$$Z_\eta(N) = \mathcal{O}\Big(N^n \log_{|\tau|} N\Big)\,.$$

This is a trivial result of Remark 4.9.2 on the preceding page.

The proof of Theorem 4.9.1 on page 69 follows the ideas used by Delange [24]. By Remark 4.9.3 we restrict ourselves to the case $\delta < n$.

We will use the following abbreviations. We omit the index η, i.e., we set $Z(N) := Z_\eta(N)$, $W := W_\eta$ and $W_j := W_{\eta,j}$, and further we set $\beta_j := \beta_{\eta,j}$, cf. Proposition 4.8.2 on page 67. By log we will denote the logarithm to the base ρ, i.e., $\log x = \log_\rho x$. These abbreviations will be used throughout the remaining section.

Proof of Theorem 4.9.1. By assumption every element of Λ is represented by a unique element of $\mathbf{NAF}_w^{\mathsf{fin},0}$. To count the occurrences of the digit η in NU, we sum up 1 over all lattice points $z \in NU \cap \Lambda$ and for each z over all digits in the corresponding w-NAF equal to η. Thus we get

$$Z(N) = \sum_{z \in NU \cap \Lambda} \sum_{j \in \mathbb{N}_0} [j\text{th digit of } w\text{-NAF of } z \text{ equals } \eta]\,.$$

The inner sum over $j \in \mathbb{N}_0$ is finite, we will choose a large enough upper bound J later in Lemma 4.9.4 on page 73.

Using

$$[j\text{th digit of } w\text{-NAF of } z \text{ equals } \eta] = \mathbb{1}_{W_j}\big(\{\Phi^{-j-w} z\}_\Lambda\big)$$

from Lemma 4.8.4 on page 69 yields

$$Z(N) = \sum_{j=0}^{J} \sum_{z \in NU \cap \Lambda} \mathbb{1}_{W_j}\big(\{\Phi^{-j-w} z\}_\Lambda\big)\,,$$

where additionally the order of summation was changed. This enables us to rewrite the sum over z as an integral

$$\begin{aligned} Z(N) &= \sum_{j=0}^{J} \sum_{z \in NU \cap \Lambda} \frac{1}{\lambda(T_z)} \int_{x \in T_z} \mathbb{1}_{W_j}\big(\{\Phi^{-j-w} x\}_\Lambda\big)\,\mathrm{d}x \\ &= \frac{1}{\lambda(T)} \sum_{j=0}^{J} \int_{x \in \lfloor NU \rceil_T} \mathbb{1}_{W_j}\big(\{\Phi^{-j-w} x\}_\Lambda\big)\,\mathrm{d}x. \end{aligned}$$

We split up the integrals into the ones over NU and others over the remaining region and get

$$Z(N) = \frac{1}{\lambda(T)} \sum_{j=0}^{J} \int_{x \in NU} \mathbb{1}_{W_j}\big(\{\Phi^{-j-w} x\}_\Lambda\big)\,\mathrm{d}x + \mathcal{F}_\eta(N)\,,$$

in which $\mathcal{F}_\eta(N)$ contains all integrals (with appropriate signs) over regions $\lfloor NU \rceil_T \setminus NU$ and $NU \setminus \lfloor NU \rceil_T$.

By substituting $x = \Phi^J y$, $\mathrm{d}x = |\det \Phi|^J \,\mathrm{d}y = \rho^{nJ}\,\mathrm{d}y$ we obtain

$$Z(N) = \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} \int_{y \in \Phi^{-J} NU} \mathbb{1}_{W_j}\big(\big\{\Phi^{J-j-w} y\big\}_\Lambda\big)\,\mathrm{d}y + \mathcal{F}_\eta(N)\,.$$

Reversing the order of summation yields

$$Z(N) = \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} \int_{y \in \Phi^{-J} NU} \mathbb{1}_{W_{J-j}}\big(\big\{\Phi^{j-w} y\big\}_\Lambda\big)\,\mathrm{d}y + \mathcal{F}_\eta(N)\,.$$

We rewrite this as

$$\begin{aligned} Z(N) = {} & \frac{\rho^{nJ}}{\lambda(T)} (J+1)\,\lambda(W) \int_{y \in \Phi^{-J} NU} \mathrm{d}y \\ & + \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} \int_{y \in \Phi^{-J} NU} \big(\mathbb{1}_W\big(\big\{\Phi^{j-w} y\big\}_\Lambda\big) - \lambda(W)\big)\,\mathrm{d}y \\ & + \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} \int_{y \in \Phi^{-J} NU} \big(\mathbb{1}_{W_{J-j}}\big(\big\{\Phi^{j-w} y\big\}_\Lambda\big) - \mathbb{1}_W\big(\big\{\Phi^{j-w} y\big\}_\Lambda\big)\big)\,\mathrm{d}y \\ & + \mathcal{F}_\eta(N)\,. \end{aligned}$$

With $\Phi^{-J} NU = \big\lfloor \Phi^{-J} NU \big\rfloor_{T,j-w} \cup \big\{\Phi^{-J} NU\big\}_{T,j-w}$ for each area of integration we get

$$Z(N) = \mathcal{M}_\eta(N) + \mathcal{Z}_\eta(N) + \mathcal{P}_\eta(N) + \mathcal{Q}_\eta(N) + \mathcal{S}_\eta(N) + \mathcal{F}_\eta(N)\,,$$

in which $\mathcal{M}_\eta$ is *"The Main Part"*, see Lemma 4.9.6 on page 74,

$$\mathcal{M}_\eta(N) = \frac{\rho^{nJ}}{\lambda(T)} (J+1)\,\lambda(W) \int_{y \in \Phi^{-J} NU} \mathrm{d}y, \tag{4.9.1a}$$

$\mathcal{Z}_\eta$ is *"The Zero Part"*, see Lemma 4.9.7 on page 74,

$$\mathcal{Z}_\eta(N) = \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} \int_{y \in \lfloor \Phi^{-J} NU \rfloor_{T,j-w}} \big(\mathbb{1}_W\big(\big\{\Phi^{j-w} y\big\}_\Lambda\big) - \lambda(W)\big)\,\mathrm{d}y, \tag{4.9.1b}$$

$\mathcal{P}_\eta$ is *"The Periodic Part"*, see Lemma 4.9.8 on page 75,

$$\mathcal{P}_\eta(N) = \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} \int_{y \in \{\Phi^{-J} NU\}_{T,j-w}} \big(\mathbb{1}_W\big(\big\{\Phi^{j-w} y\big\}_\Lambda\big) - \lambda(W)\big)\,\mathrm{d}y, \tag{4.9.1c}$$

$\mathcal{Q}_\eta$ is *"The Other Part"*, see Lemma 4.9.9 on page 77,

$$\mathcal{Q}_\eta(N) = \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} \int_{y \in \lfloor \Phi^{-J} NU \rfloor_{T,j-w}} \big(\mathbb{1}_{W_{J-j}} - \mathbb{1}_W\big)\big(\big\{\Phi^{j-w} y\big\}_\Lambda\big)\,\mathrm{d}y, \tag{4.9.1d}$$

$\mathcal{S}_\eta$ is *"The Small Part"*, see Lemma 4.9.10 on page 78,

$$\mathcal{S}_\eta(N) = \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} \int_{y \in \{\Phi^{-J} NU\}_{T,j-w}} \left(\mathbb{1}_{W_{J-j}} - \mathbb{1}_W\right)\left(\left\{\Phi^{j-w} y\right\}_\Lambda\right) \mathrm{d}y \tag{4.9.1e}$$

and $\mathcal{F}_\eta$ is *"The Fractional Cells Part"*, see Lemma 4.9.11 on page 80,

$$\begin{aligned} \mathcal{F}_\eta(N) = {} & \frac{1}{\lambda(T)} \sum_{j=0}^{J} \int_{x \in \lfloor NU \rceil_T \setminus NU} \mathbb{1}_{W_j}\left(\left\{\Phi^{-j-w} x\right\}_\Lambda\right) \mathrm{d}x \\ & - \frac{1}{\lambda(T)} \sum_{j=0}^{J} \int_{x \in NU \setminus \lfloor NU \rceil_T} \mathbb{1}_{W_j}\left(\left\{\Phi^{-j-w} x\right\}_\Lambda\right) \mathrm{d}x. \end{aligned} \tag{4.9.1f}$$

To complete the proof we have to deal with the choice of J, see Lemma 4.9.4, as well as with each of the parts in (4.9.1), see Lemmata 4.9.6 to 4.9.11 on pages 74–80. The continuity of ψ_η is checked in Lemma 4.9.12 on page 80. □

Lemma 4.9.4 (Choosing J)**.** *Let $N \in \mathbb{R}_{\geq 0}$. Then every w-NAF of $\mathbf{NAF}_w^{\mathsf{fin},0}$ with value in NU has at most $J+1$ digits, where*

$$J = \lfloor \log N \rfloor + J_0$$

with

$$J_0 = \lfloor \log d - \log B_L \rfloor + 1$$

with B_L of Proposition 4.4.2 on page 58.

Proof. Let $z \in NU$, $z \neq 0$, with its corresponding w-NAF $\boldsymbol{\xi} \in \mathbf{NAF}_w^{\mathsf{fin},0}$, and let $j \in \mathbb{N}_0$ be the largest index such that the digit ξ_j is non-zero. By using Corollary 4.4.3 on page 58, we conclude that

$$\rho^j B_L \leq \|z\| < Nd.$$

This means

$$j < \log N + \log d - \log B_L,$$

and thus we have

$$j \leq \lfloor \log N + \log d - \log B_L \rfloor \leq \lfloor \log N \rfloor + \lfloor \log d - \log B_L \rfloor + 1.$$

Defining the right hand side of this inequality as J finishes the proof. □

Remark 4.9.5. For the parameter used in the region of integration in the proof of Theorem 4.9.1 on page 69 we get

$$\left|\det\left(\Phi^{-J} N\right)\right| = \mathcal{O}(1)\,.$$

In particular, we get $\left\|\Phi^{-J} N\right\| = \mathcal{O}(1)$.

Proof. We have

$$\left|\det\left(\Phi^{-J}N\right)\right| = \left(\rho^{-J}N\right)^n.$$

With J of Lemma 4.9.4 on the preceding page we obtain

$$\rho^{-J}N = \rho^{-\lfloor \log N\rfloor - J_0}\rho^{\log N} = \rho^{\log N - \lfloor \log N\rfloor - J_0} = \rho^{\{\log N\}-J_0}.$$

Since $\rho^{\{\log N\}-J_0}$ is bounded by ρ^{1-J_0}, it is $\mathcal{O}(1)$. Therefore $\det\left(\Phi^{-J}N\right)$ is $\mathcal{O}(1)$. Since $\left\|\Phi^{-1}\right\| = \rho^{-1}$ we conclude that $\left\|\Phi^{-J}N\right\|$ is $\mathcal{O}(1)$. □

Lemma 4.9.6 (The Main Part)**.** *For* (4.9.1a) *in the proof of Theorem 4.9.1 on page 69 we get*

$$\mathcal{M}_\eta(N) = N^n\,\lambda(U)\,E\log N + N^n\,\psi_{\eta,\mathcal{M}}(\log N)$$

with a 1*-periodic function* $\psi_{\eta,\mathcal{M}}$*,*

$$\psi_{\eta,\mathcal{M}}(x) = \lambda(U)\,(J_0 + 1 - \{x\})\,E$$

and E *of Theorem 4.3.7 on page 55.*

Proof. We have

$$\mathcal{M}_\eta(N) = \frac{\rho^{nJ}}{\lambda(T)}(J+1)\,\lambda(W)\int_{y\in\Phi^{-J}NU}\mathrm{d}y.$$

As $\lambda\left(\Phi^{-J}NU\right) = \rho^{-nJ}N^n\,\lambda(U)$ we obtain

$$\mathcal{M}_\eta(N) = \frac{\lambda(W)}{\lambda(T)}(J+1)N^n\,\lambda(U)\,.$$

By taking $\lambda(W) = \lambda(T)\,E$ from (i) of Proposition 4.8.2 on page 67 and J from Lemma 4.9.4 on the preceding page we get

$$\mathcal{M}_\eta(N) = N^n\,\lambda(U)\,E\,(\lfloor\log N\rfloor + J_0 + 1)\,.$$

Finally, the desired result follows by using $\lfloor x\rfloor = x - \{x\}$. □

Lemma 4.9.7 (The Zero Part)**.** *For* (4.9.1b) *in the proof of Theorem 4.9.1 on page 69 we get*

$$\mathcal{Z}_\eta(N) = 0.$$

Proof. Consider the integral

$$I_j := \int_{y\in\lfloor\Phi^{-J}NU\rfloor_{T,j-w}} \left(\mathbb{1}_W\left(\left\{\Phi^{j-w}y\right\}_\Lambda\right) - \lambda(W)\right)\mathrm{d}y.$$

We can rewrite the region of integration as

$$\left\lfloor\Phi^{-J}NU\right\rfloor_{T,j-w} = \Phi^{-(j-w)}\left\lfloor\Phi^{j-w}\Phi^{-J}NU\right\rfloor_T = \Phi^{-(j-w)}\bigcup_{z\in R_{j-w}} T_z$$

for some appropriate $R_{j-w} \subseteq \Lambda$. Substituting $x = \Phi^{j-w}y$, $\mathrm{d}x = \rho^{n(j-w)}\,\mathrm{d}y$ yields

$$I_j = \rho^{-n(j-w)} \int_{x \in \bigcup_{z \in R_{j-w}} T_z} (\mathbb{1}_W(\{x\}_\Lambda) - \lambda(W))\,\mathrm{d}x.$$

We split up the integral and eliminate the fractional part $\{x\}_\Lambda$ by translation to get

$$I_j = \rho^{-n(j-w)} \sum_{z \in R_{j-w}} \underbrace{\int_{x \in T} (\mathbb{1}_W(x) - \lambda(W))\,\mathrm{d}x}_{=0}.$$

Thus, for all $j \in \mathbb{N}_0$ we obtain $I_j = 0$, and therefore $\mathcal{Z}_\eta(N) = 0$. □

Lemma 4.9.8 (The Periodic Part)**.** *For* (4.9.1c) *in the proof of Theorem 4.9.1 on page 69 we get*

$$\mathcal{P}_\eta(N) = N^n\,\psi_{\eta,\mathcal{P}}(\log N) + \mathcal{O}\Big(N^\delta\Big)$$

with a function $\psi_{\eta,\mathcal{P}}$*,*

$$\psi_{\eta,\mathcal{P}}(x) = \frac{\rho^{n(J_0-\{x\})}}{\lambda(T)} \sum_{j=0}^{\infty} \int_{y \in \left\{\Phi^{-\lfloor x \rfloor - J_0}\rho^x U\right\}_{T,j-w}} \big(\mathbb{1}_W\big(\big\{\Phi^{j-w}y\big\}_\Lambda\big) - \lambda(W)\big)\,\mathrm{d}y.$$

Let

$$\Phi = Q\,\mathrm{diag}\,\rho e^{i\theta_1}, \ldots, \rho e^{i\theta_n} Q^{-1},$$

where Q *is a regular matrix. If there is a* $p \in \mathbb{N}$ *such that*

$$Q\,\mathrm{diag}\,e^{i\theta_1 p}, \ldots, e^{i\theta_n p} Q^{-1} U = U, \tag{4.9.2}$$

then $\psi_{\eta,\mathcal{P}}$ *is* p*-periodic.*

Proof. Consider

$$I_j := \int_{y \in \{\Phi^{-J}NU\}_{T,j-w}} \big(\mathbb{1}_W\big(\big\{\Phi^{j-w}y\big\}_\Lambda\big) - \lambda(W)\big)\,\mathrm{d}y.$$

The region of integration satisfies

$$\big\{\Phi^{-J}NU\big\}_{T,j-w} \subseteq \partial\big(\Phi^{-J}NU\big)_{T,j-w} = \Phi^{-(j-w)} \bigcup_{z \in R_{j-w}} T_z \tag{4.9.3}$$

for some appropriate $R_{j-w} \subseteq \Lambda$.

We use the triangle inequality and substitute $x = \Phi^{j-w}y$, $\mathrm{d}x = \rho^{n(j-w)}\,\mathrm{d}y$ in the integral to get

$$|I_j| \le \rho^{-n(j-w)} \int_{x \in \bigcup_{z \in R_{j-w}} T_z} \underbrace{|\mathbb{1}_W(\{x\}_\Lambda) - \lambda(W)|}_{\le 1 + \lambda(W)}\,\mathrm{d}x.$$

After splitting up the integral and using translation to eliminate the fractional part, we get

$$|I_j| \leq \rho^{-n(j-w)} \left(1 + \lambda(W)\right) \sum_{z \in R_{j-w}} \int_{x \in T} \mathrm{d}x = \rho^{-n(j-w)} \left(1 + \lambda(W)\right) \lambda(T) \, \#(R_{j-w}) \, .$$

Using $\#(\partial(\Psi U)_T)_T = \mathcal{O}\Big(\big|\det \Psi\big|^{\delta/n}\Big)$ as assumed and (4.9.3) we gain

$$\#(R_{j-w}) = \left|\det\left(\Phi^{-J} N \Phi^{j-w}\right)\right|^{\delta/n} = \mathcal{O}\Big(\rho^{(j-w)\delta}\Big) \, ,$$

because $\left|\det\left(\Phi^{-J} N\right)\right| = \mathcal{O}(1)$, see Remark 4.9.5 on page 73, and $|\det \Phi| = \rho^n$. Thus

$$|I_j| = \mathcal{O}\Big(\rho^{\delta(j-w)-n(j-w)}\Big) = \mathcal{O}\Big(\rho^{(\delta-n)j}\Big) \, .$$

Now we want to make the summation in $\mathcal{P}_\eta$ independent from J, so we consider

$$I := \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=J+1}^{\infty} I_j$$

Again we use triangle inequality and we calculate the sum to obtain

$$|I| = \mathcal{O}\big(\rho^{nJ}\big) \sum_{j=J+1}^{\infty} \mathcal{O}\Big(\rho^{(\delta-n)j}\Big) = \mathcal{O}\Big(\rho^{nJ} \rho^{(\delta-n)J}\Big) = \mathcal{O}\Big(\rho^{\delta J}\Big) \, .$$

Note that $\mathcal{O}\big(\rho^J\big) = \mathcal{O}(N)$, so we obtain $|I| = \mathcal{O}\big(N^\delta\big)$.

Let us look at the growth of

$$\mathcal{P}_\eta(N) = \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} I_j .$$

We get

$$|\mathcal{P}_\eta(N)| = \mathcal{O}\big(\rho^{nJ}\big) \sum_{j=0}^{J} \mathcal{O}\Big(\rho^{(\delta-n)j}\Big) = \mathcal{O}\big(\rho^{nJ}\big) = \mathcal{O}(N^n) \, ,$$

using $\delta < n$.

Finally, inserting J from Lemma 4.9.4 and extending the sum to infinity, as described above, yields

$$\begin{aligned} \mathcal{P}_\eta(N) &= \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} \int_{y \in \{\Phi^{-J} N U\}_{T, j-w}} \left(\mathbb{1}_W\left(\left\{\Phi^{j-w} y\right\}_\Lambda\right) - \lambda(W)\right) \mathrm{d}y \\ &= N^n \, \psi_{\eta,\mathcal{P}}(\log N) + \mathcal{O}\Big(N^\delta\Big) \, . \end{aligned}$$

with the desired $\psi_{\eta,\mathcal{P}}$.

Now suppose (4.9.2) holds. Then

$$\begin{aligned}\Phi^{-\lfloor x\rfloor-J_0}\rho^x U &= \rho^x Q \operatorname{diag}\rho^{-\lfloor x\rfloor-J_0}e^{-i\theta_1(\lfloor x\rfloor+J_0)},\dots,\rho^{-\lfloor x\rfloor-J_0}e^{-i\theta_n(\lfloor x\rfloor+J_0)}Q^{-1}U\\ &= \rho^{\{x\}-J_0}Q\operatorname{diag}e^{-i\theta_1(\lfloor x\rfloor+J_0)},\dots,e^{-i\theta_n(\lfloor x\rfloor+J_0)}Q^{-1}U.\end{aligned}$$

Now, we can conclude that the region of integration in $\psi_{\eta,\mathcal{P}}(x)$ is p-periodic using (4.9.2). All other occurrences of x in $\psi_{\eta,\mathcal{P}}(x)$ are of the form $\{x\}$, i.e., 1-periodic, so period p is obtained. □

Lemma 4.9.9 (The Other Part)**.** *For* (4.9.1d) *in the proof of Theorem 4.9.1 on page 69 we get*

$$\mathcal{Q}_\eta(N) = N^n\psi_{\eta,\mathcal{Q}} + \mathcal{O}(N^\alpha\log N) + \mathcal{O}\Big(N^\delta\Big),$$

with

$$\psi_{\eta,\mathcal{Q}} = \frac{\lambda(U)}{\lambda(T)}\sum_{j=0}^{\infty}\frac{\beta_j}{\lambda(T)}$$

and $\alpha = n+\log\mu < n$*, where* $\mu<1$ *can be found in Theorem 4.3.7 on page 55.*

Proof. Consider

$$I_{j,\ell} := \int_{y\in\lfloor\Phi^{-J}NU\rfloor_{T,j-w}} \big(\mathbb{1}_{W_{\eta,\ell}} - \mathbb{1}_W\big)\big(\{\Phi^{j-w}y\}_\Lambda\big)\,\mathrm{d}y.$$

We can rewrite the region of integration and get

$$\big\lfloor\Phi^{-J}NU\big\rfloor_{T,j-w} = \Phi^{-(j-w)}\big\lfloor\Phi^{j-w}\Phi^{-J}NU\big\rfloor_T = \Phi^{-(j-w)}\bigcup_{z\in R_{j-w}}T_z$$

for some appropriate $R_{j-w}\subseteq\Lambda$, as in the proof of Lemma 4.9.7 on page 74. Substituting $x=\Phi^{j-w}y$, $\mathrm{d}x=\rho^{n(j-w)}\,\mathrm{d}y$ yields

$$I_{j,\ell} = \rho^{-n(j-w)}\int_{x\in\bigcup_{z\in R_{j-w}}T_z}\big(\mathbb{1}_{W_{\eta,\ell}}-\mathbb{1}_W\big)(\{x\}_\Lambda)\,\mathrm{d}x$$

and further

$$I_{j,\ell} = \rho^{-n(j-w)}\sum_{z\in R_{j-w}}\underbrace{\int_{x\in T}\big(\mathbb{1}_{W_{\eta,\ell}}-\mathbb{1}_W\big)(x)\,\mathrm{d}x}_{=\beta_\ell} = \rho^{-n(j-w)}\,\#(R_{j-w})\,\beta_\ell,$$

by splitting up the integral, using translation to eliminate the fractional part and taking β_ℓ according to (j) of Proposition 4.8.2 on page 67. From Proposition 4.7.3 on page 66 we obtain

$$\frac{\#(R_{j-w})}{\rho^{n(j-w)}} = \frac{\big|\det\big(\Phi^{-J}N\Phi^{j-w}\big)\big|}{\rho^{n(j-w)}}\frac{\lambda(U)}{\lambda(T)} + \mathcal{O}\left(\frac{\big|\det\big(\Phi^{-J}N\Phi^{j-w}\big)\big|^{\delta/n}}{\rho^{n(j-w)}}\right),$$

which can be rewritten as

$$\frac{\#(R_{j-w})}{\rho^{n(j-w)}} = \rho^{-nJ} N^n \frac{\lambda(U)}{\lambda(T)} + \mathcal{O}\Big(\rho^{(\delta-n)j}\Big)$$

because $|\det \Phi| = \rho^n$ and because $\left|\tau^{-J} N\right| = \mathcal{O}(1)$, see Remark 4.9.5 on page 73.

Now let us have a look at

$$\mathcal{Q}_\eta(N) = \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} I_{j,J-j}.$$

Inserting the result above and using $\beta_\ell = \mathcal{O}\big(\mu^\ell\big)$, see (j) of Proposition 4.8.2 on page 67, yields

$$\mathcal{Q}_\eta(N) = N^n \frac{\lambda(U)}{(\lambda(T))^2} \sum_{j=0}^{J} \beta_{J-j} + \rho^{nJ} \sum_{j=0}^{J} \mathcal{O}\Big(\rho^{(\delta-n)j}\Big)\, \mathcal{O}\big(\mu^{J-j}\big)\,.$$

Therefore, after reversing the order of the first summation, we obtain

$$\mathcal{Q}_\eta(N) = N^n \frac{\lambda(U)}{(\lambda(T))^2} \sum_{j=0}^{J} \beta_j + \rho^{nJ} \mu^J \sum_{j=0}^{J} \mathcal{O}\bigg(\Big(\mu\rho^{n-\delta}\Big)^{-j}\bigg)\,.$$

If $\mu\rho^{n-\delta} \geq 1$, then the second sum is $J\,\mathcal{O}(1)$, otherwise the sum is $\mathcal{O}\big(\mu^{-J}\rho^{(\delta-2)J}\big)$. So we obtain

$$\mathcal{Q}_\eta(N) = N^n \frac{\lambda(U)}{(\lambda(T))^2} \sum_{j=0}^{J} \beta_j + \mathcal{O}\big(\rho^{nJ}\mu^J J\big) + \mathcal{O}\Big(\rho^{\delta J}\Big)\,.$$

Using $J = \Theta(\log N)$, see Lemma 4.9.4 on page 73, and defining $\alpha = n + \log \mu$ yields

$$\mathcal{Q}_\eta(N) = N^n \frac{\lambda(U)}{(\lambda(T))^2} \sum_{j=0}^{J} \beta_j + \underbrace{\mathcal{O}\Big(N^{n+\log\mu} \log N\Big)}_{=\mathcal{O}(N^\alpha \log N)} + \mathcal{O}\Big(N^\delta\Big)\,.$$

Now consider the first sum. Since $\beta_j = \mathcal{O}\big(\mu^j\big)$, see (j) of Proposition 4.8.2 on page 67, we obtain

$$N^n \sum_{j=J+1}^{\infty} \beta_j = N^n\, \mathcal{O}\big(\mu^J\big) = \mathcal{O}(N^\alpha)\,.$$

Thus the lemma is proved, because we can extend the sum to infinity. □

Lemma 4.9.10 (The Small Part)**.** *For* (4.9.1e) *in the proof of Theorem 4.9.1 on page 69 we get*

$$\mathcal{S}_\eta(N) = \mathcal{O}(N^\alpha \log N) + \mathcal{O}\Big(N^\delta\Big)$$

with $\alpha = n + \log \mu < n$ *and* $\mu < 1$ *from Theorem 4.3.7 on page 55.*

Proof. Consider

$$I_{j,\ell} := \int_{y \in \{\Phi^{-J} NU\}_{T,j-w}} (\mathbb{1}_{W_\ell} - \mathbb{1}_W)\big(\{\Phi^{j-w} y\}_\Lambda\big)\,\mathrm{d}y.$$

Again, as in the proof of Lemma 4.9.8 on page 75, the region of integration satisfies

$$\{\Phi^{-J} NU\}_{T,j-w} \subseteq \partial\big(\Phi^{-J} NU\big)_{T,j-w} = \Phi^{-(j-w)} \bigcup_{z \in R_{j-w}} T_z \tag{4.9.4}$$

for some appropriate $R_{j-w} \subseteq \Lambda$.

We substitute $x = \Phi^{j-w} y$, $\mathrm{d}x = \rho^{n(j-w)}\,\mathrm{d}y$ in the integral to get

$$|I_{j,\ell}| = \rho^{-n(j-w)} \left| \int_{x \in \bigcup_{z \in R_{j-w}} T_z} (\mathbb{1}_{W_\ell} - \mathbb{1}_W)(\{x\}_\Lambda)\,\mathrm{d}x \right|.$$

Again, after splitting up the integral, using translation to eliminate the fractional part and the triangle inequality, we get

$$|I_{j,\ell}| \leq \rho^{-n(j-w)} \sum_{z \in R_{j-w}} \underbrace{\left| \int_{x \in T} (\mathbb{1}_{W_\ell} - \mathbb{1}_W)(x)\,\mathrm{d}x \right|}_{=|\beta_\ell|} = \rho^{-n(j-w)}\, \#(R_{j-w})\, |\beta_\ell|\,,$$

in which $|\beta_\ell| = \mathcal{O}\big(\mu^\ell\big)$ is known from (j) of Proposition 4.8.2 on page 67. Using $\#(\partial(\Psi U)_T)_T = \mathcal{O}\Big(|\det \Psi|^{\delta/n}\Big)$, Remark 4.9.5 on page 73, and (4.9.4) we get

$$\#(R_{j-w}) = \mathcal{O}\Big(\big|\det \Phi^{-J} N \Phi^{j-w}\big|^{\delta/n}\Big) = \mathcal{O}\Big(\rho^{\delta(j-w)}\Big)\,,$$

because $|\det \Phi| = \rho^n$ and $\big|\tau^{-J} N\big| = \mathcal{O}(1)$. Thus

$$|I_{j,\ell}| = \mathcal{O}\Big(\mu^\ell \rho^{(\delta-n)(j-w)}\Big) = \mathcal{O}\Big(\mu^\ell \rho^{(\delta-n)j}\Big)$$

follows by assembling all together.

Now we are ready to analyse

$$\mathcal{S}_\eta(N) = \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} I_{j,J-j}.$$

Inserting the result above yields

$$|\mathcal{S}_\eta(N)| = \frac{\rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} \mathcal{O}\Big(\mu^{J-j} \rho^{(\delta-n)j}\Big) = \frac{\mu^J \rho^{nJ}}{\lambda(T)} \sum_{j=0}^{J} \mathcal{O}\bigg(\Big(\mu \rho^{n-\delta}\Big)^{-j}\bigg)$$

and thus, by the same argument as in the proof of Lemma 4.9.9 on page 77,

$$|\mathcal{S}_\eta(N)| = \mu^J \rho^{nJ}\, \mathcal{O}\Big(J + \mu^{-J} \rho^{(\delta-n)J}\Big) = \mathcal{O}\big(\mu^J \rho^{nJ} J\big) + \mathcal{O}\Big(\rho^{\delta J}\Big)\,.$$

Finally, by using Lemma 4.9.4 on page 73 we obtain

$$|\mathcal{S}_\eta(N)| = \mathcal{O}(N^\alpha \log N) + \mathcal{O}\Big(N^\delta\Big)$$

with $\alpha = n + \log \mu$. Since $\mu < 1$, we have $\alpha < n$. □

Lemma 4.9.11 (The Fractional Cells Part)**.** *For* (4.9.1f) *in the proof of Theorem 4.9.1 on page 69 we get*

$$\mathcal{F}_\eta(N) = \mathcal{O}\Big(N^\delta \log N\Big)\,.$$

Proof. For the regions of integration in $\mathcal{F}_\eta$ we obtain

$$NU \setminus \lfloor NU \rceil_T \subseteq \lceil NU \rceil_T \setminus \lfloor NU \rfloor_T = \partial(NU)_T = \bigcup_{z \in R} T_z$$

and

$$\lfloor NU \rceil_T \setminus NU \subseteq \lceil NU \rceil_T \setminus \lfloor NU \rfloor_T = \partial(NU)_T = \bigcup_{z \in R} T_z$$

for some appropriate $R \subseteq \Lambda$ using Proposition 4.7.2 on page 65. Thus we get

$$|\mathcal{F}_\eta(N)| \le \frac{2}{\lambda(T)} \sum_{j=0}^{J} \int_{x \in \bigcup_{z \in R} T_z} \mathbb{1}_{W_j}\big(\{\Phi^{-j-w} x\}_\Lambda\big)\,\mathrm{d}x \le \frac{2}{\lambda(T)} \sum_{j=0}^{J} \sum_{z \in R} \int_{x \in T_z} \mathrm{d}x,$$

in which the indicator function was replaced by 1. Dealing with the sums and the integral, which is $\mathcal{O}(1)$, we obtain

$$|\mathcal{F}_\eta(N)| = (J+1) \# R\, \mathcal{O}(1)\,.$$

Since $J = \mathcal{O}(\log N)$, see Lemma 4.9.4 on page 73, and $\# R = \mathcal{O}\big(N^\delta\big)$, the desired result follows. □

Lemma 4.9.12. *If the ψ_η from Theorem 4.9.1 on page 69 is p-periodic for some $p \in \mathbb{N}$, then ψ_η is also continuous.*

Proof. There are two possible parts of ψ_η where a discontinuity could occur: the first is $\{x\}$ for an $x \in \mathbb{Z}$, the second is building $\{\ldots\}_{T,j-w}$ in the region of integration in $\psi_{\eta,\mathcal{P}}$.

The latter is no problem, i.e., no discontinuity, since

$$\int_{y \in \left\{\Phi^{-\lfloor x \rfloor - J_0} \rho^x U\right\}_{T,j-w}} \big(\mathbb{1}_W\big(\{\Phi^{j-w} y\}_\Lambda\big) - \lambda(W)\big)\,\mathrm{d}y$$
$$= \int_{y \in \Phi^{-\lfloor x \rfloor - J_0} \rho^x U} \big(\mathbb{1}_W\big(\{\Phi^{j-w} y\}_\Lambda\big) - \lambda(W)\big)\,\mathrm{d}y,$$

because the integral over the region $\left\lfloor \Phi^{-\lfloor x \rfloor - J_0} \rho^x U \right\rfloor_{T,j-w}$ is zero, see proof of Lemma 4.9.7 on page 74.

Now we deal with the continuity at $x \in \mathbb{Z}$. Let $m \in x + p\mathbb{Z}$, let $M = \rho^m$, and consider

$$Z_\eta(M) - Z_\eta(M-1)\,.$$

For an appropriate $a \in \mathbb{R}$ we get

$$Z_\eta(M) = aM^n \log M + M^n\,\psi_\eta(\log M) + \mathcal{O}(M^\alpha \log M) + \mathcal{O}\Big(M^\delta \log M\Big)\,,$$

and thus

$$Z_\eta(M) = aM^n m + M^n \underbrace{\psi_\eta(m)}_{=\psi_\eta(x)} + \mathcal{O}(M^\alpha m) + \mathcal{O}\Big(M^\delta m\Big)\,.$$

Further we obtain

$$\begin{aligned} Z_\eta(M-1) = a\,(M-1)^n \log(M-1) + (M-1)^n\,\psi_\eta(\log(M-1)) \\ + \mathcal{O}((M-1)^\alpha \log(M-1)) + \mathcal{O}\Big((M-1)^\delta \log(M-1)\Big)\,, \end{aligned}$$

and thus, using the abbreviation $L = \log\big(1 - M^{-1}\big)$ and $\delta \geq 1$,

$$Z_\eta(M-1) = aM^n m + M^n \underbrace{\psi_\eta(m+L)}_{=\psi_\eta(x+L)} + \mathcal{O}(M^\alpha m) + \mathcal{O}\Big(M^\delta m\Big)\,.$$

Therefore we obtain

$$\frac{Z_\eta(M) - Z_\eta(M-1)}{M^n} = \psi_\eta(x) - \psi_\eta(x+L) + \mathcal{O}\big(M^{\alpha-n} m\big) + \mathcal{O}\Big(M^{\delta-n} m\Big)\,.$$

Since $\#(MU \setminus (M-1)\,U)_T$ is clearly an upper bound for the number of w-NAFs with values in $MU \setminus (M-1)\,U$ and each of these w-NAFs has at most $\lfloor \log M \rfloor + J_0 + 1$ digits, see Lemma 4.9.4 on page 73, we obtain

$$Z_\eta(M) - Z_\eta(M-1) \leq \#(MU \setminus (M-1)\,U)_T\,(m + J_0 + 2)\,.$$

Using (b) of Proposition 4.7.3 on page 66 yields

$$Z_\eta(M) - Z_\eta(M-1) = \mathcal{O}\Big(M^\delta m\Big)\,.$$

Therefore we get

$$\psi_\eta(x) - \psi_\eta(x+L) = \mathcal{O}\Big(M^{\delta-n} m\Big) + \mathcal{O}\big(M^{\alpha-n} m\big) + \mathcal{O}\Big(M^{\delta-n} m\Big)\,.$$

Taking the limit $m \to \infty$ in steps of p, and using $\alpha < n$ and $\delta < n$ yields

$$\psi_\eta(x) - \lim_{\varepsilon \to 0^-} \psi_\eta(x+\varepsilon) = 0,$$

i.e., ψ_η is continuous at $x \in \mathbb{Z}$. □

4.10 Counting Digits in Conjunction with Hyperelliptic Curve Cryptography

As mentioned in the introduction, we are interested in numeral systems coming from hyperelliptic curve cryptography. There the base is an algebraic integer, where all conjugates have the same absolute value.

Let H be a hyperelliptic curve (or more generally an algebraic curve) of genus g defined over $\mathbb{F}_q$ (a field with q elements). The Frobenius endomorphism operates on the Jacobian variety of H and satisfies a characteristic polynomial $f \in \mathbb{Z}[T]$ of degree $2g$. This polynomial fulfils the equation

$$f(T) = T^{2g} L(1/T),$$

where $L(T)$ denotes the numerator of the zeta-function of H over $\mathbb{F}_q$, cf. Weil [107, 109]. The Riemann Hypothesis of the Weil Conjectures, cf. Weil [108], Dwork [32] and Deligne [25], states that all zeros of L have absolute value $1/\sqrt{q}$. Therefore all roots of f have absolute value $\sqrt{q}$.

Later we suppose that τ is a root of f, and we consider numeral systems with a base τ. But before, we describe getting from that setting to a lattice, which we need in Section 4.2. This is generally known and was also used in Heuberger and Krenn [50].

First consider a number field K of degree n. Denote the real embeddings of K by σ_1, $\ldots$, σ_s and the non-real complex embeddings of K by σ_{s+1}, $\overline{\sigma_{s+1}}$, $\ldots$, σ_{s+t}, $\overline{\sigma_{s+t}}$, where $\overline{\cdot}$ denotes complex conjugation and $n = s + 2t$. The *Minkowski map* $\Sigma\colon K \to \mathbb{R}^n$ maps $\alpha \in K$ to

$$(\sigma_1(\alpha), \ldots, \sigma_s(\alpha), \Re\sigma_{s+1}(\alpha), \Im\sigma_{s+1}(\alpha), \ldots, \Re\sigma_{s+t}(\alpha), \Im\sigma_{s+t}(\alpha)) \in \mathbb{R}^n.$$

Now let τ be an algebraic integer of degree n (as above, where τ was supposed to be a root of the characteristic polynomial f of the Frobenius endomorphism) and such that all its conjugates have the same absolute value $\rho > 1$. Note that the absolute value of the field norm of τ equals ρ^n. Set $K = \mathbb{Q}(\tau)$ and consider the order $\mathbb{Z}[\tau]$. We get a lattice $\Lambda = \Sigma(\mathbb{Z}[\tau])$ of degree n in the space $\mathbb{R}^n$. Application of the map $\Phi\colon \Lambda \to \Lambda$ on a lattice element should correspond to the multiplication by τ in the order, so we define Φ as block diagonal matrix by

$$\Phi := \operatorname{diag} \sigma_1(\tau), \ldots, \sigma_s(\tau), \begin{pmatrix} \Re\sigma_{s+1}(\tau) & -\Im\sigma_{s+1}(\tau) \\ \Im\sigma_{s+1}(\tau) & \Re\sigma_{s+1}(\tau) \end{pmatrix}, \ldots, \begin{pmatrix} \Re\sigma_{s+t}(\tau) & -\Im\sigma_{s+t}(\tau) \\ \Im\sigma_{s+t}(\tau) & \Re\sigma_{s+t}(\tau) \end{pmatrix}.$$

The eigenvalues of Φ are exactly the conjugates of τ, therefore all eigenvalues have absolute value ρ. For the norm $\|\cdot\|$ we choose the Euclidean norm $\|\cdot\|_2$. Then the corresponding operator norm fulfils

$$\|\Phi\| = \max\{|\sigma_j(\tau)| : j \in \{1, 2, \ldots, s+t\}\} = \rho.$$

In the same way we get $\left\|\Phi^{-1}\right\| = \rho^{-1}$.

Now let $T \subseteq \mathbb{R}^n$ be a set which tiles the $\mathbb{R}^n$ by the lattice Λ, choose w as in the set-up in Section 4.2, and let $\mathcal{D}$ be a reduced residue digit set modulo Φ^w corresponding to the tiling, cf. also Heuberger and Krenn [50]. Since our lattice Λ comes from the order $\mathbb{Z}[\tau]$ and our map Φ corresponds to the multiplication by τ map, the size of the digit set $\mathcal{D}$ is $\rho^{n(w-1)}(\rho^n - 1) + 1$, see [49] for details.

Since our set-up, see Section 4.2, is now complete, we get that Theorem 4.9.1 holds. We want to restate this for our special case of τ-adic w-NAF-expansions. This is done in Corollary 4.10.2. To prove periodicity of the function ψ_η in that corollary, we need the following lemma.

Lemma 4.10.1. *Suppose*

$$\Phi = Q \operatorname{diag} \rho e^{i\theta_1}, \dots, \rho e^{i\theta_n} Q^{-1},$$

where Q is a regular matrix and let $U = \mathcal{B}(0,1)$ be the unit ball. Then

$$Q \operatorname{diag} e^{i\theta_1}, \dots, e^{i\theta_n} Q^{-1} U = U.$$

Proof. Since Φ is normal, the matrix $Q \operatorname{diag} e^{i\theta_1}, \dots, e^{i\theta_n} Q^{-1}$ is unitary. Therefore balls are mapped to balls bijectively, which was to prove. □

Now, as mentioned above, we reformulate Theorem 4.9.1 for our τ-adic set-up. This gives the following corollary.

Corollary 4.10.2. *Let τ be an algebraic integer, where all conjugates have the same absolute value, denote the embeddings of $\mathbb{Q}(\tau)$ by $\sigma_1, \dots, \sigma_{s+t}$ as above, and define a norm by $\|z\|^2 = \sum_{i=1}^{s+t} d_i |\sigma_i(z)|^2$ with $d_1 = \dots = d_s = 1$ and $d_{s+1} = \dots = d_{s+t} = 2$.*

Let $0 \neq \eta \in \mathcal{D}$ and $N \in \mathbb{R}$ with $N > 0$. We denote the number of occurrences of the digit η in all width-w non-adjacent forms in $\mathbb{Z}[\tau]$, where the norm of its value is smaller than N, by

$$Z_\eta(N) = \sum_{\substack{z \in \mathbb{Z}[\tau] \\ \|z\| < N}} \sum_{j \in \mathbb{N}_0} [j\text{th digit of } z \text{ in its } w\text{-NAF-expansion equals } \eta].$$

Then we get

$$Z_\eta(N) = N^n \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)} E \log_\rho N + N^n \psi_\eta\left(\log_\rho N\right) + \mathcal{O}\left(N^\beta \log_\rho N\right),$$

where we have the constant of the expectation

$$E = \frac{1}{\rho^{n(w-1)}((\rho^n - 1)w + 1)},$$

cf. Theorem 4.3.7 on page 55, a function $\psi_\eta(x)$ which is 1-periodic and continuous and $\beta < n$.

Proof. We choose $U = \mathcal{B}(0,1)$ the unit ball in the $\mathbb{R}^n$. Then U is measurable, $d = 1$ and $\delta = n - 1 < n$. Further the n-dimensional Lebesgue measure of U equals $\frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}$. The condition $\#(\partial(NU)_T)_T = \mathcal{O}\left(N^\delta\right)$ can be checked easily. In the case of a quadratic τ this is done in [49]. The periodicity (and therefore continuity) of ψ_η follows from Lemma 4.10.1. We can choose $\beta = \max\{\alpha, n-1\}$. □

Chapter 5

On Linear Combinations of Units with Bounded Coefficients

This chapter contains the article [69] with the title "On Linear Combinations of Units with Bounded Coefficients and Double-Base Digit Expansions". It is joint work with Jörg Thuswaldner and Volker Ziegler. The article was accepted for publication by *Monatshefte für Mathematik* on September 5, 2012.

Abstract

Let $\mathfrak{o}$ be the maximal order of a number field. Belcher showed in the 1970s that every algebraic integer in $\mathfrak{o}$ is the sum of pairwise distinct units, if the unit equation $u+v=2$ has a non-trivial solution $u, v \in \mathfrak{o}^*$. We generalize this result and give applications to signed double-base digit expansions.

5.1 Introduction

In the 1960s Jacobson [54] asked, whether the number fields $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt{5})$ are the only quadratic number fields such that each algebraic integer is the sum of distinct units. Śliwa [92] solved this problem for quadratic number fields and showed that even no pure cubic number field has this property. These results were extended to cubic and quartic fields by Belcher [8, 9]. In particular, Belcher solved the case of imaginary cubic number fields completely by applying the following criterion, which now bears his name, cf. [9].

Belcher's Criterion. *Let F be a number field and $\mathfrak{o}$ the maximal order of F. Assume that the unit equation*

$$u+v=2, \qquad u, v \in \mathfrak{o}^*$$

has a solution $(u,v) \neq (1,1)$. Then each algebraic integer in $\mathfrak{o}$ is the sum of distinct units.

The problem of characterizing all number fields in which every algebraic integer is a sum of distinct units is still unsolved. Let us note that this problem is contained in Narkiewicz' list of open problems in his famous book [81, see page 539, Problem 18].

Recently the interest in the representation of algebraic integers as sums of units arose due to the contribution of Jarden and Narkiewicz [55]. They showed that in a given number field there does not exist an integer k, such that every algebraic integer can be written as the sum of at most k (not necessarily distinct) units. For an overview on this topic we recommend the survey paper due to Barroero, Frei, and Tichy [7]. Recently Thuswaldner and Ziegler [102] considered the following related problem. Let an order $\mathfrak{o}$ of a number field and a positive integer k be given. Does each element $\alpha \in \mathfrak{o}$ admit a representation as a linear combination $\alpha = c_1\varepsilon_1 + \cdots + c_\ell\varepsilon_\ell$ of units $\varepsilon_1, \ldots, \varepsilon_\ell \in \mathfrak{o}^*$ with coefficients $c_i \in \{1, \ldots, k\}$? This problem was attacked by using dynamical methods from the theory of digit expansions. In the present paper we address this problem again. In particular, we wish to generalize Belcher's criterion in a way to make it applicable to this problem.

In order to get the most general form, we refine the definition of the unit sum height given in [102].

Definition 5.1.1. Let F be some field of characteristic 0, Γ be a finitely generated subgroup of F^*, and $R \subset F$ be some subring of F. Assume that $\alpha \in R$ can be written as a linear combination

$$\alpha = a_1\nu_1 + \cdots + a_\ell\nu_\ell, \tag{5.1.1}$$

where $\nu_1, \ldots, \nu_\ell \in \Gamma \cap R$ are pairwise distinct and $a_1 \geq \cdots \geq a_\ell > 0$ are integers. If a_1 in the representation of α of the form (5.1.1) is chosen as small as possible, we call $\omega_{R,\Gamma}(\alpha) = a_1$ the *R-Γ-unit sum height of α*. In addition we define $\omega_{R,\Gamma}(0) := 0$ and $\omega_{R,\Gamma}(\alpha) := \infty$ if α admits no representation as a finite linear-combination of elements contained in $\Gamma \cap R$. Moreover, we define

$$\omega_\Gamma(R) = \max\{\omega_{R,\Gamma}(\alpha) \,:\, \alpha \in R\}$$

if the maximum exists. If the maximum does not exist we write

$$\omega_\Gamma(R) = \begin{cases} \omega & \text{if } \omega_{R,\Gamma}(\alpha) < \infty \text{ for each } \alpha \in R, \\ \infty & \text{if there exists } \alpha \in R \text{ such that } \omega_{R,\Gamma}(\alpha) = \infty, \end{cases}$$

where ω is a symbol (representing the cardinality of $\mathbb{N}$).

Let us note that for a number field F with the group of units Γ of an order $\mathfrak{o}$ of F we have $\omega_\Gamma(\mathfrak{o}) = \omega(\mathfrak{o})$, where $\omega(\mathfrak{o})$ is the unit sum height defined in [102].

With those notations our main result is the following.

Theorem 5.1.2. *Let $F \subset \mathbb{C}$ be a field and Γ a finitely generated subgroup of F^* with $-1 \in \Gamma$. Let R be a subring of F that is generated as a $\mathbb{Z}$-module by a finite set $\mathcal{E} \subset \Gamma \cap R$. Assume that for given integers $n \geq I \geq 2$ the equation*

$$u_1 + \cdots + u_I = n, \qquad u_1, \ldots, u_I \in \Gamma \cap R \tag{5.1.2}$$

has a solution $(u_1, \ldots, u_I) \neq (1, \ldots, 1)$. Then we have $\omega_\Gamma(R) \leq n - 1$.

The following section, Section 5.2, is devoted to the proof of Theorem 5.1.2. In the third section we apply our main theorem, Theorem 5.1.2, to some special orders of Shanks' simplest cubic fields. A special case of that theorem yields applications to double-base expansions. There we choose $F = \mathbb{Q}$, $R = \mathbb{Z}$ and $\Gamma = \langle -1, p, q \rangle$, where p and q are coprime integers. We discuss that in Section 5.4.

5.2 Proof of Theorem 5.1.2

We start this section by giving a short plan of the proof.

Plan of Proof. Let $\alpha \in R$ be arbitrary. Our goal is to find a representation of α of the form (5.1.1) in which the coefficients $a_1, \dots, a_\ell$ are all bounded by $n-1$. We first show that α can be represented as a linear combination of the form (5.1.1) with $\nu_1, \dots, \nu_\ell$ chosen in a particular way. The idea of the proof is rather simple and is based on induction over the total weight of this representation (this is the sum of all of its coefficients, see Definition 5.2.2). Start with a representation of α as above and choose a coefficient which is greater than or equal to n (if such a coefficient does not exist, we are finished). Now apply (5.1.2). This leads to a new representation of α of the form (5.1.1) whose total weight does not increase (and actually remains the same after excluding some trivial cases). This process is now repeated until we either have a representation in which all coefficients are bounded by $n-1$, or the support of the representation contains big gaps. In the first case we are finished. In the second case we can split the representation in two parts which are separated by a large gap. The total weight of each part is less than the total weight of the original representation of α. We thus use the induction hypothesis on both of them, so we get a new representation of each part with coefficients bounded by $n-1$. Now, since the gap between the supports of these two parts is large, they do not overlap after we apply (5.1.2) to them in the appropriate way and we can put them together to find a representation as desired also in this case. □

Now we start with the proof of Theorem 5.1.2. First we introduce some notations. For integers a and b we write

$$[\![a, b]\!] := \{a, a+1, \dots, b\}$$

for the integers in the interval from a to b. For tuples $\mathbf{x} = (x_1, \dots, x_M)$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_M)$ we set

$$\boldsymbol{\varepsilon}^{\mathbf{x}} := \varepsilon_1^{x_1} \dots \varepsilon_M^{x_M}.$$

Observe first that each element of R has at least one representation of the form (5.1.1). The coefficients of that representation are integers, but not necessarily smaller than n.

A. *There exists a K-th root of unity ζ, elements $\eta_1, \dots, \eta_L \in \mathcal{E}$, and multiplicatively independent elements $\varepsilon_1, \dots, \varepsilon_M \in \Gamma \cap R$, abbreviated as $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_M)$, such that*

$$u_i = \zeta^{k_i} \boldsymbol{\varepsilon}^{\mathbf{r}^{(i)}}, \qquad i \in [\![1, I]\!],$$

for some $k_1, \dots, k_I \in [\![0, K-1]\!]$ *and some* $\mathbf{r}^{(1)}, \dots, \mathbf{r}^{(I)} \in \mathbb{Z}^M$*, each* $\alpha \in R$ *can be written as*

$$\alpha = \sum_{k \in [\![0,K-1]\!]} \sum_{\ell \in [\![1,L]\!]} \sum_{\mathbf{x} \in \mathbb{Z}^M} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}}$$

with non-negative integers $a_{k,\ell,\mathbf{x}}$*, and such that no relation of the form*

$$\zeta^k \eta_i \boldsymbol{\varepsilon}^{\mathbf{x}} = \eta_j \boldsymbol{\varepsilon}^{\mathbf{y}}, \quad i \neq j$$

with integer vectors $\mathbf{x}$ *and* $\mathbf{y}$ *as exponents and* $k \in \mathbb{Z}$ *holds.*

Proof of ***A***. Let $u_1, \dots, u_I$ be as in (5.1.2). Choose a K-th root of unity $\zeta \in \Gamma \cap R$ (note that the torsion group of Γ is finite and cyclic) and multiplicatively independent $\varepsilon_1, \dots, \varepsilon_M \in \Gamma \cap R$ with $M \leq I$, such that

$$u_i = \zeta^{k_i} \varepsilon_1^{r_1^{(i)}} \dots \varepsilon_M^{r_M^{(i)}} = \zeta^{k_i} \boldsymbol{\varepsilon}^{\mathbf{r}^{(i)}} \qquad (i \in [\![1, I]\!])$$

holds for some $\mathbf{r}^{(1)}, \dots, \mathbf{r}^{(I)} \in \mathbb{Z}^M$. We set

$$r := \max \left\{ r_m^{(i)} \,:\, i \in [\![1, I]\!], m \in [\![1, M]\!] \right\} \tag{5.2.1}$$

and want to mention that we reference to that r later in this section.

Let us consider a finite subset $\{\eta_1, \dots, \eta_L\} \subset \mathcal{E}$ such that all $\alpha \in R$ can be written as a linear combination

$$\alpha = \sum_{k \in [\![0,K-1]\!]} \sum_{\ell \in [\![1,L]\!]} \sum_{\mathbf{x} \in \mathbb{Z}^M} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}}$$

with $a_{k,\ell,\mathbf{x}} \in \mathbb{Z}$ (which is possible since $\mathcal{E}$ finitely generates R as $\mathbb{Z}$-module). We can (and do) choose that finite subset such that no relation of the form

$$\zeta^k \eta_i \boldsymbol{\varepsilon}^{\mathbf{x}} = \eta_j \boldsymbol{\varepsilon}^{\mathbf{y}}, \quad i \neq j$$

with integer exponents and $k \in \mathbb{Z}$ holds.

Note that $\zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}} \in \Gamma \cap R$. Furthermore, we can choose the coefficients $a_{k,\ell,\mathbf{x}}$ to be non-negative, since, by assumption, we have $-1 \in \Gamma$, which allows us to choose the "signs" in our representation. □

From now on we suppose that ζ, $\eta_1, \dots, \eta_L$, and $\boldsymbol{\varepsilon}$ are fixed and given as in **A**. We use the following convention on representations.

Convention 5.2.1. Let $\alpha \in R$ and suppose we have a representation of α where the coefficients are denoted by $a_{k,\ell,\mathbf{x}}$ (small Latin letter with some index), i.e., α is written as

$$\alpha = \sum_{k \in [\![0,K-1]\!]} \sum_{\ell \in [\![1,L]\!]} \sum_{\mathbf{x} \in \mathbb{Z}^M} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}}$$

We denote by $A \subset \mathbb{Z}^M$ (capital Latin letter corresponding to the letter used for the coefficients) the minimal M-dimensional interval including all $\mathbf{x}$ with $a_{k,\ell,\mathbf{x}} \neq 0$. We write

$$A = [\![\underline{A}_1, \overline{A}_1]\!] \times \dots \times [\![\underline{A}_M, \overline{A}_M]\!] .$$

We omit the range of the indices k and ℓ since they are always the same. Thus α will be written as

$$\alpha = \sum_{k,\ell} \sum_{\mathbf{x}\in A} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}}.$$

An important quantity is the weight of a representation. It is defined as follows.

Definition 5.2.2. Let $\alpha \in R$ and suppose we have a representation as in **A**, i.e.,

$$\alpha = \sum_{k,\ell} \sum_{\mathbf{x}\in A} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}}.$$

with non-negative integers $a_{k,\ell,\mathbf{x}}$. We call the minimum of all

$$\sum_{k,\ell} \sum_{\mathbf{x}\in A} a_{k,\ell,\mathbf{x}}$$

among all possible representations (as above) of α the *total weight of* α and write w_α for it.

As mentioned in the plan of the proof of Theorem 5.1.2, we apply Equation (5.1.2) to an existing representation to get another one. In the following paragraph, we define that replacement step, which will then always be denoted by ✻.

✻ (Replacement Step)**.** Suppose we have a representation

$$\alpha = \sum_{k,\ell} \sum_{\mathbf{x}\in A} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}},$$

where at least one coefficient $a_{k,\ell,\mathbf{x}} \geq n$. We get a new representation by applying

$$u_1 + \cdots + u_I = n.$$

More precisely, if $u_i = \zeta^{k_i} \boldsymbol{\varepsilon}^{\mathbf{r}^{(i)}}$, then the coefficient $a_{k+k_i,\ell,\mathbf{x}+\mathbf{r}^{(i)}}$ is increased by 1 for each $i \in [\![1, I]\!]$ and $a_{k,\ell,\mathbf{x}}$ is replaced by $a_{k,\ell,\mathbf{x}} - n$.

The following statements **B** and **C** deal with two special cases.

B. *If $\alpha \in R$ with $w_\alpha < I$, then Theorem 5.1.2 holds.*

We use that statement as the basis of our induction on the total weight w.

*Proof of **B**.* Since $I \leq n$ we have $w_\alpha < n$. So the sum of all (non-negative) coefficients is smaller than n. Therefore all coefficients themselves are in $[\![0, n-1]\!]$, which proves the theorem in that special case. □

From now on suppose we have an $\alpha \in R$ with a representation

$$\alpha = \sum_{k,\ell} \sum_{\mathbf{x}\in A} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}},$$

which has minimal weight. That means, we have $w := w_\alpha$.

C. *If $I < n$, then Theorem 5.1.2 holds.*

*Proof of **C**.* Assume that there is a coefficient $a_{k,\ell,\mathbf{x}} \geq n$ in the representation of α. We apply ✻ to obtain a new representation. But because $I < n$, the new one has smaller total weight, which is a contradiction to the fact that w was chosen minimal. □

Because of **B** and **C** we suppose from now that $w \geq I$ and $I = n$. As indicated above, we prove Theorem 5.1.2 by induction on the total weight w of α. More precisely we want to prove the following claim by induction.

Claim 5.2.3. *Assume that $\alpha \in R$ has a representation*

$$\alpha = \sum_{k,\ell} \sum_{\mathbf{x} \in A} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}}$$

with non-negative integers $a_{k,\ell,\mathbf{x}}$ and with minimal total weight w. Then α has also a representation of the form

$$\alpha = \sum_{k,\ell} \sum_{\mathbf{x} \in G} g_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}}.$$

with integers $g_{k,\ell,\mathbf{x}} \in [\![0, n-1]\!]$ and where

$$G = [\![\underline{A}_1 - f(w), \overline{A}_1 + f(w)]\!] \times \cdots \times [\![\underline{A}_M - f(w), \overline{A}_M + f(w)]\!]$$

with $f(1) = 0$ and

$$f(w) = T(w)r + f(w-1) \qquad (w \in \mathbb{N}),$$

where

$$T(w) = (w + 2(w-1)f(w-1))^{Mw} K^w L^w.$$

In order to prove Theorem 5.1.2 it is sufficient to prove Claim 5.2.3. As already mentioned, we use induction on the total weight w of α. Note that the induction basis has been shown above in **B**.

Let us start by looking what happens if one applies ✻.

D. *Repeatedly applying ✻ yields pairwise "essentially different" representations of α.*

More precisely, by repeatedly applying ✻, it is not possible to get two representations

$$\alpha = \sum_{k,\ell} \sum_{\mathbf{x} \in A} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}} = \sum_{k,\ell} \sum_{\mathbf{x} \in A} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}+\mathbf{L}}$$

with some $\mathbf{L} \in \mathbb{Z}^M \setminus \{\mathbf{0}\}$.

*Proof of **D**.* Remember that we assumed $I = n$. First, let us note that we have

$$n \leq \sum_{i \in [\![1,n]\!]} |u_i|$$

because of Equation (5.1.2). Using the Cauchy-Schwarz inequality yields

$$n^2 \le \left(\sum_{i \in [\![1,n]\!]} 1 \cdot |u_i| \right)^2 \le n \sum_{i \in [\![1,n]\!]} |u_i|^2 .$$

Hence,

$$n < \sum_{i \in [\![1,n]\!]} |u_i|^2 ,$$

unless $|u_1| = \cdots = |u_n| = 1$ and $\sum_i u_i = n$, i.e., $u_1 = \cdots = u_n = 1$. Since the trivial solution has been excluded, we see that every application of ✻ makes the quantity

$$\sum_{k,\ell} \sum_{\mathbf{x} \in A} a_{k,\ell,\mathbf{x}} \left(|\varepsilon_1|^{x_1} \dots |\varepsilon_M|^{x_M} \right)^2 \tag{5.2.2}$$

larger, i.e., the quantity (5.2.2) coming from coefficients $a'_{k,\ell,\mathbf{x}}$ is larger than (5.2.2) from $a_{k,\ell,\mathbf{x}}$, where the $a'_{k,\ell,\mathbf{x}}$ are the coefficients after an application of ✻ on a representation with coefficients $a_{k,\ell,\mathbf{x}}$. Note that the $\varepsilon_1, \dots, \varepsilon_M$ are fixed, cf. statement **A**.

Hence, repeatedly applying ✻ produces pairwise disjoint representations. Moreover, we cannot get the same representation up to linear translation in the exponents twice, i.e., we cannot get representations

$$\alpha = \sum_{k,\ell} \sum_{\mathbf{x} \in A} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}} = \sum_{k,\ell} \sum_{\mathbf{x} \in A} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}+\mathbf{L}}$$

with $\mathbf{L} \in \mathbb{Z}^M \setminus \{\mathbf{0}\}$. Such a relation would imply that $\boldsymbol{\varepsilon}^{\mathbf{L}} = 1$, which is a contradiction to the assumption that the $\varepsilon_1, \dots, \varepsilon_M$ are multiplicatively independent. □

Now we look what happens after sufficiently many applications of ✻.

E. *Set*

$$T(w) := \left(w + 2(w-1)f(w-1)\right)^{Mw} K^w L^w$$

and suppose we have a representation

$$\alpha = \sum_{k,\ell} \sum_{\mathbf{x} \in A} a_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}}.$$

After at most $T(w)$ applications of ✻ we get a representation

$$\alpha = \sum_{k,\ell} \sum_{\mathbf{x} \in B} b_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}},$$

such that one of the following assertions is true:

1. *Each coefficient satisfies $b_{k,\ell,\mathbf{x}} \in [\![0, n-1]\!]$ and*

$$\overline{B}_m - \underline{B}_m \le w + 2(w-1)f(w-1)$$

holds for all $m \in [\![1, M]\!]$.

2. There exists an index m such that

$$\overline{B}_m - \underline{B}_m > w + 2(w-1)f(w-1)$$

holds.

*Proof of **E**.* Each replacement step ✻ yields an essentially different representation, see **D**, and there are at most $T(w)$ possibilities to distribute our new coefficients in an interval $[\![0, K-1]\!] \times [\![1, L]\!] \times B$ with

$$\overline{B}_m - \underline{B}_m \leq w + 2(w-1)f(w-1)$$

for each m with $1 \leq m \leq M$. Therefore after at most $T(w)$ replacement steps we are either in case 1 or in case 2 of **E**. □

F. *With the setup and notations of **E**, a possible "translation of the indices" stays small. More precisely, we have*

$$\max\{|\underline{A}_m - \underline{B}_m| \,:\, m \in [\![1, M]\!]\} \leq T(w)r,$$

and

$$\max\left\{\left|\overline{A}_m - \overline{B}_m\right| \,:\, m \in [\![1, M]\!]\right\} \leq T(w)r,$$

where r is as defined as in (5.2.1).

*Proof of **F**.* The quantity r is the maximum of all exponents in the representation of the u_i as powers of the $\varepsilon_1, \ldots, \varepsilon_M$. Thus, an application of ✻ can change the exponents, and therefore the upper and lower bounds, respectively, by at most r. We have at most $T(w)$ applications of ✻, so the statement follows. □

Now we look at the two different cases of **E**. The first one leads to a result directly, whereas in the second one we have to use the induction hypothesis to get a representation as desired.

G. *If we are in case (1) of **E**, then we are "finished".*

*Proof of **G**.* Since

$$\left|\overline{A}_m - \overline{B}_m\right| \leq T(w)r < T(w)r + f(w-1) = f(w)$$

and

$$|\underline{A}_m - \underline{B}_m| \leq T(w)r < T(w)r + f(w-1) = f(w)$$

hold for each $m \in \mathbb{N}$ we have found a representation as desired in Claim 5.2.3. □

H. *If we are in case (2) of **E**, then we can split the representation into two parts and between them there is a "large gap".*

More precisely, there is a constant c such that we can write $\alpha = \gamma + \delta$ with

$$\gamma = \sum_{k,\ell} \sum_{\substack{\mathbf{x} \in B \\ x_m < c}} b_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}} \neq 0$$

and

$$\delta = \sum_{k,\ell} \sum_{\substack{\mathbf{x}\in B \\ x_m > c+2f(w-1)}} b_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}} \neq 0.$$

*Proof of **H**.* In case 2 of **E** we have an index $m \in [\![1, M]\!]$ with

$$\overline{B}_m - \underline{B}_m \geq w + 2(w-1)f(w-1)$$

The total weight of α is w, so the representation

$$\alpha = \sum_{k,\ell} \sum_{\mathbf{x}\in B} B_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}},$$

has at most w non-zero coefficients. Therefore, by the pigeon hole principle we can find an interval J of length at least $2f(w-1)$ and with the property that all coefficients $a_{\mathbf{x},i}$ fulfilling $x_m \in J$ are zero. Therefore we can split up α as mentioned. □

I. *If we have the splitting described in **H**, then Claim 5.2.3 follows for weight w.*

*Proof of **I**.* After renaming the intervals and coefficients, we have $\alpha = \gamma + \delta$ with

$$\gamma = \sum_{k,\ell} \sum_{\mathbf{x}\in C} c_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}}$$

and

$$\delta = \sum_{k,\ell} \sum_{\mathbf{x}\in D} d_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}}.$$

Both total weights w_γ and w_δ, respectively, are smaller than $w = w_\alpha$, so we can use induction hypothesis: We get representations

$$\gamma = \sum_{k,\ell} \sum_{\mathbf{x}\in E} e_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}} \tag{5.2.3}$$

with $e_{k,\ell,\mathbf{x}} \in [\![0, n-1]\!]$ and

$$\delta = \sum_{k,\ell} \sum_{\mathbf{x}\in F} f_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}} \tag{5.2.4}$$

with $f_{k,\ell,\mathbf{x}} \in [\![0, n-1]\!]$. The upper and lower bounds of the intervals in C to E differ by at most $f(w_\gamma) \leq f(w-1)$ in each coordinate. The same is valid for the intervals of D to F. Since the intervals in C and D were separated by intervals of length at least $2f(w-1)$, therefore the intervals in E and F are disjoint. In other words, the two representations in (5.2.3) and (5.2.4) do not overlap. So we can add these two representations and obtain

$$\alpha = \sum_{k,\ell} \sum_{\mathbf{x}\in G} g_{k,\ell,\mathbf{x}} \zeta^k \eta_\ell \boldsymbol{\varepsilon}^{\mathbf{x}}$$

with $g_{k,\ell,\mathbf{x}} \in [\![0, n-1]\!]$. We have

$$\max\left\{\left|\overline{G}_m - \overline{A}_m\right| : m \in [\![1, M]\!]\right\} \leq T(w)r + f(w-1) = f(w)$$

and

$$\max\left\{\left|\underline{G}_m - \underline{A}_m\right| : m \in [\![1, M]\!]\right\} \leq T(w)r + f(w-1) = f(w),$$

which finishes the proof. □

5.3 The Case of Simplest Cubic Fields

Let a be an integer and let α be a root of the polynomial

$$X^3 - (a-1)X^2 - (a+2)X - 1.$$

Then the family of real cubic fields $\mathbb{Q}(\alpha)$ is called the family of Shanks' simplest cubic fields. These fields and the orders $\mathbb{Z}[\alpha]$ have been investigated by several authors. In particular, in a recent paper of the second and third author [102] it was shown that the unit sum height of the orders $\mathbb{Z}[\alpha]$ is 1 in case of $a = 0, 1, 2, 3, 4, 6, 13, 55$ and the unit sum height ≤ 2 in case of $a = 5$. Moreover, it was conjectured that $\omega(\mathbb{Z}[\alpha]) = 1$ for all $a \in \mathbb{Z}$.

Using our main theorem we are able to prove the following result.

Theorem 5.3.1. *We have $\omega(\mathbb{Z}[\alpha]) \leq 2$ for all $a \in \mathbb{Z}$.*

Proof. First let us note some important facts on $\mathbb{Q}(\alpha)$ and $\mathbb{Z}[\alpha]$, see for example Shanks' original paper [90]. We know that $\mathbb{Q}(\alpha)$ is Galois over $\mathbb{Q}$ with Galois group $G = \{\mathrm{id}, \sigma, \sigma^2\}$ and with $\alpha_2 = \sigma(\alpha) = -1 - \frac{1}{\alpha}$. If we set $\alpha_1 := \alpha$, then α_1 and α_2 are a fundamental system of units. Now we know enough about the structure of $\mathbb{Z}[\alpha]$ to apply Theorem 5.1.2.

If we can find three units $u_1, u_2, u_3 \in \mathbb{Z}[\alpha]^*$ such that $u_1 + u_2 + u_3 = 3$ and $u_i \neq 1$, then the theorem is a direct consequence of Theorem 5.1.2. Indeed we have

$$\begin{aligned} 3 &= \overbrace{(\alpha_1^2 + (-a+2)\alpha_1 - a)}^{=u_1} \\ &\quad + \overbrace{(-2\alpha_1^2 + (2a-1)\alpha_1 + a + 4)}^{=u_2} \\ &\quad + \overbrace{(\alpha_1^2 + (-a-1)\alpha_1 - 1)}^{=u_3} \\ &= \alpha_1\alpha_2^2 + \alpha_1^{-2}\alpha_2^{-1} + \alpha_1\alpha_2^{-1}. \end{aligned}$$

□

5.4 Application to Signed Double-Base Expansions

We start with the definition of a signed double-base expansion of an integer.

Definition 5.4.1 (Signed Double-Base Expansion)**.** Let p and q be different integers. Let n be an integer with

$$\alpha = \sum_{i \in \mathbb{N}_0, j \in \mathbb{N}_0} d_{ij} p^i q^j,$$

where $d_{ij} \in \{-1, 0, 1\}$ and only finitely many d_{ij} are non-zero. Then such a sum is called a *signed p-q-double-base expansion of α*. The pair (p, q) is called *base pair*.

A natural first question is, whether each integer has a signed double-base expansion for a fixed base pair.

If one of the bases p and q is either 2 or 3, then existence follows since every integer has a *binary representation* (base 2 with digit set $\{0,1\}$) and a *balanced ternary representation* (base 3 with digit set $\{-1,0,1\}$), respectively. To get the existence results for general base pairs, we use the following theorem, cf. [15]

Theorem 5.4.2 (Birch)**.** *Let p and q be coprime integers. Then there is a positive integer $N(p,q)$ such that every integer larger than $N(p,q)$ may be expressed as a sum of distinct numbers of the form p^iq^j all with non-negative integers i and j.*

Corollary 5.4.3. *Let p and q be coprime integers. Then each integer has a signed p-q-double-base expansion.*

Next we want to give an efficient algorithm that allows to calculate a signed double base expansion of a given integer. Birch's theorem, or more precisely the proof in [15], does not provide an efficient way to do that. However, using our main result, there is a way to compute such expansions efficiently at least for certain base pairs.

Corollary 5.4.4. *Let p and q be coprime integers with absolute value at least 3. If there are non-negative integers x and y such that*

$$2 = |p^x - q^y|\,, \tag{5.4.1}$$

then each integer has a signed p-q-double-base expansion which can be computed efficiently (there exists a polynomial time algorithm). In particular given a p-adic expansion of an integer α, one has to apply (5.4.1) at most $O(\log(\alpha)^2)$ times.

Proof. We start to prove the first part of the corollary and therefore apply Theorem 5.1.2 with $\mathbb{F} = \mathbb{Q}$, $R = \mathbb{Z}$ and Γ is the multiplicative group generated by $-1, p$ and q. Since by assumption $2 = \pm(p^x - q^y)$ we have a solution to (5.1.2) and Theorem 5.1.2 yields that p-q-double-base expansions exist.

Now let us prove the statement on the existence of a polynomial time algorithm. Assume that for the integer α the p-adic expansion

$$\alpha = a_0 + a_1 p + \cdots + a_k p^k$$

is given, with $a_0, \ldots, a_k \in [\![0, p-1]\!]$. Let us note that the weight w of this representation is at most $O(\log \alpha)$. Now the following claim yields the corollary. □

Claim 5.4.5. *Assume*

$$\alpha = \sum_{i \in [\![0,I]\!]} a_i p^i$$

with $a_i \in \mathbb{Z}$ and $I \in \mathbb{N}_0$, and set $w = \sum_{i\in[\![0,I]\!]} |a_i|$. Then, after at most $\frac{w^2-w}{2}$ replacement steps ✻ we arrive in a representation of the form

$$\alpha = \sum_{j \in [\![0,J]\!]} q^{jy} \sum_{k \in [\![0,K]\!]} b_{k,j} p^k,$$

where the $b_{k,j}$ are integers with $|b_{k,j}| \le 1$, and $J, K \in \mathbb{N}_0$.

Proof. We prove the claim by induction on w. If $w \leq 1$ the statement of the claim is obvious. Further, if all the a_i are in $\{-1, 0, 1\}$ we are done. Therefore we assume that there is at least one index i with $|a_i| > 1$.

We now apply the replacement step ✻ in the following way: If $a_i > 1$, then a_i is replaced by $a_i - 2$, if $a_i < 1$, then a_i is replaced by $a_i + 2$. After at most $w - 1$ such steps, we get a new representation of the form

$$\alpha = \sum_{i \in [\![0, I_c]\!]} c_i p^i + q^y \sum_{i \in [\![0, I_d]\!]} d_i p^i,$$

$I_c, I_d \in \mathbb{N}_0$, $c_i, d_i \in \mathbb{Z}$, such that all c_i fulfil $|c_i| \leq 1$. Note that no replacement step ✻ increases the weight w.

Now consider

$$\beta = \sum_{i \in [\![0, I_d]\!]} d_i p^i.$$

The weight of β fulfils

$$w_\beta = \sum_{i \in [\![0, I_d]\!]} |d_i| \leq w - 1,$$

since in each replacement step it is increased exactly by 1. Now, by the induction hypothesis we obtain a representation

$$\beta = \sum_{j \in [\![0, J_e]\!]} q^{jy} \sum_{k \in [\![0, K_e]\!]} e_{k,j} p^k,$$

where the $e_{k,j}$ are integers with $|e_{k,j}| \leq 1$ and $J_e, K_e \in \mathbb{N}_0$. Further, this can be done in $\frac{w_\beta^2 - w_\beta}{2}$ steps. Setting $b_{i,0} = c_i$ and $b_{i,k} = e_{i,k-1}$ for $k > 0$ yields the desired representation. Moreover, this can be done with at most

$$\frac{w_\beta^2 - w_\beta}{2} + w - 1 \leq \frac{(w-1)(w-2)}{2} + w - 1 = \frac{w(w-1)}{2}$$

applications of ✻, which finishes the proof of the claim. □

Now we want to give some examples for base pairs, where the corollary can be used.

Example 5.4.6. Let (p, q) be a *twin prime pair*, i.e., we have $q = p + 2$ and both p and q are primes. Then clearly

$$2 = q - p,$$

so, by Corollary 5.4.4, every integer has a signed p-q-double-base expansion, which can be calculated efficiently.

Example 5.4.7. Let $p = 5$ and $q = 23$. We have

$$2 = 5^2 - 23,$$

therefore every integer has a signed 5-23-double-base expansion, which can be calculated efficiently. Again Corollary 5.4.4 was used.

To see some concrete expansions, we calculated the following:

$$\begin{aligned}
995 &= -5^5 + 5^4 + 5^3 \cdot 23 - 5^2 + 5 \cdot 23 + 23^2 + 1 \\
996 &= -5^3 + 5^2 \cdot 23 + 23^2 - 5 + 23 - 1 \\
997 &= -5^3 + 5^2 \cdot 23 + 23^2 - 5 + 23 \\
998 &= 5^4 - 5^3 - 5^2 + 23^2 - 5 - 1 \\
999 &= 5^4 - 5^3 - 5^2 + 23^2 - 5 \\
1000 &= 5^4 - 5^3 - 5^2 + 23^2 - 5 + 1 \\
1001 &= -5^3 + 5^2 \cdot 23 + 23^2 + 23 - 1 \\
1002 &= -5^3 + 5^2 \cdot 23 + 23^2 + 23 \\
1003 &= 5^4 - 5^3 - 5^2 + 23^2 - 1
\end{aligned}$$

In each case we started with an initial expansion, which is obtained by a greedy algorithm: For a $v \in \mathbb{Z}$ find the closest $5^i \cdot 23^j$, change the coefficient for that base, and continue with $v - 5^i \cdot 23^j$. Then we calculated the expansion by applying the equation $2 = 5^2 - 23$ as in the proof of Theorem 5.1.2. The implementation[1] was done in Sage [97].

One can find pairs (p, q) where Corollary 5.4.4 does not work. The following remark discusses some of those pairs.

Remark 5.4.8. Consider the equation

$$2 = |p^x - q^y| \tag{5.4.2}$$

with non-negative integers x, y. A first example, where the corollary fails, is $p = 5$ and $q = 11$. Indeed, looking at Equation (5.4.2) modulo 5 yields a contradiction. Another example is $p = 7$ and $q = 13$, where looking at (5.4.2) modulo 7, yields a contradiction. A third example is $p = 7$ and $q = 11$.

So in the cases given in the remark above, as well as in a lot of other cases, we cannot use the corollary to compute a signed double-base expansion efficiently. This leads to the following question.

Question 5.4.9. Is there an efficient (polynomial time) algorithm for each base pair (p, q) to compute a signed p-q-double-base expansion for all integers?

There is also another way to use Theorem 5.1.2. For some combinations of p and q we can get a weaker result. First, we define an extension of the signed double-base expansion: we allow negative exponents in the $p^i q^j$, too.

Definition 5.4.10 (Extended Signed Double-Base Expansion)**.** Let p and q be different integers (usually coprime). Let $z \in \mathbb{Q}$. If we have

$$z = \sum_{i \in \mathbb{Z}, j \in \mathbb{Z}} d_{ij} p^i q^j,$$

[1] The source code can be found on `http://www.danielkrenn.at/belcher/`. Further a full list of expansions of the natural numbers up to 10000 can be found there.

where $d_{ij} \in \{-1, 0, 1\}$ and only finitely many d_{ij} are non-zero, then we call the sum an *extended signed p-q-double-base expansion of z.*

With that definition, we can prove the following corollary to Theorem 5.1.2.

Corollary 5.4.11. *Let p and q be coprime integers. If there are integers a, b, c, and d with $(a, b, c, d) \neq (0, 0, 0, 0)$ and such that*

$$2 = p^a q^b \pm p^c q^d, \tag{5.4.3}$$

then every element of $\mathbb{Z}[1/p, 1/q]$ has an extended signed p-q-double-base expansion which can be computed efficiently (polynomial time algorithm).

Remark 5.4.12. If we have a solution to the equation in Corollary 5.4.4, then Corollary 5.4.11 works, too. But more can be said about the existence and efficient computability of extended double-base expansions for the elements of $\mathbb{Z}[1/p, 1/q]$. If each integer has an efficient computable signed p-q-double-base expansion, then each element of $\mathbb{Z}[1/p, 1/q]$ has an extended signed p-q-double-base expansion which can be computed efficiently. This result is not difficult to prove.

Now we prove the corollary.

Proof of Corollary 5.4.11. The proof of this corollary runs along the same lines as the proof of Corollary 5.4.4.

We apply Theorem 5.1.2 with $\mathbb{F} = \mathbb{Q}$, $R = \mathbb{Z}[1/p, 1/q]$ and Γ is the multiplicative group generated by -1, p and q. Since, by assumption, $2 = \pm(p^a q^b - p^c q^d)$ we have a solution to (5.1.2), Theorem 5.1.2 yields that p-q-double-base expansions exist.

Next, we claim that we may assume p and q are odd and $p, q > 3$. Indeed assuming that $p \in \{2, 3\}$, then we can write $\alpha \in \mathbb{Z}[1/p, 1/q]$ in the form

$$\alpha = \frac{\widetilde{\alpha}}{p^{x_p} q^{x_q}}$$

with $\widetilde{\alpha} \in \mathbb{Z}$ and appropriate exponents x_p and x_q. Moreover, $\widetilde{\alpha}$ has a representation of the form

$$\widetilde{\alpha} = \sum_{i \in [\![0,k]\!]} a_i p^i$$

with $a_i \in \{-1, 0, 1\}$. However the computation of such a represenation can be done efficiently and takes polynomial time in the height $h(\alpha)$, where

$$h(n/m) = \max\{\log|n|, \log|m|, 1\}$$

provided $n, m \in \mathbb{Z}$ are coprime.

Since we may assume $p, q > 3$, we want to show next that a solution to equation (5.4.3) necessarly takes the form

$$2 = \pm p^{-a} \pm p^{-a} q^b,$$

with $a, b \geq 0$. We observe that a solution to (5.4.3) with $a, c > 0$ or $b, d > 0$ does not exist, since otherwise $p \mid 2$ or $q \mid 2$. Next we note that if $a \neq c$ ($b \neq d$ respectively) the p-adic valuation (q-adic valuation) on the right hand side of (5.4.3) would be the minimum of a and c (b and d respectively) and in view of the left hand side, this minimum must be 0. Thus any solution to equation (5.4.3) must be of one of the following forms:

$$\begin{aligned} 2 &= \pm p^a q^b \pm 1, \\ 2 &= \pm p^{-a} q^{-b} \pm p^{-a} q^{-b}, \\ 2 &= \pm p^{-a} \pm p^{-a} q^b, \end{aligned}$$

or

$$2 = \pm p^a \pm q^b,$$

where a and b are positive integers. Obviously the first two cases have no solution and the last case has been treated in Corollary 5.4.4.

Now let us write $\alpha \in \mathbb{Z}[1/p, 1/q]$ in the form

$$\alpha = \frac{a_0 + a_1 p + \cdots + a_k p^k}{q^{x_q} p^{x_p}}.$$

We are now in a similar situation as in the proof of Corollary 5.4.4. Let $w = \sum_{i=1}^{k} |a_i|$. Then by similar arguments as in Corollary 5.4.4 we find an extended signed p-q-double-base expansion of α with at most $\frac{w^2-w}{2}$ applications of $\divideontimes$. Thus we have a polynomial in $h(\alpha)$ time algorithm. □

We can use the corollary proved above to get the following examples.

Example 5.4.13. Let p be a *Sophie Germain prime* and $q = 2p + 1$. We obtain

$$2 = qp^{-1} - p^{-1}.$$

Using Corollary 5.4.11 yields that every element of $\mathbb{Z}[1/p, 1/q]$ has an efficient computable extended signed p-q-double-base expansion.

The case when p is a prime and $q = 2p - 1$ is a prime works analogously.

The end of this section is dedicated to a short discussion. All the results on efficient computability in this section needed a special representation of 2. We have given some pairs (p, q) where the methods given here do not work.

Further, one could ask, whether the representations we get have a special structure. Of particular interest would be an algorithm to get expansions with a small number of summands (small number of non-zero digits). For a given base pair (p, q) this leads to the following question

Question 5.4.14. How to compute a signed p-q-double-base expansion with minimal weight for a given integer?

A greedy approach for solving this question can be found in Berthé and Imbert [13], some further results can be found in Dimitrov and Howe [27].

Chapter 6

Sylow p-groups of Polynomial Permutations

This chapter contains the article [41] with the title "Sylow p-groups of Polynomial Permutations on the Integers mod p^n". It is joint work with Sophie Frisch. The article is submitted to *Journal of Number Theory*.

Abstract

We enumerate and describe the Sylow p-groups of the group of polynomial permutations of the integers mod p^n.

6.1 Introduction

Fix a prime p and let $n \in \mathbb{N}$. Every polynomial $f \in \mathbb{Z}[x]$ defines a function from $\mathbb{Z}_{p^n} = \mathbb{Z}/p^n\mathbb{Z}$ to itself. If this function happens to be bijective, it is called a *polynomial permutation* of $\mathbb{Z}_{p^n}$. The polynomial permutations of $\mathbb{Z}_{p^n}$ form a group $(G_n, \circ)$ with respect to composition. The order of this group has been known since at least 1921 (Kempner [61]) to be

$$|G_2| = p!(p-1)^p p^p \quad \text{and } |G_n| = p!(p-1)^p p^p p^{\sum_{k=3}^{n} \beta(k)} \quad \text{for } n \geq 3,$$

where $\beta(k)$ is the least n such that p^k divides $n!$, but the structure of $(G_n, \circ)$ is elusive. (See, however, Nöbauer [84] for some partial results). Since the order of G_n is divisible by a high power of $(p-1)$ for large p, even the number of Sylow p-groups is not obvious.

We will show that there are $(p-1)!(p-1)^{p-2}$ Sylow p-groups of G_n and describe these Sylow p-groups, see Theorem 4.5.

Some notation: p is a fixed prime throughout. A function $g\colon \mathbb{Z}_{p^n} \to \mathbb{Z}_{p^n}$ arising from a polynomial in $\mathbb{Z}_{p^n}[x]$ or, equivalently, from a polynomial in $\mathbb{Z}[x]$, is called a *polynomial function* on $\mathbb{Z}_{p^n}$. We denote by $(F_n, \circ)$ the monoid with respect to composition of polynomial functions on $\mathbb{Z}_{p^n}$, and by $(G_n, \circ)$ its group of units, the group of polynomial permutations of $\mathbb{Z}_{p^n}$.

The natural projection of polynomial functions on $\mathbb{Z}_{p^{n+1}}$ onto polynomial functions on $\mathbb{Z}_{p^n}$ we write as $\pi_n \colon F_{n+1} \to F_n$. If f is a polynomial in $\mathbb{Z}[x]$ (or in $\mathbb{Z}_{p^m}[x]$ for $m \geq n$) we denote the polynomial function on $\mathbb{Z}_{p^n}[x]$ induced by f by $[f]_{p^n}$.

The order of F_n and that of G_n have been determined by Kempner [61] in a rather complicated manner. His results were cast into a simpler form by Nöbauer [83] and Keller and Olson [60] among others. Since then there have been many generalizations of the order formulas to more general finite rings [89, 82, 20, 40, 14, 57, 58]. Also, polynomial permutations in several variables (permutations of $(\mathbb{Z}_{p^n})^k$ defined by k-tuples of polynomials in k variables) have been looked into [39, 22, 110, 106, 105, 72].

6.2 Polynomial functions and permutations

To put things in context, we recall some well-known facts, to be found, among other places, in [61, 83, 21, 60]. The reader familiar with polynomial functions on finite rings is encouraged to skip to section 3. This section does not contain new material but reviews the state of the art.

Definition 6.2.1. For p prime and $n \in \mathbb{N}$, let

$$\alpha_p(n) = \sum_{k=1}^{\infty} \left\lfloor \frac{n}{p^k} \right\rfloor \qquad \text{and} \qquad \beta_p(n) = \min\{m \mid \alpha_p(m) \geq n\}.$$

If p is fixed, we just write $\alpha(n)$ and $\beta(n)$.

Notation 6.2.2. For $k \in \mathbb{N}$, let $(x)_k = x(x-1)\dots(x-k+1)$ and $(x)_0 = 1$. We denote p-adic valuation by v_p.

Fact 6.2.3.

(1) $\alpha_p(n) = v_p(n!)$.

(2) *For* $1 \leq n \leq p$, $\beta_p(n) = np$ *and for* $n > p$, $\beta_p(n) < np$.

(3) *For all* $n \in \mathbb{Z}$, $v_p((n)_k) \geq \alpha_p(k)$; *and* $v_p((k)_k) = v_p(k!) = \alpha_p(k)$.

Proof. Easy. □

Remark 6.2.4. The sequence $(\beta_p(n))_{n=1}^{\infty}$ is obtained by going through the natural numbers in increasing order and repeating each $k \in \mathbb{N}$ $v_p(k)$ times. For instance, $\beta_2(n)$ for $n \geq 1$ is: $2, 4, 4, 6, 8, 8, 8, 10, 12, 12, 14, 16, 16, 16, 16, 18, 20, 20, \dots$.

The falling factorials $(x)_0 = 1$, $(x)_k = x(x-1)\dots(x-k+1)$, $k > 0$, form a basis of the free $\mathbb{Z}$-module $\mathbb{Z}[x]$, and representation with respect to this basis gives a convenient canonical form for a polynomial representing a given polynomial function on $\mathbb{Z}_{p^n}$.

Fact 6.2.5. *A polynomial* $f \in \mathbb{Z}[x]$, $f = \sum_k a_k\,(x)_k$, *induces the zero-function mod* p^n *if and only if* $a_k \equiv 0 \mod p^{n-\alpha(k)}$ *for all* k (*or, equivalently, for all* $k < \beta(n)$).

Proof. Induction on k using the facts that $(m)_k = 0$ for $m < k$, that $v_p((n)_k) \geq \alpha_p(k)$ for all $n \in \mathbb{Z}$, and that $v_p((k)_k) = v_p(k!) = \alpha_p(k)$. □

Corollary 6.2.6. *Every polynomial function on $\mathbb{Z}_{p^n}$ is represented by a unique $f \in \mathbb{Z}[x]$ of the form $f = \sum_{k=0}^{\beta(n)-1} a_k\,(x)_k$, with $0 \leq a_k < p^{n-\alpha(k)}$ for all k.*

Comparing the canonical forms of polynomial functions mod p^n with those mod p^{n-1} we see that every polynomial function mod p^{n-1} gives rise to $p^{\beta(n)}$ different polynomial functions mod p^n:

Corollary 6.2.7. *Let $(F_n, \circ)$ be the monoid of polynomial functions on $\mathbb{Z}_{p^n}$ with respect to composition and $\pi_n \colon F_{n+1} \to F_n$ the canonical projection.*

(1) *For all $n \geq 1$ and for each $f \in F_n$ we have $|\pi_n^{-1}(f)| = p^{\beta(n+1)}$.*

(2) *For all $n \geq 1$, the number of polynomial functions on $\mathbb{Z}_{p^n}$ is*

$$|F_n| = p^{\sum_{k=1}^{n} \beta(k)}.$$

Recall the following notation already given in the introduction.

Notation 6.2.8. We write $[f]_{p^n}$ for the function defined by $f \in \mathbb{Z}[x]$ on $\mathbb{Z}_{p^n}$.

Lemma 6.2.9. *Every polynomial $f \in \mathbb{Z}[x]$ is uniquely representable as*

$$f(x) = f_0(x) + f_1(x)(x^p - x) + f_2(x)(x^p - x)^2 + \ldots + f_m(x)(x^p - x)^m + \ldots$$

with $f_m \in \mathbb{Z}[x]$, $\deg f_m < p$, for all $m \geq 0$. Now let $f, g \in \mathbb{Z}[x]$.

(1) *If $n \leq p$, then $[f]_{p^n} = [g]_{p^n}$ is equivalent to: $f_k = g_k \bmod p^{n-k}\mathbb{Z}[x]$ for $0 \leq k < n$.*

(2) *$[f]_{p^2} = [g]_{p^2}$ is equivalent to: $f_0 = g_0 \bmod p^2\mathbb{Z}[x]$ and $f_1 = g_1 \bmod p\mathbb{Z}[x]$.*

(3) *$[f]_p = [g]_p$ and $[f']_p = [g']_p$ is equivalent to: $f_0 = g_0 \bmod p\mathbb{Z}[x]$ and $f_1 = g_1 \bmod p\mathbb{Z}[x]$.*

Note that (2) is just the special case of (1) with $n = 2$.

Proof. The canonical representation is obtained by repeated division with remainder by $(x^p - x)$, and uniqueness follows from uniqueness of quotient and remainder of polynomial division. Note that $[f]_p = [f_0]_p$ and $[f']_p = [f_0' - f_1]_p$. This gives (3).

Denote by $f \sim g$ the equivalence relation $f_k = g_k \bmod p^{n-k}\mathbb{Z}[x]$ for $0 \leq k < n$. Then $f \sim g$ implies $[f]_{p^n} = [g]_{p^n}$. There are $p^{p+2p+3p+\ldots+np}$ equivalence classes of $\sim$ and $p^{\beta(1)+\beta(2)+\beta(3)+\ldots+\beta(n)}$ different $[f]_{p^n}$. For $k \leq p$, $\beta(k) = kp$. Therefore the equivalence relations $f \sim g$ and $[f]_{p^n} = [g]_{p^n}$ coincide. This gives (1). □

We can rephrase this in terms of ideals of $\mathbb{Z}[x]$.

Corollary 6.2.10. *For every $n \in \mathbb{N}$, consider the two ideals of $\mathbb{Z}[x]$*

$$I_n = \{f \in \mathbb{Z}[x] \mid f(\mathbb{Z}) \subseteq p^n\mathbb{Z}\} \qquad \text{and} \qquad J_n = (\{p^{n-k}(x^p - x)^k \mid 0 \leq k \leq n\}).$$

Then $[\mathbb{Z}[x] \colon I_n] = p^{\beta(1)+\beta(2)+\beta(3)+\ldots+\beta(n)}$ and $[\mathbb{Z}[x] \colon J_n] = p^{p+2p+3p+\ldots+np}$. Therefore, $J_n = I_n$ for $n \leq p$, whereas for $n > p$, J_n is properly contained in I_n.

Proof. $J_n \subseteq I_n$. The index of J_n in $\mathbb{Z}[x]$ is $p^{p+2p+3p+\ldots+np}$, because $f \in J_n$ if and only if $f_k = 0 \bmod p^{n-k}\mathbb{Z}[x]$ for $0 \leq k < n$ in the canonical representation of Lemma 6.2.9. The index of I_n in $\mathbb{Z}[x]$ is $p^{\beta(1)+\beta(2)+\beta(3)\ldots+\beta(n)}$ by Corollary 6.2.7 (2) and $[\mathbb{Z}[x]\colon I_n] < [\mathbb{Z}[x]\colon J_n]$ if and only if $n > p$ by Fact 6.2.3 (2). □

Fact 6.2.11 (cf. McDonald [75])**.** *Let $n \geq 2$. The function on $\mathbb{Z}_{p^n}$ induced by a polynomial $f \in \mathbb{Z}[x]$ is a permutation if and only if*

(1) *f induces a permutation of $\mathbb{Z}_p$ and*

(2) *the derivative f' has no root mod p.*

Lemma 6.2.12. *Let $[f]_{p^n}$ and $[f]_p$ be the functions defined by $f \in \mathbb{Z}[x]$ on $\mathbb{Z}_{p^n}$ and $\mathbb{Z}_p$, respectively, and $[f']_p$ the function defined by the formal derivative of f on $\mathbb{Z}_p$. Then*

(1) *$[f]_{p^2}$ determines not just $[f]_p$, but also $[f']_p$.*

(2) *Let $n \geq 2$. Then $[f]_{p^n}$ is a permutation if and only if $[f]_{p^2}$ is a permutation.*

(3) *For every pair of functions (α, β), $\alpha\colon \mathbb{Z}_p \to \mathbb{Z}_p$, $\beta\colon \mathbb{Z}_p \to \mathbb{Z}_p$, there are exactly p^p polynomial functions $[f]_{p^2}$ on $\mathbb{Z}_{p^2}$ with $[f]_p = \alpha$ and $[f']_p = \beta$.*

(4) *For every pair of functions (α, β), $\alpha\colon \mathbb{Z}_p \to \mathbb{Z}_p$ bijective, $\beta\colon \mathbb{Z}_p \to \mathbb{Z}_p \setminus \{0\}$, there are exactly p^p polynomial permutations $[f]_{p^2}$ on $\mathbb{Z}_{p^2}$ with $[f]_p = \alpha$ and $[f']_p = \beta$.*

Proof. (1) and (3) follow immediately from Lemma 6.2.9 for $n = 2$ and (2) and (4) then follow from Fact 6.2.11. □

Remark 6.2.13. Lemma 6.2.12 (2) implies that the inverse image of G_n under $\pi_n\colon F_{n+1} \to F_n$ is G_{n+1}. We denote by $\pi_n\colon G_{n+1} \to G_n$ the restriction of π_n to G_n. Then Corollary 6.2.7 implies, for all $n \geq 2$,

$$|\ker(\pi_n)| = p^{\beta(n+1)}.$$

Corollary 6.2.14. *The number of polynomial permutations on $\mathbb{Z}_{p^2}$ is*

$$|G_2| = p!(p-1)^p p^p$$

and for $n \geq 3$ the number of polynomial permutations on $\mathbb{Z}_{p^2}$ is

$$|G_n| = p!(p-1)^p p^p p^{\sum_{k=3}^{n} \beta(k)}.$$

Proof. In the canonical representation of $f \in \mathbb{Z}[x]$ in Lemma 6.2.9, there are $p!(p-1)^p$ choices of coefficients mod p for f_0 and f_1 such that the criteria of Fact 6.2.11 for a polynomial permutation on $\mathbb{Z}_{p^2}$ are satisfied. And for each such choice there are p^p possibilities for the coefficients of f_0 mod p^2. The coefficients of f_0 mod p^2 and those of f_1 mod p then determine the polynomial function mod p^2. So $|G_2| = p!(p-1)^p p^p$. The formula for $|G_n|$ then follows from Remark 6.2.13. □

This concludes our review of polynomial functions and polynomial permutations on $\mathbb{Z}_{p^n}$. We will now introduce a homomorphic image of G_2 whose Sylow p-groups bijectively correspond to the Sylow p-groups of G_n for any $n \geq 2$.

6.3 A group between G_1 and G_2

Into the projective system of monoids $(F_n, \circ)$ we insert an extra semi-group E between F_1 and F_2 by means of monoid epimorphisms $\theta\colon F_2 \to E$ and $\psi\colon E \to F_1$ with $\psi\theta = \pi_1$.

$$F_1 \xleftarrow{\psi} E \xleftarrow{\theta} F_2 \xleftarrow{\pi_2} F_3 \xleftarrow{\pi_3} \dots$$

The restrictions of θ to G_2 and of ψ to the group of units H of E will be group-epimorphisms, so that we also insert an extra group H between G_2 and G_1 into the projective system of the G_i.

$$G_1 \xleftarrow{\psi} H \xleftarrow{\theta} G_2 \xleftarrow{\pi_2} G_3 \xleftarrow{\pi_3} \dots$$

In the following definition of E and H, f and f' are just two different names for functions. The connection with polynomials and their formal derivatives suggested by the notation will appear when we define θ and ψ.

Definition 6.3.1. We define the semi-group $(E, \circ)$ by

$$E = \{(f, f') \mid f\colon \mathbb{Z}_p \to \mathbb{Z}_p \; f'\colon \mathbb{Z}_p \to \mathbb{Z}_p\}$$

with law of composition

$$(f, f') \circ (g, g') = (f \circ g,\ (f' \circ g) \cdot g'),$$

where $(f \circ g)(x) = f(g(x))$ and $((f' \circ g) \cdot g')(x) = f'(g(x)) \cdot g'(x)$.

We denote by $(H, \circ)$ the group of units of E.

Lemma 6.3.2.

(1) *The identity element of E is* $(\mathrm{id}, 1)$*, with* id *denoting the identity function on $\mathbb{Z}_p$ and* 1 *the constant function* 1*.*

(2) *The group of units of E has the form*

$$H = \{(f, f') \mid f\colon \mathbb{Z}_p \to \mathbb{Z}_p \text{ bijective},\ f'\colon \mathbb{Z}_p \to \mathbb{Z}_p \setminus \{0\}\}.$$

(3) *The inverse of $(g, g') \in H$ is*

$$(g, g')^{-1} = (g^{-1}, \frac{1}{g' \circ g^{-1}}),$$

where g^{-1} is the inverse permutation of the permutation g and $1/a$ stands for the multiplicative inverse of a non-zero element $a \in \mathbb{Z}_p$, such that

$$(\frac{1}{g' \circ g^{-1}})(x) = \frac{1}{g'(g^{-1}(x))}$$

means the multiplicative inverse in $\mathbb{Z}_p \setminus \{0\}$ of $g'(g^{-1}(x))$.

Note that H is just a wreath product (designed to act on the left) of the permutation group S_p and a cyclic group of $p-1$ elements (here appearing as the multiplicative group of units of $\mathbb{Z}_p$).

Now for the homomorphisms θ and ψ.

Definition 6.3.3. We define $\psi\colon E \longrightarrow F_1$ by $\psi(f, f') = f$. As for $\theta\colon F_2 \to E$, given an element $[g]_{p^2} \in F_2$, set $\theta([g]_{p^2}) = ([g]_p, [g']_p)$ – this is well-defined by Lemma 6.2.12 (1).

Lemma 6.3.4.

(i) $\theta\colon F_2 \to E$ *is a monoid-epimorphism.*

(ii) The inverse image of H *under* $\theta\colon F_2 \to E$ *is* G_2*.*

(iii) The restriction of θ *to* G_2 *is a group epimorphism* $\theta\colon G_2 \to H$ *with* $|ker(\theta)| = p^p$*.*

(iv) $\psi\colon E \to F_1$ *is a monoid epimorphism and* ψ *restricted to* H *is a group-epimorphism* $\psi\colon H \to G_1$*.*

Proof. (i) follows from Lemma 6.2.12 (3) and (ii) from Fact 6.2.11. (iii) follows from Lemma 6.2.12 (4). Finally, (iv) holds because every function on $\mathbb{Z}_p$ is a polynomial function and every permutation of $\mathbb{Z}_p$ is a polynomial permutation. □

6.4 Sylow subgroups of H and G_n

We will first determine the Sylow p-groups of H. The Sylow p-groups of G_n for $n \geq 2$ then are obtained as the inverse images of the Sylow p-groups of H under the epimorphism $G_n \to H$.

Lemma 6.4.1. *Let* C_0 *be the subgroup of* S_p *generated by the* p*-cycle* $(0\,1\,2\ldots p-1)$*. Then one Sylow* p*-subgroup of* H *is*

$$S = \{(f, f') \in H \mid f \in C_0,\ f' = 1\},$$

where $f' = 1$ *means the constant function* 1*. The normalizer of* S *in* H *is*

$$N_H(S) = \{(g, g') \mid g \in N_{S_p}(C_0),\ g' \text{ a non-zero constant }\}.$$

Proof. As $|H| = p!(p-1)^p$, and S is a subgroup of H of order p, S is a Sylow p-group of H. Conjugation of $(f, f') \in S$ by $(g, g') \in H$ (using the fact that $f' = 1$) gives

$$(g, g')^{-1}(f, f')(g, g') = (g^{-1}, \frac{1}{g' \circ g^{-1}})(f \circ g, g') = (g^{-1} \circ f \circ g, \frac{g'}{g' \circ g^{-1} \circ f \circ g})$$

The first coordinate of $(g, g')^{-1}(f, f')(g, g')$ being in C_0 for all $(f, f') \in S$ is equivalent to $g \in N_{S_p}(C_0)$. The second coordinate of $(g, g')^{-1}(f, f')(g, g')$ being the constant function 1 for all $(f, f') \in S$ is equivalent to

$$\forall x \in \mathbb{Z}_p \quad g'(x) = g'(g^{-1}(f(g(x)))),$$

which is equivalent to g' being constant on every cycle of $g^{-1}fg$, which is equivalent to g' being constant on $\mathbb{Z}_p$, since f can be chosen to be a p-cycle. □

Lemma 6.4.2. *Another way of describing the normalizer of S in H is*

$$N_H(S) = \{(f, f') \in H \mid \exists k \neq 0 \ \ \forall a, b \ \ f(a) - f(b) = k(a - b); \ \ f' \text{ a non-zero constant}\}.$$

Therefore, $|N_H(S)| = p(p-1)^2$ and $[H\colon N_H(S)] = (p-1)!(p-1)^{p-2}$.

Proof. Let $\sigma = (0\,1\,2 \ldots p-1)$ and $f \in S_p$ then

$$f\sigma f^{-1} = (f(0)\, f(1)\, f(2) \ldots f(p-1))$$

Now $f \in N_{S_p}(C_0)$ if and only if, for some $1 \leq k < p$ $f\sigma f^{-1} = \sigma^k$, i.e.,

$$(f(0) \ \ f(1) \ \ f(2) \ \ldots \ f(p-1)) = (0 \ \ k \ \ 2k \ \ldots \ (p-1)k),$$

all numbers taken mod p. This is equivalent to $f(x+1) = f(x) + k$ or

$$f(x+1) - f(x) = k$$

and further equivalent to $f(a) - f(b) = k(a-b)$. Thus k and $f(0)$ determine $f \in N_{S_p}(C_0)$, and there are $(p-1)$ choices for k and p choices for $f(0)$. Together with the $(p-1)$ choices for the non-zero constant f' this makes $p(p-1)^2$ elements of $N_H(S)$. □

Corollary 6.4.3. *There are $(p-1)!(p-1)^{p-2}$ Sylow p-subgroups of H.*

Theorem 6.4.4. *The Sylow p-subgroups of H are in bijective correspondence with pairs $(C, \bar{\varphi})$, where C is a cyclic subgroup of order p of S_p, $\varphi\colon \mathbb{Z}_p \to \mathbb{Z}_p \setminus \{0\}$ is a function and $\bar{\varphi}$ is the class of φ with respect to the equivalence relation of multiplication by a non-zero constant. The subgroup corresponding to $(C, \bar{\varphi})$ is*

$$S_{(C,\bar{\varphi})} = \{(f, f') \in H \mid f \in C, \ f'(x) = \frac{\varphi(f(x))}{\varphi(x)}\}$$

Proof. Observe that each $S_{(C,\bar{\varphi})}$ is a subgroup of order p of H. Different pairs $(C, \bar{\varphi})$ give rise to different groups: Suppose $S_{(C,\bar{\varphi})} = S_{(D,\bar{\psi})}$. Then $C = D$ and for all $x \in \mathbb{Z}_p$ and for all $f \in C$ we get

$$\frac{\varphi(f(x))}{\varphi(x)} = \frac{\psi(f(x))}{\psi(x)}.$$

As C is transitive on $\mathbb{Z}_p$ the latter condition is equivalent to

$$\forall x, y \in \mathbb{Z}_p \quad \frac{\psi(x)}{\varphi(x)} = \frac{\psi(y)}{\varphi(y)},$$

which means that $\varphi = k\psi$ for a nonzero $k \in \mathbb{Z}_p$.

There are $(p-2)!$ cyclic subgroups of order p of S_p, and $(p-1)^{p-1}$ equivalence classes $\bar{\varphi}$ of functions $\varphi\colon \mathbb{Z}_p \to \mathbb{Z}_p \setminus \{0\}$. So the number of pairs $(C, \bar{\varphi})$ equals $(p-1)!(p-1)^{p-2}$, which is the number of Sylow p-groups of H, by the preceding corollary. □

In the projective system of groups

$$G_1 \xleftarrow{\psi} H \xleftarrow{\theta} G_2 \xleftarrow{\pi_2} \dots \xleftarrow{\pi_{n-1}} G_n$$

the kernel of the group epimorphism $G_n \to H$ is a finite p-group for every $n \geq 2$, because, firstly, the kernel of $\pi_{n-1}\colon G_n \to G_{n-1}$ is of order $p^{\beta(n)}$ by Remark 6.2.13, and secondly, the kernel of $\theta\colon G_2 \to H$ is of order p^p by Lemma 6.3.4 (iii). So the Sylow p-groups of G_n for $n \geq 2$ are just the inverse images of the Sylow p-groups of H:

Theorem 6.4.5. *Let $n \geq 2$. Let G_n be the group (with respect to composition) of polynomial permutations on $\mathbb{Z}_{p^n}$. There are $(p-1)!(p-1)^{p-2}$ Sylow p-groups of G_n. They are in bijective correspondence with pairs $(C, \bar{\varphi})$, where C is a cyclic subgroup of order p of S_p, $\varphi\colon \mathbb{Z}_p \to \mathbb{Z}_p \setminus \{0\}$ a function and $\bar{\varphi}$ its class with respect to the equivalence relation of multiplication by a non-zero constant. The subgroup corresponding to $(C, \bar{\varphi})$ is*

$$S_{(C,\bar{\varphi})} = \{[f]_{p^n} \in G_n \mid [f]_p \in C,\ [f']_p(x) = \frac{\varphi([f]_p(x))}{\varphi(x)}\}.$$

One particularly easy to describe Sylow p-group of G_n corresponds to a constant function φ and the subgroup C generated by $(0\,1\,2 \dots p-1)$ of S_p. It is the inverse image of S defined in Lemma 6.4.1 and consists of those polynomial functions on $\mathbb{Z}_{p^n}$ which modulo p are a power of $(0\,1\,2 \dots p-1)$, and whose derivative is constant 1 mod p.

One last remark: Each Sylow p-group of $G_1 = S_p$ is isomorphic to C_p, where C_p denotes the cyclic group of order p. Also, it is not difficult to see (using the description of G_2 in [40]) that the Sylow p-groups of G_2 are of the form $C_p \wr C_p$. It is an open question, posed by W. Herfort (personal communication), if every finite wreath product $C_p \wr C_p \wr \dots \wr C_p$ of cyclic groups of order p can be embedded in G_n for some n.

Chapter 7

Analysis of Parameters of Trees Corresponding to Huffman Codes and Sums of Unit Fractions

This chapter contains the article [51] with the title "Analysis of Parameters of Trees Corresponding to Huffman Codes and Sums of Unit Fractions". It is joint work with Clemens Heuberger and Stephan Wagner. The article was accepted for publication in the proceedings of *SIAM Meeting on Analytic Algorithmics and Combinatorics (ANALCO13)* on September 13, 2012.

Abstract

For fixed $t \geq 2$, we consider the class of representations of 1 as sum of unit fractions whose denominators are powers of t or equivalently the class of canonical compact t-ary Huffman codes or equivalently rooted t-ary plane "canonical" trees.

We study the probabilistic behaviour of the height (limit distribution is shown to be normal), the number of distinct summands (normal distribution), the path length (normal distribution), the width (main term of the expectation and concentration property) and the number of leaves at maximum distance from the root (discrete distribution).

7.1 Introduction

Let $t \geq 2$ be an integer. We consider the following combinatorial classes which turn out to be equivalent. See Figure 7.1.1 for examples.

1. Partitions of 1 into powers of t (representation of 1 as sum of unit fractions whose denominators are powers of t):

$$\mathcal{C}_{Partition} = \Big\{(x_1, \dots, x_r) \in \mathbb{Z}^r \;\Big|\; r \geq 0,\, 0 \leq x_1 \leq x_2 \leq \cdots \leq x_r,\, \sum_{i=1}^{r} \frac{1}{t^{x_i}} = 1\Big\}.$$

The external size $|(x_1, \ldots, x_r)|$ of such a representation $(x_1, \ldots, x_r)$ is defined to be the number r of summands.

2. Canonical compact t-ary Huffman codes:

$$\mathcal{C}_{Code} = \{C \subseteq \{1, \ldots, t\}^* \mid C \text{ is prefix-free, compact and canonical}\}.$$

Here,

- $\{1, \ldots, t\}^*$ denotes the set of finite words over the alphabet $\{1, \ldots, t\}$,
- a code C is said to be prefix-free if no word in C is a proper prefix of any other word in C,
- a code C is said to be compact if the following property holds: if w is a proper prefix of a word in C, then for every letter $a \in \{1, \ldots, t\}$, wa is a prefix of a word in C,
- a code C is said to be canonical if the lexicographic ordering of its words corresponds to a non-decreasing ordering of the word lengths. This condition corresponds to taking equivalence classes with respect to permutations of the alphabet (at each position in the words).

The external size $|C|$ of a code C is defined to be the cardinality of C.

If $C \in \mathcal{C}_{Code}$ with $C = \{w_1, \ldots, w_r\}$ and the property that $\mathsf{length}(w_i) \leq \mathsf{length}(w_{i+1})$ for all i, then $(\mathsf{length}(w_1), \ldots, \mathsf{length}(w_r)) \in \mathcal{C}_{Partition}$. This is a bijection between $\mathcal{C}_{Code}$ and $\mathcal{C}_{Partition}$ preserving the external size.

3. Canonical rooted t-ary trees:

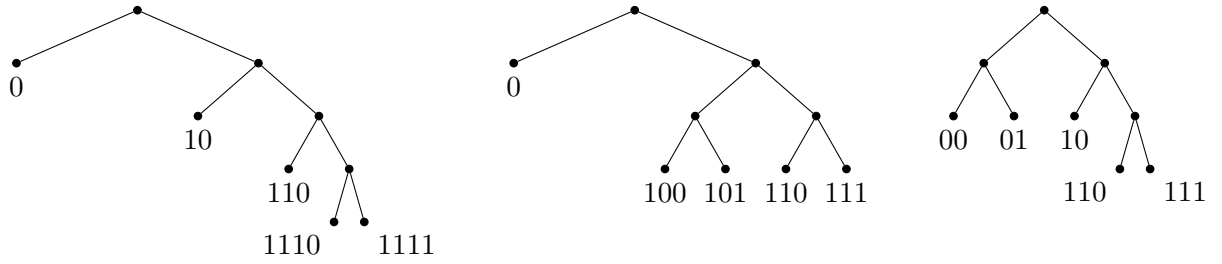
$$\mathcal{C}_{Tree} = \{T \text{ rooted } t\text{-ary plane tree} \mid T \text{ is canonical}\}.$$

Here,

- t-ary means that each vertex has no or t children,
- plane tree means that an ordering "from left to right" of the children of each vertex is specified,
- canonical means that the following holds for all k: if the vertices of depth (i.e., distance to the root) k are denoted by v_1, …, v_K from left to right, then $\deg(v_i) \leq \deg(v_{i+1})$ holds for all i.

The external size $|T|$ of a tree is given by the number of its leaves, i.e., the number of vertices of degree 1.

If $C \in \mathcal{C}_{Code}$, then a tree $T \in \mathcal{C}_{Tree}$ can be constructed such that the vertices of T are given by the prefixes of the words in C, the root is the vertex corresponding to the empty word, and the children of a proper prefix w of a code word are given from left to right by wa for $a = 1$, …, t. This is a bijection between $\mathcal{C}_{Code}$ to $\mathcal{C}_{Tree}$ preserving the external size.



$$1 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{16} \qquad 1 = \frac{1}{2} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \qquad 1 = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8}$$

Figure 7.1.1: All elements of external size 5 (and internal size 4) in $\mathcal{C}_{Tree}$, $\mathcal{C}_{Code}$ and $\mathcal{C}_{Partition}$ for $t = 2$.

Further formulations, details and remarks can be found in [35]. We will simply speak of an element in the class $\mathcal{C}$ when the particular interpretation as an element of $\mathcal{C}_{Partition}$, $\mathcal{C}_{Code}$ or $\mathcal{C}_{Tree}$ is not relevant. Our proofs will use the tree model, therefore $\mathcal{C}_{Tree}$ is abbreviated as $\mathcal{T}$.

The external size of an element in $\mathcal{C}$ is always congruent to 1 modulo $t-1$. This can easily be seen in the tree model, where the number of leaves r and the number of internal vertices n are connected by the identity

$$r = 1 + n(t-1).$$

Therefore, we will from now on consider the *internal size:* for a tree $T \in \mathcal{C}_{Tree}$ the internal size of T is the number $n(T)$ of internal vertices, for a code $C \in \mathcal{C}_{Code}$ the internal size is the number of proper prefixes of words of C, and for a partition $(x_1, \ldots, x_r) \in \mathcal{C}_{Partition}$ the internal size is defined to be $(r-1)/(t-1)$. We will omit the word "internal" and will always use the variable n to denote the size.

The asymptotics of the number of elements in $\mathcal{C}$ of size n has been studied by various authors, cf. again [35]. In that paper, building upon a generating function approach by Flajolet and Prodinger [37], the following result has been obtained:

Theorem 7.1.1 ([35]). *For $t \geq 2$, the number of elements of size n in $\mathcal{C}$ can be estimated as*

$$R\rho^{n+1} + \Theta(\rho_2^n),$$

where $\rho > \rho_2$ and R are positive real constants depending on t with asymptotic expansions (as $t \to \infty$)

$$\rho = 2 - \frac{1}{2^{t+1}} + O\left(\frac{t}{2^{2t}}\right), \quad \rho_2 = 1 + \frac{\log 2}{t} + O\left(\frac{1}{t^2}\right), \quad R = \frac{1}{8} + \frac{t-2}{2^{t+5}} + O\left(\frac{t^2}{2^{2t}}\right).$$

In fact, all O-constants can be made explicit and more terms of the asymptotic expansions in t of ρ, ρ_2 and R can be given.

The purpose of this contribution is to study the probabilistic behaviour of various parameters of a random element in $\mathcal{C}$ of size n (all elements considered to be equally likely):

1. The *height* $h(T)$ of a tree $T \in \mathcal{C}_{Tree}$ is defined to be the maximum distance of a leaf from the root. In the interpretation as a code, this is the maximum length of a code word. In a representation of 1 as a sum of unit fractions, this corresponds to the largest denominator used (more precisely, to the largest exponent of the denominator).

   The height is discussed in Section 7.3. It is asymptotically normally distributed with mean $\sim \mu_h n$ and variance $\sim \sigma_h^2 n$, where

   $$\mu_h = \frac{1}{2} + \frac{t-2}{2^{t+3}} + O\left(\frac{t^2}{2^{2t}}\right) \quad \text{and} \quad \sigma_h^2 = \frac{1}{4} + \frac{-t^2+5t-2}{2^{t+4}} + O\left(\frac{t^3}{2^{2t}}\right),$$

   cf. Theorem 7.3.1.

2. The *number of distinct summands* of a representation $(x_1, \ldots, x_r)$ of 1 as sum of unit fractions is denoted by $d(x_1, \ldots, x_r)$. In the tree model, this corresponds to the cardinality $d(T)$ of the set of depths of leaves in a tree $T \in \mathcal{C}_{Tree}$. In the code model, this is the number of distinct lengths of code words.

   The number $d(T)$ is studied in Section 7.4. It is asymptotically normally distributed with mean $\sim \mu_d n$ and variance $\sim \sigma_d^2 n$, where

   $$\mu_d = \frac{1}{2} + \frac{t-4}{2^{t+3}} + O\left(\frac{t^2}{2^{2t}}\right) \quad \text{and} \quad \sigma_d^2 = \frac{1}{4} + \frac{-t^2+9t-14}{2^{t+4}} + O\left(\frac{t^2}{2^{2t}}\right),$$

   cf. Theorem 7.4.1.

3. The *maximum number of equal summands* of a representation $(x_1, \ldots, x_r)$ of 1 as sum of unit fractions is denoted by $w(x_1, \ldots, x_r)$. In the code model, this is the maximum number of code words of equal length; in the tree model, this is the "leaf-width" $w(T)$, the maximum number of leaves on the same level.

   The number $w(T)$ is studied in Section 7.7. We prove that $\mathbb{E}(w(T)) = \mu_w \log n + O(\log\log n)$ with $\mu_w = 1/(t \log 2) + O(1/t^2)$ and a concentration property, cf. Theorem 7.7.1.

4. The *(total) path length* $\ell(T)$ of a tree $T \in \mathcal{C}_{Tree}$ is defined to be the sum of the depths of all vertices of the tree. In our context, it is perhaps most natural to consider the *external path length* $\ell_{external}(T)$, though, which is the sum of depths over all leaves of the tree, as this parameter corresponds to the sum of lengths of code words in a code $C \in \mathcal{C}_{Code}$. Likewise, the *internal path length* $\ell_{internal}(T)$ is the sum of depths over all non-leaves. Clearly, we have $\ell_{external}(T) + \ell_{internal}(T) = \ell(T)$, and the relations

   $$\ell_{external}(T) = \frac{t-1}{t}\ell(T) + n(T) \quad \text{and} \quad \ell_{internal}(T) = \frac{1}{t}\ell(T) - n(T)$$

for t-ary trees are easily proven. Therefore, all distributional results for any one of those parameters immediately cover all three. The total path length turns out to be asymptotically normally distributed as well (see Theorem 7.6.1), with mean $\sim \mu_{tpl} n^2$ and variance $\sim \sigma_{tpl}^2 n^3$. The coefficients have asymptotic expansions

$$\mu_{tpl} = \frac{t}{2}\cdot\mu_h = \frac{t}{4} + \frac{t(t-2)}{2^{t+4}} + O\left(\frac{t^3}{2^{2t}}\right) \quad \text{and} \quad \sigma_{tpl} = \frac{t^2}{12} + \frac{-t^4+5t^3+2t^2}{3\cdot 2^{t+4}} + O\left(\frac{t^5}{2^{2t}}\right).$$

The path length is studied in Section 7.6; its analysis is based on a generating function approach for the moments, combined with probabilistic arguments to obtain the central limit theorem.

5. The *number of leaves on the last level* (i.e., maximum distance from the root) of a tree $T \in \mathcal{C}_{Tree}$ is denoted by $m(T)$. This corresponds to the number of code words of maximum length and to the number of smallest summands in a representation of 1 as a sum of unit fractions.

   This parameter may appear to be the least interesting of the parameters we study. However, it is a natural technical parameter when constructing generating functions for the other parameters. From these generating functions, the probabilistic behaviour of $m(T)$ can be read off without too much effort, so we do include these results in Section 7.5.

   The limit distribution of $m(T)$ is a discrete distribution with mean $2t + o(1)$ and variance $2t^2 + o(1)$, cf. Theorem 7.5.1.

A noteworthy feature of the results listed above is the fact that the distributions we observe are quite different from those that one obtains for other probabilistic random tree models, specifically Galton–Watson trees (which include, amongst others, random t-ary trees), but also recursive trees and general families of increasing trees, see [29] for a general reference. Specifically,

- the asymptotic order of the height of a random Galton–Watson tree of order n is only $\sqrt{n}$, and it is known that the limiting distribution (which is sometimes called a Theta distribution) coincides with the distribution of the maximum of a Brownian excursion [36]. The height of random recursive trees (or other families of increasing trees) is even only of order $\log n$, and heavily concentrated around its mean, see [28].

- The path length of random Galton–Watson trees is of order $n^{3/2}$, and it follows an Airy distribution (like the area under a Brownian excursion) in the limit [100]. For recursive trees, the path length is of order $n \log n$ with a rather unusual limiting distribution [74].

- While the height of our canonical trees is greater than that of Galton–Watson trees, precisely the opposite holds for the width (as one would expect): it is of order $\sqrt{n}$ for Galton-Watson trees [30, 101], with the same limiting distribution as the height, as opposed to only $\log n$ in our setting. For recursive trees, the width is even of order $n/\sqrt{\log n}$, see [31].

Indeed, the structure of our canonical t-ary trees is comparable to that of *compositions*: counting the number of internal vertices on each level from the root, we obtain a restricted composition (see the series of papers by Bender and Canfield [10, 11, 12] on recent results concerning compositions with various local restrictions), in which each summand is at most t times the previous one. In the limit $t \to \infty$, one obtains compositions of n starting with a 1 in this way.

Last in this introduction a remark on the notations of the error terms: In all our major results those error terms have an explicit O-constant. The error functions $\varepsilon_j(\dots)$ that appear there are real functions which fulfil $|\varepsilon_j(\dots)| \leq 1$ for all values of the indicated parameters. Those constants were calculated with the computer algebra system Sage [98].

7.2 The Generating Function

The height $h(T)$, the cardinality $d(T)$ of the set of different depths of leaves and the number $m(T)$ of leaves on the last level of a tree $T \in \mathcal{T}$ of size $n = n(T)$ can be analysed by studying a multivariate generating function $H(q, u, v, w)$, where q labels the size $n(T)$, u labels the number $m(T)$ of leaves on the last level, v labels the cardinality $d(T)$ of the set of depths of leaves and w labels the height $h(T)$.

Theorem 7.2.1. *The generating function*

$$H(q, u, v, w) := \sum_{T \in \mathcal{T}} q^{n(T)} u^{m(T)} v^{d(T)} w^{h(T)}$$

can be expressed as

$$H(q, u, v, w) = a(q, u, v, w) + b(q, u, v, w) \frac{a(q, 1, v, w)}{1 - b(q, 1, v, w)} \tag{7.2.1}$$

with

$$\begin{aligned} a(q, u, v, w) &= \sum_{j=0}^{\infty} v q^{[\![j]\!]} u^{t^j} w^j \prod_{i=1}^{j} \frac{1 - v - q^{[\![i]\!]} u^{t^i}}{1 - q^{[\![i]\!]} u^{t^i}}, \\ b(q, u, v, w) &= \sum_{j=1}^{\infty} \frac{v q^{[\![j]\!]} u^{t^j} w^j}{1 - q^{[\![j]\!]} u^{t^j}} \prod_{i=1}^{j-1} \frac{1 - v - q^{[\![i]\!]} u^{t^i}}{1 - q^{[\![i]\!]} u^{t^i}}, \end{aligned} \tag{7.2.2}$$

where $[\![j]\!] := 1 + t + \cdots + t^{j-1}$.

Proof. The proof of Theorem 7.2.1 follows ideas of Flajolet and Prodinger [37], (see also [35]), which we only sketch briefly. Details can be found in Appendix 7.8. One first considers

$$H_h(q, u, v) := [w^h] H(q, u, v, w) = \sum_{\substack{T \in \mathcal{T} \\ h(T) = h}} q^{n(T)} u^{m(T)} v^{d(T)}$$

for some $h \geq 0$. A tree T' of height $h+1$ arises from a tree T of height h by replacing j of its $m(T)$ leaves on the last level by internal vertices with t succeeding leaves respectively, where $1 \leq j \leq m(T)$. If $j < m(T)$, then $d(T') = d(T) + 1$; otherwise, we have $d(T') = d(T)$. For the generating function H_h, this translates to the recursion

$$\begin{aligned} H_{h+1}(q,u,v) &= \sum_{\substack{T\in\mathcal{T}\\ h(T)=h}} \left(\sum_{j=1}^{m(T)-1} q^{n(T)+j} u^{jt} v^{d(T)+1} + q^{n(T)+m(T)} u^{m(T)t} v^{d(T)} \right) \\ &= r(q,u,v) H_h(q,1,v) + s(q,u,v) H_h(q,qu^t,v) \end{aligned} \tag{7.2.3}$$

with

$$r(q,u,v) = \frac{qu^t v}{1-qu^t}, \qquad s(q,u,v) = \frac{1-v-qu^t}{1-qu^t},$$

and initial value $H_0(q,u,v) = uv$. This further means that

$$H(q,u,v,w) = uv + wr(q,u,v)H(q,1,v,w) + ws(q,u,v)H(q,qu^t,v,w),$$

and this functional equation can be solved by iteration. One obtains

$$H(q,u,v,w) = a(q,u,v,w) + b(q,u,v,w)H(q,1,v,w),$$

and (7.2.1) results by plugging in $u = 1$ and solving for $H(q,1,v,w)$. □

Next we recall results on the singularities of $H(q,1,1,1)$, see Proposition 10 of [35].

Lemma 7.2.2 ([35])**.** *The generating function $H(q,1,1,1)$ has exactly one singularity $q = q_0$ with $|q| < 1 - \frac{0.72}{t}$. This singularity q_0 is a simple real pole. For $t \geq 4$, we have*

$$q_0 = \frac{1}{2} + \frac{1}{2^{t+3}} + \frac{t+4}{2^{2t+5}} + \frac{3t^2+23t+38}{2^{3t+8}} + \frac{7t^3}{100\cdot 2^{4t}}\varepsilon_1(t).$$

For $t \in \{2,3\}$, the values are given in Table 7.2.1. Furthermore, let

$$Q = \frac{1}{2} + \frac{\log 2}{2t} + \frac{0.06}{t^2}$$

for $t \geq 6$ and Q be given by Table 7.2.1 for $2 \leq t \leq 5$. Then q_0 is the only singularity of $H(q,1,1,1)$ with $|q| \leq q_0/Q$.

Using this result, we will be able to apply singularity analysis to all our generating functions in the coming sections.

7.3 The Height

We start our analysis with the height $h(T)$ of our canonical trees $T \in \mathcal{T}$. We show that the height is asymptotically (for large sizes $n = n(T)$) normally distributed and calculate

t	q_0	Q	t	q_0	Q
2	0.5573678720...	0.71317958	4	0.5090030531...	0.59306918
3	0.5206401166...	0.63074477	5	0.5042116835...	0.57200784

Table 7.2.1: Constants q_0 and Q for $2 \leq t \leq 5$.

its mean and variance. We will do this by means of the generating function $H(q, u, v, w)$ defined in Section 7.2.

So let us have a look at the bivariate generating function

$$H(q, 1, 1, w) = \sum_{T \in \mathcal{T}} q^{n(T)} w^{h(T)} = \frac{a(q, 1, 1, w)}{1 - b(q, 1, 1, w)}$$

for the height. We consider its denominator

$$D(q, w) := 1 - b(q, 1, 1, w) = \sum_{0 \leq j} (-1)^j w^j \prod_{i=1}^{j} \frac{q^{[\![i]\!]}}{1 - q^{[\![i]\!]}}.$$

From Lemma 7.2.2 we know that $D(q, 1)$ has a simple zero q_0. Expanding $D(q, w)$ around $(q_0, 1)$ and using Theorem IX.9 (meromorphic singularity perturbation) from the book of Flajolet and Sedgewick [38] yields the desired results for the height without much effort. They are stated precisely in the following theorem.

Theorem 7.3.1. *The height is asymptotically normally distributed. Its mean is* $\mu_h n + O(1)$ *and its variance is* $\sigma_h^2 n + O(1)$ *with*

$$\mu_h = \frac{1}{2} + \frac{t-2}{2^{t+3}} + \frac{2t^2 + 3t - 8}{2^{2t+5}} + \frac{9t^3 + 45t^2 + 2t - 88}{2^{3t+8}} + \frac{0.044t^4}{2^{4t}} \varepsilon_2(t)$$

and

$$\sigma_h^2 = \frac{1}{4} + \frac{-t^2 + 5t - 2}{2^{t+4}} + \frac{-4t^3 + 4t^2 + 27t - 7}{2^{2t+6}} + \frac{0.058t^4}{2^{3t}} \varepsilon_3(t)$$

for $t \geq 3$. *In the case* $t = 2$ *we have* $\mu_h = 0.5662757699...$ *and* $\sigma_h^2 = 0.2665499010...$.

We calculated the values of the constants μ_h and σ_h^2 numerically for $2 \leq t \leq 30$. Those values can be found in Table 7.9.1 in Appendix 7.9, where a complete proof of Theorem 7.3.1 is given as well.

7.4 The Number of Distinct Depths of Leaves

In this section we study the number of distinct depths of leaves $d(T)$ of our canonical trees $T \in \mathcal{T}$, motivated by the interpretation as the number of distinct code lengths in Huffman codes. This parameter is also asymptotically normally distributed, and the

approach is essentially the same as for the height, based on the generating function $H(q,u,v,w)$ from Section 7.2. To analyse the parameter $d(T)$, we look at the bivariate generating function

$$H(q,1,v,1) = \sum_{T\in\mathcal{T}} q^{n(T)}v^{d(T)} = \frac{a(q,1,v,1)}{1-b(q,1,v,1)}$$

for the number of distinct depths of leaves. Again, we consider its denominator

$$D(q,v) := 1-b(q,1,v,1) = 1-\sum_{1\le j}\frac{v}{1-q^{[\![j]\!]}}\prod_{i=1}^{j-1}\frac{1-v-q^{[\![i]\!]}}{1-q^{[\![i]\!]}}$$

and proceed as in the previous section. Lemma 7.2.2 tells us the existence of a simple zero q_0 of $D(q,1)$. Again, we expand the denominator $D(q,v)$ around $(q_0,1)$ and use Theorem IX.9 from the book of Flajolet and Sedgewick [38]. This results in the following theorem.

Theorem 7.4.1. *The number of distinct depths of leaves is asymptotically normally distributed. Its mean is $\mu_d n + O(1)$ and its variance is $\sigma_d^2 n + O(1)$ with*

$$\mu_d = \frac{1}{2} + \frac{t-4}{2^{t+3}} + \frac{2t^2-t-14}{2^{2t+5}} + \frac{9t^3+27t^2-76t-144}{2^{3t+8}} + \frac{0.046t^4}{2^{4t}}\varepsilon_4(t)$$

and

$$\sigma_d^2 = \frac{1}{4} + \frac{-t^2+9t-14}{2^{t+4}} + \frac{-4t^3+20t^2+3t-54}{2^{2t+6}} + \frac{0.056t^4}{2^{3t}}\varepsilon_5(t)$$

for $t \ge 2$.

Again, as in the previous section, we calculated the values of the constants μ_d and σ_d^2 numerically for $2 \le t \le 30$, and they are given in Table 7.10.1 in Appendix 7.10, where the proof of Theorem 7.4.1 is detailed as well.

7.5 The Number of Leaves on the Last Level

For analysing the parameter $m(T)$ counting the number of leaves of maximum depth (labelled by the variable u in the generating function $H(q,u,v,w)$), we note that for fixed $|u| \le 1$, the dominant simple pole q_0 of $H(q,1,1,1)$ is also the dominant singularity of $H(q,u,1,1)$ and is still a simple pole. Therefore, $m(T)$ tends to a discrete limiting distribution, we refer again to the book of Flajolet and Sedgewick [38, Section IX.2]. Note that the number $m(T)$ is always divisible by t by construction.

Theorem 7.5.1. *Let q_0 and Q be as described in Lemma 7.2.2. Set $p_m = [u^{mt}]b(q_0,u,1,1)$ for $m \ge 1$. Then, for a random tree $T \in \mathcal{T}$ of size n, we have*

$$\mathbb{P}(m(T) = mt) = p_m + O(Q^n)$$

for $m \ge 1$.

Furthermore, we have

$$\mathbb{E}(m(T)) = 2t - \frac{t^2 - t}{2^{t+1}} - \frac{t^3 + 6t^2 - 5t}{2^{2t+3}} - \frac{3t^4 + 32t^3 + 61t^2 - 56t}{2^{3t+8}} + O\left(\frac{t^5}{2^{4t}} + Q^n\right)$$

and

$$\mathbb{V}(m(T)) = 2t^2 - \frac{t^4 - 3t^2}{2^{t+1}} - \frac{t^5 + 13t^4 - 3t^3 - 17t^2}{2^{2t+3}} + O\left(\frac{t^6}{2^{3t}} + Q^n\right).$$

The proof can be found in Appendix 7.11. This theorem (slightly generalised) is a very useful tool in proving the central limit theorem for the path length in the following section.

7.6 The Path Length

This section is devoted to the analysis of the path length, as defined in the introduction. While the external path length is most natural in the setting of Huffman codes, it is more convenient to work with the total and the internal path length. As it was pointed out in the introduction, the three are essentially equivalent as they are (deterministically) related by simple linear equations.

We first use a generating functions approach to determine the asymptotic behaviour of the mean and variance. Let us define a generating function L_r for the r-th moment of the total path length as follows:

$$L_r(q, u, w) := \sum_{T \in \mathcal{T}} \ell(T)^r q^{n(T)} u^{m(T)} w^{h(T)}.$$

Note that $L_0(q, u, w) = H(q, u, 1, w)$ in the notation of the previous sections. We are specifically interested in L_1 and L_2. In analogy to the approach we used to determine a formula for $H(q, u, v, w)$, we obtain a functional equation for $L_r(q, u, w)$ by first introducing

$$L_{r,h}(q, u) = [w^h] L_r(q, u, w) = \sum_{\substack{T \in \mathcal{T} \\ h(T) = h}} \ell(T)^r q^{n(T)} u^{m(T)}.$$

Replacing j leaves of depth h by internal vertices, thus creating tj new leaves of depth $h + 1$, increases the total path length by $tj(h + 1)$. Thus we get

$$\begin{aligned} L_{1,h+1}(q, u) &= \sum_{\substack{T \in \mathcal{T} \\ h(T) = h+1}} (h+1) m(T) q^{n(T)} u^{m(T)} + \sum_{\substack{T \in \mathcal{T} \\ h(T) = h}} \sum_{j=1}^{m(T)} \ell(T) q^{n(T)+j} u^{jt} \\ &= (h+1) u \frac{\partial}{\partial u} L_{0,h+1}(q, u) + \frac{qu^t}{1 - qu^t} \left(L_{1,h}(q, 1) - L_{1,h}(q, qu^t) \right). \end{aligned}$$

Define, for the sake of convenience, the linear operators $\Phi_u = u\frac{\partial}{\partial u}$, $\Phi_w = w\frac{\partial}{\partial w}$ and $\Phi_q = q\frac{\partial}{\partial q}$ acting on our generating functions. Then we obtain

$$L_1(q,u,w) = \Phi_u\Phi_w L_0(q,u,w) + \frac{qu^t w}{1-qu^t}\left(L_1(q,1,w) - L_1(q,qu^t,w)\right).$$

Likewise, one gets a functional equation for $L_2(q,u,w)$:

$$L_2(q,u,w) = 2\Phi_u\Phi_w L_1(q,u,w) - \Phi_u^2\Phi_w^2 L_0(q,u,w) + \frac{qu^t w}{1-qu^t}\left(L_2(q,1,w) - L_2(q,qu^t,w)\right).$$

Both functional equations can be solved by means of iteration in the same way as the functional equation for the generating function $H(q,u,v,w)$ that we used in previous sections, see Appendix 7.12 for details. In order to determine the asymptotic behaviour of mean and variance, one only needs to find the expansion around the dominating singularity q_0 and apply singularity analysis. The main term of the mean is easy to guess: assuming that the vertices are essentially uniformly distributed along the entire height, it is natural to conjecture that $\ell(T)$ is typically around $tn(T)h(T)/2$ and thus of quadratic order. This is indeed true, and the variance turns out to be of cubic order (terms of degree 4 cancel, as one would expect). The details are rather lengthy and given in the appendix.

In order to prove convergence to the Gaussian distribution, a different, more probabilistic approach is needed. Standard theorems from analytic combinatorics no longer apply since the path length grows faster than, for example, the height, so that mean and variance no longer have linear order.

We number the internal vertices of a random canonical t-ary tree of size n from 1 to n in a natural top-to-bottom, left-to-right way, starting at the root. Let $X_{k,n}$ denote the depth of the k-th internal vertex in a random tree $T \in \mathcal{T}$ of order n. Moreover, set $Y_{k,n} = X_{k+1,n} - X_{k,n} \in \{0,1\}$. In words, $Y_{k,n}$ is 1 if the $(k+1)$-th internal vertex has greater distance from the root than the k-th, and 0 otherwise. It is clear that the height can be expressed as

$$h(T) = 1 + \max_k X_{k,n} = 1 + X_{n,n} = 1 + \sum_{k=1}^{n-1} Y_{k,n},$$

which would indeed be an alternative approach to the central limit theorem for the height. More importantly, though, the internal path length can also be expressed in terms of the random variables $Y_{k,n}$:

$$\ell_{internal}(T) = \sum_{k=1}^{n} X_{k,n} = \sum_{k=1}^{n}\sum_{j=1}^{k-1} Y_{j,n} = \sum_{j=1}^{n-1}(n-j)Y_{j,n}.$$

Now

$$n^{-1}\ell_{internal}(T) = \sum_{j=1}^{n-1}\frac{n-j}{n}Y_{j,n}$$

can be seen as a sum of $n-1$ bounded random variables $Z_{j,n} = \frac{n-j}{n} Y_{j,n}$. An advantage of this decomposition over other possible decompositions (e.g., by counting the number of vertices at different depths) is that the number of variables is not random. The $Z_{j,n}$ are neither identically distributed (which is not a major issue) nor independent. Fortunately, however, they are almost independent in that they satisfy a so-called "strong mixing condition". Let $\mathcal{F}_{s_1}$ be the σ-algebra induced by the random variables $Z_{1,n}, Z_{2,n}, \ldots, Z_{s_1,n}$, and let $\mathcal{G}_{s_2}$ be the σ-algebra induced by the random variables $Z_{s_2,n}, Z_{s_2+1,n}, \ldots, Z_{n-1,n}$. There exist constants κ and λ such that

$$|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| \leq \kappa e^{-\lambda(s_2 - s_1)} \tag{7.6.1}$$

for all $1 \leq s_1 < s_2 \leq n$ and all events $A \in \mathcal{F}_{s_1}$ and $B \in \mathcal{G}_{s_2}$. The main idea is simple: events $A \in \mathcal{F}_{s_1}$ describe the shape of the random tree T up to the s_1-th internal vertex, while events $B \in \mathcal{G}_{s_2}$ describe the shape of the random tree T from the s_2-th internal vertex on. The probabilities of such events can be calculated by means of the generating function approach explained in Section 7.2, and the exponential error terms that one obtains through this approach (as in Theorem 7.5.1) yield the estimate (7.6.1) above. A more detailed explanation can be found in the appendix once again.

Once the stated mixing condition has been proven, one can apply general central limit theorems for sums of random variables with strong mixing conditions, here specifically a result of Sunklodas [99, Theorem 1]. Putting everything together, we get

Theorem 7.6.1. *The total path length (as well as the internal and external path lengths) is asymptotically normal distributed. Its mean is asymptotically $\mu_{tpl} n^2 + O(n)$ and its variance is asymptotically $\sigma_{tpl}^2 n^3 + O(n^2)$ with*

$$\mu_{tpl} = \frac{t}{2}\mu_h = \frac{t}{4} + \frac{t^2 - 2t}{2^{t+4}} + \frac{2t^3 + 3t^2 - 8t}{2^{2t+6}} + \frac{9t^4 + 45t^3 + 2t^2 - 88t}{2^{3t+9}} + O\left(\frac{t^4}{2^{4t}}\right)$$

and

$$\sigma_{tpl}^2 = \frac{t^2}{12} + \frac{-t^4 + 5t^3 + 2t^2}{3 \cdot 2^{t+4}} + \frac{-4t^5 + 4t^4 + t^3 + 14t^2}{3 \cdot 2^{2t+6}} + O\left(\frac{t^6}{2^{3t}}\right)$$

for $t \geq 2$.

7.7 The Width

In this final section, we consider the width $w(T)$, the maximum number of leaves on the same level, for which we have the following theorem:

Theorem 7.7.1. *For a random $T \in \mathcal{T}$ of size n, we have*

$$\mathbb{E}(w(T)) = \mu_w \log n + O(\log \log n),$$

where μ_w is given by

$$\mu_w = \frac{1}{-(t-1)\log q_0} = \frac{1}{t\log(2)} + \frac{1}{t^2\log(2)} + \frac{1}{t^3\log(2)} + \frac{1}{t^4\log(2)} + \frac{1}{t^5\log(2)} + \frac{2}{t^6}\varepsilon_6(t)$$

t	μ_w	t	μ_w	t	μ_w
2	1.7107...	9	0.1804...	16	0.0961...
3	0.7660...	10	0.1603...	17	0.0901...
4	0.4936...	11	0.1442...	18	0.0848...
5	0.3650...	12	0.1311...	19	0.0801...
6	0.2902...	13	0.1202...	20	0.0759...
7	0.2411...	14	0.1109...	21	0.0721...
8	0.2063...	15	0.1030...	22	0.0686...

Table 7.7.1: Values of μ_w for $2 \le t \le 22$.

for $t \ge 23$. For $2 \le t \le 22$, the values of μ_w are given in Table 7.7.1.

Furthermore, we have the concentration property

$$\mathbb{P}(|w(T) - \mu_w \log n| \ge 3\mu_w \log\log n) = O\left(\frac{1}{\log n}\right). \tag{7.7.1}$$

Once again, we only sketch the idea of the proof here, details can be found in Appendix 7.13.

We consider the trees with width bounded by K. The corresponding generating function $W_K(q) = \sum_{\substack{T\in\mathcal{T}\\ w(T)\le K}} q^{n(T)}$ can be constructed by a suitable transfer matrix, and we quantify the obvious convergence of $W_K(q)$ to $H(q,1,1,1)$. The dominant singularity q_K of $W_K(q)$ is estimated by truncating the infinite positive eigenvector of an infinite transfer matrix corresponding to $H(q,1,1,1)$ and applying methods from Perron-Frobenius theory. Then the probability $\mathbb{P}(w(T) \le K)$ can be extracted from $W_K(q)$ using singularity analysis. Our key estimate states that the singularity q_K converges exponentially to q_0, from which the main term of the expectation as well as the concentration property are obtained quite easily. A more precise result on the distribution of the width would depend on a better understanding of the behaviour of q_K as $K \to \infty$, which seems to be quite complicated.

7.8 Supplement to Section 7.2, "The Generating Function"

Proof of Theorem 7.2.1. As it was already mentioned in Section 7.2, we first consider

$$H_h(q,u,v) := [w^h]H(q,u,v,w) = \sum_{\substack{T\in\mathcal{T}\\ h(T)=h}} q^{n(T)} u^{m(T)} v^{d(T)}$$

for some $h \ge 0$.

A tree T' of height $h+1$ arises from a tree T of height h by replacing j of its $m(T)$ leaves on the last level by internal vertices with t succeeding leaves respectively, where $1 \le j \le m(T)$. If $j = m(T)$, then all old leaves become internal vertices, so that

$d(T') = d(T)$; otherwise, at least one of them becomes a new leaf, meaning that we have a new level that contains one or more leaves, hence $d(T') = d(T)+1$. For the generating function, this translates to the following functional equation:

$$\begin{aligned} H_{h+1}(q,u,v) &= \sum_{\substack{T\in\mathcal{T}\\ h(T)=h}} \left(\sum_{j=1}^{m(T)-1} q^{n(T)+j} u^{jt} v^{d(T)+1} + q^{n(T)+m(T)} u^{m(T)t} v^{d(T)} \right) \\ &= \sum_{\substack{T\in\mathcal{T}\\ h(T)=h}} q^{n(T)} v^{d(T)} \left(qu^t v \frac{1-(qu^t)^{m(T)}}{1-qu^t} + (1-v)(qu^t)^{m(T)} \right) \\ &= r(q,u,v) H_h(q,1,v) + s(q,u,v) H_h(q,qu^t,v), \end{aligned} \tag{7.8.1}$$

where we set

$$r(q,u,v) = \frac{qu^t v}{1-qu^t}, \qquad s(q,u,v) = \frac{1-v-qu^t}{1-qu^t}.$$

Note that the initial value is given by $H_0(q,u,v) = uv$. Now set

$$\mathcal{D}_0 := \{(q,u,v,w) \in \mathbb{C}^4 \mid |q| < 1/5, |u| \le 1, |v-1| < 1/5, |w| \le 1\}.$$

We note that if $(q,u,v,w) \in \mathcal{D}_0$, we have

$$|r(q,u,v)| \le \frac{3}{10}, \qquad |s(q,u,v)| \le \frac{1}{2}.$$

This and (7.8.1) imply that $|H_h(q,u,v)| \le (4/5)^h$ holds for $h \ge 0$ and $(q,u,v,w) \in \mathcal{D}_0$. This implies that $H(q,u,v,w) = \sum_{h\ge 0} H_h(q,u,v) w^h$ converges uniformly for $(q,u,v,w) \in \mathcal{D}_0$.

Multiplying (7.8.1) by w^{h+1} and summing over all $h \ge 0$ yields the functional equation

$$H(q,u,v,w) = uv + wr(q,u,v)H(q,1,v,w) + ws(q,u,v)H(q,qu^t,v,w). \tag{7.8.2}$$

We iterate this functional equation and obtain

$$\begin{aligned} H(q,u,v,w) = a_k(q,u,v,w) + b_k(q,u,v,w)H(q,1,v,w) \\ + c_k(q,u,v,w)H(q,q^{[\![k+1]\!]}u^{t^{k+1}},v,w) \end{aligned} \tag{7.8.3}$$

for $k \ge 0$ with

$$\begin{aligned} a_k(q,u,v,w) &= v \sum_{j=0}^{k} q^{[\![j]\!]} u^{t^j} w^j \prod_{i=0}^{j-1} s(q,q^{[\![i]\!]}u^{t^i},v), \\ b_k(q,u,v,w) &= \sum_{j=0}^{k} r(q,q^{[\![j]\!]}u^{t^j},v) w^{j+1} \prod_{i=0}^{j-1} s(q,q^{[\![i]\!]}u^{t^i},v), \\ c_k(q,u,v,w) &= w^{k+1} \prod_{i=0}^{k} s(q,q^{[\![i]\!]}u^{t^i},v). \end{aligned}$$

Let now

$$\mathcal{D} = \{(q,u,v,w) \in \mathbb{C}^4 \mid |q| < |u|^{1-t}, |w| \cdot |1-v| < 1\}.$$

For $(q,u,v,w) \in \mathcal{D}$, we have $\lim_{k\to\infty} q^{[\![k]\!]}u^{t^k} = 0$ and $\lim_{k\to\infty} ws(q, q^{[\![k]\!]}u^{t^k}, v) = |w| \cdot |v-1| < 1$ and the limits

$$a(q,u,v,w) := \lim_{k\to\infty} a_k(q,u,v,w) = v\sum_{j=0}^{\infty} q^{[\![j]\!]}u^{t^j}w^j \prod_{i=0}^{j-1} s(q, q^{[\![i]\!]}u^{t^i}, v),$$

$$b(q,u,v,w) := \lim_{k\to\infty} b_k(q,u,v,w) = \sum_{j=0}^{\infty} r(q, q^{[\![j]\!]}u^{t^j}, v)w^{j+1} \prod_{i=0}^{j-1} s(q, q^{[\![i]\!]}u^{t^i}, v)$$

exist. As $\lim_{k\to\infty} c_k(q,u,v,w) = 0$ for these $(q,u,v,w) \in \mathcal{D}$, the limit of (7.8.3) for $k \to \infty$ is

$$H(q,u,v,w) = a(q,u,v,w) + b(q,u,v,w)H(q,1,v,w). \tag{7.8.4}$$

Setting $u = 1$ in (7.8.4) yields (7.2.1). □

We also state a simplified expression and a functional equation for $b(q,u,v,w)$ in the case $v = 1$, $w = 1$:

Lemma 7.8.1. *We have*

$$b(q,u,1,1) = \sum_{j=1}^{\infty}(-1)^{j-1}\prod_{i=1}^{j}\frac{q^{[\![i]\!]}u^{t^i}}{1-q^{[\![i]\!]}u^{t^i}} = \frac{qu^t}{1-qu^t}(1-b(q,qu^t,1,1)).$$

In particular, the coefficient $[u^j]b(q,u,1,1)$ vanishes if j is not a multiple of t.

Proof of Lemma 7.8.1. This is an immediate consequence of (7.2.2). □

7.9 Supplement to Section 7.3, "The Height"

This section is, as the title reveals, a supplement to our discussion of the height. It contains numerically calculated values for the constants of Theorem 7.3.1 and the proof of this theorem. We start with the latter. Note that a brief sketch of the proof was already given in Section 7.3.

Proof of Theorem 7.3.1. Throughout this proof the notations of Section 7.3 are used. Further, we make use of Theorem IX.9 of Flajolet and Sedgewick [38] and apply that theorem to the function $H(q,1,1,w)$.

Recall the notation $D(q,w)$ as the denominator of $H(q,1,1,w)$ and let q_0 be the zero of $D(q,1)$ according to Lemma 7.2.2. Set

$$c_{ij} = \frac{\partial^{i+j}}{\partial q^i \partial w^j} D(q,w)\Big|_{q=q_0,\, w=1}.$$

Then the expectation of $h(T)$ is asymptotically normally distributed and we can obtain the mean $\mu_h n + O(1)$ with

$$\mu_h = \frac{c_{01}}{c_{10} q_0},$$

and the variance to $\sigma_h^2 n + O(1)$ with

$$\sigma_h^2 = \frac{c_{01}^2 c_{20} q_0 + c_{01} c_{10}^2 q_0 - 2\, c_{01} c_{10} c_{11} q_0 + c_{02} c_{10}^2 q_0 + c_{01}^2 c_{10}}{c_{10}^3 q_0^2}.$$

To calculate the coefficients c_{ij} we need derivatives of $D(q, w)$. In order to avoid working with infinite sums, we use the approximations

$$D_K(q, w) := \sum_{0 \le k < K} (-1)^k w^k \prod_{j=1}^{k} \frac{q^{[\![j]\!]}}{1 - q^{[\![j]\!]}}.$$

Lemma 7.9.1 shows that the error made by using those approximations is small. For the calculations themselves, Sage [98] was used. □

Lemma 7.9.1. *Let $i \in \{0, 1, 2\}$ and $j \in \mathbb{N}_0$, and let $q \in \mathbb{C}$ with $1/2 \le |q| \le 1/r_3$, where $r_3 = 1 + \frac{\log 2}{t} - \frac{\log 2 - \log^2 2}{2t^2}$. Then*

$$\frac{\partial^{i+j}}{\partial q^i \partial w^j} \big(D(q, w) - D_4(q, w)\big)\Big|_{w=1} = O\left(\frac{1}{2^{t^2}}\right).$$

Proof. The result was shown for $i \in \{0, 1\}$ and $j = 0$ in [35]. Here we follow the proof of Lemma 9 of that article. We first note that it is sufficient to show the result for $j = 0$, since that derivative results in a polynomial in k, which is asymptotically smaller than the factor t^k which appears.

Now set

$$f_j(q) := \frac{q^{[\![j]\!]}}{1 - q^{[\![j]\!]}}.$$

We obtain

$$\frac{\partial}{\partial q}\left(\prod_{j=1}^{k} f_j(q)\right) = \frac{1}{q} \prod_{j=1}^{k} f_j(q) \left(\sum_{j=1}^{k} \frac{[\![j]\!]}{1 - q^{[\![j]\!]}}\right)$$

for its first derivative and

$$\frac{\partial^2}{\partial q^2}\left(\prod_{j=1}^{k} f_j(q)\right) = \frac{1}{q^2} \prod_{j=1}^{k} f_j(q) \left(\left(\sum_{j=1}^{k} \frac{[\![j]\!]}{1 - q^{[\![j]\!]}}\right)^2 - \sum_{j=1}^{k} \frac{[\![j]\!]}{1 - q^{[\![j]\!]}} + \sum_{j=1}^{k} \frac{[\![j]\!]^2 q^{[\![j]\!]}}{\left(1 - q^{[\![j]\!]}\right)^2}\right)$$

for its second. As in [35], we can find the bounds

$$\left|\prod_{j=1}^{k} f_j(1/z)\right| \le \frac{t}{2^{-1 + t(k-1)/2 + (k-3)t^2}}$$

and

$$\left|\sum_{j=1}^{k} \frac{[\![j]\!]}{1-(1/z)^{[\![j]\!]}}\right| \le 4kt^k.$$

Therefore, we also deduce that

$$\left|\sum_{j=1}^{k} \frac{[\![j]\!]^2 q^{[\![j]\!]}}{\left(1-q^{[\![j]\!]}\right)^2}\right| \le \left(\sum_{j=1}^{k} \frac{[\![j]\!]}{1-(1/|z|)^{[\![j]\!]}}\right)^2 \le (4kt^k)^2.$$

This yields the bound

$$\begin{aligned}\left|\frac{\partial^{i+j}}{\partial q^i \partial w^j}\big(D(q,w)-D_4(q,w)\big)\Big|_{w=1}\right| &\le |z|^2 \sum_{k=4}^{\infty} \frac{t}{2^{(k-3)t^2+(k-1)t/2-1}}\left(2(4kt^k)^2+4kt^k\right)\\ &\le \sum_{k=4}^{\infty} \frac{k^2t^{2k+1}}{2^{(k-3)t^2+(k-1)t/2-9}} \le \sum_{k=4}^{\infty} \frac{c}{2^{(k-3)t^2}}\end{aligned}$$

for some positive constant c. Since the last sum in the previous inequality is $O(2^{-t^2})$, the result follows. □

The end of this section contains the following: For $t \le 30$ we calculated the constants of Theorem 7.3.1 numerically. The computer algebra software Sage [98] was used for this purpose. The results can be found in Table 7.9.1.

7.10 Supplement to Section 7.4, "The Number of Distinct Depths of Leaves"

Similar to the previous supplementary section, this section contains explicitly calculated values for the constants of Theorem 7.4.1 and a detailed proof of this theorem, following the proof sketch that was given in Section 7.4. We start with the latter. The ideas used are very similar to the ones in the analysis of the height.

Proof of Theorem 7.4.1. Throughout this proof the notations of Section 7.4 are used. Again, as with the heights, we make use of Theorem IX.9 of Flajolet and Sedgewick [38] and apply that theorem to the function $H(q,1,v,1)$.

Again, we use the notation $D(q,v)$ for the denominator of $H(q,1,v,1)$ and let q_0 be the zero of $D(q,1)$ according to Lemma 7.2.2. We expand $D(q,v)$ around $(q_0,1)$ and can then calculate the main term of mean and variance from the coefficients of that series. The required formulas can be found in the proof of Theorem 7.3.1 in Appendix 7.9.

Again, to calculate the coefficients we need derivatives of $D(q,v)$ and we use the approximations

$$D_K(q,v) := 1 - \sum_{1\le k<K} \frac{v}{1-q^{[\![k]\!]}} \prod_{j=1}^{k-1} \frac{1-v-q^{[\![j]\!]}}{1-q^{[\![j]\!]}}.$$

t	μ_h	σ_h^2
2	0.5662757699172865	0.2665499010273937
3	0.5330981433252730	0.2636253024859229
4	0.5216132420088969	0.2465916388296734
5	0.5137644953351326	0.2404182925457133
6	0.5084950082063058	0.2396633993739495
7	0.5051047365215813	0.2411570855092153
8	0.5030001253275541	0.2432575483836213
9	0.5017308605343554	0.2452173961787763
10	0.5009832278618641	0.2467757623911674
11	0.500551313637743	0.2479077234990245
12	0.5003057656286383	0.2486821135530906
13	0.5001680187030247	0.2491894707701658
14	0.5000916023570357	0.2495111461587043
15	0.5000496052425100	0.2497099052572736
16	0.5000267068978588	0.2498301915991255
17	0.5000143062444377	0.2499017551259219
18	0.5000076297101404	0.2499437283128117
19	0.5000040532034994	0.2499680504612380
20	0.5000021457914275	0.2499819989727347
21	0.5000011324949086	0.2499899266916567
22	0.500000596048271	0.2499943971277963
23	0.5000003129248821	0.2499969005482699
24	0.5000001639129082	0.2499982938141369
25	0.500000085681714	0.2499990649513116
26	0.5000000447034934	0.2499994896349970
27	0.5000000232830670	0.2499997224658077
28	0.5000000121071942	0.2499998495913860
29	0.500000006286428	0.2499999187421003
30	0.5000000032596291	0.2499999562278376

Table 7.9.1: Numerical values of the constants in mean and variance of the height for small values of t, cf. Theorem 7.3.1. It would be possible to calculate the values with even higher accuracy.

Lemma 7.10.1 shows that the error in this approximation is small. Again, for the calculations themselves, Sage [98] was used. □

Lemma 7.10.1. *Let $i, j \in \{0, 1, 2\}$, and let $q \in \mathbb{C}$ with $1/2 \le |q| \le 1/r_3$, where $r_3 = 1 + \frac{\log 2}{t} - \frac{\log 2 - \log^2 2}{2t^2}$. Then*

$$\frac{\partial^{i+j}}{\partial q^i \partial v^j}\big(D(q,v) - D_4(q,v)\big)\Big|_{v=1} = O\left(\frac{1}{2^{t^2}}\right).$$

Proof. The proof is similar to the proof of Lemma 7.9.1. □

The end of this section contains numerically calculated values for the constants of Theorem 7.3.1. We used the computer algebra software Sage [98], and the results can be found in Table 7.10.1.

7.11 Supplement to Section 7.5, "The Number of Leaves on the Last Level"

Proof of Theorem 7.5.1. Let $q_1 = 1 - \frac{0.72}{t}$. Then singularity analysis shows that the probability generating function $p_n(u)$ of $m(T)$ is given by

$$p_n(u) = b(q_0, u, 1, 1) + O(Q^n),$$

uniformly for $|u| \le 1$.

The limiting distribution follows from [38, Theorem IX.2]. Expectation and variance follow upon differentiating $b(q_0, u, 1, 1)$ with respect to u and inserting the asymptotic expression for q_0. □

7.12 Supplement to Section 7.6, "The Path Length"

Here we provide some more details of our analysis of the total (internal, external) path length, starting with the generating functions. Recall that we defined the generating function $L_r(q, u, w)$ for the r-th moment of the total path length:

$$L_r(q, u, w) = \sum_{T \in \mathcal{T}} \ell(T)^r q^{n(T)} u^{m(T)} w^{h(T)}.$$

In particular, $L_0(q, u, w)$ is the ordinary generating function for all trees, where u marks the number of leaves on the highest level and w the height. From the recursive characterisation of canonical trees, we got the identity

$$L_0(q, u, w) = u + \frac{qu^t}{1 - qu^t}\left(L_0(q, 1, w) - L_0(q, qu^t, w)\right),$$

from which we obtained, by means of iteration, an explicit formula for L_0, namely

$$L_0(q, u, w) = a_0(q, u, w) + b(q, u, w) L_0(q, 1, w)$$

t	μ_d	σ_d^2
2	0.4042366935349558	0.2491723144610512
3	0.4868358747318154	0.2900504810033636
4	0.5024585834463688	0.2741245386044700
5	0.5050331954313614	0.2607084552774208
6	0.5043408269340329	0.2530808413030350
7	0.5030838633817897	0.2495578056054625
8	0.5020050053196333	0.2483362931739360
9	0.5012375070905983	0.2482103208441572
10	0.5007377066674932	0.2485046286268309
11	0.5004288693844008	0.2488904008073738
12	0.5002446296853791	0.2492332759318571
13	0.5001374740872935	0.2494951950687874
14	0.5000763363460676	0.2496791536316180
15	0.5000419739265400	0.2498015045792620
16	0.5000228916911940	0.2498797960254888
17	0.500012398761189	0.2499284618053178
18	0.500006676000353	0.2499580344990146
19	0.5000035763570187	0.2499756801559131
20	0.5000019073704041	0.2499860521721408
21	0.5000010132849795	0.2499920724820041
22	0.5000005364434586	0.2499955296224207
23	0.500000283122517	0.2499974965964656
24	0.5000001490117357	0.2499986067389993
25	0.500000078231130	0.2499992288642147
26	0.5000000409782024	0.2499995753167091
27	0.5000000214204217	0.2499997671693008
28	0.5000000111758715	0.2499998728744530
29	0.5000000058207663	0.2499999308492945
30	0.5000000030267984	0.2499999625142652

Table 7.10.1: Values of the constants in mean and variance of the number of distinct depths of leaves for small values of t, cf. Theorem 7.4.1. It would be possible to calculate the values with even higher accuracy.

and in particular,

$$L_0(q,1,w) = \frac{a_0(q,1,w)}{1-b(q,1,w)},$$

where

$$a_0(q,u,w) = \sum_{j=0}^{\infty}(-1)^j w^j q^{[\![j]\!]} u^{t^j} \prod_{i=1}^{j} \frac{q^{[\![i]\!]} u^{t^i}}{1-q^{[\![i]\!]} u^{t^i}}$$

and

$$b(q,u,w) = \sum_{j=1}^{\infty}(-1)^{j-1} w^j \prod_{i=1}^{j} \frac{q^{[\![i]\!]} u^{t^i}}{1-q^{[\![i]\!]} u^{t^i}}.$$

Likewise, the functional equations one obtains for L_1 and L_2 can be solved by means of iteration: one has

$$L_1(q,u,w) = \Phi_u \Phi_w L_0(q,u,w) + \frac{qu^t}{1-qu^t}\left(L_1(q,1,w) - L_1(q,qu^t,w)\right),$$

and thus

$$L_1(q,u,w) = a_1(q,u,w) + b(q,u,w) L_1(q,1,w),$$

and in particular

$$L_1(q,1,w) = \frac{a_1(q,1,w)}{1-b(q,1,w)}$$

with

$$a_1(q,u,w) = \sum_{j=0}^{\infty}(-1)^j w^j (\Phi_u \Phi_w L_0)(q, q^{[\![j]\!]} u^{t^j}, w) \prod_{i=1}^{j} \frac{q^{[\![i]\!]} u^{t^i}}{1-q^{[\![i]\!]} u^{t^i}}.$$

Finally,

$$L_2(q,u,w) = 2\Phi_u \Phi_w L_1(q,u,w) - \Phi_u^2 \Phi_w^2 L_0(q,u,w) + \frac{qu^t}{1-qu^t}\left(L_2(q,1,w) - L_2(q,qu^t,w)\right),$$

and thus

$$L_2(q,u,w) = a_2(q,u,w) + b(q,u,w) L_2(q,1,w),$$

and in particular

$$L_2(q,1,w) = \frac{a_2(q,1,w)}{1-b(q,1,w)}$$

with

$$a_2(q,u,w) = \sum_{j=0}^{\infty}(-1)^j w^j \left(2(\Phi_u \Phi_w L_1)(q, q^{[\![j]\!]} u^{t^j}, w) - (\Phi_u^2 \Phi_w^2 L_0)(q, q^{[\![j]\!]} u^{t^j}, w)\right) \prod_{i=1}^{j} \frac{q^{[\![i]\!]} u^{t^i}}{1-q^{[\![i]\!]} u^{t^i}}.$$

Substituting back, we get an explicit expression for $L_1(q,1,w)$:

$$\begin{aligned}L_1(q,1,w) = {} & \frac{a_0(q,1,w)(\Phi_w b)(q,1,w)}{(1-b(q,1,w))^3}\sum_{j=0}^{\infty}(-1)^j w^j \prod_{i=1}^{j}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u b)(q,q^{[\![j]\!]},w)\\ & + \frac{a_0(q,1,w)}{(1-b(q,1,w))^2}\sum_{j=0}^{\infty}(-1)^j w^j \prod_{i=1}^{j}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u\Phi_w b)(q,q^{[\![j]\!]},w)\\ & + \frac{(\Phi_w a_0)(q,1,w)}{(1-b(q,1,w))^2}\sum_{j=0}^{\infty}(-1)^j w^j \prod_{i=1}^{j}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u b)(q,q^{[\![j]\!]},w)\\ & + \frac{1}{1-b(q,1,w)}\sum_{j=0}^{\infty}(-1)^j w^j \prod_{i=1}^{j}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u\Phi_w a_0)(q,q^{[\![j]\!]},w).\end{aligned}$$

The dominant term in this sum is the first one, with a triple pole at the dominant singularity q_0. The second and third term, however, are also relevant in the calculation of the variance, where one further term in the asymptotic expansion is needed in view of the inevitable cancellation in the main term. Singularity analysis immediately yields the asymptotic behaviour of the mean: since the pole is of cubic order, the order of the mean is quadratic, i.e., it is asymptotically equal to $\mu_{tpl}n^2$, where the constant μ_{tpl} is given by

$$\mu_{tpl} = \frac{(\Phi_w b)(q_0,1,1)}{2(\Phi_q b)(q_0,1,1)^2}\sum_{j=0}^{\infty}(-1)^j(\Phi_u b)(q_0,q_0^{[\![j]\!]},1)\prod_{i=1}^{j}\frac{q_0^{[\![i]\!]}}{1-q_0^{[\![i]\!]}}.$$

Plugging in the definition of b as a sum, it is possible to simplify this further: one has

$$(\Phi_u b)(q,u,1) = \sum_{k=1}^{\infty}(-1)^{k-1}\prod_{h=1}^{k}\frac{q^{[\![h]\!]}u^{t^h}}{1-q^{[\![h]\!]}u^{t^h}}\sum_{h=1}^{k}\frac{t^h}{1-q^{[\![h]\!]}u^{t^h}}$$

by logarithmic differentiation and thus

$$\begin{aligned}(\Phi_u b)(q,q^{[\![j]\!]},1) &= \sum_{k=1}^{\infty}(-1)^{k-1}\prod_{h=1}^{k}\frac{q^{[\![h]\!]+t^h[\![j]\!]}}{1-q^{[\![h]\!]+t^h[\![j]\!]}}\sum_{h=1}^{k}\frac{t^h}{1-q^{[\![h]\!]+t^h[\![j]\!]}}\\ &= \sum_{k=1}^{\infty}(-1)^{k-1}\prod_{i=j+1}^{j+k}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}\sum_{h=1}^{k}\frac{t^h}{1-q^{[\![h+j]\!]}}\end{aligned}$$

since $[\![h]\!]+t^h[\![j]\!] = [\![h+j]\!]$ by definition. Plugging in, we find

$$\mu_{tpl} = \frac{(\Phi_w b)(q_0,1,1)}{2(\Phi_q b)(q_0,1,1)^2}\sum_{j=0}^{\infty}\sum_{k=1}^{\infty}(-1)^{j+k-1}\prod_{i=1}^{j+k}\frac{q_0^{[\![i]\!]}}{1-q_0^{[\![i]\!]}}\sum_{h=1}^{k}\frac{t^h}{1-q_0^{[\![h+j]\!]}}.$$

Substituting $\ell = j + k$ and interchanging the order of summation, we arrive at

$$\begin{aligned}\mu_{tpl} &= \frac{(\Phi_w b)(q_0,1,1)}{2(\Phi_q b)(q_0,1,1)^2}\sum_{\ell=1}^{\infty}(-1)^{\ell-1}\prod_{i=1}^{\ell}\frac{q_0^{[\![i]\!]}}{1-q_0^{[\![i]\!]}}\sum_{k=1}^{\ell}\sum_{h=1}^{k}\frac{t^h}{1-q_0^{[\![h+\ell-k]\!]}}\\ &= \frac{(\Phi_w b)(q_0,1,1)}{2(\Phi_q b)(q_0,1,1)^2}\sum_{\ell=1}^{\infty}(-1)^{\ell-1}\prod_{i=1}^{\ell}\frac{q_0^{[\![i]\!]}}{1-q_0^{[\![i]\!]}}\sum_{r=1}^{\ell}\sum_{h=1}^{r}\frac{t^h}{1-q_0^{[\![r]\!]}}\\ &= \frac{(\Phi_w b)(q_0,1,1)}{2(\Phi_q b)(q_0,1,1)^2}\sum_{\ell=1}^{\infty}(-1)^{\ell-1}\prod_{i=1}^{\ell}\frac{q_0^{[\![i]\!]}}{1-q_0^{[\![i]\!]}}\sum_{r=1}^{\ell}\frac{t[\![r]\!]}{1-q_0^{[\![r]\!]}}.\end{aligned}$$

Noting now that

$$(\Phi_q b)(q,1,1) = \sum_{\ell=1}^{\infty}(-1)^{\ell-1}\prod_{i=1}^{\ell}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}\sum_{r=1}^{\ell}\frac{[\![r]\!]}{1-q^{[\![r]\!]}},$$

which can be seen by another logarithmic differentiation, we can replace the sum in the expression for μ_{tpl} above by $t\cdot(\Phi_q b)(q_0,1,1)$, which finally yields

$$\mu_{tpl} = \frac{t}{2}\cdot\frac{(\Phi_w b)(q_0,1,1)}{(\Phi_q b)(q_0,1,1)},$$

and the fraction is precisely μ_h since the generating function of the mean height is

$$\frac{a_0(q,1,1)(\Phi_w b)(q,1,1)}{(1-b(q,1,1))^2} + \frac{(\Phi_w a_0)(q,1,1)}{1-b(q,1,1)},$$

of which the first term dominates (yet another application of singularity analysis). This means that we have proven the identity $\mu_{tpl} = t\mu_h/2$.

For the variance, one also needs the asymptotic behaviour of $L_2(q,1,1)$ at the dominant singularity. Only the terms of pole order 4 and 5 (i.e., highest and second-highest) are

needed: they are

$$
\begin{aligned}
L_2(q,1,1) &= \frac{6a_0(q,1,1)(\Phi_w b)(q,1,1)^2}{(1-b(q,1,1))^5}\left(\sum_{j=0}^{\infty}(-1)^j w^j\prod_{i=1}^{j}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u b)(q,q^{[\![j]\!]},1)\right)^2\\
&+\frac{4a_0(q,1,1)(\Phi_w b)(q,1,1)^2}{(1-b(q,1,1))^4}\sum_{j=0}^{\infty}(-1)^j\prod_{i=1}^{j}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}\left([\![j+1]\!](\Phi_u^2 b)(q,q^{[\![j]\!]},1)+\sum_{r=1}^{j}\frac{t^{[\![r]\!]}}{1-q^{[\![r]\!]}}(\Phi_u b)(q,q^{[\![j]\!]},1)\right)\\
&+\frac{8a_0(q,1,1)(\Phi_w b)(q,1,1)}{(1-b(q,1,1))^4}\sum_{j=0}^{\infty}(-1)^j\prod_{i=1}^{j}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u b)(q,q^{[\![j]\!]},1)\sum_{k=0}^{\infty}(-1)^k\prod_{i=1}^{k}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u\Phi_w b)(q,q^{[\![k]\!]},1)\\
&+\frac{6(\Phi_w a_0)(q,1,1)(\Phi_w b)(q,1,1)}{(1-b(q,1,1))^4}\left(\sum_{j=0}^{\infty}(-1)^j w^j\prod_{i=1}^{j}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u b)(q,q^{[\![j]\!]},1)\right)^2\\
&+\frac{2a_0(q,1,1)(\Phi_w b)(q,1,1)^2}{(1-b(q,1,1))^4}\left(\sum_{j=0}^{\infty}(-1)^j w^j\prod_{i=1}^{j}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u b)(q,q^{[\![j]\!]},1)\right)^2\\
&+\frac{2a_0(q,1,1)(\Phi_w b)(q,1,1)}{(1-b(q,1,1))^4}\sum_{j=0}^{\infty}(-1)^j\prod_{i=1}^{j}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u b)(q,q^{[\![j]\!]},1)\sum_{k=1}^{\infty}(-1)^k k\prod_{i=1}^{k}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u b)(q,q^{[\![k]\!]},1)\\
&-\frac{2a_0(q,1,1)(\Phi_w b)(q,1,1)^2}{(1-b(q,1,1))^4}\sum_{j=0}^{\infty}(-1)^j\prod_{i=1}^{j}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u^2 b)(q,q^{[\![j]\!]},1)
\end{aligned}
$$

Applying singularity analysis to the highest- and second-highest order terms of both L_1 and L_2 yields the variance: the terms of order n^4 cancel (as one would expect), and one finds that the variance is asymptotically $\sigma_{tpl}^2 n^3$, where

$$
\begin{aligned}
\sigma_{tpl}^2 &= \frac{F(q_0)^2(\Phi_q^2 b)(q_0,1,1)}{(\Phi_q b)(q_0,1,1)^5}-\frac{F(q_0)(\Phi_q F)(q_0)}{(\Phi_q b)(q_0,1,1)^4}\\
&-\frac{(\Phi_w b)(q_0,1,1)^2}{3(\Phi_q b)(q_0,1,1)^3}\sum_{j=0}^{\infty}(-1)^j\prod_{i=1}^{j}\frac{q_0^{[\![i]\!]}}{1-q_0^{[\![i]\!]}}(\Phi_u^2 b)(q_0,q_0^{[\![j]\!]},1)\\
&+\frac{2(\Phi_w b)(q_0,1,1)^2}{3(\Phi_q b)(q_0,1,1)^3}\sum_{j=0}^{\infty}(-1)^j\prod_{i=1}^{j}\frac{q_0^{[\![i]\!]}}{1-q_0^{[\![i]\!]}}\left([\![j+1]\!](\Phi_u^2 b)(q_0,q_0^{[\![j]\!]},1)+\sum_{r=1}^{j}\frac{t^{[\![r]\!]}}{1-q^{[\![r]\!]}}(\Phi_u b)(q_0,q_0^{[\![j]\!]},1)\right)\\
&+\frac{(\Phi_w b)(q_0,1,1)}{3(\Phi_q b)(q_0,1,1)^3}\sum_{j=0}^{\infty}(-1)^j\prod_{i=1}^{j}\frac{q_0^{[\![i]\!]}}{1-q_0^{[\![i]\!]}}(\Phi_u b)(q_0,q_0^{[\![j]\!]},1)\sum_{k=0}^{\infty}(-1)^k\prod_{i=1}^{k}\frac{q_0^{[\![i]\!]}}{1-q_0^{[\![i]\!]}}(\Phi_u\Phi_w b)(q_0,q_0^{[\![k]\!]},1)\\
&+\frac{(\Phi_w b)(q_0,1,1)^2}{3(\Phi_q b)(q_0,1,1)^3}\left(\sum_{j=0}^{\infty}(-1)^j\prod_{i=1}^{j}\frac{q_0^{[\![i]\!]}}{1-q_0^{[\![i]\!]}}(\Phi_u b)(q_0,q_0^{[\![j]\!]},1)\right)^2\\
&+\frac{(\Phi_w b)(q_0,1,1)}{3(\Phi_q b)(q_0,1,1)^3}\sum_{j=0}^{\infty}(-1)^j\prod_{i=1}^{j}\frac{q_0^{[\![i]\!]}}{1-q_0^{[\![i]\!]}}(\Phi_u b)(q_0,q_0^{[\![j]\!]},1)\sum_{k=1}^{\infty}(-1)^k k\prod_{i=1}^{k}\frac{q_0^{[\![i]\!]}}{1-q_0^{[\![i]\!]}}(\Phi_u b)(q_0,q_0^{[\![k]\!]},1)
\end{aligned}
$$

and the function $F(q)$ is given by

$$F(q)=(\Phi_w b)(q,1,1)\sum_{j=0}^{\infty}(-1)^j\prod_{i=1}^{j}\frac{q^{[\![i]\!]}}{1-q^{[\![i]\!]}}(\Phi_u b)(q,q^{[\![j]\!]},1).$$

We determined numerical values of these constants as in the previous sections, they are given in Table 7.12.1.

t	μ_{tpl}	σ^2_{tpl}
2	0.5746406730225036	0.636553899565319
3	0.7996893802701904	0.9538514746097371
4	1.043226570739454	1.424940599745666
5	1.284411238386164	2.078739994014109
6	1.525485024618925	2.926628748193911
7	1.767866577825535	3.972171302166417
8	2.012000501310217	5.210807673614956
9	2.257788872404600	6.634216448921346
10	2.504916139309320	8.23405080979501
11	2.753032225007583	10.00388584911538
12	3.001834593771830	11.93967669304990
13	3.251092121569661	14.03939441023803
14	3.500641216499250	16.30239232264572
15	3.750372039318825	18.72881526046276
16	4.000213655182871	21.31916858572890
17	4.250121603077721	24.07405283275217
18	4.500068667391264	26.99402565883372
19	4.750038505433244	30.07954611947160
20	5.000021457914275	33.33096498586472
21	5.250011891196540	36.74853674754146
22	5.500006556530974	40.33243901952973
23	5.750003598636143	44.08279205182939
24	6.000001966954898	47.99967527333957
25	6.250001071021418	52.08314008372820
26	6.500000581145415	56.33321917051825
27	6.750000314321405	60.74993301181408
28	7.000000169500719	65.33329426993452
29	7.250000091153200	70.08331068457584
30	7.500000048894436	74.99998693820047

Table 7.12.1: Values of the constants in mean and variance of the total path length t, cf. Theorem 7.6.1. It would be possible to calculate the values with even higher accuracy.

Finally, let us describe in some more detail how the central limit theorem is obtained. Recall that $X_{k,n}$ is the (random) depth of the k-th vertex, and that $Y_{k,n} = X_{k+1,n} - X_{k,n}$. The internal path length is given by

$$\ell_{internal}(T) = \sum_{j=1}^{n-1} (n-j) Y_{j,n},$$

and thus

$$n^{-1} \ell_{internal}(T) = \sum_{j=1}^{n-1} \frac{n-j}{n} Y_{j,n}.$$

Setting $Z_{j,n} = \frac{n-j}{n} Y_{j,n}$, we obtain a decomposition for the random variable $n^{-1}\ell_{internal}(T)$:

$$n^{-1} \ell_{internal}(T) = \sum_{j=1}^{n-1} Z_{j,n}.$$

The point behind this rescaling is that the $Z_{j,n}$ are bounded now, so that they have bounded third absolute moments (and generally bounded moments of any order), which is one of the conditions to make Theorem 1 of Sunklodas [99] applicable. Another condition is that the variance of the sum grows at least linearly, which is satisfied in view of our considerations above (the variance of $\ell(T)$ is of cubic order, so the variance of the rescaled random variable is still of linear order). In Sunklodas' paper, the variables are also assumed to have expectance 0, which we could of course achieve by subtracting the mean from each $Z_{j,n}$.

The main criterion is the strong mixing inequality that was already mentioned in Section 7.6. Let two events $A \in \mathcal{F}_{s_1}$ in the σ-algebra generated by $Z_{1,n}, \ldots, Z_{s_1,n}$ and $B \in \mathcal{G}_{s_2}$ in the σ-algebra generated by $Z_{s_2,n}, Z_{s_2+1,n}, \ldots, Z_{n-1,n}$ be given. The event A consists of a collection of possible shapes of the random tree T up to the s_1-th vertex v_{s_1}, and likewise B consists of a collection of possible shapes of the random tree T from the s_2-th vertex v_{s_2} onwards.

Let H_0 be the number of vertices with label $> s_1$ on the same level as v_{s_1}, and let H_1 be the number of vertices on the following level. For any possible shape that is allowed in the event A, there is only a limited number of possibilities for H_0 and H_1. Likewise, we define K_0 to be the number of vertices on the same level as v_{s_2}, but with lower label, and K_1 the number of vertices on the previous level. The part between the levels of v_{s_1} and v_{s_2} (excluding the levels on which these two vertices are located) can be regarded as a canonical *forest*, which is defined like a canonical tree, but with H_1 different roots and K_1 different leaves on the last level.

It is not complicated to modify our generating functions approach that we used to obtain Theorem 7.5.1 to the case of several roots. Let the generating function for this purpose be $H_h(q,u)$, where h is the number of roots, q marks the size and u the number of leaves on the last level. Then it follows that

$$H_h(q,u) = a_h(q,u) + \frac{a_h(q,1) b(q,u)}{1 - b(q,1)},$$

where

$$a_h(q,u) = \sum_{j=0}^{\infty} (-1)^j q^{h[\![j]\!]} u^{ht^j} \prod_{i=1}^{j} \frac{q^{[\![i]\!]} u^{t^i}}{1 - q^{[\![i]\!]} u^{t^i}},$$
$$b(q,u) = \sum_{j=1}^{\infty} (-1)^{j-1} \prod_{i=1}^{j} \frac{q^{[\![i]\!]} u^{t^i}}{1 - q^{[\![i]\!]} u^{t^i}}.$$

The number of canonical t-ary forests with h roots, k leaves on the last level and r internal vertices is $[q^r u^k] H_h(q,u)$. Singularity analysis yields a distributional result analogous to Theorem 7.5.1, with an error term that is even uniform in h (note that $a_h(q,u)$ is bounded as a function of h in the relevant region!), but unfortunately not in k: one has

$$\frac{[q^r u^{mt}] H_h(q,u)}{[q^r] H_h(q,1)} = p_m (1 + O(Q_1^{-m} Q_2^r)),$$

where p_m is defined as in Theorem 7.5.1 and $0 < Q_1, Q_2 < 1$. However, p_m decreases exponentially in m as well, which we can use to our advantage: it is also true that

$$\frac{[q^r u^{mt}] H_h(q,u)}{[q^r] H_h(q,1)} = O(Q_3^m)$$

for some real number $0 < Q_3 < 1$.

Note that $[q^r u^k] H_h(q,u)$ gives the number of ways to fill a "gap" of r vertices, starting with h roots and ending with k leaves. This can be applied to the part of our tree T between the vertices v_{s_1} and v_{s_2}, the part between the root v_1 and v_{s_2} (where we just set $h = 1$) as well as the part from v_{s_1} to v_n (where we can sum over all k, which amounts to taking $[q^r] H_h(q,1)$).

The estimate above implies the following: the event that K_1, the number of vertices on the level before v_{s_2}, is greater than Mt, has probability $O(Q_3^{\delta(s_2-s_1)})$ if $M = \delta(s_2 - s_1)$ for some suitably chosen δ. Conditioned on the event that this is not the case, however, the difference of the probability of $A \cap B$ and the product of the probabilities of A and B is small:

$$\mathbb{P}(A \cap B | K_1 \le Mt) = \mathbb{P}(A | K_1 \le Mt) \mathbb{P}(B | K_1 \le Mt) \Big(1 + O(Q_1^{-\delta(s_2-s_1)} Q_2^{s_2-s_1})\Big).$$

Combining the two, we arrive at

$$|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| = O\Big(Q_1^{-\delta(s_2-s_1)} Q_2^{s_2-s_1} + Q_3^{\delta(s_2-s_1)}\Big),$$

and if δ is chosen sufficiently small, but fixed, then both terms decrease exponentially in $s_2 - s_1$, proving the strong mixing condition and thus the central limit theorem.

7.13 Supplement to Section 7.7, "The Width"

This appendix is devoted to the proof of Theorem 7.7.1.

Apart from the width $w(T)$, we also need the "inner width" $w^*(T)$ defined to be

$$w^*(T) := \max_{0 \le k < h(T)} L_T(k)$$

for a recursive construction, where $L_T(k)$ denotes the number of leaves at level k. By definition, the inner width $w^*(T)$ does not take the leaves on the last level into account.

For $K > 0$, we are interested in the generating function

$$W_K(q) := \sum_{\substack{T \in \mathcal{T} \\ w(T) \le K}} q^{n(T)}.$$

We represent $W_K(q)$ in terms of the generating functions

$$W_{K,r} := \sum_{\substack{T \in \mathcal{T} \\ w^*(T) \le K \\ m(T) = tr}} q^{n(T)}$$

for $r \ge 0$ such that

$$W_K(q) = 1 + \sum_{r=1}^{\lfloor K/t \rfloor} W_{K,r}.$$

Here, the summand 1 corresponds to the tree of order 1. For all other trees, the number $m(T)$ of leaves on the last level is clearly a multiple of t.

In order to compute $W_{K,r}$ recursively, we will do so for $1 \le r \le N(K)$ with $N(K) := \lceil K/(t-1) \rceil - 1$. Thus we consider the column vector

$$\mathbf{W}_K(q) := (W_{K,1}(q), \ldots, W_{K,N(K)}(q))^T.$$

We consider the "transfer matrix"

$$M_K(q) := \left(q^r \left[\frac{r}{t} \le s \le \frac{r+K}{t} \right] \right)_{\substack{1 \le r \le N(K) \\ 1 \le s \le N(K)}},$$

where the Iversonian notation[1]

$$[\text{expr}] = \begin{cases} 1 & \text{if expr is true,} \\ 0 & \text{if expr is false} \end{cases}$$

popularised by Graham, Knuth and Patashnik [45] has been used.

We now express $\mathbf{W}_K(q)$ in terms of $M_K(q)$:

[1] Keep in mind that we also use square brackets for extracting coefficients: $[q^n]Q(q)$ gives the nth coefficient of the power series Q.

Lemma 7.13.1. *For $K \geq t$, we have*

$$\mathbf{W}_K(q) = (I - M_K(q))^{-1} \begin{pmatrix} q \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{7.13.1}$$

Proof. As in the proof of Theorem 7.2.1, a tree T' of height $h+1 \geq 2$, inner width at most K and $m(T') = rt$ arises from a tree T of height h, inner width at most K and $m(T) = st$ by replacing r of the st leaves of T on the last level by inner vertices with t succeeding leaves each. We obviously have $r \leq st$. In order to ensure that $w^*(T') \leq K$, we have to ensure that $st - r \leq K$. We rewrite these two inequalities as

$$\frac{r}{t} \leq s \leq \frac{r+K}{t}. \tag{7.13.2}$$

If we have $r \leq N(K)$, we have $r < K/(t-1)$ and therefore $s < K/(t-1)$ by (7.13.2), i.e., $s \leq N(K)$. This justifies our choice of $N(k)$. This construction yields s new inner vertices in T'.

There is only one tree T' of height < 2, inner width at most K and $m(T') = rt$, namely the star of order $t+1$ for $r = 1$ which has one internal vertex (the root).

Translating these considerations into the language of generating functions yields

$$W_{K,r}(q) = q[r=1] + \sum_{s=1}^{N(k)} q^r \left[\frac{r}{t} \leq s \leq \frac{r+K}{t}\right] W_{k,s}(q).$$

Rewriting this in vectorial form yields (7.13.1). □

In order to get asymptotic expressions for the coefficients of $\mathbf{W}_K$, we have to find the singularities of $(I - M_K(q))^{-1}$ as a meromorphic function in q. A value q is a singularity of $(I - M_K(q))^{-1}$ if and only if it is a zero of the determinant $\det(I - M_K(q))$, which holds if and only if 1 is an eigenvalue of $M_K(q)$. In the next lemma, we collect a few results connecting $M_K(q)$ with Perron–Frobenius theory.

Lemma 7.13.2. *Let $K \geq t$ and $q > 0$. Then*

1. *the matrix $M_K(q)$ is a non-negative, irreducible, primitive matrix;*
2. *the function $q \mapsto \lambda_{\max}(M_K(q))$ mapping q to the spectral radius of $M_K(q)$ is a strictly increasing function from $(0,\infty)$ to $(0,\infty)$;*
3. *if $M_K(q)x \leq x$ or $M_K(q)x \geq x$ holds componentwise for some positive vector x, then $\lambda_{\max}(M_K(q)) \leq 1$ or $\lambda_{\max}(M_K(q)) \geq 1$, respectively.*

Proof. 1. The matrix $M_K(q)$ is non-negative by definition. We note that $\frac{r}{t} \leq r-1$ holds for all $r \geq 2$ and $r+1 \leq \frac{r+K}{t}$ holds for all $r < N(K)$. This implies that all subdiagonal, diagonal and superdiagonal elements of $M_K(q)$ are positive. Thus $M_K(q)$ is irreducible. As all diagonal elements are positive, it is also primitive.

2. By Perron–Frobenius theory, the spectral radius is the largest eigenvalue. For $k \geq 1$, set $a_k(q) = (1,\dots,1)M_K(q)^k(1,\dots,1)^T$ and assume that $q_1 < q_2$. As $a_k(q)$ is $q^{kN(K)}$ times a polynomial in q with positive integer coefficients, we have $a_k(q_2) > (q_2/q_1)^{kN(K)}a_k(q_1)$. This implies that $\lim_{k\to\infty} a_k(q_2)/a_k(q_1) = +\infty$.

   On the other hand, $a_k(q_j) \sim c_j\lambda_{\max}(M_K(q_j))^k$ for $j \in \{1,2\}$ and suitable positive constants c_1, c_2. As

$$+\infty = \lim_{k\to\infty}\frac{a_k(q_2)}{a_k(q_1)} = \lim_{k\to\infty}\frac{c_2}{c_1}\left(\frac{\lambda_{\max}(M_K(q_2)))}{\lambda_{\max}(M_K(q_1)))}\right)^k,$$

   we conclude that $\lambda_{\max}(M_K(q_2)) > \lambda_{\max}(M_K(q_1))$.

3. Assume that $M_K(q)x \leq x$ for some positive x. Iterating this equation and multiplying with x^T from the left yields

$$x^T M_K(q)^k x \leq x^T x x$$

   for all $k \geq 1$. As $x^T M_K(q)^k x \sim c\lambda_{\max}(M_K(q))^k$ for some positive constant c and $k \to \infty$, we conclude that $\lambda_{\max}(M_K(q)) \leq 1$.

   The same argument can be used for the case $M_K(q)x \geq x$, too.

□

We consider the infinite matrix

$$M_\infty(q) := \left(q^r\left[\frac{r}{t} \leq s\right]\right)_{\substack{1\leq r\\ 1\leq s}}$$

and the infinite determinant $\det(I - M_\infty(q))$ which is defined to be the limit of the principal minors $\det([r=s] - q^r\left[\frac{r}{t} \leq s\right])_{\substack{1\leq r\leq N\\ 1\leq s\leq N}}$ when N tends to ∞, cf. Eaves [33]. For $|q| < 1$, this infinite determinant converges by Eaves' sufficient condition.

We now show that the infinite determinant is indeed the denominator of the generating function $H(q,1,1,1)$.

Lemma 7.13.3. *We have*

$$\det(I - M_\infty(q)) = 1 - b(q,1,1,1)$$

where $b(q,u,1,1)$ is given in Lemma 7.8.1.

Proof. When expanding the infinite determinant, we take the 1 on the diagonal in almost all rows and some other entry in rows $a_1 < a_2 < \cdots < a_k$ for some k. These other entries have to come from $-M_\infty(q)$. Extracting the sign for these rows, we get

$$\begin{aligned}\det(I - M_\infty(q)) &= \sum_{k\geq 0}(-1)^k \sum_{1\leq a_1<a_2<\cdots<a_k} \det(q^{a_i}[a_i \leq ta_j])_{1\leq i,j\leq k}\\ &= \sum_{k\geq 0}(-1)^k \sum_{1\leq a_1<a_2<\cdots<a_k} q^{a_1+\cdots+a_k}\det([a_i \leq ta_j])_{1\leq i,j\leq k}.\end{aligned}$$

We trivially have $a_i \le ta_j$ for $j \ge i$, so all entries above and on the diagonal of $([a_i \le ta_j])_{1\le i,j\le k}$ are 1. If $a_2 \le ta_1$, the first and the second row of $([a_i \le ta_j])_{1\le i,j\le k}$ are identical, so the determinant vanishes. Therefore, we only have to consider summands with $a_2 > ta_1$. In this case, we clearly have $a_i > ta_1$ for all $i \ge 2$, i.e., the first column of $([a_i \le ta_j])_{1\le i,j\le k}$ is $(1, 0, \dots, 0)^T$. Repeating this argument, we see that only summands with $a_{j+1} > ta_j$ for $1 \le j < k$ contribute to the determinant. In this case, the matrix $([a_i \le ta_j])_{1\le i,j\le k}$ equals $([j \ge i])_{1\le i,j\le k}$ and has determinant 1.

We therefore obtained the representation

$$\det(I - M_\infty(q)) = \sum_{k\ge 0}(-1)^k \sum_{\substack{a_1,\dots,a_k \\ \forall j: a_{j+1} > ta_j}} q^{a_1+\dots+a_k}.$$

With the change of variables $a_1 =: b_k$ and $a_{j+1} - ta_j =: b_{k-j}$ for $1 \le j < k$, we obtain

$$\begin{aligned}\det(I - M_\infty(q)) &= \sum_{k\ge 0}(-1)^k \sum_{b_1,\dots,b_k\ge 1} q^{b_1[\![1]\!]+\dots+b_k[\![k]\!]} \\ &= \sum_{k\ge 0}(-1)^k \prod_{j=1}^{k}\Big(\sum_{b_j\ge 1}(q^{[\![j]\!]})^{b_j}\Big) = 1 - b(q,1,1,1).\end{aligned}$$

□

If K tends to infinity, we do expect $W_K(q)$ to tend to $H(q,1,1,1)$, as the restriction on the width becomes meaningless. We will need a slightly stronger result: we also need convergence of the numerator and the denominator of $W_K(q)$ given by (7.13.1) and Cramer's rule to the numerator $a(q,1,1,1)$ and the denominator $1 - b(q,1,1,1)$ of $H(q,1,1,1)$, respectively. We prove this in two steps: first, we prove that the numerator and the denominator of $W_K(q)$ given by (7.13.1) and Cramer's rule tend to the corresponding infinite determinants.

Lemma 7.13.4. *For $|q| < 1$, we have*

$$\det(I - M_K(q)) = \det(I - M_\infty(q)) + O(q^{K/(2t)}).$$

The same conclusion holds when the s-th column of both $I - M_K(q)$ and $I - M_\infty(q)$ are replaced by the vector $(q, 0, \dots)^T$ with $K-1$ and infinitely many zeroes, respectively. The estimate still holds for derivatives with respect to q.

Proof. The infinite determinant $\det(I - M_\infty(q))$ consists of summands

$$\pm\prod_{s\in S} q^{\pi(s)} = \pm q^{\sum_{s\in S}\pi(s)}$$

where $\pi : \mathbb{N} \to \mathbb{N}$ is a bijection such that there are only finitely many non-fixed points s of π and S is a finite subset of $\mathbb{N}$ containing all non-fixed points of π. Note that the complement of S corresponds to those columns where 1 has been chosen on the

diagonal in the expansion of the determinant. Not all (π, S) will actually occur due to the Iversonian expression in the definition of $M_\infty(q)$.

For every $k \in \mathbb{N}$, there is a bijection from the set

$$\Big\{(\pi, S) \mid \pi : \mathbb{N} \to \mathbb{N} \text{ bijective}, S \subseteq \mathbb{N} \text{ finite such that } \{s \in \mathbb{N} \mid \pi(s) \neq s\} \subseteq S \text{ and } \sum_{s \in S} \pi(s) = k\Big\}$$

to the set

$$\Big\{(x_1, \ldots, x_j) \in \mathbb{N}^j \mid j \in \mathbb{N}, \sum_{i=1}^{j} x_i = k \text{ with pairwise distinct } x_i\Big\}$$

of compositions of k with distinct parts: the set S can be recovered as the set of summands in the composition, the permutation π can be recovered from the order of the summands.

As there are at most $\exp(2\sqrt{k}\log k)$ compositions of k with distinct parts by a result of Richmond and Knopfmacher [87], there are at most that many summands $\pm q^k$ in the infinite determinant $\det(I - M_\infty(q))$.

The difference between $\det(I - M_\infty(q))$ and $\det(I - M_K(q))$ consists of those summands which do not choose the 1 on the diagonal in some row $> N(K)$ or choose some column s in some row r with $s > (r + K)/t$. In the latter case, the 1 on the diagonal cannot be chosen in row s, so that the exponent of q in this summand is at least $r + s > K/t$. So all summands in the difference are of the form $\pm q^k$ for some $k \geq K/t$. By the triangle inequality and the above estimates, we obtain

$$|\det(I - M_\infty(q)) - \det(I - M_K(q))| \leq \sum_{k \geq K/t} \exp(2\sqrt{k}\log k) q^k = O(q^{K/(2t)}).$$

The argument does not change if the s-th column of both matrices is replaced by the column vector $(q, 0, \ldots, 0)^T$.

Differentiating the determinant can be done term by term. The error term does not change as the bound $O(q^{K/(2t)})$ is weak enough. □

The second step in the proof of the convergence of numerator and denominator of $W_K(q)$ consists of the following simple lemma.

Lemma 7.13.5. *The denominator* $\det(I - M_K(q))$ *of* $W_K(q)$ *converges to* $1 - b(q, 1, 1, 1)$ *with error* $O(q^{K/(2t)})$. *The numerator* $\det(I - M_K(q))W_K(q)$ *of* $W_K(q)$ *converges to* $a(q, 1, 1, 1)$ *with the same error. The same is true for derivatives with respect to* q.

Proof. The first statement is simply the combination of Lemmata 7.13.4 and 7.13.3.

As a formal power series, $W_K(q)$ converges to $H(q, 1, 1, 1)$ as $[q^n]W_K(q) = [q^n]H(q, 1, 1, 1)$ holds for $n \leq (K - 1)/(t - 1)$, as a tree with n internal states has at most $1 + n(t - 1)$ leaves and therefore width at most $1 + n(t - 1)$.

As $1 - b(q,1,1,1)$ has no root with $|q| < 1/2$ by Lemma 7.2.2, $W_K(q)$ converges to $H(q,1,1,1)$ for $|q| < 1/2$. As the denominator is already known to converge to the denominator $1 - b(q,1,1,1)$ of $H(q,1,1,1)$, we conclude that the numerators (which are already known to converge to some infinite determinant) actually have to converge to $a(q,1,1,1)$.

Taking derivatives with respect to q does not change the argument by Lemma 7.13.4. □

In order to obtain information on the roots of $\det(I - M_K(q))$ and therefore the singularities of $\mathbf{W}_K(q)$, we approximate the Perron–Frobenius eigenvector of $M_K(q)$ by the eigenvector of the infinite matrix $M_\infty(q)$. In the following lemma it turns out that we actually met this infinite eigenvector earlier.

Lemma 7.13.6. *For $r \geq 1$, we have*

$$q^r \Big(1 - \sum_{j=1}^{\lceil r/t \rceil - 1} [u^{jt}] b(q,u,1,1)\Big) = [u^{rt}] b(q,u,1,1). \tag{7.13.3}$$

In particular, $(p_r)_{r\geq 1}$ as defined in Theorem 7.5.1 is a right eigenvector of $M_\infty(q_0)$ to the eigenvalue 1, i.e.,

$$M_\infty(q_0) \cdot (p_r)_{r\geq 1} = (p_r)_{r\geq 1}. \tag{7.13.4}$$

Proof. Multiplying the left hand side of (7.13.3) with u^{rt} and summing over $r \geq 1$ yields

$$\begin{aligned}
\frac{qu^t}{1-qu^t} - \sum_{\substack{r\geq 1\\ j\geq 1\\ jt<r}} (qu^t)^r [u^{jt}] b(q,u,1,1) &= \frac{qu^t}{1-qu^t} - \sum_{j=1}^{\infty} [u^{jt}] b(q,u,1,1) \sum_{r=jt+1}^{\infty} (qu^t)^r \\
&= \frac{qu^t}{1-qu^t} - \frac{qu^t}{1-qu^t} \sum_{j=1}^{\infty} (qu^t)^{jt} [u^{jt}] b(q,u,1,1) \\
&= \frac{qu^t}{1-qu^t} (1 - b(q,qu^t,1,1)) = b(q,u,1,1),
\end{aligned}$$

which concludes the proof of (7.13.3).

Setting $q = q_0$ in (7.13.3) and noting that $1 = b(q_0,1,1,1) = \sum_{r\geq 1} p_r$ yields (7.13.4). □

We now use the fact that $(p_r)_{r\geq 1}$ is an eigenvector of $M_\infty(q)$ to derive bounds for its entries.

Proposition 7.13.7. *All p_r, $r \geq 1$, are positive and we have*

$$\frac{1}{r} q_*^r \ll_t p_r \ll_t r^2 q_*^r.$$

with

$$q_* = q_0^{1+\frac{1}{t-1}}.$$

Proof. By Theorem 7.5.1, the p_r are limits of probabilities and therefore non-negative.

By the eigenvalue equation (7.13.4), we have

$$p_r \geq q_0^r p_{\lceil r/t \rceil}$$

for all $r \geq 1$. Iterating this, we get

$$\begin{aligned} p_r &\geq q_0^{\sum_{j=0}^{\lfloor \log_t r \rfloor - 1} \lceil r/t^j \rceil} p_{\lceil r/t^{\lfloor \log_t r \rfloor} \rceil} \gg_t q_0^{\sum_{j=0}^{\lfloor \log_t r \rfloor - 1} (1 + r/t^j)} \\ &\geq q_0^{\log_t r + \sum_{j=0}^{\infty} r/t^j} = r^{\log_t q_0} q_0^{r(1+1/(t-1))}. \end{aligned}$$

As $q_0 \geq 1/t$ by Lemma 7.2.2, we have $\log_t q_0 \geq -1$ and the lower bound follows.

To prove the upper bound, we proceed in two steps. In a first step, we note that the eigenvalue equation (7.13.4) together with the fact that $\sum_{r \geq 1} p_r = 1$ yields the weaker upper bound

$$p_r = q_0^r \sum_{s \geq \lceil r/t \rceil} p_s \leq q_0^r \sum_{s \geq 1} p_s = q_0^r.$$

In a second step, we use induction on r and assume that $p_s \leq c s^2 q_*^s$ for $s < r$ for some constant c depending on t. Then the eigenvalue equation (7.13.4) yields

$$\begin{aligned} p_r &\leq q_0^r \sum_{s \geq \lceil r/t \rceil} p_s \leq c q_0^r \sum_{\lceil r/t \rceil \leq s < r} s^2 q_*^s + q_0^r \sum_{r \leq s} q_0^s \leq c q_0^r \sum_{\lceil r/t \rceil \leq s} s^2 q_*^s + \frac{1}{1-q_0} q_0^{2r} \\ &= c q_0^r \left(\frac{\lceil r/t \rceil^2}{1-q_*} + \frac{2 q_* \lceil r/t \rceil}{(1-q_*)^2} + \frac{q_*(1+q_*)}{(1-q_*)^3} \right) q_*^{\lceil r/t \rceil} + \frac{1}{1-q_0} q_0^{2r} \\ &\leq c q_0^r \left(\frac{(r+t)^2}{t^2(1-q_*)} + \frac{2 q_* (r+t)}{t(1-q_*)^2} + \frac{q_*(1+q_*)}{(1-q_*)^3} \right) q_*^{r/t} + \frac{1}{1-q_0} q_0^{2r}. \end{aligned}$$

As $t^2(1-q_*) > 1$ for $t \geq 2$ (cf. Lemma 7.2.2), we obtain

$$p_r \leq c r^2 q_0^r q_*^{r/t} = c r^2 q_0^{r\left(1 + \frac{1}{t}\left(1 + \frac{1}{t-1}\right)\right)} = c r^2 q_*^r$$

for sufficiently large r.

□

Lemma 7.13.8. *The generating function $W_K(q)$ has a unique singularity q_K with $|q_K| \leq 0.6$ for $K \geq c_1$ for a suitable positive constant c_1 depending on t. It is a simple pole and a zero of* $\det(I - M_K(q))$. *Furthermore*

$$q_0 + c_2 \frac{1}{K} q_0^{K/(t-1)} \leq q_K \leq q_0 + c_3 K^2 q_0^{K/(t-1)}$$

for suitable positive constants c_2, c_3 depending on t.

Proof. In the following, c_4, c_5, ... denote suitable constants depending on t.

As $H(q,1,1,1)$ has a unique pole q with $|q| \le 0.6$ by Lemma 7.2.2 and numerator and denominator of $W_K(q)$ tend to the numerator and denominator of $H(q,1,1,1)$ respectively by Lemma 7.13.5, $W_K(q)$ also has a a unique pole with $|q| \le 0.6$ for sufficiently large K.

We set $x_K = (p_1, \dots, p_{N(K)})^T$. If we find a q such that $M_K(q)x_K \ge x_K$, then Lemma 7.13.2 implies that $\lambda_{\max}(M_K(q)) \ge 1$ and $q_K < q$.

We therefore consider the r-th row of $M_K(q)x_K$ for some $1 \le r \le N(K)$. We have

$$\begin{aligned}(M_K(q)x_K)_r &= q^r \sum_{\frac{r}{t} \le s \le \frac{r+K}{t}} p_s \ge q^r \sum_{\frac{r}{t} \le s < \frac{r+K}{t}} p_s = q^r \left(\frac{p_r}{q_0^r} - \frac{p_{r+K}}{q_0^{r+K}} \right) \\ &= p_r \left(\frac{q}{q_0} \right)^r \left(1 - \frac{p_{r+K}}{p_r q_0^K} \right)\end{aligned}$$

by the eigenvalue equation (7.13.4). By Proposition 7.13.7, we have

$$\frac{p_{r+K}}{p_r q_0^K} \le c_4 r (r+K)^2 \frac{q_*^{r+K}}{q_*^r q_0^K} = c_4 r (r+K)^2 q_0^{K/(t-1)} \le c_5 K^3 q_0^{K/(t-1)}.$$

Therefore, we have

$$\sqrt[r]{1 - \frac{p_{r+K}}{p_r q_0^K}} = \frac{1}{\left(1 - \frac{p_{r+K}}{p_r q_0^K}\right)^{-1/r}} \ge \frac{1}{1 + \frac{2p_{r+K}}{r p_r q_0^K}} \ge \frac{1}{1 + c_6 K^2 q_0^{K/(t-1)}}.$$

This means that for $q = q_0 + c_7 K^2 q_0^{K/(t-1)}$, we have $M_K(q)x_K \ge x_K$, as requested.

The proof of the lower bound runs along the same lines.

□

Proof of Theorem 7.7.1. By singularity analysis, we have

$$\mathbb{P}(w(T) \le K) = (1 + O(0.6^{K/2t})) \left(\frac{q_K}{q_0} \right)^{-n-1} (1 + O(0.99^n))$$

for $K \ge c_8$.

We now estimate

$$\mathbb{E}(w(T)) = \sum_{K \ge 0} (1 - \mathbb{P}(w(T) \le K)). \tag{7.13.5}$$

We use the abbreviation $S := 1/q_0^{t-1} > 1$.

First, we consider the summands of (7.13.5) with $S^K \le n/\log^2 n$. By Lemma 7.13.8, we have

$$\left(\frac{q_K}{q_0} \right)^n \ge \left(1 + c_9 \frac{1}{S^K \log_S n} \right)^n \ge \left(1 + c_{10} \frac{\log n}{n} \right)^n \ge c_{10} \log n.$$

We conclude that these summands of (7.13.5) contribute $\log_S n + O(\log\log n)$. In particular, the above estimates imply that

$$\mathbb{P}(w(T) - \log_S n \le -2\log_S \log n) = O(1/\log n). \tag{7.13.6}$$

Now, we consider the summands of (7.13.5) with $n/\log^2 n < S^K \le n\log^3 n$. These are $O(\log\log n)$ summands with each trivially contributing at most 1, so the total contribution is $O(\log\log n)$.

Next, we consider the summands of (7.13.5) with $n\log^3 n < S^K \le n^{4t\log S}$. We now have

$$\frac{q_k}{q_0} \le 1 + c_{11}\frac{\log^2 n}{S^K} \le 1 + c_{11}\frac{1}{n\log n}$$

and therefore

$$\mathbb{P}(w(T) \le K) \ge (1 + O(n^{-|\log_S 0.6|/(2t)}))\exp\left(-n\log\left(\frac{q_k}{q_0}\right)\right) \ge 1 - c_{12}\frac{1}{\log n}.$$

The total contribution of these summands is therefore $O(1)$. In particular, the above estimates imply that

$$\mathbb{P}(w(T) - \log_S n \ge 3\log_S \log n) = O(1/\log n). \tag{7.13.7}$$

Next, we consider the summands of (7.13.5) with $n^{4t\log S} < S^K \le S^{tn}$. This time, we have

$$\frac{q_k}{q_0} \le 1 + c_{13}\frac{n^2}{n^4}$$

and therefore

$$\mathbb{P}(w(T) \le K) = (1 + O(n^{-2|\log 0.6|}))\exp\left(-n\log\left(\frac{q_k}{q_0}\right)\right) \ge 1 - c_{14}\frac{1}{n}.$$

The total contribution of these summands is therefore $O(1)$.

Finally, we note that all summands with $K > tn$ vanish: any tree with n internal nodes has at most width tn.

Collecting all terms, we obtain

$$\mathbb{E}(w(T)) = \log_S n + O(\log\log n) = \frac{\log n}{-(t-1)\log q_0} + O(\log\log n).$$

Combining (7.13.6) and (7.13.7) immediately yields the concentration property (7.7.1). □

Bibliography

[1] David W. Ash, Ian F. Blake, and Scott A. Vanstone, *Low complexity normal bases*, Discrete Appl. Math. **25** (1989), no. 3, 191–210.
(Cited on page 2.)

[2] Roberto Avanzi, Clemens Heuberger, and Helmut Prodinger, *Minimality of the Hamming weight of the τ-NAF for Koblitz curves and improved combination with point halving*, Selected Areas in Cryptography: 12th International Workshop, SAC 2005, Kingston, ON, Canada, August 11–12, 2005, Revised Selected Papers (B. Preneel and S. Tavares, eds.), Lecture Notes in Comput. Sci., vol. 3897, Springer, Berlin, 2006, pp. 332–344.
(Cited on pages 5, 12, and 6.)

[3] ———, *Scalar multiplication on Koblitz curves. Using the Frobenius endomorphism and its combination with point halving: Extensions and mathematical analysis*, Algorithmica **46** (2006), 249–270.
(Cited on pages 5, 12, and 6.)

[4] ———, *Arithmetic of supersingular Koblitz curves in characteristic three*, Tech. Report 2010-8, Graz University of Technology, 2010, `http://www.math.tugraz.at/fosp/pdfs/tugraz_0166.pdf`, also available as Cryptology ePrint Archive, Report 2010/436, `http://eprint.iacr.org/`.
(Cited on pages 3, 17, 27, 28, 11, and 21.)

[5] Roberto Maria Avanzi, *A Note on the Signed Sliding Window Integer Recoding and a Left-to-Right Analogue*, Selected Areas in Cryptography: 11th International Workshop, SAC 2004, Waterloo, Canada, August 9-10, 2004, Revised Selected Papers (H. Handschuh and A. Hasan, eds.), Lecture Notes in Comput. Sci., vol. 3357, Springer-Verlag, Berlin, 2004, pp. 130–143.
(Cited on pages 5, 12, 16, 6, and 10.)

[6] Michael Barnsley, *Fractals everywhere*, Academic Press, Inc, 1988.
(Cited on pages 61, 55, and 56.)

[7] Fabrizio Barroero, Christopher Frei, and Robert Tichy, *Additive unit representations in global fields - a survey*, Publ. Math. Debrecen **79** (2011), no. 3-4, 291–307.
(Cited on pages 86 and 80.)

[8] Paul Belcher, *Integers expressible as sums of distinct units*, Bull. Lond. Math. Soc. **6** (1974), 66–68.
(Cited on pages 85 and 79.)

[9] ———, *A test for integers being sums of distinct units applied to cubic fields*, J. Lond. Math. Soc., II. Ser. **12** (1976), 141–148.
(Cited on pages 85 and 79.)

[10] Edward A. Bender and E. Rodney Canfield, *Locally restricted compositions. I. Restricted adjacent differences*, Electron. J. Combin. **12** (2005), Research Paper 57, 27 pp.
(Cited on page 114.)

[11] ———, *Locally restricted compositions. II. General restrictions and infinite matrices*, Electron. J. Combin. **16** (2009), no. 1, Research Paper 108, 35 pp.
(Cited on page 114.)

[12] ———, *Locally restricted compositions III. Adjacent-part periodic inequalities*, Electron. J. Combin. **17** (2010), no. 1, Research Paper 145, 9 pp.
(Cited on page 114.)

[13] Valérie Berthé and Laurent Imbert, *Diophantine approximation, Ostrowski numeration and the double-base number system*, Discrete Mathematics and Theoretical Computer Science **11:1** (2009), 153–172.
(Cited on pages 99 and 93.)

[14] Manjul Bhargava, *P-orderings and polynomial functions on arbitrary subsets of Dedekind rings*, J. Reine Angew. Math. **490** (1997), 101–127.
(Cited on pages 102 and 96.)

[15] Bryan J. Birch, *Note on a problem of Erdős*, Proc. Cambridge Philos. Soc. **55** (1959), 370–373.
(Cited on pages 95 and 89.)

[16] Ian F. Blake, V. Kumar Murty, and Guangwu Xu, *Efficient algorithms for Koblitz curves over fields of characteristic three*, J. Discrete Algorithms **3** (2005), no. 1, 113–124.
(Cited on pages 3, 19, 40, 13, and 34.)

[17] ———, *A note on window τ-NAF algorithm*, Inform. Process. Lett. **95** (2005), 496–502.
(Cited on pages 3, 40, and 34.)

[18] ———, *Nonadjacent radix-τ expansions of integers in Euclidean imaginary quadratic number fields*, Canad. J. Math. **60** (2008), no. 6, 1267–1282.
(Cited on pages 3, 40, and 34.)

[19] Ian F. Blake, Gadiel Seroussi, and Nigel P. Smart, *Elliptic curves in cryptography*, London Mathematical Society Lecture Note Series, vol. 265, Cambridge University Press, 1999.
(Cited on page 2.)

[20] Joel V. Brawley and Gary L. Mullen, *Functions and polynomials over Galois rings*, J. Number Theory **41** (1992), no. 2, 156–166.
(Cited on pages 102 and 96.)

[21] Leonard Carlitz, *Functions and polynomials* (mod p^n), Acta Arith. **9** (1964), 67–78.
(Cited on pages 102 and 96.)

[22] Zhibo Chen, *On polynomial functions from* $\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_r}$ *to* $\mathbb{Z}_m$, Discrete Math. **162** (1996), no. 1-3, 67–76.
(Cited on pages 102 and 96.)

[23] Mathieu Ciet, Tanja Lange, Francesco Sica, and Jean-Jacques Quisquater, *Improved algorithms for efficient arithmetic on elliptic curves using fast endomorphisms*, Advances in cryptology — EUROCRYPT 2003. International conference on the theory and applications of cryptographic techniques, Warsaw, Poland, May 4–8, 2003. Proceedings (E. Biham, ed.), Lecture Notes in Comput. Sci., vol. 2656, Springer, Berlin, 2003, pp. 388–400.
(Cited on page 4.)

[24] Hubert Delange, *Sur la fonction sommatoire de la fonction "somme des chiffres"*, Enseignement Math. (2) **21** (1975), 31–47.
(Cited on pages 6, 71, and 65.)

[25] Pierre Deligne, *La conjecture de Weil. I*, Inst. Hautes Études Sci. Publ. Math. (1974), no. 43, 273–307.
(Cited on pages 3, 48, 82, 42, and 76.)

[26] Whitfield Diffie and Martin E. Hellman, *New directions in cryptography*, IEEE Transactions on Information Theory **22** (1976), no. 6, 644–654.
(Cited on page 1.)

[27] Vassil S. Dimitrov and Everett W. Howe, *Lower bounds on the lengths of double-base representations*, Proc. Amer. Math. Soc. **139** (2011), no. 10, 3423–3430.
(Cited on pages 99 and 93.)

[28] Michael Drmota, *The height of increasing trees*, Ann. Comb. **12** (2009), no. 4, 373–402.
(Cited on page 113.)

[29] ———, *Random trees*, SpringerWienNewYork, Vienna, 2009.
(Cited on page 113.)

[30] Michael Drmota and Bernhard Gittenberger, *On the profile of random trees*, Random Structures Algorithms **10** (1997), no. 4, 421–451.
(Cited on page 113.)

[31] Michael Drmota and Hsien-Kuei Hwang, *Profiles of random trees: correlation and width of random recursive trees and binary search trees*, Adv. in Appl. Probab. **37** (2005), no. 2, 321–341.
(Cited on page 113.)

[32] Bernard Dwork, *On the rationality of the zeta function of an algebraic variety*, Amer. J. Math. **82** (1960), 631–648.
(Cited on pages 3, 48, 82, 42, and 76.)

[33] Reuben E. Eaves, *A sufficient condition for the convergence of an infinite determinant.*, SIAM J. Appl. Math. **18** (1970), 652–657.
(Cited on page 138.)

[34] Gerald A. Edgar, *Measure, topology, and fractal geometry*, second ed., Undergraduate Texts in Mathematics, Springer-Verlag, New York, 2008.
(Cited on pages 52, 61, 64, 47, 55, 56, 58, and 59.)

[35] Christian Elsholtz, Clemens Heuberger, and Helmut Prodinger, *The number of Huffman codes, compact trees, and sums of unit fractions*, to appear in IEEE Trans. Inf. Theory, earlier version available at arXiv:1108.5964v1 [math.CO].
(Cited on pages 111, 114, 115, and 124.)

[36] Philippe Flajolet, Zhicheng Gao, Andrew Odlyzko, and Bruce Richmond, *The distribution of heights of binary trees and other simple trees*, Combin. Probab. Comput. **2** (1993), no. 2, 145–156.
(Cited on page 113.)

[37] Philippe Flajolet and Helmut Prodinger, *Level number sequences for trees*, Discrete Math. **65** (1987), no. 2, 149–156.
(Cited on pages 111 and 114.)

[38] Philippe Flajolet and Robert Sedgewick, *Analytic combinatorics*, Cambridge University Press, Cambridge, 2009.
(Cited on pages 116, 117, 123, 125, and 127.)

[39] Sophie Frisch, *When are weak permutation polynomials strong?*, Finite Fields Appl. **1** (1995), no. 4, 437–439.
(Cited on pages 102 and 96.)

[40] ———, *Polynomial functions on finite commutative rings*, Advances in commutative ring theory (Fez, 1997), Lecture Notes in Pure and Appl. Math., vol. 205, Dekker, New York, 1999, pp. 323–336.
(Cited on pages 102, 108, and 96.)

[41] Sophie Frisch and Daniel Krenn, *Sylow p-groups of polynomial permutations on the integers mod p^n*, arXiv:1112.1228v1 [math.NT], 2011.
(Cited on pages 101 and 95.)

[42] László Germán and Attila Kovács, *On number system constructions*, Acta Math. Hungar. **115** (2007), no. 1-2, 155–167.
(Cited on pages 4, 10, 46, and 40.)

[43] Daniel M. Gordon, *A survey of fast exponentiation methods*, J. Algorithms **27** (1998), 129–146.
(Cited on pages 5, 12, 16, 6, and 10.)

[44] Peter J. Grabner, Clemens Heuberger, and Helmut Prodinger, *Distribution results for low-weight binary representations for pairs of integers*, Theoret. Comput. Sci. **319** (2004), 307–331.
(Cited on page 6.)

[45] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik, *Concrete mathematics. A foundation for computer science*, second ed., Addison-Wesley, 1994.
(Cited on pages 56, 69, 136, 51, and 64.)

[46] Christian Günther, Tanja Lange, and Andreas Stein, *Speeding up the arithmetic on Koblitz curves of genus two*, Selected areas in cryptography (Waterloo, ON, 2000), Lecture Notes in Comput. Sci., vol. 2012, Springer, Berlin, 2001, pp. 106–117.
(Cited on page 4.)

[47] Clemens Heuberger, *Redundant τ-adic expansions II: Non-optimality and chaotic behaviour*, Math. Comput. Sci. **3** (2010), 141–157.
(Cited on pages 5, 12, and 6.)

[48] Clemens Heuberger and Daniel Krenn, *Optimality of the width-w non-adjacent form: General characterisation and the case of imaginary quadratic bases*, arXiv:1110.0966v1 [math.NT], 2011.
(Cited on pages 9, 49, 3, and 43.)

[49] ———, *Analysis of width-w non-adjacent forms to imaginary quadratic bases*, to appear in J. Number Theory (2012), earlier version available at arXiv:1009.0488v2 [math.NT].
(Cited on pages 4, 6, 17, 19, 20, 40, 52, 55, 59, 60, 61, 62, 66, 67, 69, 83, and 84.)

[50] Clemens Heuberger and Daniel Krenn, *Existence and optimality of w-non-adjacent forms with an algebraic integer base*, to appear in Acta Math. Hungar. (2012), earlier version available at arXiv:1205.4414v1 [math.NT].
(Cited on pages 39, 55, 82, and 83.)

[51] Clemens Heuberger, Daniel Krenn, and Stephan Wagner, *Analysis of parameters of trees corresponding to Huffman codes and sums of unit fractions*, 2013 Proceedings of the Tenth Workshop on Analytic Algorithmics and Combinatorics, to appear.
(Cited on page 109.)

[52] Clemens Heuberger and Helmut Prodinger, *Analysis of alternative digit sets for nonadjacent representations*, Monatsh. Math. **147** (2006), 219–248.
(Cited on pages 5, 6, 12, 62, and 56.)

[53] Thomas W. Hungerford, *Algebra*, Graduate Texts in Mathematics, vol. 73, Springer, 1996.
(Cited on page 15.)

[54] Bernard Jacobson, *Sums of distinct divisors and sums of distinct units*, Proc. Am. Math. Soc. **15** (1964), 179–183.
(Cited on pages 85 and 79.)

[55] Moshe Jarden and Władysław Narkiewicz, *On sums of units*, Monatsh. Math. **150** (2007), no. 4, 327–336.
(Cited on pages 86 and 80.)

[56] Jonathan Jedwab and Chris J. Mitchell, *Minimum weight modified signed-digit representations and fast exponentiation*, Electron. Lett. **25** (1989), 1171–1172.
(Cited on pages 5, 12, 16, 6, and 10.)

[57] Jian Jun Jiang, Guo Hua Peng, Qi Sun, and Qifan Zhang, *On polynomial functions over finite commutative rings*, Acta Math. Sin. (Engl. Ser.) **22** (2006), no. 4, 1047–1050.
(Cited on pages 102 and 96.)

[58] Jianjun Jiang, *A note on polynomial functions over finite commutative rings*, Adv. Math. (China) **39** (2010), no. 5, 555–560.
(Cited on pages 102 and 96.)

[59] Don Johnson, Alfred Menezes, and Scott Vanstone, *The elliptic curve digital signature algorithm (ecdsa)*, Tech. report, Certicom, available at `http://www.certicom.com/pdfs/whitepapers/ecdsa.pdf`.
(Cited on page 1.)

[60] Gordon Keller and Frank R. Olson, *Counting polynomial functions* (mod p^n), Duke Math. J. **35** (1968), 835–838.
(Cited on pages 102 and 96.)

[61] Aubrey J. Kempner, *Polynomials and their residue systems*, Trans. Amer. Math. Soc. **22** (1921), 240–266, 267–288.
(Cited on pages 101, 102, 95, and 96.)

[62] Donald E. Knuth, *Seminumerical algorithms*, third ed., The Art of Computer Programming, vol. 2, Addison-Wesley, 1998.
(Cited on pages 10 and 4.)

[63] Neal Koblitz, *CM-curves with good cryptographic properties*, Advances in cryptology—CRYPTO '91 (Santa Barbara, CA, 1991) (J. Feigenbaum, ed.), Lecture Notes in Comput. Sci., vol. 576, Springer, Berlin, 1992, pp. 279–287.
(Cited on pages 3 and 5.)

[64] ———, *A course in number theory and cryptography*, 2nd ed., Graduate Texts in Mathematics, Springer, 1994.
(Cited on page 3.)

[65] ———, *An elliptic curve implementation of the finite field digital signature algorithm*, Advances in cryptology—CRYPTO '98 (Santa Barbara, CA, 1998), Lecture Notes in Comput. Sci., vol. 1462, Springer, Berlin, 1998, pp. 327–337.
(Cited on page 3.)

[66] Béla Kovács and Attila Pethő, *Number systems in integral domains, especially in orders of algebraic number fields*, Acta Sci. Math. (Szeged) **55** (1991), 287–299.
(Cited on pages 43 and 37.)

[67] David W. Kravitz, *Digital signature algorithm*, 1991, U.S. Patent 5,231,668.
(Cited on page 1.)

[68] Daniel Krenn, *Analysis of the width-w non-adjacent form in conjunction with hyperelliptic curve cryptography and with lattices*, arXiv:1209.0618v1 [math.NT], 2012.
(Cited on pages 51 and 45.)

[69] Daniel Krenn, Jörg Thuswaldner, and Volker Ziegler, *On linear combinations of units with bounded coefficients and double-base digit expansions*, to appear in Monatsh. Math., arXiv:1205.4833v1 [math.NT], 2012.
(Cited on pages 85 and 79.)

[70] Markus Kröll, *Optimality of digital expansions to the base of the Frobenius endomorphism on Koblitz curves in characteristic three*, Tech. Report 2010-09, Graz University of Technology, 2010, available at `http://www.math.tugraz.at/fosp/pdfs/tugraz_0167.pdf`.
(Cited on page 5.)

[71] T. Lange, *Koblitz curve cryptosystems*, Finite Fields Appl. **11** (2005), 200–229.
(Cited on pages 3 and 4.)

[72] Nian Ping Liu and Jian Jun Jiang, *Polynomial functions in n variables over a finite commutative ring*, Sichuan Daxue Xuebao **46** (2009), no. 1, 44–46.
(Cited on pages 102 and 96.)

[73] M. Lothaire, *Algebraic combinatorics on words*, Encyclopedia of Mathematics and its Applications, vol. 90, Cambridge University Press, Cambridge, 2002.
(Cited on pages 11 and 12.)

[74] Hosam M. Mahmoud, *Limiting distributions for path lengths in recursive trees*, Probab. Engrg. Inform. Sci. **5** (1991), no. 1, 53–59.
(Cited on page 113.)

[75] Bernard R. McDonald, *Finite rings with identity*, Dekker, 1974.
(Cited on pages 104 and 98.)

[76] Willi Meier and Othmar Staffelbach, *Efficient multiplication on certain nonsupersingular elliptic curves*, Advances in cryptology—CRYPTO '92 (Santa Barbara, CA, 1992) (Ernest F. Brickell, ed.), Lecture Notes in Comput. Sci., vol. 740, Springer, Berlin, 1993, pp. 333–344.
(Cited on page 5.)

[77] Atsuko Miyaji, Takatoshi Ono, and Henri Cohen, *Efficient elliptic curve exponentiation*, Information and communications security. 1st international conference, ICICS '97, Beijing, China, November 11–14, 1997. Proceedings (Yongfei Han, Tatsuaki Okamoto, and Sihan Qing, eds.), Lecture Notes in Comput. Sci., vol. 1334, Springer-Verlag, 1997, pp. 282–290.
(Cited on page 2.)

[78] James A. Muir and Douglas R. Stinson, *New minimal weight representations for left-to-right window methods*, Topics in Cryptology — CT-RSA 2005 The Cryptographers' Track at the RSA Conference 2005, San Francisco, CA, USA, February 14–18, 2005, Proceedings (A. J. Menezes, ed.), Lecture Notes in Comput. Sci., vol. 3376, Springer, Berlin, 2005, pp. 366–384.
(Cited on pages 5, 12, and 6.)

[79] ———, *Minimality and other properties of the width-w nonadjacent form*, Math. Comp. **75** (2006), 369–384.
(Cited on pages 3, 5, 12, 16, 6, and 10.)

[80] Volker Müller, *Fast multiplication on elliptic curves over small fields of characteristic two*, J. Cryptology **11** (1998), no. 4, 219–234.
(Cited on page 4.)

[81] Władysław Narkiewicz, *Elementary and Analytic Theory of Algebraic Numbers*, Monografie matematyczne, no. 54, PWN - Polish Scientific Publishers, Warsaw, 1974.
(Cited on pages 86 and 80.)

[82] Alexandr A. Nechaev, *Polynomial transformations of finite commutative local rings of principal ideals*, Math. Notes **27** (1980), 425–432, transl. from Mat. Zametki 27 (1980) 885-897, 989.
(Cited on pages 102 and 96.)

[83] Wilfried Nöbauer, *Gruppen von Restpolynomidealrestklassen nach Primzahlpotenzen*, Monatsh. Math. **59** (1955), 194–202.
(Cited on pages 102 and 96.)

[84] ———, *Polynomfunktionen auf primen Restklassen*, Arch. Math. (Basel) **39** (1982), no. 5, 431–435.
(Cited on pages 101 and 95.)

[85] Braden Phillips and Neil Burgess, *Minimal weight digit set conversions*, IEEE Trans. Comput. **53** (2004), 666–677.
(Cited on pages 5, 12, 16, 6, and 10.)

[86] George W. Reitwiesner, *Binary arithmetic*, Advances in computers, vol. 1, Academic Press, New York, 1960, pp. 231–308.
(Cited on pages 2, 3, 5, 12, 16, 6, and 10.)

[87] Bruce Richmond and Arnold Knopfmacher, *Compositions with distinct parts*, Aequationes Math. **49** (1995), no. 1–2, 86–97.
(Cited on page 140.)

[88] Ronald L. Rivest, Adi Shamir, and Leonard M. Adleman, *Cryptographic communications system and method*, 1977, U.S. Patent 4,405,829.
(Cited on page 1.)

[89] Ivo G. Rosenberg, *Polynomial functions over finite rings*, Glas. Mat. **10** (1975), no. 30, 25–33.
(Cited on pages 102 and 96.)

[90] Daniel Shanks, *The simplest cubic fields*, Math. Comp. **28** (1974), 1137–1152.
(Cited on pages 94 and 88.)

[91] Joseph H. Silverman, *The arithmetic of elliptic curves*, Graduate Texts in Mathematics, vol. 106, Springer, New York, 1992.
(Cited on page 3.)

[92] Jan Śliwa, *Sums of distinct units*, Bull. Acad. Pol. Sci. **22** (1974), 11–13.
(Cited on pages 85 and 79.)

[93] Nigel P. Smart, *Elliptic curve cryptosystems over small fields of odd characteristic*, J. Cryptology **12** (1999), no. 2, 141–151.
(Cited on pages 3 and 4.)

[94] Jerome A. Solinas, *An improved algorithm for arithmetic on a family of elliptic curves*, Advances in Cryptology — CRYPTO '97. 17th annual international cryptology conference. Santa Barbara, CA, USA. August 17–21, 1997. Proceedings (B. S. Kaliski, jun., ed.), Lecture Notes in Comput. Sci., vol. 1294, Springer, Berlin, 1997, pp. 357–371.
(Cited on pages 2, 3, 5, 12, 19, 40, 45, 6, 13, 34, and 39.)

[95] ———, *Efficient arithmetic on Koblitz curves*, Des. Codes Cryptogr. **19** (2000), 195–249.
(Cited on pages 2, 3, 5, 12, 16, 19, 40, 45, 6, 10, 13, 34, and 39.)

[96] William A. Stein et al., *Sage Mathematics Software (Version 4.7.1)*, The Sage Development Team, 2011, `http://www.sagemath.org`.
(Cited on pages 38 and 30.)

[97] ———, *Sage Mathematics Software (Version 4.8)*, The Sage Development Team, 2012, `http://www.sagemath.org`.
(Cited on pages 97 and 91.)

[98] ———, *Sage Mathematics Software (Version 5.2)*, The Sage Development Team, 2012, `http://www.sagemath.org`.
(Cited on pages 114, 124, 125, and 127.)

[99] Jonas Kazys Sunklodas, *The rate of convergence in the central limit theorem for strongly mixing random variables*, Litovsk. Mat. Sb. **24** (1984), no. 2, 174–185.
(Cited on pages 120 and 134.)

[100] Lajos Takács, *On the total heights of random rooted trees*, J. Appl. Probab. **29** (1992), no. 3, 543–556.
(Cited on page 113.)

[101] ———, *Limit distributions for queues and random rooted trees*, J. Appl. Math. Stochastic Anal. **6** (1993), no. 3, 189–216.
(Cited on page 113.)

[102] Jörg Thuswaldner and Volker Ziegler, *On linear combinations of units with bounded coefficients*, Mathematika **57** (2011), no. 2, 247–262.
(Cited on pages 86, 94, 80, and 88.)

[103] Christiaan van de Woestijne, *The structure of Abelian groups supporting a number system (extended abstract)*, Actes des rencontres du CIRM **1** (2009), no. 1, 75–79.
(Cited on pages 10 and 4.)

[104] Andrew Vince, *Replicating tessellations*, SIAM J. Discrete Math. **6** (1993), no. 3, 501–521.
(Cited on pages 42, 43, 36, and 37.)

[105] Qi Jiao Wei and Qi Fan Zhang, *On permutation polynomials in two variables over* $\mathbb{Z}/p^2\mathbb{Z}$, Acta Math. Sin. (Engl. Ser.) **25** (2009), no. 7, 1191–1200.
(Cited on pages 102 and 96.)

[106] Qijiao Wei and Qifan Zhang, *On strong orthogonal systems and weak permutation polynomials over finite commutative rings*, Finite Fields Appl. **13** (2007), no. 1, 113–120.
(Cited on pages 102 and 96.)

[107] André Weil, *Variétés abéliennes et courbes algébriques*, Actualités scientifiques et industrielles, no. 1064, Hermann & Cie, 1948.
(Cited on pages 3, 48, 82, 42, and 76.)

[108] ———, *Numbers of solutions of equations in finite fields*, Bull. Amer. Math. Soc. **55** (1949), 497–508.
(Cited on pages 3, 48, 82, 42, and 76.)

[109] ———, *Courbes algébriques et variétés abéliennes*, Hermann, 1971.
(Cited on pages 3, 48, 82, 42, and 76.)

[110] Qifan Zhang, *Polynomial functions and permutation polynomials over some finite commutative rings*, J. Number Theory **105** (2004), no. 1, 192–202.
(Cited on pages 102 and 96.)