Aalto University
School of Science
Degree Programme in Engineering Physics and Mathematics

Alex Karrila

Number-theoretic and Geometric Lattice Code Design for Secure and Reliable Wireless Communications

Master's Thesis
Espoo, October 20, 2015

Supervisor:	Professor Camilla Hollanti
Advisor:	Professor Camilla Hollanti

Aalto University
School of Science
Degree Programme in Engineering Physics and Mathematics

ABSTRACT OF
MASTER'S THESIS

Author: Alex Karrila

Title:
Number-theoretic and Geometric Lattice Code Design for Secure and Reliable Wireless Communications

Date:	October 20, 2015	**Pages:**	vi + 85
Major:	Mathematics	**Code:**	Mat-1
Supervisor:	Professor Camilla Hollanti		
Advisor:	Professor Camilla Hollanti		

In data transmissions over wireless channels, the signal quality is weakened by random fading and noise of the electric field. This intrinsic property of the channel poses a challenge as the transmitted messages should be decodable at the receiver. On the other hand, it can be utilized for physical-layer security, in which the correct decoding probability drastically decreases when the signal quality weakens, hence securing the message from unintended receivers farther away. In this thesis, we study the design of lattices for lattice codes with an emphasis on lattice coset codes mostly in the Rayleigh fast fading channel model. Good lattice codes, *i.e.*, solutions to the legitimate receiver's problem are known based on number-theoretic lattice constructions, whereas the design of lattice coset codes providing also physical-layer security is an open problem.

We begin with a review of basic information theory, providing existence results and performance bounds on codes. Then, we specialize in lattice codes and lattice coset codes in wireless channels, deriving probability bounds for the legitimate receiver's error probability and the eavesdropper's correct decoding probability. In terms of these bounds, algebraic lattice constructions based on field extensions perform well, and for such lattices the bounds yield number-theoretic optimization problems. We study algebraic number theory extensively in order to have the tools to construct algebraic lattices and formulate and compute the probability bounds in terms of the properties of a given field extension. Finally, we compute the number-theoretic invariants and the eavesdropper's probability bound for algebraic lattices to assess and geometrize the different number-theoretic approaches that have been suggested to predict the eavesdropper's correct decoding probability for lattice coset codes.

Keywords: Additive white gaussian noise (AWGN) channels, algebraic number theory, algebraic lattices, inverse norm sum, lattice codes, lattice coset codes, lattice design, physical-layer security, Rayleigh block fading channels, Rayleigh fast fading channels, reliability, wireless channels, wiretap channels

Language: English

Aalto-yliopisto
Perustieteiden korkeakoulu
Teknillisen fysiikan ja matematiikan koulutusohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä: Alex Karrila

Työn nimi:
Turvallisten ja luotettavien koodihilojen lukuteoreettinen ja geometrinen suunnittelu langattomassa viestinnässä

Päiväys:	20. lokakuuta 2015	**Sivumäärä:**	vi + 85
Pääaine:	Matematiikka	**Koodi:**	Mat-1
Valvoja:	Professori Camilla Hollanti		
Ohjaaja:	Professori Camilla Hollanti		

Langattomassa viestinnässä signaalinlaatua heikentävät sähkömagneettisten aaltojen satunnaissironta sekä taustakohina. Tämän erityispiirteen vuoksi viestinnän luotettavuuden takaaminen on langattomien kanavien perusongelma. Toisaalta sähkökentän häipymistä ja kohinaa voidaan hyödyntää fyysisen kerroksen salausmenetelmissä, joissa viestintä suunnitellaan sellaiseksi, että vastaanottajan oikean dekoodauksen todennäköisyys romahtaa signaalin laadun heikentyessä tarpeeksi. Tällöin kaukana oleva salakuuntelija ei pysty tulkitsemaan viestiä. Diplomityössä tutkitaan viestintähilojen suunnittelua hilakoodeja ja erityisesti hilojen jäännösluokkakoodeja varten pääasiassa nopeasti häipyvän Rayleigh-kanavan mallissa. Luotettavan viestinnän takaaville hilakoodeille tunnetaan lukuteoreettisia kostruktioita, kun taas myös fyysisen kerroksen salauksen takaavien hilojen jäännösluokkakoodien suunnittelu on avoin ongelma.

Diplomityö aloitetaan kertaamalla informaatioteorian perustuloksia, jotka koskevat koodien olemassaoloa ja tiedonsiirtokapasiteettia. Tämän jälkeen erikoistutaan langattoman viestinnän kanavamalleihin sekä hilakoodeihin ja jäännösluokkakoodeihin. Näissä tapauksissa johdetaan ylärajat tarkoitetun vastaanottajan virhetodennäköisyydelle sekä salakuuntelijan oikean dekoodauksen todennäköisyydelle. Todennäköisyysrajojen perusteella lukukuntalaajennuksiin perustuvat algebralliset hilat suoriutuvat hyvin, ja tällaisten hilojen suunnittelu on lukuteoreettinen ongelma. Algebrallista lukuteoriaa tutkitaan laajasti ja saadaan algebrallisten hilojen konstruktio sekä työkalut viestinnän vertailukriteerien muotoiluun ja laskentaan lukuteoreettisin keinoin. Lopuksi lasketaan lukuteoreettiset invariantit sekä salakuuntelijan todennäköisyysraja joukolle algebrallisia hiloja. Tämän perustella arvioidaan ja geometrisoidaan salakuunteluongelmaan ehdotettuja jäännösluokkakoodien lukuteoreettisia hilasuunnittelukriteerejä.

Asiasanat: additiivisen valkoisen gaussisen kohinan kanavat, algebralliset hilat, algebrallinen lukuteoria, blokkihäipyvät Rayleigh-kanavat, fyysisen kerroksen salausmenetelmät, hilakoodit, hilasuunnittelu, hilojen jäännösluokkakoodit, käänteisnormisumma, langaton viestintä, nopeasti häipyvät Rayleigh-kanavat, salakuuntelukanavat, viestinnän luotettavuus

Kieli: Englanti

Acknowledgements

I wish to thank my partner Vera-Maria and my daughter Varpu for keeping my thoughts off the math at home and my feet steadily on the ground during the course of this work. I also thank everyone with whom I have had the pleasure of discussing mathematics, in particular my advisor Prof. Camilla Hollanti, Dr. David Karpuk, and Niko Väisänen, who have been involved in research related to this work, and other members of the Research Group in Algebra, Number Theory, and Applications. For financing the work of the research group, we gratefully acknowledge Aalto university, Magnus Ehrnrooth Foundation, Academy of Finland grants #276031, #282938, #268364, and #283262, and the support from the European Science Foundation under the ESF COST Action IC1104.

Espoo, October 20, 2015

Alex Karrila

Contents

Chapter 1

Introduction

1.1 Background

Most of today's communications happens through wireless channels, and the amount of transmitted data and connected devices constantly increases, posing a challenge to the designers of wireless systems. Wireless channels are open in nature, which makes them particularly vulnerable to eavesdropping on the one hand, and on the other hand to the weakening of the signal quality both due to noise from external sources and fading of the electromagnetic waves. The distinctive challenges and increasing demand make the study of wireless communications one of the most important branches of contemporary information theory, both industrial and academic.

In practically any information transmission scheme there are three main objectives: efficiency, reliability, and security. This means that the transmission of data should be fast and economically efficient, the data should be received correctly with high probability, and it should be hard for an unintended receiver to read the transmitted messages. As a starting point for practical designs, it is highly conventional to use lattice coding in wireless channels for efficiency and reliability reasons. This means roughly that m subsequent values of the transmitted electric field can be regarded as the components of a vector belonging to a lattice in $\mathbb{R}^m$. This thesis considers a few mathematical lattice design problems arising from the reliability and security aims in the code design for wireless channels.

Having chosen lattice coding, the first lattice design problem addressed is to find a lattice that yields as reliable lattice codes as possible, given certain efficiency constraints. The answer depends of course on how fading and noise are modelled, and good lattices are known in low dimensions for the simplest channel models, the Gaussian and the Rayleigh fast fading channels, due to geometric and number-theoretic methods. In this thesis, we review the known designs and in an ongoing related work not included in this text, we generalize the geometric designs to slightly more complicated channel models [17].

The second optimization problem, into which more effort is put in this thesis, is related to the security objective. Traditionally, security has been ensured by cryptography, which will however turn out to be useless if the eavesdropper knows the key, and secretly distributing the key is not necessarily straightforward in wireless communications. Alternatively, cryptography based on products of large primes such as RSA could be cracked by quantum computers. A possible solution to these problems that has gained increasing interest is physical-layer security, in which the code is designed so that the weakening of the signal-quality will drastically decrease the receiver's ability to decode correctly. We study lattice design related to a physical-layer security strategy called lattice coset coding.

When this thesis work was initiated, it had been suggested that secure lattice coset codes could be designed by number-theoretic means and some possible design criteria had been suggested, related to the invariants of the underlying algebraic field extension called discriminant and regulator, and the prime ramification of the extension. The original purpose of the thesis was to introduce the necessary information theory and the number theory, with an emphasis on the prime ramification, and then run numerical simulations to compare the suggested design criteria.

In the course of studying lattice coset codes numerically, it turned out that a probability bound on which the proposed number-theoretic designs were based might be too loose in the sense that it does not necessarily correlate with the actual performance of the lattice. Based on this, we studied what kind of an alternative probability bound could be tighter, and gave a heuristic geometric lattice design based on the tighter bound. The examination of the earlier bound and number-theoretic design is considered in this thesis with the negative result, whereas the proposed new designs and the geometric design are considered in [15, 16] and mainly referred to in this thesis. This structural choice was made since the approach of the article is somewhat different from what we had adopted at the beginning of the thesis project.

1.2 Organization

In a very large scale, we tackle the lattice design problems by first deriving expressions that we use as design criteria for lattices and showing that good lattice constructions involve extensions of the rational field $\mathbb{Q}$. Then, we study field extensions for a while from a purely mathematical point of view, and in the end obtain a number-theoretic construction of lattices. Finally, we put the two pieces together and formulate the lattice design criteria as questions about field extensions in the case of algebraic lattices, and study algebraic lattice codes with some computational examples.

Our approach is divided between the chapters as follows. In Section 2, we present some very basic concepts of information theory that describe how good solutions are achievable for the given communication problems in a very general setting. Section 3 introduces the practical setup with wireless channel models and lattice coding strategies, from which the main lattice design problems and objective functions arise. Section 4 considers algebraic number theory, starting and finishing with applications but otherwise written from a purely mathematical point of view, with an emphasis on a phenomenon called prime ramification. Finally, in Section 5, we are ready to numerically evaluate codes based on algebraic lattice constructions.

As this thesis balances between mathematics and information theory, even results that might seem elementary to a specialist of either field are stated, typically without proof to save space, which in turn might be confusing for an unaccustomed reader. Almost all omitted proofs are nevertheless doable for an undergraduate student and, in addition, references are given.

Chapter 2

Motivation: communication problems

In this section, we present the real-life optimization problems of the thesis on an informal level. We also review some very basic information theory to give a quantization for the real-life problems, provide some fundamental bounds for the solutions and, finally, motivate the choice of the so-called coset codes, which are of interest later in this work. We refer to [5, 18] for details on information theory.

2.1 An informal introduction

2.1.1 The reliability problem

In its most general form, the reliability problem asks for a way to transmit information with a high speed and small error probability. Intuitively, there seems to be a trade-off; if one is allowed to take risks, then one should also be able to transmit information faster. Indeed, as we shall soon see, given a *channel model* there is an upper bound for the *data rate* at which arbitrarily small *error rates* can be achieved, called the *channel capacity.* In this result the rate is measured in bits per channel use, not bits per second. From the point of view of this thesis, the main content of this result is that optimizing reliability makes sense and good solutions exist.

In a practical approach, achieving reliability in wireless communications can be roughly divided in *physical layer* and *error correction.* The task of *physical-layer code design* is to fix a way to convert data from the wireless channel to a device that guarantees reasonably small *decoding delays* and a fixed amount of transmitted bits per channel use, *i.e.*, a beneficial conversion

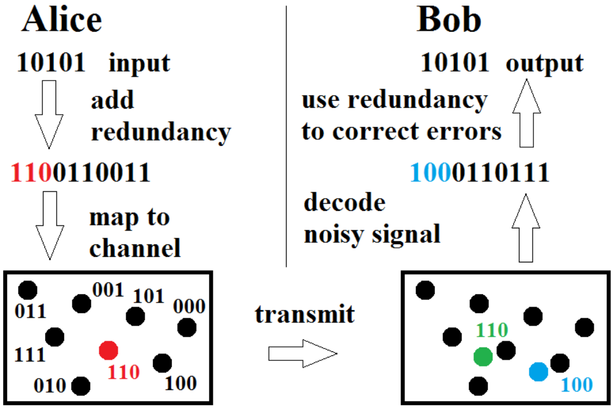

Figure 2.1: A schematic illustration of error correction and physical-layer reliability design.

of units from bits per channel use to bits per second. Then, the objective is to minimize the *decoding error rate* occuring in the conversion from channel to device. In particular, physical-layer code design alone never achieves arbitrarily small error rates since decoding errors necessarily occur, and physical-layer design does not directly address *error correction* by adding *redundancy* in the transmitted data. Instead, a succesful physical-layer design gaurantees a low decoding error rate so that less redundancy needs to be added to the transmitted data in the *error correction codes* to achieve low *bit error rates*, hence in the end contributing to a fast and reliable communication.

We only consider the physical-layer reliability problem of lattice codes, in which subsequent values of the transmitted electric field are the coordinates of a lattice point. The regular structure of the lattice allows fast decoding, *i.e.*, in practise finding the lattice point closest to the vector of received noisy electric-field values. Then, we minimize the vector decoding error rate. The interplay of error-correction and lattice coding in an example where Alice transmits data to Bob is illustrated schematically in Fig. 2.1. With this illustration at hand, we point out that also the *bit labelling* affects the reliability of a lattice code: the bit vectors should be mapped to the lattice points so that geometrically nearby lattice points correspond to nearby bit vectors. This guarantees a benefial conversion from *vector decoding error rates* to *bit decoding error rates*. Bit labelling is not addressed in this thesis, and we talk about correct and wrong decoding always on the vector level.

2.1.2 Information security

There are two main approaches to securing information, *cryptography* and *physical-layer security*. The former one is based on changing the *plaintext* into *ciphertext* by some injective function before sending it. The inverse of this function is kept secret or difficult to find so that only the legitimate receiver can convert the ciphertext into plaintext again. The alternative approach, physical-layer security, is based on the assumption that an intruder receives the messages in lower quality, due to noise and in wireless communications also fading. Then, the *coding system*, *i.e.*, the way the information is encoded into the channel (*e.g.*, into $0-1$ impulses in a wire or electric-field values in wireless communications) is chosen so that the messages are blurred already by a moderate amount of noise and fading. Cryptography and physical-layer security can of course be applied simultaneuosly. The security part of this work considers only physical-layer security.

2.1.3 The wiretap problem

The wiretap scheme, which is the most common setup for physical-layer security problems, was introduced by A. D. Wyner [31]. It can be roughly described as follows. An encoded message is transmitted to its legitimate receiver, while a wiretapper intercepts a version of this message with an inferior signal-quality. The wiretapper is assumed to know the decryption key, if any, but the transmitter and receiver can choose a suitable coding system in order to maximize the wiretapper's confusion due to the poor signal quality.

In addition to maximizing the confusion, the transmitter and receiver have to ensure that the probability of the receiver's correct decoding is high and the communication is fast, similarly to the reliability problem. This generic scheme, even though named as if tapping a wire, is naturally highly applicable and even more relevant in wireless communications, where any outsider can intercept the transmitted messages, and secretly delivering cryptation keys is difficult. The main theorem of Wyner's paper is, informally stated, that with the very few assumptions on the setup, the legitimate receiver can gain information with asymptotically zero error probability and asymptotically nonzero data rate, whilst the wiretapper's received information is asymptotically zero. Hence, in a wiretap setup, finding a good coding system is just a matter of innovation with a fundamental limit only on the data rate, but not on the reliability and security.

In what follows, we also use the established terminology where the sender is called Alice, the legitimate receiver Bob, and the wiretapper is called Eve or the eavesdropper.

2.2 Some basic concepts of information theory

In this subsection, a bit of information theory is presented in order to formalize the setup and yield an understanding of Wyner's result on a formal level, as well as to motivate for the mathematical problems studies later on. The results in this subsection are only utilized implicitly in the rest of this work, in the sense that they guarantee the existence of an optimal code design.

2.2.1 Entropy and information

We begin with two essential definitions.

Definition 1. *The (Shannon) entropy* of a discrete $\mathcal{X}$-valued random variable X with probability distribution P is

$$H(X) = -\sum_{x \in \mathcal{X}, P(x) > 0} P(x) \log P(x). \tag{2.1}$$

The conditional (Shannon) entropy of a discrete $\mathcal{X}$-valued random variable X given a discrete $\mathcal{Y}$-valued random variable Y is

$$H(X|Y) = -\sum_{y \in \mathcal{Y}} P(y) \sum_{x \in \mathcal{X}, P(x|y) > 0} P(x|y) \log P(x|y). \tag{2.2}$$

The name "conditional entropy of X given Y" is maybe somewhat misleading to an unaccustomed reader in the sense that no value of Y is given. A more precise but longer formulation would be, *e.g.*, "the expectation of the conditional entropy of X over all given Y".

Information theoretists like to emphasize the base of logarithms being 2, not e. Since for any $x, a, b \in \mathbb{R}_+$ we have $\log_b x = \log_b a \log_a x$ the different interpretations of the log operator will only scale the Shannon entropy and (as will be seen soon) the information by a constant. Hence, the base is actually irrelevant for a mathematician. However, for a computer scientist, the point is that if the base is two, then the entropy (conditional entropy) gives the optimal compression of the information X (given Y). To illustrate this naively, if Alice observes a realization of the $\mathcal{X}$-valued random variable X with a known distribution and wants to tell Bob about her observation by transmitting a sequence of zeroes and ones, she can encode more probable realizations into shorter sequences so that the smallest expected amount of zeroes and ones she needs is $H(X)$. This is known as *Shannon's source coding theorem.* Similarly, if Alice and Bob both know the realization of Y, and Alice knows the conditional distribution of X given Y, then Alice can communicate the realization of X in analogously optimized sequences of zeroes and ones that also depend on the realization of Y, with optimal expected sequence length $H(X|Y)$. To give yet another description, the (conditional) entropy tells how many bits the information of the realization of a random variable X is worth (given Y). In more general, letting the logarithms be of base $q \in \mathbb{Z}_{\geq 2}$ instead of 2, the entropy yields the optimal compression in q-ary alphabets instead of binary.

Remark 2. A combination of a finite number of discrete random variables is a discrete random variable. Hence, Def. 1 actually also defines entropies with several conditions, *e.g.* $H(X|Y, Z)$.

We continue by listing some of the properties of the Shannon entropy.

Lemma 3. *The entropy has the following properties:*

i) The entropy of a probability distribution is always convergent (also for countably infinite $\mathcal{X}$), and $0 \leq H(X)$ for any random variable X. The equality holds if and only if $P(X)$ is a delta distribution.

ii) Given a finite set $\mathcal{X}$, $H(X)$ is maximized if and only if the probability distribution $P(x)$ is uniform.

iii) Conditioning reduces entropy, i.e., $H(X|Y) \leq H(X)$ for any random variables X and Y. The equality holds if and only if X and Y are independent.

iv) The reduction in entropy caused by conditioning does not depend on which variable is unknown and which gives the condition, i.e., $H(X) - H(X|Y) = H(Y) - H(Y|X)$.

The parts *iii − iv* of Lemma 3 allow us to give the following quantification of information.

Definition 4. *The mutual information* of the random variables X and Y is

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y;X). \tag{2.3}$$

Similarly, *the conditional mutual information* of X and Y given Z is

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(Y|Z) - H(Y|X,Z) = I(Y;X|Z). \tag{2.4}$$

Recalling the interpretation of entropy as optimal compression, the mutual information tells how much the knowledge of Y reduces the optimal compression of X. Hence, it can be roughly regarded as an information theoretic measure of correlation; it is symmetric, and the heavier the "correlation" of X and Y, the larger this reduction in bits is. Conversely, by Lemma 3, the reduction in bits is zero if and only if X and Y are independent.

There is one more lemma concerning the entropy functional that is useful for discrete memoryless channels. This is actually an easier special case of the chain rule for entropy [See, *e.g.*, [18], Appendix A].

Lemma 5. *Let $X_1, ..., X_N$ be independent random variables. Then, the entropy of any single random variable $H(X_i)$ and that of $\mathbf{X}^N = (X_1, ..., X_N)$ are related by*

$$H(\mathbf{X}^N) = \sum_{i=1}^{N} H(X_i).$$

2.2.2 Discrete memoryless channels

From a general information-theoretic point of view, a channel is an entity that transports information. Channels are subject to distortions, and hence, if the input alphabet is $\mathcal{X}$ and the output $\mathcal{Y}$, an abstract channel is described by *transition probabilities* $p(y|x)$. In discrete channels, the input and output alphabets are finite. The channel is called memoryless if the transition probabilities are the same for all channel uses, not depending, *e.g.*, on previously transmitted or received alphabetic values. In the language of electromagnetism this means that the coherence time of the channel is less than the time interval between the transmissions. For example, Shannon's noisy-channel coding theorem and Wyner's results on the wiretap channel [31] concern discrete memoryless channels. The fact that memorylessness simplifies our computations is depicted by Lemma 5, stating essentially that entropies, and hence also informations, related to subsequent independent channel uses can be simply summed, and thus all importat information-theoretic computations only need to consider a single channel use.

2.2.3 Channel capacity

Definition 6. The *channel capacity* C of a discrete memoryless channel with given alphabetic transition probabilities is the best possible information about a random input alphabet X that the channel can yield to a receiver, *i.e.*,

$$C := \sup_{\text{all PDFs } P(x)} I(X;Y).$$

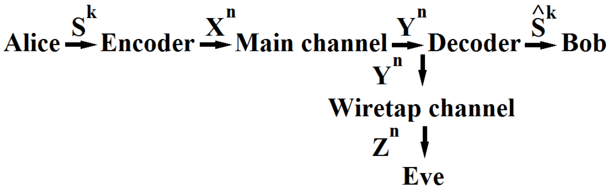

Figure 2.2: Wyner's wiretap channel.

The channel capacity also gives the maximum rate (in bits per channel use) at which information can be transmitted with an arbitrarily low error probability. This is called *Shannon's noisy-channel coding theorem*. Whereas the source coding theorem quantizes the optimal compression, this theorem should be regarded as a quantization of the optimal error-correction: given the transition probabilities, how much actual data can we transmit per channel use and, conversely, how large a proportion of the transmission must be covered by redundant content to combat errors. We are not going to use this theorem directly, but it is one of the fundamental results of information theory, stating that every model of a discrete memoryless channel has a fundamental limit for the efficiency of information transition. Likewise, it is a fundamental principle in information theory to first compress source data and then add redundancy for transmission. These steps are called *source coding* and *channel coding*, respectively.

2.2.4 Wyner's theorem

We can now state Wyner's result on the wiretap channel. Consider a channel model as depicted in Fig. 2.2, with both channels discrete and memoryless. Assume that k values of source alphabets from set $\mathcal{S}$ are compressed into n values channel alphabets. The subsequent values of source alphabets are assumed i.i.d. with "intrinsic" entropy $H(S) := H$. Then, the rate R at which source data is transmitted is on average $R = Hk/n$ bits per channel use. In bits per source alphabet, the information that the eavesdropper falls short, called *equivocation*, is $H(\mathbf{S}^k|\mathbf{Z}^n)/k$. (Recall that $H(\mathbf{S}^k|\mathbf{Z}^n)$ is the optimal compression of the random variable $\mathbf{S}^k$ given $\mathbf{Z}^n$.) A rate-equivocation pair (R, d) is *achievable*, if there exist channel coding systems for which aforementioned formulae for rate and equivocation can reach arbitrarily close to R and d, respectively, with an arbitrarily small decoding error probability for the legitimate receiver. Here "decoding error probability" is quantized as the average of error probabilities of each component of $\mathbf{S}^k$, *i.e.*, $1/k \sum_\ell P$(Bob's guess $\hat{\mathbf{S}^k}$ differs from $\mathbf{S}^k$ in ℓ^{th} component). Finally, let C_M be the capacity of the main channel from Alice to Bob. Then, the fundamental result of wiretap channels is stated as follows.

Theorem 7. *The achievable pairs* (R, d) *are given by*

$$\{0 \leq R \leq C_M, 0 \leq d \leq H, Rd \leq H \sup_{\textit{all PDFs } P(x)} I(X;Y|Z)\}. \tag{2.5}$$

Note that the supremum only depends on the channel model, so in the (R, d) plane, the achievable points are the intersection of a rectangle and the subgraph of a hyperbola $Rd \leq$ constant. The upper-right corner of this domain, with large rate and equivocation, is particularly interesting. Motivated by this, the largest rate R at which (R, H) is achievable is called the *secrecy capacity* C_s — it is the best possible data rate that can be achieved with arbitrarily little information leaking to the eavesdropper. The secrecy capacity is positive, satisfying $0 < C_s \leq C_M$, so in particular Wyner's theorem states that for any wiretap channel model, secure communication based on only physical-layer security is possible. This is a very fundamental existence result that can be informally crystallized in the words "physical-layer security makes sense".

2.2.5 Coset Coding

The information-theoretic foundation of coset coding presented in [25] is based on the following simplified setup. Alice transmits a vector of k alphabetic values $(x_1, ...x_k)$, of which Eve correctly receives a subset $(x_{e_1}, ..., x_{e_\mu})$ of $\mu < k$ alphabets x_{e_j} at some indices e_j but obtains no knowledge of the other components x_j. Then, if the transmission arrangement contains "too much information" in the sense that several different $\mathbf{x}^k$ decode to the same input alphabet, the arrangement can be chosen such that Eve's information is negligible.

Example 8. Assume that Alice encodes one bit into a two-bit vector, *i.e.*, $\mathcal{S}^k = \mathcal{S} = \mathbb{Z}_2$ and $\mathcal{X}^n = \mathcal{X}^2 = \mathbb{Z}_2 \times \mathbb{Z}_2$. She uses parity encoding $f : \mathcal{S} \to \mathcal{X}^2$ based on random choice,

$$\begin{aligned} f(0) &= \text{choose}\{(0,0),(1,1)\} \\ f(1) &= \text{choose}\{(0,1),(1,0)\}, \end{aligned}$$

and a parity decoding,

$$f^{-1}((a,b)) = a + b, \qquad a, b, a+b \in \mathbb{Z}_2.$$

Now, it is elementary to check that if Eve only receives one channel alphabet, *e.g.*, $(\text{not received}, 0)$, the original encoded bit is equally likely to have been 0 or 1, so Eve obtains no information.

A lattice analogue of modulo classes considered in the example above are modulo classes of sublattices — lattices are considered rigorously in the next section but the reader probably has an intuitive picture already at this point. Hence, the redundancy is added to the channel alphabet by taking the alphabet of a lattice Λ_b code to be the classes of Λ_b/Λ_e, where Λ_e is a sublattice. For the utility of lattice coset coding, we refer to, *e.g.*, [24]. For a report of implementation in wireless channels, see [20].

Chapter 3

Lattice codes in wireless channels

In this section, we study lattice codes and lattice coset codes in three basic models for wireless communications. The section is structured as follows. First, we present some basic lattice theory and the gaussian, Rayleigh fast fading and block fading channel models. Then, we derive the probability bounds for the eavesdropper's correct decision probability and the legitimate receiver's error probability for the respective channel models. These probabilities provide the objective functions of the optimization problem that this work considers.

3.1 Lattices

3.1.1 Basic concepts

We begin with a definition.

Definition 9. A *lattice* is a discrete additive subgroup of $\mathbb{R}^n$.

Any point in a lattice $\Lambda \subset \mathbb{R}^n$ can be expressed in terms of a *generator matrix* $M \in \mathbb{R}^{n \times m}$ as follows

$$\Lambda = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = M\omega, \omega \in \mathbb{Z}^m\}.$$

We assume that the columns of M are linearly independent over $\mathbb{Z}$ and hence, the *lattice coordinates* ω of a lattice point are unique and m is the *dimension* of the lattice. If $m = n$, the lattice is of *full rank.* A *sublattice* of a lattice of dimension m in $\mathbb{R}^n$ is an additive subgroup. It has a generator matrix MZ, where $Z \in \mathbb{Z}^{m \times k}$. Here k is the dimension of the sublattice and for a square matrix Z [29],

$$|\Lambda_b/\Lambda_e| = |\det Z|.$$

The *fundamental parallellotope* F_Λ of a lattice Λ is defined as

$$F_\Lambda = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = M\omega, \omega \in (0,1)^n\}.$$

The *volume* $\mathrm{Vol}(\Lambda)$ of the lattice Λ is the volume of the fundamental parallellotope, given by

$$\mathrm{Vol}(\Lambda) = |\det M|$$

for full-rank lattices. For a general rank m,

$$\mathrm{Vol}(\Lambda) = \sqrt{\det(\mathbf{b}_i \cdot \mathbf{b}_j)_{i,j=1}^m},$$

where $\mathbf{b}_i$ are the generator vectors. The volume of a lattice does not depend on the choice of the fundamental parallellotope since if Λ has two different generator matrices M and M', then the latter generates a sublattice, $M' = MZ$ with $|\Lambda/\Lambda| = 1 = |\det Z|$ and hence, $|\det M| = |\det M'|$. The matrix $(\mathbf{b}_i \cdot \mathbf{b}_j)_{i,j=1}^m$ is called the Gram matrix and it determines the geometry of the lattice basis.

Remark 10. Differing from some information theory references, here vectors are identified with *column* matrices and the lattice generator vectors with the *columns* of the generator matrix M.

Definition 11. The *dual lattice* $\Lambda^\star$ of a full-rank lattice Λ generated by M is the one generated by

$$M^{-T} := (M^{-1})^T = (M^T)^{-1}.$$

The dual lattice is well-defined, in the sense that it does not depend on the choice of the generator matrix; again if $M' = MZ$ with $|\Lambda/\Lambda| = 1 = |\det Z|$, then

$$M'^{-T} = M^{-T}Z^{-T}.$$

Now, by Cramer's rule for matrix inversion, Z^{-T} is an integer matrix and hence M'^{-T} generates a sublattice $\Lambda'^\star$ of $\Lambda^\star$. Furthermore, $1 = |\det I| = |\det Z \det Z^{-T}| = |\det Z^{-T}| = |\Lambda^\star/\Lambda'^\star|$, so actually M'^{-T} and M^{-T} generate the same lattice.

Theorem 12 (The Poisson formula for lattices)**.** *Let Λ be a full-rank lattice with generator M and let $f : \mathbb{R}^n \to \mathbb{C}$ be a continuous function with $\int_{\mathbf{x}\in\mathbb{R}^n} |f(\mathbf{x})| d^n x < \infty$ and $\sum_{\mathbf{t}\in\Lambda^\star} |\hat{f}(\mathbf{t})| < \infty$ such that the partial sums of $\sum_{\mathbf{t}\in\Lambda} |f(\mathbf{t}+\mathbf{u})|$ converge uniformly whenever $\mathbf{u}$ is restricted onto a compact set. Then,*

$$\sum_{\mathbf{t}\in\Lambda} f(\mathbf{t}) = |\det M|^{-1} \sum_{\mathbf{t}\in\Lambda^\star} \hat{f}(\mathbf{t})$$

where the Fourier transform is defined as

$$\hat{f}(\mathbf{t}) = \int_{\mathbf{y}\in\mathbb{R}^n} e^{-i2\pi\mathbf{y}\cdot\mathbf{t}} f(\mathbf{y})dy.$$

Proof. The proof is given in [7]. We point out that the condition on the continuity of f is essential. □

3.1.2 Geometric properties

Lattices allow a variety of different mathematical approaches, ranging from group theory to complex analysis and geometry. In this thesis, we mainly consider the geometric approach.

Definition 13. The Voronoi cell $\mathcal{V}(\mathbf{t})$ of a lattice point $\mathbf{t} \in \Lambda$ is defined as

$$\mathcal{V}(\mathbf{t}) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{w}\| > \|\mathbf{x} - \mathbf{t}\| \ \forall\ \mathbf{t} \neq \mathbf{w} \in \Lambda\}.$$

Example 14. The Voronoi cells of the hexagonal lattice in $\mathbb{R}^2$ are open regular hexagons. The Voronoi cells of a square lattice in $\mathbb{R}^n$ are open squares.

The following lemma concerning Voronoi cells is easy to prove, but it is given here for completeness. The lemma could be directly generalized to consider bounded sets that tile $\mathbb{R}^n$ under translations of $\mathbf{t} \in \Lambda$ or lattices that are not of full rank, but neither one of these generalizations is necessary later on.

Lemma 15. *All Voronoi cells become $\mathcal{V}(\mathbf{0})$ under translation, $\mathcal{V}(\mathbf{t}) - \mathbf{t} = \mathcal{V}(\mathbf{0})$, so they have equal size. Furthermore, if the lattice is of full rank, the size is given by*

$$\mu(\mathcal{V}(\mathbf{0})) = \int_{\mathbf{y}\in\mathcal{V}(\mathbf{0})} d^n y = \mathrm{Vol}(\Lambda) = |\det M|.$$

Proof sketch. The first part follows easily from the self-similarity of the lattice.

For the second part, it is easy to show that a Voronoi cell or a fundamental parallellotope has a finite diameter. Then, consider all the $N(R)$ lattice points in $B(0, R)$. The union of cells (parallellotopes) corresponding to these parallellotopes is between the sets $B(0, R - d)$ and $B(0, R + d)$, where d is the diameter of the cells (parallellotopes). Hence,

$$\mathrm{Vol}(B(0, r - d)) \leq N(R)\mathrm{Vol}(D) \leq \mathrm{Vol}(B(0, R + d)),$$

where D is the parallellotope or the cell. These inequalities give upper and lower bounds to $\mathrm{Vol}(F)/\mathrm{Vol}(\mathcal{V}(\Lambda))$. Letting $R \to \infty$, we notice that the ratio is one. □

Problem 16. *Throughout this thesis, we call the sphere-packing problem the problem of finding the densest lattice packing of non-intersecting spheres of equal radius in $\mathbb{R}^n$.*

Equivalently, since from the definition it follows each Voronoi cell intersects with exactly one sphere in the packing, the solution to the sphere-packing problem in $\mathbb{R}^n$ maximizes $\mathrm{Vol}(B)/\mathrm{Vol}(\Lambda)$, where B is the insphere of the Voronoi cell, hence with diameter (radius) equal to (half of) the minimal vector length of the lattice.

The following two definitions are more related to the theory of lattice codes than lattice theory, but let us still present them here.

Definition 17. The *ℓ-product distance* of (lattice) vectors $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{x}' = (x'_1, \ldots, x'_n) \in \Lambda$ differing in exactly ℓ components is defined as $d_p^\ell(\mathbf{x}, \mathbf{x}') = \prod_{x_i \neq x'_i} |x_i - x'_i|$. The quantity $\ell(\mathbf{x}, \mathbf{x}')$ is called the *modulation diversity* of $\mathbf{x}$ and $\mathbf{x}'$. The minimum of ℓ over the lattice Λ,

$$\delta = \min_{\mathbf{x}, \mathbf{x}' \in \Lambda, \mathbf{x} \neq \mathbf{x}'} \ell(\mathbf{x}, \mathbf{x}'),$$

is referred to as the *diversity* of the lattice Λ. If $\delta = n$, Λ is said to have *full diversity.*

Definition 18. For full-diversity lattices, the *minimum product distance* is defined as $d_{p,min}(\Lambda) = \min_{0 \neq \mathbf{x} \in \Lambda} \prod_{i=1}^n |x_i|$.

Remark 19. Let us illustrate the interpretation of the geometric concepts presented here with a simplified example. In wireless communications, information is coded into magnitudes of electric field components. Let us assume that Alice and Bob agree on a transmitting arrangement in which n subsequent electric field values, regarded as a vector in $\mathbb{R}^n$, always form lattice Λ points $\mathbf{x}$. The receiver measures a random vector $\mathbf{Y} = \mathbf{x} + \mathbf{N}$ corrupted by a random noise $\mathbf{N}$, and his natural guess for the original lattice point $\mathbf{x}$ is the one with $\mathbf{Y} \in \mathcal{V}(\mathbf{x})$. (Here we implicitly assumed that the noise vector $\mathbf{N}$ has zero mean and its probability density is spherically symmetric and radially decreasing.) With this decoding, Bob decodes the vector $\mathbf{x}$ correctly iff $\mathbf{N} \in \mathcal{V}(\mathbf{0})$. Still assuming that the noise is spherically symmetric with a radially decreasing probability density, the sphere-packing radius then roughly quantizes the probability of this event, *i.e.*, the reliability of the code given by this lattice.

Let us next study fading (wireless) channel models, in which the transmitted vectors are streched by a random diagonal fading matrix $\mathrm{diag}(h_1, ..., h_n)$, where the entries $h_i \geq 0$ are in the simplest case independent and identically distributed (i.i.d.). Hence, Bob receives $\mathbf{Y} = \text{diag}(h_1, ..., h_n)\mathbf{x}+\mathbf{N}$, where it is typically assumed that Bob knows $\text{diag}(h_1, ..., h_n)$ but not $\mathbf{N}$. By the preceding channel model, we would then like the lattice $\text{diag}(h_1, ..., h_n)\Lambda$ to have a dense sphere packing, equivalently, long minimal vectors. Then, diversity discribes roughly how resistant vector lengths are to such fading, and hence the diversity of a lattice describes the resistance of its sphere-packing radius. For example, let $\mathbf{x} = (\sqrt{n}, 0, ..., 0)$ and $\mathbf{x}' = (1, ..., 1)$ so that $\|\mathbf{x}\| = \|\mathbf{x}'\| = \sqrt{n}$. Then, it is elementary to verify that $\mathbb{E}(\|\text{diag}(\mathbf{h})\mathbf{x}\|^2) = \mathbb{E}((\|\text{diag}(\mathbf{h})\mathbf{x}'\|^2))$ and that $\text{Var}(\|\text{diag}(\mathbf{h})\mathbf{x}\|^2) = n^2\text{Var}(h_1^2)$ whereas $\text{Var}(\|\text{diag}(\mathbf{h})\mathbf{x}'\|^2) = n\text{Var}(h_1^2)$, so the length of a diverse vector is considerably more stable.

3.2 Channel models

3.2.1 The gaussian channel

In the following three subsections, we establish three different channel models. The gaussian model is a general noisy channel model, but for concreteness, we motivate its use in the special case of wireless communications.

Assume that the information is encoded into values of a sinusoidal electric field, which, containing a phase and a magnitude, can be collected into a vector $\mathbf{u}$ of $\mathbb{C}^m$. The received vector is blurred by an additive zero-mean gaussian noise. Hence, the received vector $\mathbf{w} \in \mathbb{C}^m$ is given componentwise by

$$w_i = u_i + v_i',$$

where $v_i' \in \mathbb{C}$ represents the noise. The zero-mean gaussian form of the noise as an electric-field value, given by $\Re v_i'$, is motivated by the central limit theorem. Furthermore, since the noise certainly does not depend on the phase of the transmitted wave, its complex representative v_i' should hence be a complex gaussian variable. The independence of the noise terms is an assumption equivalent to the memorylessness of the channel.

All this can however be simplified to a real channel model. Consider a transmitted vector $\mathbf{u} \in \mathbb{C}^n$ and received $\mathbf{w} \in \mathbb{C}^n$. We can naturally identify these with vectors $\mathbf{x} \in \mathbb{R}^{2m}$ ($\mathbf{y} \in \mathbb{R}^{2m}$) by choosing $\Re u_1 = x_1$, $\Re u_2 = x_2$, ..., $\Im u_1 = x_{n+1}$ etc. ($\Re w_1 = y_1$ etc.). Thus, we will in continuation study a real model, where the received vector $\mathbf{y} \in \mathbb{R}^{2m}$ is given by

$$y_i = x_i + v_i,$$

and v_i are now i.i.d. zero-mean gaussian r.v. of variance σ^2.

Let us yet consider briefly quantizing the “channel quality” of such a channel. In the *oscillating dipole model*, which is the simplest model for an antenna, the squared length of the electric field vector is proportional to the power flux carried by the wave. (For a mathematical formulation see, *e.g.*, [12].) With the vague notion of squares representing energy flux density and the intuitive guess that this density determines the channel quality, the natural measure for the channel quality is $\mathbb{E}\{v_i^2\} = \sigma^2$. In the wiretap setup, Bob and Eve experience different noises and, by the basic assumption, *i.e.*, Eve having an inferior signal quality, the mathematical formulation is then $\sigma_b^2 < \sigma_e^2$.

3.2.2 The Rayleigh block fading channel

The block fading model is a generalization of the gaussian one to include the fading of the transmitted electric field. Then, after fading and additional noise, the received vector $\mathbf{w}'$ is given componentwise by

$$w_i' = h_i u_i' + v_i',$$

where h_i is a complex fading coefficient and v_i' the complex noise term. We make the following assumptions.

i) The variables h_i and v_i are complex zero-mean Gaussian random variables with variances σ_h^2 and σ^2, respectively. (The gaussian form of h_i is motivated by the large number of scatterers and the central limit theorem.) The sending frequency is low enough to make the random variables v_i and h_i i.i.d.

ii) Both Eve and Bob have a perfect channel state information (CSI), *i.e.*, they know the complex fading coefficients h_i.

After removing the phase of h_i (assumption (ii)), the received message is

$$w_i = |h_i| u_i + v_i,$$

where $|h_i|$ is a real r.v. with parameter σ_h Rayleigh-distribution with the probability density function

$$r(x) = \frac{x}{\sigma_h} \exp\left(\frac{x^2}{2\sigma_h^2}\right). \tag{3.1}$$

and v_i is a complex zero-mean Gaussian r.v. with variance σ^2 (assumption (i)).

Next, we move over to a real model almost as in the gaussian case by choosing $\Re u_i = x_{i,1}$, $\Im u_i = x_{i,2}$, so that $\mathbf{X} \in \mathbb{R}^{m\times 2}$. Similarly, choose $\Re w_i = y_{i,1}$ and $\Im w_i = y_{i,2}$ Then, the received point of $\mathbf{Y} \in \mathbb{R}^{m\times 2}$ is given by

$$y_{i,j} = |h_i| x_{i,j} + v_{i,j}, \quad j \in \{1, 2\},$$

where $v_{i,j}$ is a real gaussian noise with variance σ^2. In matrix notation, this will simplify to

$$\mathbf{Y} = \text{diag}|h_i|\mathbf{X} + \mathbf{V}, \tag{3.2}$$

where diag$|h_i|$ is a diagonal matrix of $\mathbb{R}^{m\times m}$ with $|h_i|$ in the i^{th} diagonal entry. To have a broader applicability, assume that not 1 but instead $L/2$, where L is even, transmitted complex numbers u_i' are multiplied by each fading coefficient h_i. Physically, this is a simplification of a situation where the electric field scatterers are quasi-static. Then, the above assumptions yield the same equation (3.2) but with $\mathbf{X}, \mathbf{Y}, \mathbf{V} \in \mathbb{R}^{m\times L}$. By channel interleaving (see [23]) we can omit the assumption of L being even. This is the general block-fading model.

Analogous to the AWGN channel, the natural way to describe channel quality is given from the energy densities. The relevant signal has expected square $\mathbb{E}\{\|\text{diag}|h_i|\mathbf{X}\|^2\} \propto \mathbb{E}\{|h_i|^2\}$ and the noise a square proportional to σ^2. Thus, the ratio of the energies related to the actual signal and the noise is described by $\mathbb{E}\{h_1^2\}/\mathbb{E}\{v_1^2\} \propto \sigma_h^2/\sigma^2$. For a rigorous but less intuitive motivation for this measure of channel quality, the reader can verify that Eve's and Bob's probability bounds given later in Section 3.3.4 only depend on σ_h^2/σ^2, given the lattice and sending constellation.

3.2.3 The Rayleigh fast fading channel

The Rayleigh fast fading channel is a variant of the block fading channel. From a mathematician's point of view, it is a simplification, but in an engineering sense, it is more complicated; we add a third assumption.

iii) The transmission arrangement utilizes a channel interleaver (see Ref. [23], Section 2.1) so as to make the fading coefficients h_i of one codeword i.i.d.

Mathematically, this means that the Rayleigh fast fading channel is described by the equation (3.2) of block fading channels but with $L = 1$. The channel quality is desribed by σ_h^2/σ^2. We shall mainly consider the Rayleigh fast fading channels in this thesis, but we point out already now that it can be expected that lattice designs for Rayleigh fast fading channels can be generalized to block fading channels, as we shall see in Section 3.3.3.

3.3 Lattice codes and coset codes

3.3.1 Codes

In the previous subsections we concluded that the channel alphabet of all the previously considered channel models can be regarded as points in $\mathbb{R}^m$. The next question is how to choose such points.

Recall that if Alice transmits $\mathbf{x}$, Bob receives $\mathbf{y}$ given componentwise by $y_i = |h_i|x_i + v_i$, and knows $|h_i|$ ($h_i = 1$ for the AWGN channel). Hence, Bob's first task is to find the best guess for a signal point $\mathbf{x}$, given $\mathbf{y}$. Since all models assume that the additive noise has a spherical symmetry, Bob's natural guess is the signal point $\hat{\mathbf{x}}$ such that $\operatorname{diag}(|h_i|)\hat{\mathbf{x}}$ is closest to $\mathbf{y}$. Thus, the signal vectors should allow efficient computational closest-point search. A good choice is then to take the signalling points $\mathbf{x}$ to be a subset of a lattice. As we shall see, also the faded vectors $\operatorname{diag}(|h_i|)\hat{\mathbf{x}}$ form a lattice then.

In coset coding, as motivated by information theory, one still transmits points from the lattice Λ_b but decodes all representatives of a modulo class Λ_b/Λ_e of a sublattice to the same input alphabet. Conversely, input is encoded in uniform random representatives. In practise, the signalling constellation is a finite part of the lattice, so uniform distributions make sense.

Regarding lattice codes in wireless channels, one should yet point out the *power constraint.* The conventional way to formulate a power constraint is to choose the sending region to be a neighbourhood of the origin, typically spherical or hypercubic.

3.3.2 On generalizing lattice code results from gaussian to block fading channels

Since block fading channels are a generalization of gaussian channels, we will typically want to generalize results considering gaussian channels to block fading channels. Hence, let us vectorize the transmitted matrix $\mathbf{X} \in \mathbb{R}^{m\times L}$ by stacking its column vectors to $\operatorname{vec}(\mathbf{X}) \in \mathbb{R}^{mL}$. Next, the noise is added upon a faded message. Hence, we have to take into account the fading, in which the vectorized information received by Bob or Eve becomes multiplied by a block diagonal matrix in $\mathbb{R}^{mL\times mL}$ whose non-zero submatrices at the diagonal block are diagonal matrices $\operatorname{diag}(|h_i|)$, *i.e.*,

$$\begin{aligned}\mathrm{vec}(\mathbf{Y}) &= \begin{pmatrix} \mathrm{diag}(|h_i|) & & \\ & \ddots & \\ & & \mathrm{diag}(|h_i|) \end{pmatrix} \mathrm{vec}(\mathbf{X}) + \mathrm{vec}(\mathbf{V}) \\ &= \mathrm{diag}(\mathrm{diag}(|h_i|), ..., \mathrm{diag}(|h_i|))\mathrm{vec}(\mathbf{X}) + \mathrm{vec}(\mathbf{V}).\end{aligned}$$

The latter line is to be regarded as a definition for a notation.

Next, if the vectorized transmitted matrix $\mathrm{vec}(\mathbf{X})$ is given by lattice coordinates $\mathbf{u}$ in the dense lattice Λ_b, *i.e.*, $\mathrm{vec}(\mathbf{X}) = M\mathbf{u}$ with M being the generator matrix of Λ_b, then the received lattice point is given by

$$\mathrm{diag}(\mathrm{diag}(|h_i|), ..., \mathrm{diag}(|h_i|))M\mathbf{u} = M_h\mathbf{u}.$$

Hence, we can regard the received points as belonging to a "skewed" lattice $\Lambda_{h,e}$ generated by M_h. Similarly, the generator matrix MZ of the sparse lattice Λ_e is skewed to M_hZ. We emphasize that the derivation of this subsection, as well as the next one, does not assume a particular distribution for the fading coefficients h_i, but only that they are known to the receiver.

3.3.3 Geometry of lattices in fast and block fading channels

As another generalization procedure between channel models, we point out a heuristic that allows us to generalize lattice designs from Rayleigh fast fading channels to T-block fading channels. Consider code lattices Λ_R and Λ_B (either in lattice or coset code purposes) in the respective channels, generated by M and the block diagonal matrix $\mathrm{diag}(M, ..., M)$, respectively. Then, the faded lattices $\Lambda_{R,h}$ and $\Lambda_{B,h}$ are generated by $\mathrm{diag}(|h_i|)M$ and the block diagonal matrix $\mathrm{diag}(\mathrm{diag}(|h_i|)M, ..., \mathrm{diag}(|h_i|)M)$. Hence, in particular, $\Lambda_{B,h}$ consists of T orthogonal copies $\Lambda_{R,h} \times ... \times \Lambda_{R,h}$.

Next, by the previous subsection, a lattice Λ_R in the Rayleigh fast fading channel is equivalent to the random lattice $\Lambda_{R,h}$ in the AWGN channel. Typically, desings for the AWGN channel are highly geometric due to the spherical symmetry. For example, the reliability of a lattice code boils down to its sphere-pacing density. Now, if we know how to design Λ_R, then $\Lambda_{R,h}$ will (probabilistically, as h_i's are random) have a certain geometric behaviour, for example a dense sphere packing. Then, it is often easy to make the lattice $\Lambda_{B,h} = \Lambda_{R,h} \times ... \times \Lambda_{R,h}$ also satisfy the same geometric criteria.

3.3.4 Probability bounds

Even though the fundamental results of information theory are almost all based on the concepts of entropy and information, it is not always easy to design codes based on these quantities. In the case of lattice coset codes, it is conventional to only consider the legitimate receiver's error probability (REP) and the eavesdropper's correct decision probability (ECDP). We give here bounds for these probabilities. The bounds are proven and their tightness is discussed in the appendices of this thesis, and the connection of the ECDP and the eavesdropper's information is discussed formally in [19] and on an intuitive level in [16].

For Bob's decoding error, we have the following upper bounds that are asymptotically tight at good signal quality: for the AWGN channel

$$P_{e,b} \;\leq\; \frac{1}{2} \sum_{\mathbf{0} \neq \mathbf{w} \in \Lambda_b} e^{-\|\mathbf{w}\|^2/8\sigma_b^2}, \tag{3.3}$$

and for the Rayleigh block fading channel

$$P_{e,b} \leq \frac{1}{2} \sum_{\mathbf{0} \neq \mathbf{w} \in \Lambda_b} \prod_{i=1}^{m} \frac{1}{1+\gamma_b \|\mathbf{W}_i\|^2/4}, \tag{3.4}$$

where $\gamma_b = \sigma_{h,b}^2/\sigma_b^2$ depicts Bob's channel quality, $\mathbf{W}$ is the matrix form of $\mathbf{w}$ and $[\text{matrix}]_i$ denotes the i^{th} row vector. As a special case of the above, for the Rayleigh fast fading channel,

$$P_{e,b} \leq \frac{1}{2} \sum_{\mathbf{0} \neq \mathbf{w} \in \Lambda_b} \prod_{i=1}^{m} \frac{1}{1+\gamma_b w_i^2/4}. \tag{3.5}$$

For the ECDP, we have the following bounds that are asymptotically tight at poor signal quality: for the AWGN channel,

$$P_{c,e} \leq \sum_{\mathbf{t} \in \Lambda_e} \mathrm{Vol}(\Lambda_b) g_n(\mathbf{t}), \tag{3.6}$$

where $g_n(\mathbf{w}) = e^{-\|\mathbf{w}\|^2/(2\sigma_e^2)}/(2\pi\sigma_e^2)^{n/2}$ is the standard n-dimensional spherical zero-mean Gaussian density function with variance σ_e^2. For the Rayleigh block fading channel,

$$P_{c,e} \leq \Gamma(L/2+1)^m \mathrm{Vol}(\Lambda_b) \left(\frac{\gamma_e}{\pi}\right)^{Lm/2} \sum_{\mathrm{vec}(\mathbf{X}) \in \Lambda_e} \prod_{i=1}^{m} \frac{1}{(1+\|\mathbf{X}_i\|^2\gamma_e)^{L/2+1}}. \tag{3.7}$$

Here $\gamma_e = \frac{\sigma_{h,e}^2}{\sigma_e^2}$ is again the channel quality, Γ is the standard gamma function and $\mathbf{X}_i$ is the i^{th} row of $\mathbf{X}$. Again, as a special case of the above, for Rayleigh fast fading channels we have

$$P_{c,e} \leq \frac{\text{Vol}(\Lambda_b)}{2^m}\gamma_e^{m/2} \sum_{\mathbf{x}\in\Lambda_e} \prod_{i=1}^{m} \frac{1}{(1+x_i^2\gamma_e)^{3/2}}. \tag{3.8}$$

Physical-layer design of reliablility and security would require minimizing all the six objective functions above. However, lattice designs for Bob's probability bounds (3.3) and (3.5) are known, and hence also at least moderately good constructions for Eq. (3.4) of block fading channels follow by the generalization presented in Section 3.3.3. For Eve, we address minimizing the bound (3.6) in [15], so the remaining main task of this thesis is to minimize the bound (3.5). Again, block fading channels are then addressed via the generalization of Section 3.3.3.

It should be pointed out that all these series are conventionally truncated over the finite signalling region, although the ECDP bounds are not completely rigorous if truncated. Physically, this truncation is equivalent to neglecting the boundary effects of the finiteness of the constellation, which should be legitimate for any interesting signal quality of the eavesdropper.

As a second remark, we point out how these bounds call our attention to full-diversity lattices. The detailed computations are in appendix A.3.4. Consider the objective of minimizing the REP (3.5). This minimization is typically considered when the REP is on the order of 10^{-5}–10^{-3}, and hence, it is relevant to only consider the limit $\gamma_b \to \infty$. In this limit, it is obvious that full-diversity lattices will perform best, since only then all terms of the sum will decrease at a rate γ_b^{-m}. Then, having chosen a full-diversity lattice Λ_b, the sublattice Λ_e will also necessarily be of full diversity, so both the reliability and the security problem lead to the study of full-diversity lattices. In addition, the 1's in the denominator of (3.5) can be neglected in the limit $\gamma_b \to \infty$ (see Eq. (A.8) in the Appendices), and the dominating term of the series is determined by the minimum product distance of the lattice, which we want to maximize in order to minimize the REP.

Finally, we point out that also sum (3.8) has traditionally been approximated by dropping the ones in the denominator. This is called the inverse norm sum (INS), since for *algebraic lattices* it becomes a sum of *algebraic norms* (both definitions are given in the next section and the INS in Eq. (5.1) later on). This has yielded some interesting number-theoretic problems, which were the starting point of this thesis. Nevertheless, the problem with this approximation is that the bounds (3.7) and (3.8) are asymptotic at poor signal quality $\gamma_e \to 0^+$, and dropping the ones is asymptotic at good signal quality $\gamma_e \to \infty$. As mentioned in the introduction, our numerical computations suggest that the tighter bounds given above do not seem to correlate very well with this conventional approximation.

There is also a heuristic "engineer's explanation" to why we should not drop the ones in the denominators in Eq. (3.8). Namely, from the derivation in Appendices A.2.1–A.2.2 we may notice that each term in the series (3.7) and (3.8) depicts the probability that the vector $\hat{\mathbf{x}}$ to which Eve decodes $\mathbf{x}$ satisfies $\hat{\mathbf{x}} - \mathbf{x} = \mathbf{w}$. On the other hand, as described earlier, coset coding is only useful when Eve's detection resolution for the vectors $\mathbf{x}$ is approximately the resolution of Λ_e. Hence, in particular, shifts of vectors $\mathbf{w} \in \Lambda_e$ should occur with non-negligible probability. Consequently, we should not aim at comparing the asymptotics or decay rates of expressions (3.7)–(3.8) at large γ_e but instead cosider moderate values. An approximative formula for how large a valuie of γ_e is "moderate" is given in Appendix B, and for unit-volume lattices in low dimensions one will find approximately $0.5 \leq \gamma_e \leq 10$. Short vectors of unit-volume lattices will satisfy $x_i^2 \ll 1$ for some component i, and hence for $0.5 \leq \gamma_e \leq 10$ we cannot substitute $(1 + x_i^2\gamma_e)^{3/2} \approx (x_i^2\gamma_e)^{3/2}$. The inverse norm sum is examined numerically in Section 5.

Chapter 4

Algebraic number theory

This section considers algebraic number theory from a purely mathematical perspective. To give a hint of the link between this and the previous sections, we give the following example, showing that all constructions of lattices of diversity larger than one necessarily require considering field extensions of $\mathbb{Q}$.

Example 20. Let $\Lambda \subset \mathbb{R}^n$ be a lattice generated by $M \in \mathbb{Q}^{n\times n}$. Then, Λ intersects all the axes and is hence of diversity one. For let $\mathbf{m}$ be the j^{th} column vector of M^{-1}. By Cramer's rule, we have that $M^{-1} \in \mathbb{Q}^{n\times n}$, so $\mathbf{m}$ can be scaled to an integer vector $\omega = q\mathbf{m}$, $q \in \mathbb{Z}$. Then, the lattice Λ point $M\omega$ satisfies

$$(M\omega)_k = q\sum_l M_{k,l}M^{-1}_{l,j} = q\delta_{k,j},$$

so $M\omega$ lies on the j^{th} coordinate axis.

4.1 Algebraic structures and tools

In this subsection, we study finite algebraic field extensions. The main results concern their algebraic structure and substructures, *e.g.*, the ring of algebraic integers and the unit group as well as field extension invariants, *i.e.*, the discriminant and the regulator. None of the results in this section are particularly difficult, but some proofs are only cited in order to save space. The results are presented in an order that allows constructing the theory along the lines of [29]. The reader is assumed to be familiar with basic group, ring, and module theory as well as Galois theory applied to number fields. Good references for these topics are, *e.g.*, [28], [22], and [30].

4.1.1 Algebraic numbers and integers

This subsection introduces the two main algebraic structures, *i.e.*, the field of algebraic numbers and the ring of algebraic integers.

Definition 21. The set of *algebraic numbers* $\mathbb{A}$ is the set of all roots of polynomials over $\mathbb{Q}$.

Theorem 22. *$\mathbb{A}$ a field.*

Definition 23. The set *algebraic integers* $\mathbb{B}$ is the set of all roots of monic polynomials over $\mathbb{Z}$.

Lemma 24. *A complex number θ is in $\mathbb{B}$ if and only if the $\mathbb{Z}$-module generated by its powers ${}_{\mathbb{Z}}\langle 1, \theta, \theta^2, ...\rangle$ is finitely generated.*

Theorem 25. *Any root of a monic polynomial over $\mathbb{B}$ is in $\mathbb{B}$.*

Theorem 26. *$\mathbb{B}$ is a ring.*

Definition 27. Let K be a finite extension of $\mathbb{Q}$. The intersection ring $K \cap \mathbb{B} := \mathfrak{O}_K$ is called the ring of integers of K.

Notation 28. Here and in what follows we will encounter notations with subscripts K that emphasize that a field extension $K : \mathbb{Q}$ is considered. If there is no danger of confusion, these subscrips might be dropped.

Lemma 29. *For any $\alpha \in K$, there exists a nonzero $c \in \mathbb{Z}$ s.t. $c\alpha \in \mathfrak{O}_K$.*

Corollary 30. *K has a $\mathbb{Q}$-basis consisting of algebraic integers.*

4.1.2 Discriminants, field polynomials, norms, and traces

In this subsection, we will give the definitions and the basic properties of discriminants, norms, and traces, which are standard tools in considerations of given algebraic extensions.

However, we start with a lemma and a theorem that will later on make several proofs essentially easier.

Lemma 31. *Let $K : \mathbb{Q}$ be a finite extension. Then, there exists $\theta \in K$ such that $K = \mathbb{Q}(\theta)$.*

Theorem 32. *Let $K = \mathbb{Q}(\theta)$ be algebraic with $[K : \mathbb{Q}] = n$. Then, there are exactly n monomorphisms $\sigma_i : K \to \mathbb{C}$. These monomorphisms satisfy $\sigma_i|_{\mathbb{Q}} = id$ and they can hence be written as $\sigma_i : \theta \mapsto \theta_i$, where θ_i is some root of the minimal polynomial of θ.*

Remark 33. From the identity $\sigma_i|_{\mathbb{Q}} = \text{id}$ and the homomorphism property we have directly that a root in K of any $p(t) \in \mathbb{Q}[t]$ is mapped to some K-root of p by σ_i. In particular, σ_i maps $\mathfrak{O}_K$ to algebraic integers.

Definition 34. The *discriminant* $\Delta(\alpha_1, ..., \alpha_n)$ *of a* $\mathbb{Q}$*-basis* $\{\alpha_1, ..., \alpha_n\}$ of an extension $K : \mathbb{Q}$ is defined as

$$\Delta(\alpha_1, ..., \alpha_n) = (\det \mathbf{\Sigma})^2, \qquad \text{where } \mathbf{\Sigma}_{ij} = \sigma_j(\alpha_i). \tag{4.1}$$

Lemma 35. *If the basis* $\{\alpha_1, ..., \alpha_n\}$ *of* K *over* $\mathbb{Q}$ *is expressed by another basis* $\{\beta_1, ..., \beta_n\}$ *as*

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \mathbf{M} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \qquad \textit{where } \mathbf{M} \in \mathbb{Q}^{n\times n}, \tag{4.2}$$

then the discriminants are related by

$$\Delta(\alpha_1, ..., \alpha_n) = |\mathbf{M}|^2 \Delta(\beta_1, ..., \beta_n). \tag{4.3}$$

Lemma 36. *The discriminant of any basis is rational and non-zero.*

Definition 37. The *field polynomial* $f_\alpha(t)$ of $\alpha \in K$ is defined as

$$f_\alpha(t) = \prod_{i=1}^{n} (t - \sigma_i(\alpha)). \tag{4.4}$$

Lemma 38. *The field polynomial* $f_\alpha(t)$ *is a power of the minimal polynomial of* α.

Definition 39. The *norm* $N_K(\alpha)$ and *trace* $T_K(\alpha)$ of $\alpha \in K$ are defined as

$$N_K(\alpha) = \prod_{i=1}^{n} \sigma_i(\alpha) \qquad \text{and} \qquad T_K(\alpha) = \sum_{i=1}^{n} \sigma_i(\alpha).$$

Lemma 40. *The norm is multiplicative,* $N_K(\alpha_1\alpha_2) = N_K(\alpha_1)N_K(\alpha_2)$*, and the trace is* $\mathbb{Q}$*-linear,* $T_K(a_1\alpha_1 + a_2\alpha_2) = a_1T_K(\alpha_1) + a_2T_K(\alpha_2)$ *for all* $a_1, a_2 \in \mathbb{Q}$.

Lemma 41. *If* $\alpha \in \mathfrak{O}_K$*, then* $N_K(\alpha), T_K(\alpha) \in \mathbb{Z}$.

4.1.3 The ring of integers

In this section, we consider the algebraic structure that is of the utmost interest in this work, namely the ring of integers of a finite algebraic field extension. We prove that the ring of integers of any finite extension $K : \mathbb{Q}$ is a free abelian group. Furthermore, we give the definition of our first field invariant, the discriminant, and develop a machinery which will allow us to find a $\mathbb{Z}$-basis for this group.

Theorem 42. *Assume $[K : \mathbb{Q}] = n < \infty$. Then $\mathfrak{O}_K$ is a free abelian group of rank n.*

Corollary 43. *Any $\mathbb{Z}$-basis of $\mathfrak{O}_K$ is a $\mathbb{Q}$-basis of K.*

Proof. Firstly, by the free abelian property, the elements of the $\mathbb{Z}$-basis are linearly independent over $\mathbb{Z}$ and hence over $\mathbb{Q}$. Secondly, by Lemma 29, any $\alpha \in K$ can be expressed as a $\mathbb{Q}$-linear combination of a $\mathbb{Z}$-basis of $\mathfrak{O}_K$. □

Combining Theorem 42 with what we already know about discriminants, we can obtain some useful information.

Corollary 44. *The discriminant of a $\mathbb{Z}$-basis of $\mathfrak{O}_K$ does not depend on the choice of basis.*

Proof. The change-of-basis matrices between any two $\mathbb{Z}$-bases of a free abelian group are integer matrices with determinant one. Hence, by the discriminant relation in Lemma 35, the discriminants of any two $\mathbb{Z}$-bases are equal. □

Due this corollary, we define the following

Definition 45. The *discriminant Δ_K of a finite extension $K : \mathbb{Q}$* is the discriminant of any $\mathbb{Z}$-basis of $\mathfrak{O}_K$.

Lemma 46. *For any finite extension $K : \mathbb{Q}$, $\Delta_K \in \mathbb{Z} \setminus \{0\}$.*

Knowing quite a lot about the ring of integers $\mathfrak{O}_K$, the next step is to show that it can be found explicitly.

Algorithm 47. *There is an algorithm with the following properties. The algorithm is given any basis of K consisting of elements in $\mathfrak{O}_K$ and the monomorphisms $\sigma_i : K \to \mathbb{C}$. In a finite number of steps, the algorithm will find a $\mathbb{Z}$-basis of $\mathfrak{O}_K$.*

4.1.4 Dirichlet Unit Theorem and the regulator

In this subsection, we study the multiplicative unit group $(\mathfrak{O}_K^*, \cdot, 1)$ of $\mathfrak{O}_K$. The algebraic structure of this group is determined by the Dirichlet Unit Theorem, and the density of the units is measured by the regulator, our second field extension invariant.

Theorem 48 (Dirichlet Unit Theorem)**.** *Let $K : \mathbb{Q}$ be a finite algebraic extension with s real monomorphisms $K \to \mathbb{R}$ and $2t$ truly complex*[1] *monomorphisms $K \to \mathbb{C}$. Then, the multiplicative unit group $(\mathfrak{O}_K^*, \cdot, 1)$ of $\mathfrak{O}_K$ has $s + t - 1$ fundamental units η_j such that any $x \in \mathfrak{O}_K^*$ can be expressed as $x = \zeta\eta_1^{r_1}...\eta_{s+t-1}^{r_{s+t-1}}$, where $\zeta \in \mathfrak{O}_K^*$ is some complex root of unity and the exponents r_j are unique.*

Remark 49. The roots of unity are the torsion subgroup T of $\mathfrak{O}_K^*$, so a group-theoretical formulation of the theorem would be that the quotient group $(\mathfrak{O}_K^*/T, \cdot, 1)$ is free abelian of rank $(s + t - 1)$.

Remark 50. The Dirichlet unit theorem only tells the number of the fundamental units. Finding them is another non-trivial task.

Definition 51. The *logarithmic embedding matrix* $\mathbf{A} \in \mathbb{R}^{(s+t)\times(s+t-1)}$of Dirichlet units $\eta_1, ..., \eta_{s+t-1}$ is given elementwise by

$$\mathbf{A}_{ij} = d_i \log |\sigma_i(\eta_j)|, \tag{4.5}$$

where $d_i = 1$ if σ_i is a real monomorphism and $d_i = 2$ otherwise. The embeddings σ_i are labelled so that there are no conjugate pairs in $1 \leq i \leq s + t$.

Lemma 52. *The value of $|\det \mathbf{A}_{i_0}|$ for a submatrix $\mathbf{A}_{i_0} \in \mathbb{R}^{(s+t-1)\times(s+t-1)}$ obtained by deleting the i_0^{th} row of $\mathbf{A}$ does not depend on the choice of i_0, neither on the system of fundamental units or the choice of indexing of the embeddings and the fundamental units.*

Proof sketch. The norm of a unit is plus or minus one, so

$$\sum_{i=0}^{s+t} d_i \log |\sigma_i(\eta_j)| = \log(|N_K(\eta_j)|) = 0. \tag{4.6}$$

Hence, all the columns of $\mathbf{A}$, considered as vectors in $\mathbb{R}^{s+t}$, lie on the zero-mean hyperplane $H := \{\mathbf{x} \in \mathbb{R}^{s+t} | \sum_{i=0}^{s+t} x_i = 0\}$. Geometrically, deleting

[1] Note that the truly complex monomorphisms exist in conjugate pairs, so they must be even in number.

the i_0^{th} row of $\mathbf{A}$ and corresponds to orthogonally projecting the column vectors to the subspace $x_{i_0} = 0$ in $\mathbb{R}^{s+t}$. Consider the $(s+t)$-dimensional volume V of the parallellotope in $\mathbb{R}^{s+t}$ spanned by the column vectors of $\mathbf{A}$ and $\mathbf{e}_{i_0}$. Then, $V = |\det \mathbf{A}_{i_0}| =$[$(s+t-1)$-dimensional volume of the unit parallellotope spanned by the column vectors in H]$\cos(\angle(H, \mathbf{e}_{i_0}))$. The plane H is symmetric with respect to the axes, so $\cos(\angle(H, \mathbf{e}_{i_0}))$ is independent of i_0 □

This lemma ensures that the following definition is not ambiguous in the sense that it would depend on our indexing i of monomorphisms σ_i and logarithmic embeddings l_i.

Definition 53. The *regulator* R_K of a finite algebraic extension $K : \mathbb{Q}$ is defined as $|\det \mathbf{A}_1|$.

Remark 54. As calculated in the proof of Lemma 52, $R_K =$ is proportional to the $(s+t-1)$-dimensional volume of the unit parallellotope spanned by the column vectors of $\mathbf{A}$ in H. Hence, the regulator measures the density of the additive abelian group lattice generated by the fundamental units in the logarithmic space.

The above interpretation is quantized by the following theorem [8]. We point out that contrary to all other omitted proofs in this section, the proof of this theorem is not easy unless we skip the term $(cR_K^{2/r})^{-1}$ in the exponent.

Theorem 55. *Let K be a degree n extension of $\mathbb{Q}$ with s real and $2t$ (non-real) complex embeddings $\sigma_i : K \to \mathbb{C}$, and let $r := s+t-1 \geq 2$. Then, as $q \to \infty$,*

$$\begin{aligned} &\#\{x \in \mathfrak{O}_K^\times : \max_{1\leq i\leq n} |\sigma_i(x)| < q\} \\ = \quad & \frac{w(r+1)}{R_K r!}(\log q)^r + \mathcal{O}(\log q)^{r-1-(cR_K^{2/r})^{-1}}, \end{aligned}$$

where w is the number of roots of unity in K, R_K is the regulator and $c = 6 \cdot 2 \cdot 10^{12} n^{10}(1 + 2\log n)$ is a constant depending only on the degree of the extension $K/\mathbb{Q}$.

4.2 Ideals and factorizations

In this chapter, we consider the problem of factorizing elements of the ring of integers $\mathfrak{O}_K$ of a finite algebraic extension $K : \mathbb{Q}$. This problem became relevant in solving diophantine equations (or, actually, often in proving the non-existence of a solution) and, in particular, in attempting to prove the famous Fermat's last theorem. However, in our case, it will have its motivation in bounding the number of elements in $\mathfrak{O}_K$ of a given norm. It will turn out, that a unique factorization into irreducibles is not always possible, but it is always possible to factorize ideals.

As in the previous sections, many proofs are only cited.

4.2.1 Introduction: non-unique factorization in the ring of integers

Definition 56. Consider a field extension $K : \mathbb{Q}$. Then, $a \in \mathfrak{O}_K$ is *irreducible* if all its factorizations $a = bc$, where $b, c \in \mathfrak{O}_K$, satisfy $b \in \mathfrak{O}_K^*$ or $c \in \mathfrak{O}_K^*$. The element a is *prime* if $ab = cd$, where $b, c, d \in \mathfrak{O}_K$ implies $ab' = c$ or $ab' = d$, with some $b' \in \mathfrak{O}_K$.

It can be shown that a prime element is always irreducible, but the converse does not hold. Another important difference between the two concepts is that a prime factorization is unique, whereas a factorization into irreducibles is not. Both of these properties are illustrated by the following example.

Example 57. Consider $K = \mathbb{Q}(\sqrt{-5})$, with $\mathfrak{O}_K = \mathbb{Z}[\sqrt{-5}]$, and the factorization of $6 \in \mathfrak{O}_K$. Now, it is easy to come up with two alternative factorizations

$$6 = (1 - \sqrt{-5})(1 + \sqrt{-5}) = 2 \cdot 3.$$

We are going to show that both these factorizations are irreducible and hence, the factorization into irreducibles is not unique. Furthermore, we show that the irreducible element $(1 + \sqrt{-5})$ is not prime.

The norm in K given by

$$N(a + b\sqrt{-5}) = a^2 + 5b^2.$$

Hence, the norms of the respective factors above are $6, 6, 4$, and 9. Recall that the norm is multiplicative by Lemma 40 and integral by Lemma 41. Now, assume that any of the factors in the two factorizations has a factorization, *e.g.*, $2 = (a + b\sqrt{-5})(c + d\sqrt{-5})$, where $a, b, c, d \in \mathbb{N}$. It is easy to show that the integers $6, 6, 4$, and 9 have no nontrivial factors of the form $N(a + b\sqrt{-5}) = a^2 + 5b^2$, where $a, b \in \mathbb{N}$. Hence, we must have $N(c + d\sqrt{-5}) = 1$, equivalently, $(c + d\sqrt{-5})$ is a unit. Then, by definition, the factors above are irreducible.

For irreducible elements not being necessarily prime, consider again the above two factorizations. If $(1 + \sqrt{-5})$ was prime, then we shold have $2 =$

$x(1+\sqrt{-5})$ or $3 = x(1+\sqrt{-5})$ with some $x \in \mathfrak{O}_K$. But, using norm multiplicativity, this would require $4 = 6N(x)$ or $9 = 6N(x)$, both impossible for $x \in \mathfrak{O}_K$.

As a conclusion, a general ring of integers $\mathfrak{O}_K$ does not exhibit the pleasant properties of $\mathbb{Z} = \mathfrak{O}_{\mathbb{Q}}$, where all irreducibles are primes and, equivalently, all factorizations into irreducibles are unique up to order and units. Rings of integers or more generally any *domains* with this $\mathbb{Z}$-like property are called *unique factorization domains* (UFDs). Furthermore, by now we have no good tools for finding a factorization into irreducibles or even proving that a given factorization is irreducible. Working with norms as in the example above will in general lead to diophantine equations that might be very difficult to solve. It will turn out that to proceed to at least partially answer these questions, the theory of ideals will be useful.

4.2.2 General theory: ideal products, fractional and prime ideals

This subsection considers a more general theory, which will later on be applied to algebraic number fields. The aim of this section is to develop the theory of ideal products and prove that in an algebraic structure called a *Dedekind domain*, a generalization of rings of integers, ideal products exhibit the unique factorization property not generally present in the rings of integers, hence motivating the concept of *prime ideals.*

We point out that the concept of Dedekind domains and the theorem on the existence of unique prime-ideal factorization are in a way superfluous. This in the sense that the existence of a unique prime-ideal factorization in rings of integers will also follow from the theorem and algorithm given later that will find the prime-ideal decomposition. The abstract approach is still given for mainly two reasons. First, the explicit construction of the prime-ideal decomposition takes some ten pages, and it would be strange to work on such a lengthy construction if there were no guarantee of the results actually existing. Second, the theorem and definition are given in an abstract setting and they answer to the question *why* a unique prime-ideal decomposition exists in rings of integers, and in what other algebraic structures prime-ideal decompositions will exist.

We start with some general theory of rings and ideals.

Definition 58. The *product ideal* $\mathfrak{ab}$ of two ideals $\mathfrak{a}$ and $\mathfrak{b}$ of a commutative ring is defined as

$$\mathfrak{ab} = \langle \{ab | a \in \mathfrak{a}, b \in \mathfrak{b}\}\rangle$$

Remark 59. If generator-representations $\mathfrak{a} = \langle \cup_{i \in I} a_i \rangle$ and $\mathfrak{b} = \langle \cup_{j \in J} b_j \rangle$ are at hand, then the product ideal can be expressed as $\mathfrak{ab} = \langle \cup_{i \in I, j \in J} a_i b_j \rangle$, and its elements are of the form $a_1 b_1 + ... + a_n b_n$.

Definition 60. A *prime ideal* $\mathfrak{a}$ has the property $\mathfrak{bc} \subseteq \mathfrak{a} \Rightarrow \mathfrak{b} \subseteq \mathfrak{a}$ or $\mathfrak{c} \subseteq \mathfrak{a}$.

Theorem 61. *Let $\mathfrak{a}$ be an ideal of a commutative ring R. Then,*

i) $\mathfrak{a}$ is maximal if and only if $R/\mathfrak{a}$ is a field.

ii) $\mathfrak{a}$ is prime if and only if $R/\mathfrak{a}$ is an integral domain.

Proof. The proof is given in [29, Lemma 5.1], or in [22]. □

The importance of the preceding theorem is that it is typically easier to examine the algebraic structures of the quotients that the ideals themselves. As an example, it yields directly the following sufficient and easy-to-check condition for an ideal to be prime, which would not be as easy to establish directly from the definition of a prime ideal.

Corollary 62. *A maximal ideal of a commutative ring is prime.*

Definition 63. A ring R is *noetherian* if every ideal of R is finitely generated.

Lemma 64. *The following are equivalent for a ring R:*

i) R is noetherian

ii) all ascending chains of strict ideal inclusions $I_1 \subsetneq I_2 \subsetneq ...$ are finitely long

iii) all families of ideals have a maximal element w.r.t. the inclusion relation.

Proof. The direction (iii) ⇒ (ii) is easy. The converse (ii) ⇒ (iii) follows by applying Zorn's lemma on a family of ideals. Hence, we have to prove (i) ⇔ (ii).

For (ii) ⇒ (i), assume the ascending chain condition. Then, let I be any ideal and $x_i \in I$ a set of elements defined inductively such that $x_i \notin \langle x_1, ..., x_{i-1} \rangle$. Then, by the ascending chain condition, the ascending chain $\langle x_1 \rangle \subsetneq \langle x_1, x_2 \rangle \subsetneq ...$ cannot be continued infinitely. Hence, there exists n such that $I \setminus \langle x_1, ...x_n \rangle = \emptyset$, proving that I is finitely generated.

For (i) ⇒ (ii), assume that the ring is noetherian and take an ascending chain $I_1 \subsetneq I_2 \subsetneq ....$ Then, by the basic ideal test, the union of these ascending ideals is an ideal, hence finitely generated, $\cup_{i \geq 1} I_i = \langle x_1, ..., x_m \rangle$. But then, for $1 \leq k \leq m$ there exist $j(k)$ such that $x_j \in I_{j(k)}$ and thus the chain must end at $\max_{1 \leq k \leq m} j(k)$. □

Theorem 65. *In a noetherian commutative ring, every element has a factorization into irreducibles.*

Proof sketch. If an element x could be iteratively factorized into non-units infinitely many times as $x = x_1y_1 = x_1x_2y_2 = ...$, then this would provide an infinite ascending chain of ideals $\langle x\rangle \subsetneq \langle y_1\rangle \subsetneq \langle y_2\rangle \subsetneq ...$ □

For the remainder of this subsection, we consider the special case of rings called Dedekind domains. We point out the following notations.

Notation 66. Throughout this subsection, we assume $\mathfrak{O}$ to be a Dedekind domain and K its field of fractions. Ideals are denoted with gothic lower-case letters such as $\mathfrak{a}$, $\mathfrak{b}$, $\mathfrak{c}$. The letter $\mathfrak{p}$ is reserved for prime ideals. Fractional ideals are denoted by capital gothic such as $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$ (except $\mathfrak{O}$). Multiplication of a set and an element is defined as $c\mathfrak{a} = \cup_{a\in\mathfrak{a}} ca$.

Definition 67. An integral domain $\mathfrak{O}$ with the field of fractions K is a *Dedekind domain* if it satisfies all the following conditions:

i) noetherianity

ii) if $\alpha \in K$ is a root of a monic polynomial $p(t) \in K[t] \cap \mathfrak{O}[t]$, then $\alpha \in \mathfrak{O}$

iii) every non-zero prime ideal of $\mathfrak{O}$ is maximal.

Remark 68. Together with Theorem 65, condition (i) implies that a factorization into irreducibles exists in Dedekind domains and hence, as we shall soon see, in rings of integers.

Condition (iii) above is the converse of the general result in Corollary 62, so it could also be stated and is typically used in the form “in Dedekind domains, maximality and primality of a non-trivial ideal are equivalent”.

The following theorem gaurantees that Dedekind domains are of interest in algebraic number theory.

Theorem 69. *Let $K : \mathbb{Q}$ be a finite extension. Then, $\mathfrak{O}_K$ is a Dedekind domain.*

We give the proof here. The proof is based on the following lemma.

Lemma 70. *Assume $[K : \mathbb{Q}] = n$. Then, all non-zero ideals I of $\mathfrak{O}_K$ are free additive abelian groups of rank n.*

Proof. $\mathfrak{O}_K$ is a free abelian group by Lemma 42, and a subgroup of a free abelian is always free abelian. The rank is since if $0 \neq x \in I$, then the ideal $\langle x \rangle = x\mathfrak{O}_K$ is an additive subgroup of I. But if $\{\omega_i\}_{i=1}^n$ is an integer basis of $\mathfrak{O}_K$, then $\{x\omega_i\}_{i=1}^n$ is easily proven to be a $\mathbb{Z}$-basis of $x\mathfrak{O}_K$. Hence, I has the subgroup $x\mathfrak{O}_K$ which is of maximal rank in $\mathfrak{O}_K$ and is thus itself also a maximal-rank subgroup of $\mathfrak{O}_K$. □

Proof of Theorem 69. Condition (i) follows directly from Lemma 70, since any ideal is generated by its n-element $\mathbb{Z}$-basis.

Condition (ii) is a re-statement of Lemma 25.

Condition (iii): let I be a non-zero prime ideal of $\mathfrak{O}_K$, hence of maximal rank by Lemma 70. By basic theory of free abelian groups (see, *e.g.*, [29, Theorem 1.7]), the qoutient rings of maximal-rank ideals such as $\mathfrak{O}_K/I$ is finite. On the other hand, by Theorem 61 (i), $\mathfrak{O}_K/I$ is a domain. But all finite domains are fields, so by Theorem 61 (ii), I is maximal. □

Definition 71. An $\mathfrak{O}$-submodule $\mathfrak{A}$ of K is a *fractional ideal*, if there exists $0 \neq c \in \mathfrak{O}$ such that $c\mathfrak{A} \subseteq \mathfrak{O}$.

Remark 72. A useful equivalent formulation of the definition is that $\mathfrak{A} \subseteq K$ is a fractional ideal if and only if $\mathfrak{A} = c^{-1}\mathfrak{b}$ for some ideal $\mathfrak{b}$ of $\mathfrak{O}$ and $c \in \mathfrak{O}$. Then, $\mathfrak{A}' \subseteq \mathfrak{A}$ implies $\mathfrak{A}' = c^{-1}\mathfrak{b}'$ where $\mathfrak{b}' \subseteq \mathfrak{b}$.

Definition 73. The product of two fractional ideals $\mathfrak{A} = c^{-1}\mathfrak{b}$ and $\mathfrak{A}' = c'^{-1}\mathfrak{b}'$ is defined using the product of ideals as

$$\mathfrak{A}\mathfrak{A}' = (cc')^{-1}\mathfrak{b}\mathfrak{b}'.$$

The following lemma is very easy, but it will be used repeatedly, which is why we point it out in the beginning.

Lemma 74. *A fractional ideal of K is an ideal of $\mathfrak{O}$ if and only if it is contained in $\mathfrak{O}$.*

Proof. 'Only if' is obvious. 'If' is since the set $\mathfrak{A} = c^{-1}\mathfrak{b}$ is an additive group because $\mathfrak{b}$ is one, and for any $x \in \mathfrak{O}_K$, we have $x\mathfrak{A} = c^{-1}x\mathfrak{b} \subseteq c^{-1}\mathfrak{b} = \mathfrak{A}$. □

The following theorem is the main result of this subsection.

Theorem 75. *Let $\mathfrak{O}$ be a Dedekind domain and $\mathfrak{d}$ an ideal of $\mathfrak{O}$. Then, there exists a decomposition of $\mathfrak{d}$ into a product of finitely many prime ideals. This decomposition is unique up to order.*

Proof. For the ease of reading, the proof is given in steps. The general idea is to define the inverse of an ideal and show that for prime ideals, $\mathfrak{p}\mathfrak{p}^{-1} = \mathfrak{O}$. Then, one proves the existence of a prime-ideal decomposition of $\mathfrak{d}$ by counter-assumption and the use of the maximal and prime ideal $\mathfrak{d} \subset \mathfrak{p}$. These are done in steps (i), (v), and (vi), respectively, with steps (ii)–(iv) containing lemmas and (vi) proving the uniqueness of the prime-ideal decomposition.

i) *Define an inverse of an ideal* $\mathfrak{a} \subseteq \mathfrak{O}$ *as* $\mathfrak{a}^{-1} = \{x \in K | x\mathfrak{a} \subseteq \mathfrak{O}\} \subseteq K$. *Then, if* $\mathfrak{a} \subseteq \mathfrak{b}$, *we have* $\mathfrak{b}^{-1} \subseteq \mathfrak{a}^{-1}$. *Furthermore,* $\mathfrak{a}^{-1}$ *is a fractional ideal, and so is* $\mathfrak{a}^{-1}\mathfrak{b}$ *for any ideal* $\mathfrak{b}$ *of* $\mathfrak{O}$

The first property follows directly from the definition. The second one follows since $\mathfrak{a}^{-1}$ is easily proven to be an $\mathfrak{O}$-submodule of K and, taking any $a \in \mathfrak{a}$, we have $a\mathfrak{a}^{-1} \subseteq \mathfrak{O}$. The third part is identical to the second one.

ii) *If* $\mathfrak{a} \neq \mathfrak{O}$, *we have a strict inclusion* $\mathfrak{O} \subsetneq \mathfrak{a}^{-1}$.

Before proving this, we need the following lemma.

iii) *Given any non-trivial ideal* $\mathfrak{a}$, *there exist finitely many prime ideals* $\mathfrak{p}_1, ..., \mathfrak{p}_r$ *such that* $\mathfrak{p}_1 ... \mathfrak{p}_r \subseteq \mathfrak{a}$

If $\mathfrak{a}$ is prime, we are done. Otherwise, assume on the contrary that for some non-prime ideals $\mathfrak{a}_i$ there does not exist such prime ideals. Since the ring $\mathfrak{O}$ is noetherian, using Lemma 64 (iii), we can take take a maximal ideal $\mathfrak{a}$ amongst such ideals $\mathfrak{a}_i$. By the non-primality, there exist $\mathfrak{b}$ and $\mathfrak{c}$ such that $\mathfrak{b}\mathfrak{c} \subseteq \mathfrak{a}$, $\mathfrak{b} \not\subseteq \mathfrak{a}$, $\mathfrak{c} \not\subseteq \mathfrak{a}$. Then, it follows that the ideal $(\mathfrak{a}+\mathfrak{b})(\mathfrak{a}+\mathfrak{c})$ is contained in $\mathfrak{a}$, which is in turn strictly contained in both $(\mathfrak{a}+\mathfrak{b})$ and $(\mathfrak{a}+\mathfrak{c})$. By the maximality of $\mathfrak{a}$, the latter two contain some prime-ideal products,

$$\begin{aligned} \mathfrak{p}_1 ... \mathfrak{p}_s &\subseteq (\mathfrak{a}+\mathfrak{b}) \\ \mathfrak{p}_{s+1} ... \mathfrak{p}_r &\subseteq (\mathfrak{a}+\mathfrak{c}), \end{aligned}$$

but this would imply

$$\Rightarrow \mathfrak{p}_1 ... \mathfrak{p}_r \subseteq (\mathfrak{a}+\mathfrak{b})(\mathfrak{a}+\mathfrak{c}) \subseteq \mathfrak{a},$$

a contradiction.

We now return to the proof of part (ii).

Recall that primality and maximality of an ideal are equivalent in Dedekind domains. Hence, $\mathfrak{a} \subseteq \mathfrak{p}$ for some prime ideal $\mathfrak{p}$ and thus by part (i), $\mathfrak{p}^{-1} \subseteq \mathfrak{a}^{-1}$. Hence, it suffices to prove that we have a strict inclusion $\mathfrak{O} \subset \mathfrak{p}^{-1}$.

Now, take $a \in \mathfrak{p}$ and hence $\langle a \rangle \subseteq \mathfrak{p}$. By part (ii), we have prime ideals

$$\mathfrak{p}_1 ... \mathfrak{p}_r \subseteq \langle a \rangle \subseteq \mathfrak{p}.$$

But then, using inductively the primality of $\mathfrak{p}$, we can assume $\mathfrak{p}_1 \subseteq \mathfrak{p}$, and hence by maximality of $\mathfrak{p}_1$, we have $\mathfrak{p}_1 = \mathfrak{p}$. If we now choose r to be the smallest possible amount of ideals such that

$$\mathfrak{p}_1 ... \mathfrak{p}_r \subseteq \langle a \rangle,$$

then it follows that

$$\mathfrak{p}_2 ... \mathfrak{p}_r \not\subseteq \langle a \rangle.$$

(If $r = 1$, then use $\mathfrak{O} \not\subseteq \langle a \rangle$.) Now choose any $b \in \mathfrak{p}_2 ... \mathfrak{p}_r \setminus \langle a \rangle$, so $a^{-1}b \notin \mathfrak{O}$. Then, $a^{-1}b\mathfrak{p} \subseteq a^{-1}\mathfrak{p}_1 ... \mathfrak{p}_r = a^{-1}\langle a \rangle = \mathfrak{O}$, and hence by definition $a^{-1}b \in \mathfrak{p}^{-1} \setminus \mathfrak{O}$. This completes the proof of part (ii).

iv) *Assume that $\theta \in K$ satisfies $\theta\mathfrak{a} \subseteq \mathfrak{a}$ for some non-zero ideal $\mathfrak{a}$ of $\mathfrak{O}$. Then, $\theta \in \mathfrak{O}$.*

By noetherianity, $\mathfrak{a} = \langle a_1, ..., a_m \rangle$. Then, since $\theta\mathfrak{a} \subseteq \mathfrak{a}$, there is a matrix $A \in \mathfrak{a}^{m \times m} \subset K^{m \times m}$ such that

$$\begin{pmatrix} a_1\theta \\ \vdots \\ a_m\theta \end{pmatrix} = A \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}.$$

But since the vector $(a_1, ..., a_m)$ is non-zero, considering this as linear algebra of the field K implies that

$$\det(A - \theta) = 0.$$

This charachteristic equation is monic with coefficients in $\mathfrak{O}$, and hence $\theta \in \mathfrak{O}$ by the axoim (ii) of Dedekind domains.

v) *For a prime ideal $\mathfrak{p}$, $\mathfrak{p}\mathfrak{p}^{-1} = \mathfrak{O}$.*

The product $\mathfrak{p}\mathfrak{p}^{-1}$ is by part (i) a fractional ideal and by definition contained in $\mathfrak{O}$. Then, it is an ideal by Lemma 74. Since $1 \in \mathfrak{p}^{-1}$, we have the chain of ideals

$$\mathfrak{p} \subseteq \mathfrak{p}\mathfrak{p}^{-1} \subseteq \mathfrak{O},$$

and by the maximality of $\mathfrak{p}$, one equality holds. But if it was the first one, then by part (iv) we had $\mathfrak{p}^{-1} \subseteq \mathfrak{O}$, contradicting part (ii), so we must have $\mathfrak{p}\mathfrak{p}^{-1} = \mathfrak{O}$

vi) *All ideals* $\mathfrak{d}$ *can be written as a product of prime ideals.*

Assume that this was not true, and, using noetherianity, take a maximal ideal $\mathfrak{d}$ amongst those contradicting the existence of prime-ideal decomposition. We now construct a suitable ideal stricly larger that $\mathfrak{d}$. First, we find the ideal by starting from $\mathfrak{d} \subsetneq \mathfrak{p}$ for some maximal and hence prime $\mathfrak{p}$ (which is a prime-ideal decomposition itself and hence not $\mathfrak{d}$). Using part (i),

$$\mathfrak{d}\mathfrak{p}^{-1} \subseteq \mathfrak{d}\mathfrak{d}^{-1} \subseteq \mathfrak{O},$$

where the latter inclusion is directly from the definition of $\mathfrak{d}^{-1}$. But using part (i) again, $\mathfrak{d}\mathfrak{p}^{-1}$ is a fractional ideal and by the above contained in $\mathfrak{O}$, hence an ideal by Lemma 74.

Next, we prove the strict inclusion $\mathfrak{d} \subsetneq \mathfrak{d}\mathfrak{p}^{-1}$. By part (ii), there exists an element $\theta \in \mathfrak{p}^{-1} \setminus \mathfrak{O}$. Applying part (iv) on this element, we have $\theta\mathfrak{d} \nsubseteq \mathfrak{d}$, so $\mathfrak{d}\mathfrak{p}^{-1} \nsubseteq \mathfrak{d}$. Since on the other hand $1 \in \mathfrak{p}^{-1}$, it follows that $\mathfrak{d} \subseteq \mathfrak{d}\mathfrak{p}^{-1}$ and finally $\mathfrak{d} \subsetneq \mathfrak{d}\mathfrak{p}^{-1}$.

Now that we have constructed the larger ideal $\mathfrak{d}\mathfrak{p}^{-1}$, we use the counter-assumption. By the maximality of $\mathfrak{d}$, the ideal $\mathfrak{d}\mathfrak{p}^{-1}$ has a prime-ideal decomposition,

$$\mathfrak{d}\mathfrak{p}^{-1} \quad = \quad \mathfrak{p}_1 ... \mathfrak{p}_r.$$

Multiplying by $\mathfrak{p}$ and using part (v),

$$\mathfrak{d}\mathfrak{p}^{-1}\mathfrak{p} = \mathfrak{d}\mathfrak{O} = \mathfrak{d} = \mathfrak{p}_1 ... \mathfrak{p}_r \mathfrak{p},$$

a contradiction. Hence, any ideal $\mathfrak{d}$ of $\mathfrak{O}$ indeed possesses a prime-ideal decomposition.

We now have the existence of a prime-ideal decomposition. For the uniqueness, we need the following highly useful lemma, which is labelled separately for later reference.

Lemma 76. *Ideal inclusion* $\mathfrak{b} \subseteq \mathfrak{a}$ *and factorization* $\mathfrak{b} = \mathfrak{a}\mathfrak{c}$*, denoted* $\mathfrak{a}|\mathfrak{b}$*, are equivalent.*

Proof. The direction $\mathfrak{b} = \mathfrak{a}\mathfrak{c} \Rightarrow \mathfrak{b} \subseteq \mathfrak{a}$ is immediate from the definition of an ideal.

The converse direction is proved by considering the fractional ideal $\mathfrak{b}\mathfrak{a}^{-1} \subseteq \mathfrak{a}\mathfrak{a}^{-1} \subseteq \mathfrak{O}$, which is an ideal by Lemma 74. Then, taking $\mathfrak{c} = \mathfrak{b}\mathfrak{a}^{-1}$ yields immediately $\mathfrak{b} = \mathfrak{a}\mathfrak{c}$ □

Remark 77. Lemma 76 means that in Dedekind domains, the definition of a prime ideal can be equivalently formulated as "$\mathfrak{p}$ is prime if it has the property that $\mathfrak{p}|\mathfrak{ab} \Rightarrow \mathfrak{p}|\mathfrak{a}$ or $\mathfrak{p}|\mathfrak{b}$". This is what one would intuitively call primality.

We now return to the final part of Theorem 75.

vii) *The prime-ideal decomposition is unique up to order.*

By the product definition of prime ideals in Remark 77, each prime ideal must appear in both decompositions. □

We point out that it could be proven that the inverse ideal $\mathfrak{a}^{-1}$ as defined above satisfies $\mathfrak{a}\mathfrak{a}^{-1} = \mathfrak{O}$ also for non-prime ideals $\mathfrak{a}$ [see [29], Proof 5.5 (vi)]. However, from the existence of prime-ideal decomposition and inverses of prime ideals, we have the following result.

Theorem 78. *The set $\mathcal{F}$ of fractional ideals is an abelian group $(\mathcal{F}, \cdot, \mathfrak{O}_K)$.*

Also the following lemma is useful and easy to prove using Lemma 76.

Corollary 79. *The smallest ideal in the sense of inclusions (equivalently , the one having the most prime-ideal factors) of all ideals dividing the two ideals $\mathfrak{a}$ and $\mathfrak{b}$ is denoted by $\gcd(\mathfrak{a}, \mathfrak{b})$ and given by $\mathfrak{a} + \mathfrak{b}$. The largest ideal in the sense of inclusions (the one having the least prime-ideal factors) divisible by the ideals $\mathfrak{a}, \mathfrak{b}$ is denoted $\operatorname{lcm}(\mathfrak{a}, \mathfrak{b})$ and is given by $\mathfrak{a} \cap \mathfrak{b}$.*

This corollary implies immediately the following theorem, which is not important in this work, but a classic of algebraic number theory.

Theorem 80 (Chinese remainder theorem)**.** *Let $\mathfrak{a}$ and $\mathfrak{b}$ be two coprime ideals of $\mathfrak{O}$. Then, for every $x \in \mathfrak{O}$ there exists $b \in \mathfrak{b}$ such that $x \equiv b$ mod $\mathfrak{a}$. Furthermore, the set of solutions $b \in \mathfrak{b}$ are the $\mathfrak{ab}$-translates of one solution.*

Proof. By the corollary above, we have $\mathfrak{O} = \gcd(\mathfrak{a}, \mathfrak{b}) = \mathfrak{a} + \mathfrak{b}$ since the ideals are coprime. But $\mathfrak{O} = \mathfrak{a} + \mathfrak{b}$ is a re-statement of the first part. Similarly, the second part follows from $\mathfrak{ab} = \operatorname{lcm}(\mathfrak{a}, \mathfrak{b}) = \mathfrak{a} \cap \mathfrak{b}$. □

We yet return to rings of integers and consider very briefly the concept of an ideal norm.

Definition 81. The norm of a non-zero ideal $\mathfrak{a}$ of $\mathfrak{O}_K$ is defined as $N(\mathfrak{a}) = |\mathfrak{O}_K/\mathfrak{a}|$.

Remark 82. By Lemma 70 and the basic theory of free abelian groups, the ideal norm is always finite.

Lemma 83. *The norm of an ideal is multiplicative, i.e., $N(\mathfrak{a}\mathfrak{b}) = N(\mathfrak{a})N(\mathfrak{b})$.*

Proof. This is given in [29, Theorem 5.12]. □

Corollary 84. *An ideal $\mathfrak{a}$ with $N(\mathfrak{a}) \in \mathbb{P}$ is prime.*

Lemma 85. *The ideal norm of a principal ideal in $\mathfrak{O}_K$ is the element norm of its generator, $N(\langle x \rangle) = |N_K(x)|$.*

Proof. This is given in [29, Corollary 5.10]. □

4.2.3 The Dedekind zeta function

We return to algebraic number fields. As a model problem related to the problems we study the Dedekind zeta function. For the final purposes of this thesis work, the zeta function will have no other use than simply motivating the study of ideal norms and prime ideals at this point. However, in a larger perspective, the functions has relevance from the point of view of the very same communication problem. It has even been suggested (cf. [11]), that the Dedekind zeta function could be used for optimizing the ECDP in algebraic lattices. However, the Dedekind zeta function seems not to serve as a very good estimate for the ECDP. Very recently, some more elaborate but related estimates based on the related ideal class zeta function have been established [13].

Definition 86. The *Dedekind zeta function* ζ_K of a finite algebraic field extension $K : \mathbb{Q}$ is defined as

$$\zeta_K(s) = \sum_{\mathfrak{a}} \frac{1}{N(\mathfrak{a})^s},$$

where the summation goes over all ideals.

The following lemmas will first show the connection of the Dedekind zeta function to the prime ideals considered in the preceding chapter and to the ideals $p\mathfrak{O}_K$ generated by rational primes p, which will be considered in the next chapter. Together, these lemmas will also allow proving the convergence of the zeta function.

Lemma 87. *The Dedekind zeta function has an Euler product representation. The convergence of these two representations is equivalent and they are given as*

$$\zeta_K(s) = \prod_{\mathfrak{p}} \left(1 - \frac{1}{N(\mathfrak{p})^s}\right)^{-1},$$

where the product goes over all prime ideals.

Proof. The proof is analogous to proving the Euler product form of the Riemann zeta function. For details, we refer to, *e.g.*, [21]. □

We state and prove the following lemma first in a general setting since it is a more general property.

Lemma 88. *Let $\mathfrak{O}_K$ be a commutative ring and $\mathfrak{p}$ its prime ideal with $[\mathfrak{O}_K/\mathfrak{p}] < \infty$. Then,*

i) there is exactly one prime p such that $\langle p\rangle \subseteq \mathfrak{p}$,

ii) The norm of the ideal is a power this prime, $[\mathfrak{O}_K/\mathfrak{p}] = p^k$,

iii) If $\mathfrak{O}_K$ is a free abelian group of rank n, then $k \leq n$.

Proof. i) By Theorem 61, $\mathfrak{O}_K/\mathfrak{p}$ is a domain, and a finite domain is a field. Now, given $q \in \mathbb{P}$, $\langle q\rangle \subseteq \mathfrak{p}$ would imply that every element of $\mathfrak{O}_K/\mathfrak{p}$ would have an additive order dividing q. But the additive order of every non-zero element is the characteristic of the field, hence a prime p.

ii)By Cauchy's theorem for abelian groups, if $[\mathfrak{O}_K/\mathfrak{p}]$ had any prime factor q other than p, then in $\mathfrak{O}_K/\mathfrak{p}$ there would exist an element with additive order q, contradicting well-definedness of the characteristic p.

iii) $\mathfrak{p}/p\mathfrak{O}_K$ is an additive subgroup of $\mathfrak{O}_K/p\mathfrak{O}_K$, which has the order p^n in rank n free abelian groups $\mathfrak{O}_K$. Hence, by Lagrange's theorem, we have $k \leq n$. □

Using Lemmas 42 and 76, the preceding Lemma is easily seen to take the following form.

Lemma 89. *Let $K : \mathbb{Q}$ be an algebraic extension of degree n and $\mathfrak{p}$ a prime ideal of $\mathfrak{O}_K$. Then, the ideal norm $N(\mathfrak{p})$ is a power of a rational prime p^k, $1 \leq k \leq n$. Furthermore, p is the only rational prime such that $\mathfrak{p}|\langle p\rangle$.*

This seemingly innocent lemma has remarkable implications.

Corollary 90. *For ideals of $\mathfrak{O}_K$ with $[K : \mathbb{Q}] = n$, the following hold:*

i) The prime ideals form equivalence classes according to which rational-prime generated ideals they divide. In each such equivalence class, assume $\mathfrak{p}_i$ with the norm $N(\mathfrak{p}_i) = p^{f_i}$ factors $\langle p\rangle = \prod_{i=1}^{g} \mathfrak{p}_i^{e_i}$ with multiplicity e_i. Then, $\sum_{i=1}^{g} f_i e_i = n$.

ii) The prime ideals of $\mathfrak{O}_K$ are countably infinite in number.

As a continuation to part (i) of Cor. 90, we make the following terminological definition.

Definition 91. Consider the prime-ideal factorization of a rational-prime generated ideal $\langle p \rangle$ of $\mathfrak{O}_K$ with $[K : \mathbb{Q}] = n$. We say that the rational prime p

i) *ramifies*, if some of the prime ideals dividing it has multiplicity larger than one,

ii) *ramifies totally*, if it is an n^{th} power of a rational-prime normed prime ideal, $\langle p \rangle = \mathfrak{p}^n$ with $N(\mathfrak{p}) = p$,

iii) *splits*, if it has at least two distinct prime-ideal factors,

iv) *splits totally*, if it is a product of n distinct rational-prime normed prime ideals, $\langle p \rangle = \prod_{i=1}^{n} \mathfrak{p}_i$ with $N(\mathfrak{p}_i) = p$,

v) is *inert*, if $\langle p \rangle$ is a prime ideal itself.

With the previously introduced results, the following result becomes easy.

Corollary 92. *The Dedekind zeta series converges for $\Re s > 1$.*

Proof. For any complex series, absolute convergence implies convergence. Hence, note that the absolute-value series, can be expressed as

$$\sum_{\mathfrak{a}} \left| \frac{1}{N(\mathfrak{a})^s} \right| = \sum_{\mathfrak{a}} \frac{1}{N(\mathfrak{a})^{\Re s}} = \zeta_K(\Re s).$$

This means that actually convergence for real $s > 1$ implies convergence for all complex s with $\Re s > 1$.

Consider now the contribution to the Euler representation by prime ideals whose norm is a power of a fixed rational prime p. Using the notation of part (i) of Corollary 90, this is given by

$$\prod_{N(\mathfrak{p})|p} \left(1 - \frac{1}{N(\mathfrak{p})^s}\right)^{-1} = \prod_{i=1}^{g} \left(1 - \frac{1}{p^{f_i s}}\right)^{-1}.$$

Here $1 \leq g \leq n$ and $1 \leq f_i \leq n$. Hence, we can give an upper bound by taking a maximal number of maximal factors all larger than one, given by

$$\begin{aligned} \left| \prod_{i=1}^{g} \left(1 - \frac{1}{p^{f_i s}}\right)^{-1} \right| &\leq \left(1 - \max_i \left| \frac{1}{p^{s f_i}} \right| \right)^{-n} \\ &\leq \left(1 - \frac{1}{p^{s(\min_i f_i)}}\right)^{-n} \\ &= \left(1 - \frac{1}{p^s}\right)^{-n}. \end{aligned}$$

By Cor. 90, part (i) the Euler representation can be upper bounded by taking into account this upper bound for all $p \in \mathbb{P}$. This gives

$$\begin{aligned} |\zeta_K(s)| &= \prod_{p\in\mathbb{P}} \left| \prod_{N(\mathfrak{p})|p} \left(1 - \frac{1}{N(\mathfrak{p})^s}\right)^{-1} \right| \\ &\leq \prod_{p\in\mathbb{P}} \left(1 - \frac{1}{p^s}\right)^{-n} \\ &= \zeta(s)^n, \end{aligned}$$

where ζ is the ordinary Riemann zeta function, whose Euler product was used in the last step. Since the Riemann zeta function converges for $\Re s > 1$, as is easily proven using the harmonic series, this proves the claim. □

From the proof of the preceding corollary we extract the following definition which is useful in numerical evaluations of the Dedekind zeta function.

Definition 93. The local Euler factor $L_{p,K}$ of the Dedekind zeta function for an extension $K : \mathbb{Q}$ at prime p is the function $\mathbb{C} \to \mathbb{C}$

$$L_{p,K}(s) := \prod_{N(\mathfrak{p})|p} \left(1 - \frac{1}{N(\mathfrak{p})^s}\right)^{-1},$$

so that whenever the zeta function converges, it has the convergent product representation

$$\zeta_K(s) = \prod_{p\in\mathbb{P}} L_{p,K}(s).$$

4.3 Finding the prime ideals explicitly

In this section, we give a theorem and an algorithm sufficient to yield all prime ideals of a ring of integers, based on the prime-ideal decomposition of a rational-prime generated ideal $p\mathfrak{O}_K$ of $\mathfrak{O}_K$. The theorem given in the first subsection applies for all but finitely many p and gives the generators of $p\mathfrak{O}_K$. For the remaining cases, the algorithm given in the second subsection will construct a basis of $p\mathfrak{O}_K$ computationally efficiently.

4.3.1 Almost all cases: Dedekind's theorem

After obtaining some knowledge on the convergence of the Dedekind zeta function, the step towards any numerical evaluations is to find the prime ideals of any equivalence class of Cor. 90, part (i). The next theorem is also due to Dedekind, and given a finite field extension $K : \mathbb{Q}$, it yields the prime-ideal decomposition of any rational-prime generated ideal for all but pinitely many rational primes p.

Theorem 94 (Dedekind)**.** *Let $K = \mathbb{Q}(\theta)$ be an algebraic extension with $\theta \in \mathbb{B}$ and p a prime not dividing*[2] $[\mathfrak{O}_K : \mathbb{Z}[\theta]]$. *Consider the prime-ideal factorization $\langle p \rangle = \prod_{i=1}^{g} \mathfrak{p}_i^{e_i}$ of $\langle p \rangle$ with $N(\mathfrak{p}_i) = p^{f_i}$. Denoting the minimal polynomial of θ by $m(t) \in \mathbb{Z}[t]$ and its decomposition into distinct irreducible factors in $\mathbb{Z}_p[t]$ as $m(t) \equiv \prod_{i=1}^{g_p} \pi_i(t)^{\epsilon_i}$. Then, with a suitable indexing of the ideals $\mathfrak{p}_i$ the following hold.*

i) The number of prime factors $\pi_i(t)$ equals that of the prime factors $\mathfrak{p}_i$, i.e., $g = g_p$.

ii) The exponents f_i are given by $f_i = \partial \pi_i(t)$.

iii) The multiplicity of a prime-ideal factor equals that of an irreducible polynomial factor, i.e., $e_i = \epsilon_i$.

iv) The generator-representation of the prime ideals in this indexing is $\mathfrak{p}_i = \langle \Pi_i(\theta), p \rangle$, where $\Pi_i(t) \in \mathbb{Z}[t]$ is any polynomial such that $\Pi_i(t) \equiv \pi_i(t) (\mathrm{mod}\ p)$

Along with the statement of this theorem, some remarks on the assumptions and applicability tests are useful.

Remark 95. By the lemmas 31 and 29, the statenent $K = \mathbb{Q}(\theta), \theta \in \mathbb{B}$ is not an assumption but simply a notational matter whenever $[K : \mathbb{Q}] < \infty$. Similarly, theorems 75 and 69 guarantee that the ideal $\langle p \rangle$ has a unique prime-ideal factorization. Contrast to these, the assumption that p does not divide $[\mathfrak{O}_K : \mathbb{Z}[\theta]]$ does provide a real restriction. Our subsequent efforts will be on trying to circumvent this problem. Later in this subsection, we will provide both sufficient and equivalent conditions for this to hold. Interestingly, neither one of these requires knowledge of $\mathfrak{O}_K$ (which is cumbersome to calculate by hand). In the next subsection, we will prove a general algorithm for finding the prime-ideal generators.

Next, we are about to prove Theorem 94. Before the actual proof we yet state the necessary ring isomorphism theorem for the convenience of the reader.

[2] This ought to be considered as a quotient of additive groups; $\mathbb{Z}[\theta]$ is not an ideal but a subring.

Theorem 96 (Third Isomorphism Theorem of rings). *Let R be a ring and I and J its two-sided ideals with $I \subseteq J$. Then,*

$$R/I \Big/ J/I \cong R/J$$

and the isomorphism map $R/I \Big/ J/I \to R/J$ is given explicitly by $(r+I) + J/I \mapsto r + J$.

Proof of Theorem 94. We will split the proof into steps for the ease od reading. Throughout the proof, denote $l = [\mathfrak{O}_K : \mathbb{Z}[\theta]]$ and reduction modulo p by an overbar. Pricipal ideals of a ring R generated by $r \in R$ are denoted rR instead of $\langle r \rangle$ as usually in this text, since we will work with several different rings.

Before starting the steps, we will modify the statement of the thorem. Note that $\mathfrak{p}_i$ factors, equivalently contains $p\mathfrak{O}_K$ and is prime, equivalently maximal. Hence using Theorem 61 (i),

$$\begin{array}{rcl} \mathfrak{p}_i \text{ is a maximal ideal of } \mathfrak{O}_K & \Leftrightarrow & \mathfrak{O}_K/\mathfrak{p}_i \text{ is a field} \\ \text{(Third Isomorphism Theorem)} & \cong & \mathfrak{O}_K/p\mathfrak{O}_K \Big/ \mathfrak{p}_i/p\mathfrak{O}_K, \text{ hence also a field} \\ & \Leftrightarrow & \mathfrak{p}_i/p\mathfrak{O}_K \text{ is a maximal ideal of } \mathfrak{O}_K/p\mathfrak{O}_K. \end{array}$$

Hence, our objects of interest are the maximal ideals of $\mathfrak{O}_K/p\mathfrak{O}_K$. This and analogous calculations are in continuation called the canonical and maximality-preserving one-to-one correspondence between the ideals J/I of R/I and the ideals J of R containing I.

i) *We have an isomorphism $\mathbb{Z}[\theta]/p\mathbb{Z}[\theta] \cong \mathfrak{O}_K/p\mathfrak{O}_K$.*

For an equivalence class $\alpha + p\mathbb{Z}[\theta]$ of $\mathbb{Z}[\theta]/p\mathbb{Z}[\theta]$, there is a ring homomorphism $\varphi : \alpha + p\mathbb{Z}[\theta] \mapsto \alpha + p\mathfrak{O}_K \in \mathfrak{O}_K/p\mathfrak{O}_K$. This homomorphism does not depend on which representative α is chosen, since $p\mathbb{Z}[\theta] \subseteq p\mathfrak{O}_K$.

We claim that this is an isomorphism. We start by showing injectivity. Take $\alpha + p\mathbb{Z}[\theta] \in \ker(\varphi)$. Then, the representative $\alpha \in \mathbb{Z}[\theta]$ satisfies $\alpha = px \in p\mathfrak{O}_K$ for some $x \in \mathfrak{O}_K$. But then, the element $x + \mathbb{Z}[\theta]$ of the additive group $\mathfrak{O}_K/\mathbb{Z}[\theta]$ has an order dividing p since $p(x + \mathbb{Z}[\theta]) = \alpha + \mathbb{Z}[\theta] = 0 + \mathbb{Z}[\theta]$. By Lagrange's theorem for (additive) groups, the order of an element of $\mathfrak{O}_K/\mathbb{Z}[\theta]$ must divide $[\mathfrak{O}_K/\mathbb{Z}[\theta]]$. Hence, this order is 1. But then $x \in \mathbb{Z}[\theta]$ and hence $\alpha = px \in p\mathbb{Z}[\theta]$ and finally $\alpha + p\mathbb{Z}[\theta] = 0 + p\mathbb{Z}[\theta]$, so our arbitrary element of the kernel is the zero element.

Next, we show surjectivity. Take any $x \in \mathfrak{O}_K$. Then, by Lagrange's theorem for the additive group $\mathfrak{O}_K/\mathbb{Z}[\theta]$, we have $l(x+\mathbb{Z}[\theta]) = 0+\mathbb{Z}[\theta]$, *i.e.*, $lx \in \mathbb{Z}[\theta]$. On the other hand, since $p \not| l$, l has an inverse $l' + p\mathbb{Z}$ in $\mathbb{Z}_p$, *i.e.*, $ll' = 1 + kp, k \in \mathbb{Z}$. But now we can take $l'lx \in \mathbb{Z}[\theta]$ and map it to $\varphi(l'lx + \mathbb{Z}[\theta]) = x + kpx + p\mathfrak{O}_K = x + \mathfrak{O}_K$. This shows the surjectivity.

ii) *We have the isomorphism* $\mathfrak{O}_K/p\mathfrak{O}_K \cong \mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t]$.

This follows using $\mathbb{Z}[\theta] \cong \mathbb{Z}[t]/m(t)\mathbb{Z}[t]$ as follows. First, note that

$$p(\mathbb{Z}[\theta]) \cong p(\mathbb{Z}[t]/m(t)\mathbb{Z}[t]) = \frac{p\mathbb{Z}[t] + m(t)\mathbb{Z}[t]}{m(t)\mathbb{Z}[t]}. \tag{4.7}$$

So from part (i) we obtain

$$\begin{aligned}
\mathfrak{O}_K/p\mathfrak{O}_K &\cong \mathbb{Z}[\theta]/p\mathbb{Z}[\theta] \\
&\cong \mathbb{Z}[t]/m(t)\mathbb{Z}[t] \Big/ p(\mathbb{Z}[t]/m(t)\mathbb{Z}[t]) \\
&= \mathbb{Z}[t]/m(t)\mathbb{Z}[t] \Big/ (p\mathbb{Z}[t] + m(t)\mathbb{Z}[t])/(m(t)\mathbb{Z}[t]) \\
\text{(Third Isomorphism Theorem)} &\cong \mathbb{Z}[t]/(p\mathbb{Z}[t] + m(t)\mathbb{Z}[t]) \\
\text{(Third Isomorphism Theorem)} &\cong \mathbb{Z}[t]/p\mathbb{Z}[t] \Big/ (p\mathbb{Z}[t] + m(t)\mathbb{Z}[t])/p\mathbb{Z}[t] \\
&\cong \mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t].
\end{aligned}$$

iii) *The maximal ideals of* $\mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t]$ *are those generated by the equivalence classes* $\pi_i(t) + \overline{m}(t)\mathbb{Z}_p[t]$ *irreducible factors* $\pi_i(t)$ *of* $\overline{m}(t)$.

There is a canonical and maximality-preserving one-to-one correspondence between the ideals of $\mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t]$ and the ideals of $\mathbb{Z}_p[t]$ containing $\overline{m}(t)\mathbb{Z}_p[t]$. We regard it as known from the theory of polynomial rings that the maximal ideals of $\mathbb{Z}_p[t]$ containing $\overline{m}(t)\mathbb{Z}_p[t]$ are exactly $\pi_i(t)\mathbb{Z}_p[t]$. Their canonical correspondents in $\mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t]$ are

$$(\pi_i(t)\mathbb{Z}_p[t] + \overline{m}(t)\mathbb{Z}_p[t])/\overline{m}(t)\mathbb{Z}_p[t] = \pi_i(t)(\mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t]).$$

For the rest of the proof we assume that the prime ideals are indexed so that the isomorphic image of $\mathfrak{p}_i/p\mathfrak{O}_K$ in $\mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t]$ is $\pi_i(t)(\mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t])$. This is what is meant by 'a suitable indexing' in the statement of the theorem.

iv) *The number of prime factors* $\pi_i(t)$ *equals that of the prime factors* $\mathfrak{p}_i$, *i.e.*, $g = g_p$.

This follows directly from (iii).

v) *The exponents are given by $f_i = \partial\pi_i(t)$.*

We have the following chain of isomorphisms

$$\begin{array}{rcl} & & \mathfrak{O}_K/\mathfrak{p}_i \\ \text{(Third Isomorphism Theorem)} & \cong & \mathfrak{O}_K/p\mathfrak{O}_K \Big/ \mathfrak{p}_i/p\mathfrak{O}_K. \end{array}$$

Next, using parts (ii) and (iii) of the proof and recalling that by the Third Isomorphism Theorem, $\mathfrak{p}_i/p\mathfrak{O}_K$ is a maximal ideal of $\mathfrak{O}_K/p\mathfrak{O}_K$, we obtain

$$\begin{array}{rcl} \mathfrak{O}_K/\mathfrak{p}_i & \cong & \mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t] \Big/ (\pi_i(t)\mathbb{Z}_p[t] + \overline{m}(t)\mathbb{Z}_p[t])/\overline{m}(t)\mathbb{Z}_p[t] \\ & \cong & \mathbb{Z}_p[t]/(\pi_i(t)\mathbb{Z}_p[t] + \overline{m}(t)\mathbb{Z}_p[t]) \\ & = & \mathbb{Z}_p[t]/\pi_i(t)\mathbb{Z}_p[t], \end{array}$$

Where the second step applied the Third Isomorphism Theorem again, and the third steps uses the fact that $\pi_i(t)$ is a factor of $\overline{m}(t)$. Finally, using this isomorphism of the rings, we immediately obtain $p^{f_i} = |\mathfrak{O}_K/\mathfrak{p}_i| = |\mathbb{Z}_p[t]/\pi_i(t)\mathbb{Z}_p[t]| = p^{\partial\pi_i(t)}$.

vi) *The multiplicity e_i of a prime-ideal factor equals that of its isomorphic image in $\mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t]$, i.e., the multiplicity ϵ_i of an irreducible polynomial factor.*

Recall that py part (iii), under the isomorphism between $\mathfrak{O}_K/p\mathfrak{O}_K$ and $\mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t]$, the maximal ideals $\mathfrak{p}_i/p\mathfrak{O}_K$ become $\pi_i(t)\mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t]$. Now, denote $\overline{m}(t) = \prod_{i=1}^{g} \pi_i(t)^{\epsilon_i}$ and consider

$$\begin{array}{rcl} (\mathfrak{p}_i/p\mathfrak{O}_K)^k & \cong & \langle \pi_i(t) + \overline{m}(t)\mathbb{Z}_p[t]\rangle^k \\ & = & \langle \pi_i(t)^k + \overline{m}(t)\mathbb{Z}_p[t]\rangle \\ & = & \langle \gcd(\pi_i(t)^k, \overline{m}(t)) + \overline{m}(t)\mathbb{Z}_p[t]\rangle \\ & = & \langle \pi_i(t)^{\min\{k,\epsilon_i\}} + \overline{m}(t)\mathbb{Z}_p[t]\rangle \end{array}$$

Hence, the ideal $\mathfrak{p}_i/p\mathfrak{O}_K$ has ϵ_i distinct powers. (Note that all the ϵ_i powers of $\pi_i(t)$ are different modulo $\overline{m}(t)$.)

On the other hand, using the definition of a product ideal, $(\mathfrak{p}_i/p\mathfrak{O}_K)^k$ can be expressed by its generators in $\mathfrak{O}_K/p\mathfrak{O}_K$, giving

$$\begin{aligned}(\mathfrak{p}_i/p\mathfrak{O}_K)^k &= \langle \cup_{\{a_j\}_{j=1}^k \subset \mathfrak{p}_i} \prod_{j=1}^{k} (a_j + p\mathfrak{O}_K)\rangle \\ &= \langle \cup_{\{a_j\}_{j=1}^k \subset \mathfrak{p}_i} (\prod_{j=1}^{k} a_j + p\mathfrak{O}_K)\rangle\end{aligned}$$

The $\mathfrak{O}_K$-representatives of the generators are

$$\begin{aligned} & \cup_{b\in p\mathfrak{O}_K} \cup_{\{a_j\}_{j=1}^k \subset \mathfrak{p}_i} (\prod_{j=1}^{k} a_j + b) \\ \text{(Definition of ideal sum)} &= \mathfrak{p}_i^k + p\mathfrak{O}_K \\ \text{(Lemma 79)} &= \gcd(\mathfrak{p}_i^k, p\mathfrak{O}_K) \\ &= \mathfrak{p}_i^{\min\{e_i,k\}}.\end{aligned}$$

Hence, the product ideal can be expressed as

$$\begin{aligned}(\mathfrak{p}_i/p\mathfrak{O}_K)^k &= \langle \mathfrak{p}_i^{\min\{e_i,k\}}/p\mathfrak{O}_K\rangle \\ \text{(the generators are an ideal of } \mathfrak{O}_K/p\mathfrak{O}_K) &= \mathfrak{p}_i^{\min\{e_i,k\}}/p\mathfrak{O}_K.\end{aligned}$$

So the ideal $\mathfrak{p}_i/p\mathfrak{O}_K$ has e_i distinct powers. Thus, $e_i = \epsilon_i$.

vii) *The generator-representation of the prime ideals dividing $p\mathfrak{O}_K$ are $\mathfrak{p}_i = \langle \Pi_i(\theta), p\rangle$.*

We have the generators of the maximal ideals of the ring $\mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t]$ isomorphic to $\mathfrak{O}_K/p\mathfrak{O}_K$. All we have to do is to go backwards the chain of isomorphisms. We recall that all isomorphisms except for the last one and the two first ones (which are the first one and the two last ones now that we work backwards) are due to the Third Isomorphism Theorem

$$\begin{tikzcd}
\pi_i(t)+\overline{m}(t)\mathbb{Z}_p[t] \arrow[d, mapsto] & \in & \mathbb{Z}_p[t]/\overline{m}(t)\mathbb{Z}_p[t] \arrow[d] \\
(\Pi_i(t)+p\mathbb{Z}[t])+[(p\mathbb{Z}[t]+m(t)\mathbb{Z}[t])/p\mathbb{Z}[t]] \arrow[d, mapsto] & \in & \mathbb{Z}[t]/p\mathbb{Z}[t]\Big/(p\mathbb{Z}[t]+m(t)\mathbb{Z}[t])/p\mathbb{Z}[t] \arrow[d] \\
\Pi_i(t)+(p\mathbb{Z}[t]+m(t)\mathbb{Z}[t]) \arrow[d, mapsto] & \in & \mathbb{Z}[t]/(p\mathbb{Z}[t]+m(t)\mathbb{Z}[t]) \arrow[d] \\
(\Pi_i(t)+m(t)\mathbb{Z}[t])+ (p\mathbb{Z}[t]+m(t)\mathbb{Z}[t])/m(t)\mathbb{Z}[t] \arrow[d, mapsto] & \in & \mathbb{Z}[t]/m(t)\mathbb{Z}[t]\Big/(p\mathbb{Z}[t]+m(t)\mathbb{Z}[t])/m(t)\mathbb{Z}[t] \arrow[d] \\
\Pi_i(\theta)+p\mathbb{Z}[\theta] \arrow[d, mapsto] & \in & \mathbb{Z}[\theta]/p\mathbb{Z}[\theta] \arrow[d] \\
\Pi_i(\theta)+p\mathfrak{O}_K & \in & \mathfrak{O}_K/p\mathfrak{O}_K.
\end{tikzcd}$$

Hence, we have $\mathfrak{p}_i/p\mathfrak{O}_K = \langle \Pi_i(\theta)+p\mathfrak{O}_K\rangle$. This can be lifted from $\mathfrak{O}_K/p\mathfrak{O}_K$ to $\mathfrak{O}_K$ as $\mathfrak{p}_i = \langle \Pi_i(\theta), p\rangle$.

□

Lemma 97. *If p^2 does not divide $\Delta(1,\theta,...,\theta^{n-1})$, then p cannot divide $[\mathfrak{O}_K : \mathbb{Z}[\theta]]$.*

Proof. Denote the $\mathbb{Z}$-basis of $\mathfrak{O}_K$ by $\beta_1,...,\beta_n$, giving a unique coefficient matrix $\mathbf{M} \in \mathbb{Z}^{n\times n}$ such that

$$\begin{pmatrix} 1 \\ \theta \\ \vdots \\ \theta^{n-1} \end{pmatrix} = \mathbf{M} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}. \tag{4.8}$$

But by the Lemma 35 and the definition of Δ_K this implies $\Delta(1,\theta,...,\theta^{n-1}) = (\det \mathbf{M})^2\Delta_K$. Next, from the theory of free abelian groups, $[\mathfrak{O}_K : \mathbb{Z}[\theta]] = \det \mathbf{M}$ and by Lemma 46, $\Delta_K \in \mathbb{Z}\setminus\{0\}$. Hence, $\Delta(1,\theta,...,\theta^{n-1}) = [\mathfrak{O}_K : \mathbb{Z}[\theta]]^2\Delta_K$ is a product of two integers and if p divides $[\mathfrak{O}_K : \mathbb{Z}[\theta]]$ then p^2 divides $\Delta(1,\theta,...,\theta^{n-1})$. □

The condition above for p not dividing $[\mathfrak{O}_K : \mathbb{Z}[\theta]]$ is sufficient but not necessary. However, the preceding lemma gives a finite number of primes p that can possibly divide $[\mathfrak{O}_K : \mathbb{Z}[\theta]]$, after which the following equivalence-form condition is available. The following theorem it a computationally efficient condition, and it can also be used to yield a computationally efficient way to find the ring of integers $\mathfrak{O}_K$.

Notation 98. Define the following notation for the next theorem: $m(x) \in \mathbb{Z}[x]$ is the monic polynomial of θ and $^{-}$ denotes reduction modulo p. Let $\pi_i \in \mathbb{Z}_p[x]$ be the distinct irreducible factors of $\bar{m}(x)$, *i.e.*,

$$\bar{m}(x) = \prod_{i=1}^{g} \pi_i(x)^{e_i}.$$

Let $\Pi_i(x) \in \mathbb{Z}[x]$ be any monic polynomials such that $\Pi_i(x) \equiv \pi_i(x) \pmod{p}$. Define

$$\begin{aligned} l(x) &= \prod_{i=1}^{g} \Pi_i(x) \\ h(x) &= \prod_{i=1}^{g} \Pi_i(x)^{e_i-1} \end{aligned}$$

so that $l(x)h(x)$ is a monic lift of $\bar{m}(x)$. Finally, set

$$f(x) = \frac{l(x)h(x) - m(x)}{p} \in \mathbb{Z}[x]. \tag{4.9}$$

Theorem 99 ([3], Theorem 6.1.4.(2))**.** *Consider a finite field extension $K = \mathbb{Q}(\theta)$ of $\mathbb{Q}$, where $\theta \in \mathfrak{O}_K$. With the notation defined above, a prime p divides $[\mathfrak{O}_K : \mathbb{Z}[\theta]]$ if and only if $\gcd(\bar{f}(x), \bar{l}(x), \bar{h}(x)) \neq 1$ in $\mathbb{Z}_p[x]$.*

4.3.2 Special cases: the Buchmann-Lenstra algorithm

Dedekind's theorem provides an easy way to compute the prime ideals dividing $\langle p \rangle$ as long as p does not factor $[\mathfrak{O}_K : \mathbb{Z}[\theta]]$. However, it might of course happen that $p | [\mathfrak{O}_K : \mathbb{Z}[\theta]]$. There might still exist, a lazy mathematician could argue, a value $\theta' \in \mathfrak{O}_K$ such that $\mathfrak{O}_K = \mathbb{Z}[\theta']$, but this trick will not always work.

Example 100. (This is an example due to Dedekind and reproduced by Stein in [27], Section 5.3.1, with computer algebra calculations.) Let θ be a root of $x^3 + x^2 - 2x + 8$ and $K = \mathbb{Q}(\theta)$. Then, the ideal $\langle 2 \rangle$ of $\mathfrak{O}_K$ has three distinct prime-ideal factors. Now, if there existed any $\theta' \in \mathfrak{O}_K$ such that the ideal $\langle 2 \rangle$ could be computed by Dedekind's theorem, then the minimal polynomial m of θ' should split into three distinct and hence first-degree factors over $\mathbb{Z}_2$, but $\mathbb{Z}_2$ only has two distinct first-degree polynomials.

Hence the only way to deal with this is to really find a method other that that presented in Theorem 94. The final answer to the question of how to find the prime-ideal decomposition for the problematic primes is algorithmic and due to not Dedekind but Buchmann and Lenstra.

Algorithm 101 (Buchmann-Lenstra)**.** *Let $K = \mathbb{Q}(\theta)$ be a finite extension. For any ideal $\langle p \rangle$ of $\mathfrak{O}_K$, there is an algorithm to find the two-generator representations and the exponents of the ideals in the prime-ideal decomposition of $\langle p \rangle$, given only the minimal polynomial of θ and the $\mathbb{Z}$-basis of $\mathfrak{O}_K$.*

We point out that for evaluating the Dedekind zeta series, this algorithm is an overkill. Two ways to optimize the numerical computations are presented in Remark 108. However, this algorithm yields a complete theory with not much extra work.

By the following theorem, we denote extensions of $\mathbb{Z}_p$ by $\mathbb{Z}_p(\alpha)$, but actually all the computations in our proof of the algorithm do not require the knowledge that there is a particular primitive element.

Theorem 102 (Primitive element theorem)**.** *Let $F : E$ be a finite field extension. Then, there is an element $\alpha \in F$ such that $F = E(\alpha)$ if and only if the number of intermediate fields $F : H : E$ is finite.*

In addition, the following definition and lemma are separated from the proof of the algorithm.

Definition 103. We define the p-radical I_p of $\mathfrak{O}_K$ as

$$I_p = \{x \in \mathfrak{O}_K | x^m \in p\mathfrak{O}_K \text{ for some } m \in \mathbb{N}\}.$$

Lemma 104. *Denote the prime-ideal pactorization of $\langle p \rangle$ as $\langle p \rangle = \prod_{i=1}^{g} \mathfrak{p}_i^{e_i}$. Then, I_p is the ideal product of the distinct prime ideals, $I_p = \prod_{i=1}^{g} \mathfrak{p}_i$.*

Proof. Let us first show the inclusion $I_p \subseteq \prod_{i=1}^{g} \mathfrak{p}_i$. Take $x \in I_p$. Then by definition $x^m \in \langle p \rangle$ and since ideal inclusion and factorization are equivalent, we have $\mathfrak{p}_i \supseteq \langle p \rangle$ for all prime factors, so $x \in \cap_{i=1}^{g} \mathfrak{p}_i$. Finally, since the prime ideals $\mathfrak{p}_i$ are distinct, Cor. 79 yields $\cap_{i=1}^{g} \mathfrak{p}_i = \prod_{i=1}^{g} \mathfrak{p}_i \ni x$.

Now we prove the inclusion $I_p \supseteq \prod_{i=1}^{g} \mathfrak{p}_i$. Take any $x \in \prod_{i=1}^{g} \mathfrak{p}_i$. Next, take $m = \max_{1 \leq i \leq g} e_i$ and $x^m \in \prod_{i=1}^{g} \mathfrak{p}_i^m = \langle p \rangle \prod_{i=1}^{g} \mathfrak{p}_i^{m-e_i} \subseteq \langle p \rangle$. Hence, $x^m \in \langle p \rangle$. □

Remark 105. The preceding lemma implies in particular that I_p is an ideal and that $I_p^{\max_i e_i} \subseteq \langle p \rangle$ and consequently $x \in I_p \Leftrightarrow x^n \in \langle p \rangle$ (note that $\max_i e_i \leq n$ so $x^n \in I_p^{\max_i e_i} \subseteq \langle p \rangle$).

One more easy lemma should be extracted from the proof of the algorithm.

Lemma 106. *Let $\mathfrak{a}$ be any ideal factoring $\langle p \rangle$. Then, all non-zero elements of $\mathfrak{O}_K / \mathfrak{a}$ and $\mathfrak{a} / \langle p \rangle$ have an additive order of p.*

Proof. Firtsly, since $\mathfrak{a} \supseteq \langle p \rangle = p\mathfrak{O}_K$ by Lemma 76, then in particular for any $[a] = a + \mathfrak{a} \in \mathfrak{O}_K/\mathfrak{a}$ we have $p[a] = pa + \mathfrak{a} = [0]_{\mathfrak{O}_K/\mathfrak{a}}$ since $pa \in p\mathfrak{O}_K \subseteq \mathfrak{a}$. But by Lagrange's theorem for the additive group $\mathfrak{O}_K/\mathfrak{a}$ this implies that $\mathrm{ord}([a])|p$. Hence, using the primality of p, all elements other than $[0]_{\mathfrak{O}_K/\mathfrak{a}}$ have order p. The proof for $\mathfrak{a}/\langle p \rangle$ is identical. □

This lemma and Remark 105 allow us to show the following computational result.

Algorithm 107. *Given $\mathfrak{O}_K$, there is a computationally efficient way to find I_p.*

Proof. Lemma 106 implies that the $\mathbb{Z}$-basis of $\mathfrak{O}_K$ is the $\mathbb{Z}_p$-basis of $\mathfrak{O}_K/\langle p \rangle$. On the other hand, in commutative rings R where $p = 0$, $R \ni x \mapsto x^p \in R$ is a ring homomorphism $R \to R$, called the Frobenius homomorphism. Hence, $x \mapsto x^p$ is a homomorphism $\mathfrak{O}_K/\langle p \rangle \to \mathfrak{O}_K/\langle p \rangle$ and in particular, it has a $\mathbb{Z}_p^{n\times n}$-matrix representation in the $\mathbb{Z}_p$. This representation is efficient to compute, since it only requires considering the basis elements.

Next, by Remark 105, the quotient $I_p/\langle p \rangle$ is the kernel of the map $\mathfrak{O}_K/\langle p \rangle \ni x \mapsto x^{lp} \in \mathfrak{O}_K/\langle p \rangle$, where $lp \geq n$. But this map has a matrix representation as the l^{th} power of the Frobenius matrix. Hence, $I_p/\langle p \rangle$ is in the basis of $\mathfrak{O}_K$ given by the kernel of this matrix. □

Proof of Algorithm 101. (This algorithm is given in [3], Secs. 6.2.2-6.2.5, with an emphasis on computational matters. Not all subalgorithms such as ideal division and multiplication are discribed here, but the theoretical part is given more explicitly. In practise, the ideal divisions and multiplications are done in the p^n-element ring $R := \mathfrak{O}_K/\langle p \rangle$, since there is a one-to-one maximality-preserving correspondence between the ideals of $\mathfrak{O}_K$ containing $\langle p \rangle$ and the ideals. By Lemma 106, all basis of the different quotient groups considered in this proof can be constructed of elements of $\mathfrak{O}_K$, just recalling that they have additive order p.) For the sake of clarity, the algorithm is divided into steps here.

i) *We can calculate explicitly the bases for a sequence of ideals H_j such that if $\langle p \rangle = \prod_{i=1}^{g} \mathfrak{p}_i^{e_i}$, then $H_j = \prod_{e_i=j} \mathfrak{p}_i$.*

For this part, start from finding the basis for I_p as described in Lemma 107. Next, define the sequences of ideals (for which the product formulation is obtained using Lemmas 104 and 79)

$$K_j = I_p^j + \langle p \rangle = \prod_{i=1}^{g} \mathfrak{p}_i^{\min(e_i,j)}$$

and

$$J_j = K_j K_{j-1}^{-1} = \prod_{e_i \geq j} \mathfrak{p}_i$$

and finally,

$$H_j = J_j J_{j+1}^{-1} = \prod_{e_i = j} \mathfrak{p}_i.$$

Hence, with ideal multiplication and division algorithms with basis output available, this involves no complications.

ii) *The ring $\mathfrak{O}_K/H_j$ (and hence $R/(H_j/\langle p\rangle)$ by the Third Isomorphism Theorem) is isomorphic to a product of fields $\mathbb{Z}_p(\alpha_1)\times ...\times\mathbb{Z}_p(\alpha_k)$, where α_i are roots of irreducible polynomials over $\mathbb{Z}_p$ and k is the number of prime ideals factoring H_j.*

First off, to ease notation, assume that the prime ideals $\mathfrak{p}_i$ factoring H_j are labelled $i = 1, ..., k$ for the remainder of this proof.

Now, for an element of $\mathfrak{O}_K$, denote α, for its homomorphic image in $\mathfrak{O}_K/H_j$ denote $[\alpha]$ and in $\mathfrak{O}_K/\mathfrak{p}_i$ denote $[\alpha]_i$. Consider the map $\varphi : \mathfrak{O}_K/H_j \to \mathfrak{O}_K/\mathfrak{p}_1 \times ... \times \mathfrak{O}_K/\mathfrak{p}_k$ given by $\varphi([\alpha]) = ([\alpha]_1, ..., [\alpha]_k)$. (Since $\mathfrak{p}_i \supseteq H_j$, this does not depend on the choice of representative α, so the map is well-defined.) This is an isomorphism; it is clearly homomorphic and it is injective since its kernel is $\{[\alpha] | \alpha \in \cap_{i=1}^k \mathfrak{p}_i = \prod_{i=1}^k \mathfrak{p}_i = H_j\} = \{0_{\mathfrak{O}_K/H_j}\}$. Finally, it is surjective since the cardinality of the target domain equals that of the domain: $\prod_{i=1}^k |\mathfrak{O}_K/\mathfrak{p}_\mathfrak{i}| = \prod_{i=1}^k N(\mathfrak{p}_i) = N(\prod_{i=1}^k \mathfrak{p}_i) = N(H_j) = |\mathfrak{O}_K/H_j|$. (Alternatively, the surjectivity can be shown using the Chinese Remainder Theorem. Even more, the Chinese Remainder Theorem is actually equivalent to the statement that $\varphi : \mathfrak{O}_K/\mathfrak{p}_1...\mathfrak{p}_k \to \mathfrak{O}_K/\mathfrak{p}_1 \times ... \times \mathfrak{O}_K/\mathfrak{p}_k$ as defined above is an isomorphism.)

To complete this step, note that $\mathfrak{O}_K/\mathfrak{p}_i$ is a finite field by Theorem 61(a), and by Lemma 106 its subfield generated by $[1]_i$ is isomorphic to $\mathbb{Z}_p$. Since a finite field has a finite number of subsets (subfields), we can denote, using Theorem 102 that $\mathfrak{O}_K/\mathfrak{p}_i \cong \mathbb{Z}_p(\alpha_i)$, where α_i is a root of an irreducible polynomial over $\mathbb{Z}_p$ (in an abstract sense, so this is not an element of $\mathfrak{O}_K$).

iii) *Given a $\mathbb{Z}_p$-basis of $R/(H_j/\langle p\rangle)$ (which is quite easily computable as the $\mathbb{Z}_p$-bases of R and $H_j/\langle p\rangle$ are known), there is a computationally efficient way to find $\mathbb{Z}_p$-bases for the isomorphic images of the different fields in the representation given in part (ii).*

Denote

$$A := R/(H_j/\langle p\rangle) \cong \mathbb{Z}_p(\alpha_1) \times ... \times \mathbb{Z}_p(\alpha_k).$$

First off, we provide a test that shows, whether A is a field itself. Consider the map $\psi : A \to A$ given by $\psi(x) = x^p - x = (x^{p-1} - 1)x$. This is linear for rings where $p = 0$ such as A, so we can find its matrix representation in the basis of A. Now, consider ψ in the isomorphic ring $\mathbb{Z}_p(\alpha_1) \times ... \times \mathbb{Z}_p(\alpha_k)$. By Lagrange's theorem for the group $(\mathbb{Z}_p^*, \cdot, 1)$, $(x^{p-1} - 1)x = 0$ for all the p elements of $\mathbb{Z}_p$, so $x^p - x = (x-1)...(x-p)$. Hence, this polynomial has no other zeroes in any of the component fields $\mathbb{Z}_p(\alpha_i)$. Finally, in the notation of $\mathbb{Z}_p(\alpha_1) \times ... \times \mathbb{Z}_p(\alpha_k) \cong A$, we have, using linearity of ψ and the kernels of each component,

$$\ker(\psi) \cong \operatorname{span}_{\mathbb{Z}_p}\{(1,0,...,0),...,(0,...,0,1)\}.$$

Hence, $\dim_{\mathbb{Z}_p} \ker(\psi) = k$. This allows computing k efficiently in the numerical matrix computations.

Next, assume $k \geq 2$. The strategy is to show that using an idempotent element $\epsilon \in A \setminus \{0,1\}$ such that $\epsilon^2 = \epsilon$, we can always find a (non-trivial) splitting of A in the sense

$$A \cong (\mathbb{Z}_p(\alpha_{i_1}) \times ... \times \mathbb{Z}_p(\alpha_{i_{k-l}})) \times (\mathbb{Z}_p(\alpha_{i_{k-l+1}}) \times ... \times \mathbb{Z}_p(\alpha_{i_k})).$$

Hence, working inductively, we can always find the isomorphic images of each $\mathbb{Z}_p(\alpha_i)$. To prove the validity of our strategy, recall first that fields have no non-trivial ideals, so all ideals of A are in the isomorphic ring $\mathbb{Z}_p(\alpha_1) \times ... \times \mathbb{Z}_p(\alpha_k)$ of the from

$$\mathbb{Z}_p(\alpha_{i_1}) \times ... \times \mathbb{Z}_p(\alpha_{i_{k-l}}) \times \{0\} \times ... \times \{0\}.$$

Now, take the two ideals ϵA and $(1-\epsilon)A$ of A, where ϵ is the idempotent. We claim that these ideals will provide the non-trivial splitting in the sense that

$$\begin{aligned} \epsilon A &\cong (\mathbb{Z}_p(\alpha_{i_1}) \times ... \times \mathbb{Z}_p(\alpha_{i_{k-l}})) \times \{0\} \times ... \times \{0\} \\ (1-\epsilon)A &\cong \{0\} \times ... \times \{0\} \times (\mathbb{Z}_p(\alpha_{i_{k-l+1}}) \times ... \times \mathbb{Z}_p(\alpha_{i_k})). \end{aligned}$$

First, the ideals ϵA and $(1-\epsilon)$ are non-trivial, since they are not zero and not equal to A since neither ϵ nor $(1-\epsilon)$ is a unit by $\epsilon(1-\epsilon) = 0$. Second, any $a \in A$ can be written as $a = \epsilon a + (1-\epsilon)a$, so in the view of $\mathbb{Z}_p(\alpha_1) \times ... \times \mathbb{Z}_p(\alpha_k)$, each component $\mathbb{Z}(\alpha_i)$ must be non-zero in at least one of the ideals. Third, $\epsilon A \cap (1-\epsilon)A = \{0\}$, since if $\epsilon a_1 = (1-\epsilon)a_2$, then $\epsilon a_1 = \epsilon^2 a_1 = (\epsilon - \epsilon^2)a_2 = 0$. Hence, in the view of $\mathbb{Z}_p(\alpha_1) \times ... \times \mathbb{Z}_p(\alpha_k)$, each component must be non-zero in at most one of the two ideals. Hence, the ideals indeed privide a splitting. The bases of these ideals are obtained by multiplying the basis elements of A and removing $\mathbb{Z}_p$-linearly dependent elements.

Now it remains to find an idempotent explicitly. Again in the notation of $\mathbb{Z}_p(\alpha_1) \times ... \times \mathbb{Z}_p(\alpha_k)$, 1_A becomes $(1, 1, ..., 1)$. Now, consider $x_A \in \ker(\psi) \setminus \operatorname{span}_{\mathbb{Z}_p}\{1_A\}$, whose correspondent in $\mathbb{Z}_p(\alpha_1) \times ... \times \mathbb{Z}_p(\alpha_k)$ is $x = (\beta_1, ..., \beta_k)$, where $\beta_i \in \mathbb{Z}_p$ and not all equal. The minimal polynomial $m : A \to A$ of x_A, is hence

$$m(t) = \operatorname{lcm}((t - \beta_1 1_A), ..., (t - \beta_k 1_A)) = \prod_{\text{all distinct } \beta_i} (t - \beta_i 1_A)$$

since, in the view of $\mathbb{Z}_p(\alpha_1) \times ... \times \mathbb{Z}_p(\alpha_k)$, every element in the lcm operator makes one of the k components of $m(x_A)$ become zero. Now, it is obvious that whatever factorization $m(t) = m_1(t)m_2(t)$ is considered, the factors are coprime polynomials of $\mathbb{Z}_p[t]$. Hence, by the polynomial gcd algorithm, we can find $U(t), V(t) \in \mathbb{Z}_p[t]$ such that

$$U(t)m_1(t) = 1 - V(t)m_2(t).$$

Finally, choosing ϵ by the above polynomials and the kernel element x_A as

$$\epsilon = U(x_A)m_1(x_A) \in A,$$

we have

$$\epsilon^2 = U(x_A)m_1(x_A)(1 - V(x_A)m_2(x_A)) = \epsilon - U(x_A)V(x_A)m_1(x_A)m_2(x_A) = \epsilon,$$

since $m_1(x_A)m_2(x_A) = m(x_A) = 0$.

iv) *Having found the $\mathbb{Z}_p$-bases for the isomorphic images of the different fields in $A := R/(H_j/\langle p\rangle) \cong \mathbb{Z}_p(\alpha_1) \times ... \times \mathbb{Z}_p(\alpha_k)$, we can find the contribution of $\mathfrak{p}_i|H_j$ to the Dedekind zeta function and the $\mathfrak{O}_K$-generators of the maximal ideals factoring H_j.*

Let us first consider what is needed for approximating the Dedekind zeta function. Note that ideals of $\mathbb{Z}_p(\alpha_1) \times ... \times \mathbb{Z}_p(\alpha_k)$ are products of ideals of each field and that the ideals of a field are trivial. Hence maximal ideals $I_i := (\mathfrak{p}_i/\langle p\rangle)/(H_j/\langle p\rangle)$ of A become in the isomophism

$$\{0\} \times \mathbb{Z}_p(\alpha_2) \times ... \times \mathbb{Z}_p(\alpha_k),$$

or

$$\mathbb{Z}_p(\alpha_1) \times \{0\} \times \mathbb{Z}_p(\alpha_3) \times ... \times \mathbb{Z}_p(\alpha_k)$$

etc. For evaluating the Dedekind zeta function, note that by a furious use of the Third Isomorphism Theorem,

$$\begin{aligned} & \mathfrak{O}_K/\mathfrak{p}_i && (4.10)\\ \cong\ & \mathfrak{O}_K/\langle p\rangle \Big/ \mathfrak{p}_i/\langle p\rangle && (4.11)\\ \cong\ & \left(\mathfrak{O}_K/\langle p\rangle \Big/ H_j/\langle p\rangle\right) \Bigg/ \left(\mathfrak{p}_i/\langle p\rangle \Big/ H_j/\langle p\rangle\right) && (4.12)\\ =\ & A/I_i \cong \mathbb{Z}_p(\alpha_i) && (4.13) \end{aligned}$$

so $N(\mathfrak{p}_i) = |\mathfrak{O}_K/\mathfrak{p}_i| = |\mathbb{Z}_p(\alpha_i)| = p^d$, where d is the $\mathbb{Z}_p$-dimension of the ideal of A isomorphic to $\mathbb{Z}_p(\alpha_i)$. Hence, having the $\mathbb{Z}_p$-bases of these ideals from step (iii), we are done.

For finding the prime ideals of $\mathfrak{O}_K$ explicitly, note that one generating set of $I_i = (\mathfrak{p}_i/\langle p\rangle)/(H_j/\langle p\rangle)$ is the union of the bases of all the fields in the field product representation of I_i. This can be lifted to a generating set of $\mathfrak{p}_i/\langle p\rangle$ as an additive group by choosing representatives of I_i in $\mathfrak{O}_K/\langle p\rangle$ and taking union with the basis of $H_j/\langle p\rangle$. Repeating this idea, the generating set of $\mathfrak{p}_i/\langle p\rangle$ as an additive group lifted to $\mathfrak{O}_K$ together with the basis of $p\mathfrak{O}_K$ generates the additive group (and hence also the ideal) $\mathfrak{p}_i$ in $\mathfrak{O}_K$. Then, Ref. [3] presents Algorithm 4.7.10 that reduces the number of generators of a prime ideal. This algorithm would once again require much theory not presented in this work. We point out that in a theoretical approach, one could find a relatively small set of generators for the ideal $\mathfrak{p}_i$ by using the primitive elements of the isomorphic images of $\mathbb{Z}_p(\alpha_i)$. However, this would be a long proof and computationally inefficient.

□

Remark 108. For numerical evaluations of the Dedekind zeta function, it is actually not necessary to compute steps (iii)–(iv) of the proof, see Ref. [3], exercise 8 for Section 6. Nevertheless, this algorithm is somewhat shorter to prove with the tools at hand.

Another (less efficient) shortcut for Dedekind zeta evaluations that will not essentially change the proof is not to define the ideals H_j, J_j, and K_j or worry about ideal multiplications and divisions but only operate on the p-radical $I_p = \prod_{\mathfrak{p}_i|\langle p\rangle} \mathfrak{p}_i$. Then, just replace H_j by I_p in steps (ii)–(iv) and obtain the values of $|\mathfrak{O}_K/\mathfrak{p}_i|$ as suggested in the first paragraph of step (iv). A way to view this shortcut is that we skip something in the algorithm and the information we lose is the powers e_i, which are irrelevant for computing the local Euler factors.

4.4 Algebraic lattices

As a final word on algebraic number theory, we return to the connection between design of full-diversity lattices and field extensions. We start from a definition.

Definition 109. Let $K : \mathbb{Q}$ be a finite algebraic extension of degree n with integral basis $\{\omega_i\}_{i=1}^n$, real embeddings $\{\sigma_i\}_{i=1}^s$, and truly complex embeddings $\{\sigma_i, \overline{\sigma_i}\}_{i=s+1}^{s+t}$. Then, *the algebraic lattice of* K is the lattice in $\mathbb{R}^n$ generated by

$$M = \begin{pmatrix} \sigma_1(\omega_1) & \dots & \sigma_1(\omega_n) \\ & \vdots & \\ \sigma_s(\omega_1) & \dots & \sigma_s(\omega_n) \\ \Re\sigma_{s+1}(\omega_1) & \dots & \Re\sigma_{s+1}(\omega_n) \\ & \vdots & \\ \Re\sigma_{s+t}(\omega_1) & \dots & \Re\sigma_{s+t}(\omega_n) \\ \Im\sigma_{s+1}(\omega_1) & \dots & \Im\sigma_{s+1}(\omega_n) \\ & \vdots & \\ \Im\sigma_{s+t}(\omega_1) & \dots & \Im\sigma_{s+t}(\omega_n) \end{pmatrix}.$$

Remark 110. The lattice generated by this matrix is independent of the choice of integer basis ω_i, and re-labelling the embeddings σ_i permutes the axes.

It is also worth pointing out that this indeed is a full-rank lattice. First, the colums of M are linearly independent over $\mathbb{Z}$ as required in the definition of lattice generator matrices. For a proof, assume there exist lattice coordinates $\{\alpha_i\}_{i=1}^n$ such that $(M\boldsymbol{\alpha})_j = \sigma_j(\sum_{i=1}^n \alpha_i\omega_i) = 0$ for all $j \leq s$ and $(M\boldsymbol{\alpha})_{s+j} + i(M\boldsymbol{\alpha})_{s+t+j} = \sigma_j(\sum_{i=1}^n \alpha_i\omega_i) = 0$ for all $1 \leq j \leq t$. But $\ker(\sigma_j)$ is trivial, so this implies that $\alpha_i = 0$ for all i. Second, the additive subgroup generated by M is discrete since a suitable product in the coordinates of $M\boldsymbol{\alpha}$ yields the algebraic norm of $\sum_{i=1}^n \alpha_i\omega_i$, and for $\boldsymbol{\alpha} \neq \mathbf{0}$ the norm is a non-zero integer, so lattice vectors $M\boldsymbol{\alpha}$ cannot be arbitrarily short.

Notation 111. Throughout this and the next section, K, ω_i, σ_j, s, t, M, and n will be as in the preceding definition if not otherwise stated.

Recalling the objective of full diversity (see Definition 17 and the asymptotic probability bounds in Section 3.3.4), the following two results will turn our attention to algebraic lattices of totally real field extensions. As a by-product, it is a counterpart of Example 20, showing that algebraic field extensions allow designing lattices of any desired diversity.

Lemma 112. *Let $K : \mathbb{Q}$ have s real and $2t$ truly complex embeddings. Then, the algebraic lattice of K has diversrity $s + t$.*

Proof. Let us represent the points of the algebraic lattice with their lattice coordinates $\{\alpha_i\}_{i=1}^n$. Recall that this representation is unique and in particular, $M\boldsymbol{\alpha} = \mathbf{0} \Leftrightarrow \boldsymbol{\alpha} = \mathbf{0}$. The diversity L is given by $\min_{\boldsymbol{\alpha}\neq\mathbf{0}} l(\mathbf{0}, M\boldsymbol{\alpha})$.

First, we have $L \geq s + t$ since for a real embedding $\sigma_j, 1 \leq j \leq s$, we have $(M\boldsymbol{\alpha})_j = \sigma_j(\sum_{i=1}^n \alpha_i\omega_i)$, which is non-zero for $\boldsymbol{\alpha} \neq \mathbf{0}$. For a complex embedding $\sigma_{s+j}, 1 \leq j \leq t$, at least one of $(M\boldsymbol{\alpha})_{s+j} = \Re\sigma_j(\sum_{i=1}^n \alpha_i\omega_i)$ and $(M\boldsymbol{\alpha})_{s+t+j} = \Im\sigma_j(\sum_{i=1}^n \alpha_i\omega_i)$ is non-zero. Hence, for any $\boldsymbol{\alpha} \neq \mathbf{0}$, at least $s + t$ components of $M\boldsymbol{\alpha}$ are non-zero.

Second, we have $L \leq s + t$ since taking $\boldsymbol{\alpha}$ such that $\sum_{i=1}^n \alpha_i\omega_i = 1$ we have $\sigma_j(\sum_{i=1}^n \alpha_i\omega_i) = 1$ for all j and hence $M\boldsymbol{\alpha} = (1, ..., 1, 0, ...0)$ has $s + t$ non-zero components. □

Corollary 113. *An algebraic lattice is of full diversity if and only if it is totally real, i.e., all the roots of the generating polynomial are real.*

Remark 114. It is worth pointing out that generalized notions of full diversity exist for different lattice code problems, with their respective algebraic constructions. For example, a full-diversity matrix lattice is a discrete additive group of matrices (whose vectorizations hence form a lattice) where no matrix has a zero row. The probability bounds for block fading channels motivate the study of such matrix lattices. Then, full-diversity matrix lattices of two-column matrices in vectorized form can be constructed from algebraic lattices of totally complex extensions; each row of a matrix corresponding to a vector $M\boldsymbol{\alpha}$ is $(\Re\sigma_j(x), \Im\sigma_j(x))$, where $\sum_{i=1}^n \alpha_i\omega_i$, and hence non-zero.

It is also possible to consider relative number-field extensions, *e.g.*, $\mathbb{Q}(\alpha, i) : \mathbb{Q}(i)$, where $\mathbb{Q}(\alpha)$ is totally real. Then we define full diversity as the relative extension degree. Both of these examples are however out of the scope of this thesis.

For the rest of this section, let us study the connection of the number-theoretic properties of totally real extensions and the geometric properties of their lattices. Let $K : \mathbb{Q}$ be such a field extension and Λ the lattice. First, using Lemma 15, and the definition of Δ_K, we immediately have

$$\mathrm{Vol}(\Lambda) = |\det(M)| = \sqrt{\Delta_K}.$$

Second, the the coordinate product of a point in Λ with lattice coordinates $\boldsymbol{\alpha}$ is given by

$$\prod_{j=1}^n \sigma_j(\sum_{i=1}^n \alpha_i\omega_i) = N_{K/\mathbb{Q}}(\sum_{i=1}^n \alpha_i\omega_i).$$

Recalling that $\{\omega_i\}_{i=1}^n$ is an integral basis, this is an non-zero integer for $\boldsymbol{\alpha} \neq \mathbf{0}$. Hence all lattice points $\mathbf{x} \in \Lambda$ lie on hyperbolic naps $\prod_{i=1}^n x_i \in \mathbb{Z}$. Since the algebraic norm of 1 is 1,

$$d_{p,min}(\Lambda) = 1$$

for any field extension $K : \mathbb{Q}$. Hence, in a unit-volume scaling of Λ, $d_{p,min}$ is inversely proportional to $\text{Vol}(\Lambda) = \sqrt{\Delta_K}$.

Third, the unique shortest non-zero vectors of Λ are $\pm(1, ..., 1)$ corresponding to $\pm 1 \in K$. This is since $(\pm 1, ..., \pm 1)$ are the unique shortest vectors of the innermost hyperboloids $\prod_{i=1}^n x_i = \pm 1$, and since $\sigma|_{\mathbb{Q}} = \text{id}_{\mathbb{Q}}$ for all embeddings σ, only $\pm(1, ..., 1)$ of the points $(\pm 1, ..., \pm 1)$ belong to the lattice Λ. Again, in a unit-volume scaling of Λ, the minimal vector-length is hence inversely proportional to $\text{Vol}(\Lambda) = \sqrt{\Delta_K}$.

Fourth, the asymptotic density of lattice points on the innermost hyperboloids $N_K(x) = \pm 1$, equivalently, the density of lattice points corresponding to $\mathfrak{O}_K^*$, is by Theorem 55 controlled by the regulator of the underlying number field. For the rest of the hyperboloids, Lemma 85 shows that, *e.g.*, if the prime 2 is inert, then the nap $\prod_{i=1}^n x_i = 2$ is empty.

Chapter 5

Algebraic lattice codes and coset codes

Having the necessary tools of algebraic number theory and information theory at our hands, we are now at a position to design algebraic lattices for lattice codes and coset codes.

5.1 The reliability problem

In Section 3.3.4 we deduced that the reliability problem in Rayleigh fast fading channels asymptotically boils down to maximizing the minimum product distance, which for a unit-volume scaling of the algebraic lattice of a totally real extension is equivalent to minimizing the discriminant. Unluckily, this is as far as number-theory helps us in this problem — there are no known general methods for designing number fields with given discriminants. Nevertheless, one can of course use tables of number fields to proceed using algebraic lattices. Alternatively, there are also more advanced methods to design rotations of $\mathbb{Z}^n$ based on ideals of algebraic lattices [23]. Note that all algebraic lattices have the unique minimal vectors $\pm(1, ..., 1)$, so no direct algebraic lattice construction will yield full-diversity rotations of $\mathbb{Z}^n$.

The validity of designing lattices based on the minimum product has been tested by simulations in numerous publications; see , *e.g.*, [23]. For a recent contribution, [14] found, among other results, the optimal rotation of $\mathbb{Z}^2$ by simulating all possible rotations for different signalling constellations. It seemed that independently of the considered signalling constellation, as γ_b grows, the optimal algebraic rotation turns to yield the optimal angle. This is of course natural in the sense that the optimal algebraic rotation minimizes the minimum product distance, which in turn is dominates the asymptotic expression for the REP at $\gamma_b \to \infty$. It should nevertheless be mentioned that at low SNRs, some other rotations performed slightly better than the algebraic ones, with the Hadamard rotations obtained as a low-SNR asymptotics both theoretically and computationally. Hadamard rotations seem to perform well also in slightly different fading models [17].

As a conclusion, we refer to a vast amount of literature and numerical data to state that algebraic methods based on the minimum product distance yield very good code lattices for Rayleigh fast fading channels with a good signal quality, equivalently, large γ_b. As a consequence, good coset codes should also be searched amongst full-diversity lattices.

5.2 The security problem

5.2.1 Number-theoretic objectives

Let us here briefly formulate the number-theoretic problems derived from the ECDP estimates of Section 3.3.4. These criteria pose the lattice design problem in a number-theoretic way, and they were the starting point of this thesis. It is worth to warn already at this stage that by our numerical computations, it seems that the *inverse norm sum* (INS) (5.1) below turned out not to be a good approximation of the ECDP in the sense that it seems to correlate with the actual performance of the lattices well enough only to distinguish the lattices that are poor for the applications, whereas the interesting cases cannot be ordered by the INS estimate. This will be considered in detail in [16] and subsequent numerical computations if this thesis.

Let us now derive the number-theoretic criteria. As already mentioned in Section 3.3.4, dropping the ones in the denominators of Eq. (3.8), except in the term for the lattice point $\mathbf{x} = \mathbf{0}$, we obtain a number-theoretic problem. With the tools of algebraic lattices from the preceding section at our hands, taking Λ_b to be the algebraic lattice, (3.8) then becomes (see also [1])

$$P_{c,e} \leq \frac{\sqrt{\Delta_K}}{2^m} \left(\gamma_e^{m/2} + \gamma_e^{-m} \sum_{\mathbf{0} \neq \mathbf{x} \in \Lambda_e \cap \mathcal{R}} \prod_{i=1}^{m} \frac{1}{|N_K(x)|^3} \right), \tag{5.1}$$

where x is the algebraic integer corresponding to the point $\mathbf{x}$ of the sublattice Λ_e, and the series has been truncated over a finite sending region $\mathcal{R}$. If the series is not truncated, it diverges since the Dirichlet Unit Theorem 48 guarantees that any totally real number field, except $\mathbb{Q}$ itself, has infinitely many units $x \in \mathfrak{O}_K^*$ with $N_K(x) = \pm 1$. This formula and our examinations on the properties of totally real number fields directly imply that we would like to minimize the factor Δ_K in front of the sum, maximize the regulator to have points with $N_K(x) = \pm 1$ sparsely located and have small primes inert to have the hyperboloids $N_K(x) = \pm 2, \pm 3$ empty.

Motivated by this, we study here the effect of the discriminant, prime ramification, and the regulator of the underlying totally real number field on the ECDP approximates of the algebraic lattice code. We do not use Eq. (5.1) for the reasons mentioned above, but choose instead Eq. (3.8), which is asymptotic at $\gamma_e \to 0^+$ and yields values less than one. In order to keep the length of the thesis under control, we mainly refer to subsequent work [16] that followed this thesis for a geometric explanation to the behaviour of the ECDP.

5.2.2 Numerical computations

We are now at the stage of being ready to compare the importance of the three design criteria for algebraic lattices of totally real algebraic number fields. The strategy of the computations is as follows. First, we generated (pseudo-)randomly integer polynomials of degree four and picked those that are irreducible and generate totally real number fields. Then, the discriminant, regulator, and the prime ramification can be examined by the tools presented in the sections considering algebraic number theory. These computations were not implemented from scratch, since there are several free softwares for algebraic number theory, of which we chose to use PARI gp [4]. After obtaining the number-theoretic quantities, any numerical software can be used to compute the probability estimates (3.8) and the inverse norm sums (5.1) of any finite constellation. We transferred the results from PARI gp to MATLAB for this sum evaluation and plotting.

As for technical details, all lattices Λ_b were scaled to unit volume, a sublattice $\Lambda_e = 2\Lambda_b$ was chosen, and the sending region was chosen to be the radius 15 origin-centered ball. The large enough radius guarantees that the usefulness of lattice coset codes is not ruined by the boundary effects, and, as was also confirmed computationally, that the average of $\|\mathbf{x}\|^2$, describing the average transmission power, and the number of code points does not vary significantly. Computationally efficient enumeration of the lattice points in such a ball given a generator matrix M is not trivial. Here it was done based on the eigenvalues of the Gram matrix $M^T M$ of the lattice, but we point out that this approach is not feasible in much larger dimensions. Then, one would need to implement, *e.g.*, a *sphere decoder algorithm* (see [23]) to find the short vectors. Signal quality values that we are interested in were decided based on the approximative formulae derived in Appendix B. This would suggest the interval $0.55 \leq \gamma_e \leq 7.5$. This choice of range seems quite

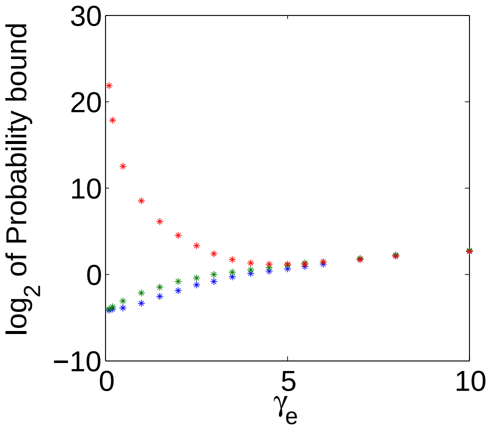

Figure 5.1: Base two logarithms of probability bounds for code from randomly generated algebraic lattices as a function of γ_e. Red: smallest values of the INS bound (5.1) amongst the lattices. Green: largest values of the tighter bound (3.8). Blue: smallest values of the tighter bound. Largest values of the bound (5.1) were considerably larger.

good in the sense that at the upper bound, the tighter ECDP bound also yields one, and at the lower bound, the ECDP estimate yields approximately $|\Lambda_b/\Lambda_e|$, *i.e.*, the eavesdropper receives a uniform random coset class.

On the validity of the inverse norm sum approximation

To illustrate the problems of the INS Eq.(5.1), we plot the smallest INS amongst our 474 randomly generated algebraic lattices as a function of γ_e; see Fig. 5.1. In no case is this quantity smaller than 1. This was predicted in the discussion in Section 3.3.4, since the INS is a valid approximation of the bound (3.8) in a good signal quality limit $\gamma_e \to \infty$, but the formula (3.8) in turn is derived based on a poor signal quality limit of the ECDP in AWGN channels (see Appendices A.2.1-A.2.2 for the derivation and Section 3.3.4 for discussion). As an unfortunate result, the INS never approximates the ECDP. It could of course happen that the INS *correlates* with the ECDP even if it is not a good *approximation*. This possibility will be discussed later with numerical data. Regarding the generality of our computations, other indices $[\Lambda_b/\Lambda_e]$ can be obtained by sclaing Λ_b, which will only affect the $\mathrm{Vol}(\Lambda_b)$ factor in both compared probability estimates. Other dimensions were not computed, but one can expect similar behaviour.

Let us yet compare the two probability bounds. The reasoning that the INS (5.1) approximates (3.8) only when (3.8) is not valid anymore is very much supported by the results in Fig. 5.1. Consider first large values of γ_e. The tighter estimate (3.8) yields values larger than one and equal to those of the lowest INS for good signal qualities, approximately $\gamma_e \geq 6$, which is already near the upper bound of the range of interest $0.55 \leq \gamma_e \leq 7.5$. At values $\gamma_e \approx 7.5$ Eve decodes the lattice Λ_b correctly, and hence coset coding is not interesting (see Appendix B). For small γ_e, contrary to the INS, the bound (3.8) tends to the uniform random limit $|\Lambda_b/\Lambda_e| = 2^{-4}$, in particular yielding a tight estimate for most of the interesting range of γ_e. The INS tends to infinity.

For more discussion on the INS we refer to [16]. We point out also that the tighter ECDP bound (3.8), is motivated in [16] in terms of *ordering* the lattices according to the ECDP for all interesting values of γ_e. In addition, it is proven that for small $\gamma_e = \sigma_h^2/\sigma^2$, the bound (3.8) is asymptotic with an error $\mathcal{O}(\sigma_h^2/\sigma^2)$ and will consequently yield values less than one.

Effect of the number-theoretic invariants on the ECDP

Even though the interest for the study of the number-theoretic design criteria, *i.e.*, minimizing the discriminant, maximizing the regulator and having small inert primes, is somewhat decreased by the fact that they are most visible in the problematic inverse norm sum approximation, we study here briefly their effect on the more accurate ECDP approximate (3.8). Naturally, they are also interesting in their own right, as well as the INS, from a purely mathematical point of view. Furthermore, Bob's decoding error probability bound (3.4) yields an inverse norm sum type problem at its asymptotically tight limit $\gamma_b \to \infty$ if we consider full-diversity matrix lattices constructed from totally complex extensions as discussed in Remark 114.

The probability bound (3.8), denoted $P_{c,e,upper}$, is plotted as a function of the discriminant, regulator, and the first inert prime of the underlying number field in Fig. 5.2. As it is obvious from the plots, there is a strong correlation between the discriminant and the ECDP approximate. The prime ramification seems not to affect the ECDP in any way, which could also be verified by computing the experimental means and variances of the ECDP estimate over the families of field extensions with different prime ramifications. Finally, there is a correlation between the regulator and the ECDP estimate, but it is "the wrong way around"; if the density of units had a significant effect on the estimate, low ECDP estimates should be obtained for *large* regulator values instead of *small*. It seems that this correlation

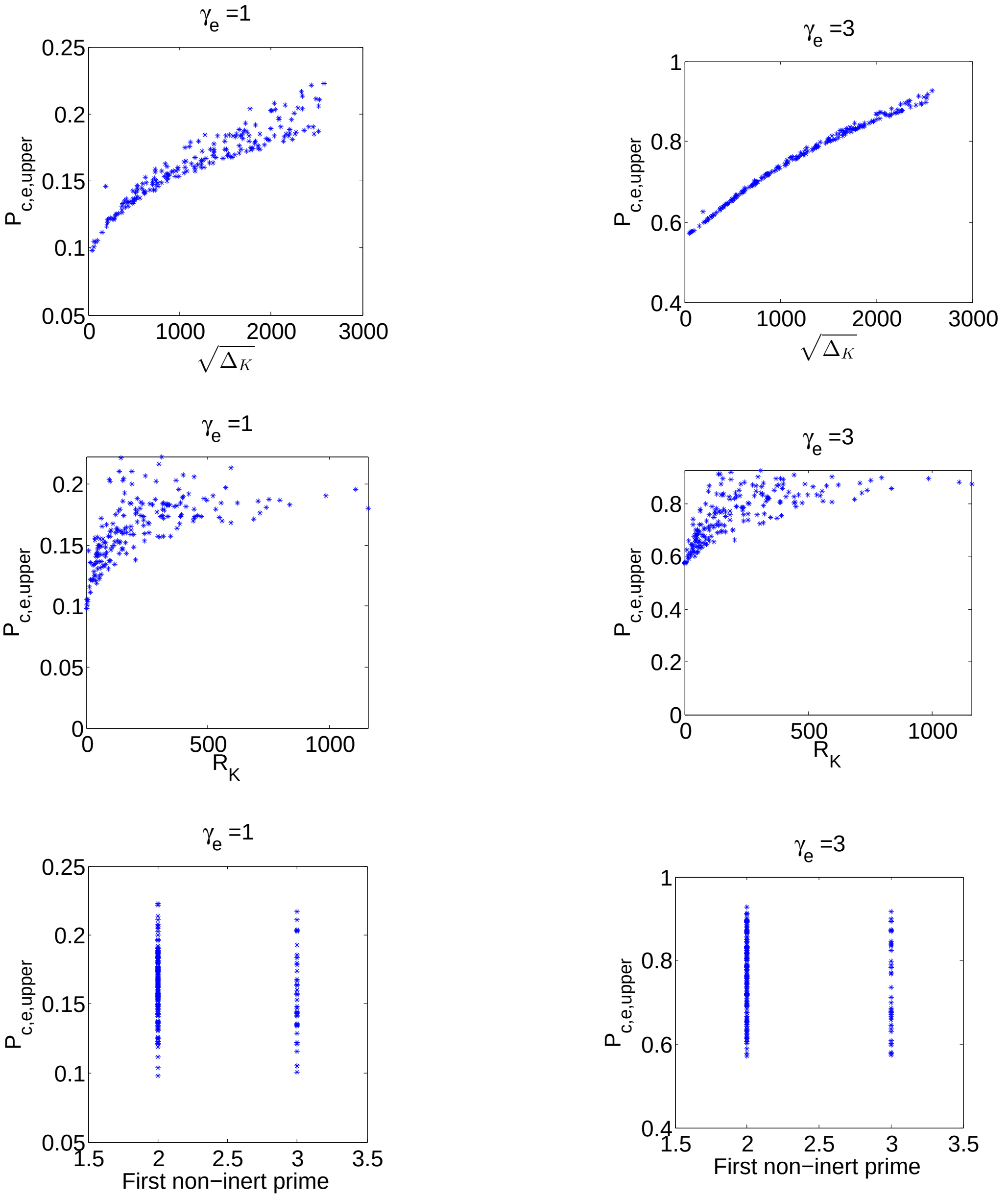

Figure 5.2: The ECDP estimates (3.8) for the channel parameter values $\gamma_e = 1$ and $\gamma_e = 3$ for lattice codes based on randomly generated field extensions as a function of the discriminant, regulator, and prime ramification of the underlying extension. Similar plots were obtained for other relevant values of γ_e.

of the regulator and the ECDP bound is rather due to the interplay of the discriminant and the regulator than a direct effect of regulator, since there are various upper bounds for the discriminant in terms of regulator (see, *e.g.*, [26] and references therein) and the discriminant seems to predict the ECDP estimate well. Similar insight is also implied by an inverse norm type approximation in [13].

As a remark related to the plots in Fig. 5.2, we point out that there is one field extension that behaves somewhat differently from the others in the discriminant plots. This is due to repeated components in the generator vectors, whose geometric effect is once again considered in [16]. The fact that repeated components really cause such behaviour can be easily verified, *e.g.*, by studying biquadratic extensions.

To give a heuristic explanation, the fact that the effect of the discriminant seems dominant is not particularly surprising since the regulator and the prime ramification dictate the appearance of the lattice points with the different algebraic norms, but the algebraic norms only appear in the problematic inverse norm sum approximation. Contrary to these, the discriminant has a geometric interpretation as a power of the inverse of the minimal vector-length of the lattice. As is familiar from lattice coding, the minimal vector-length, equivalently, the sphere-packing diameter, describes well the properties of lattices in AWGN channels. This also holds for the security of coset codes [15, 16]. Recalling from Remark 19 the interpretation of lattice diversity as resistance of vector-lengths to fading, a full-diversity lattice with a dense sphere packing is hence expected to pertain a reasonably dense sphere-packing in fading, after which the channel is effectively gaussian.

On the geometric design of full-diversity lattices

Let us here discuss the statement that the effect of the discriminant on the number-theoretic lattices is due to its geometric interpretation. The results summarized in this section are motivated in detail with partial analytic results and more computational examples in [16].

Now, since the shortest vectors of any algebraic lattice are $\pm(1,...,1)$, a large discriminant and hence a large volume of the lattice cell means that the lattice has a poor sphere-packing density. Conceptually this is to say that the lattice is “flat”: for any choice of a generator system (a lattice cell) containing $(1,...,1)$, there must be long vectors. It is worth pointing out at this point that the rigorous opposite of flat lattices are *well-rounded* lattices, *i.e.*, lattices whose minimal vectors span $\mathbb{R}^n$. There is an emerging study of number-theoretic constructions of well-reounded lattices [9, 10] that might in continuation be relevant also for this physical layer security problem.

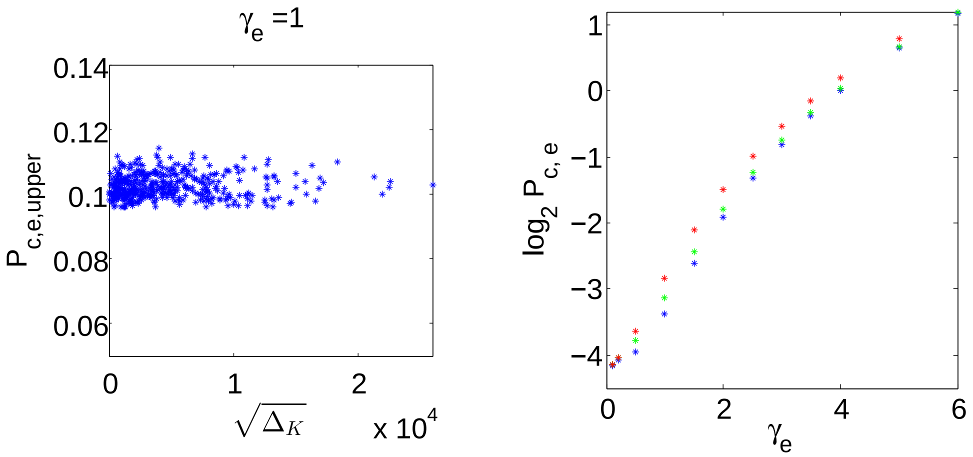

Figure 5.3: Left: the bound (3.8), denoted $P_{c,e,upper}$, as a function of the discriminant for non-flat sublattices of algebraic lattices of randomly generated totally real extensions.
Right: base two logarithm of $P_{c,e,upper}$ as a fuction of γ_e for $\mathbb{Z}^4$ (red), the worst (green) and the best (blue) full-diversity sublattices of the random algebraic lattices.

Based on the interpretation of flat lattices, we implemented a heuristic algorithm that finds a "less flat" sublattice of the original lattice. The algorithm starts from an *LLL-reduced* basis of the algebraic lattice, *i.e.*, roughly speaking a basis of near-orthogonal and short vectors. Then, the shortest generator vectors of the reduced basis are scaled by a suitable power of 2 to obtain the basis of a sublattice whose near-orthogonal generators have lengths varying at most by a factor of 2. Intuitively, such a sublattice should not be as flat as the original lattice, and as a sublattice of a full-diversity lattice, it is of full diversity. However, we cannot by theoretical means access the sphere-packing diameter of the lattice anymore, so we either have to solve the shortest vectors computationally or trust the intuition on flat lattices. ECDP bounds for such lattices are plotted in Fig. 5.3. The algorithm and lattice reduction are also described more formally and in detail in [16].

After this sublattice procedure the ECDP seems not to depend on the discriminant anymore, as can be seen from the first plot of Fig. 5.3. We also notice that the ECDP estimate varies by only some 10% over the family of lattices. Since we are only dealing with an estimate, it is hence not straightforward to distinguish which of the lattices perform best without direct simulations. Furthermore, comparing the values of the ECDP in this and

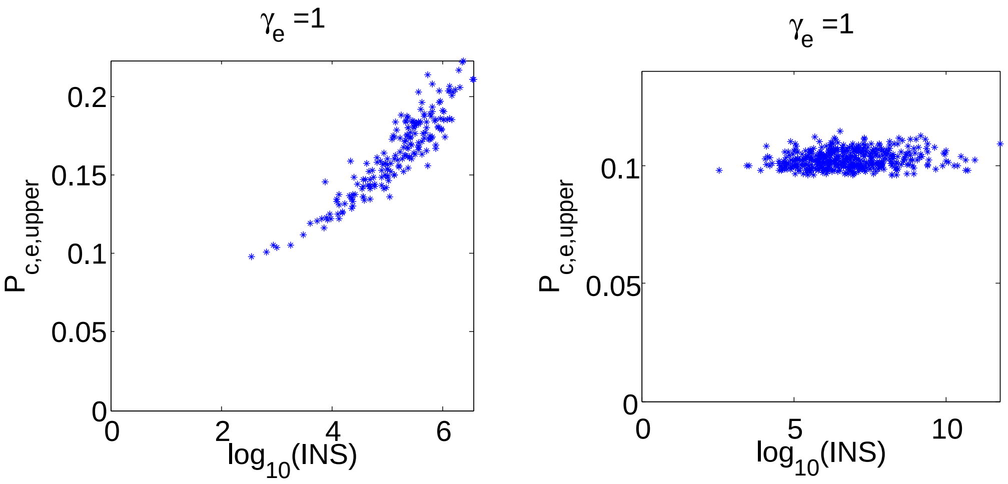

Figure 5.4: The 10-base logarithm of the INS estimate and the tighter probability estimate (3.8), denoted $P_{c,e,upper}$.
Left: algebraic lattices of randomly generated totally real extensions.
Right: the non-flat sublattices

the previous example in Fig. 5.2, we can see that the sublattice procedure vastly improved the ECDP so that all lattices perform as well as the best algebraic lattices. In the second plot of Fig. 5.3, we compared the ECDP estimates of the sublattices and $\mathbb{Z}^4$, finding that the full-diversity lattices perform consistently better.

At this point, let us briefly go back to the INS, whose correlation with the ECDP estimate remained unknown. For lattices that arise directly from number fields, the INS is known to correlate with the discriminant (see [6] for numerical evidence and [13] for an analytic estimate). Since the ECDP estimate also correlates with the discriminant, one would expect a correlation of the two estimates. This is indeed confirmed by our numerical results; see Fig. 5.4. However, after the sublattice algorithm described above, the INS estimate (restricting the summation to the sublattice) shows no correlation with the ECDP; see again Fig. 5.4. It should also be noted that the sublattices that perform better than the original lattices in terms of the tighter bound can attain considerably higher INS values than the original lattices. As a conclusion of our comparison of probability estimates, it would be preferable to use the tighter probability bound (3.8) for lattice design, and the correlation of the INS and the tighter ECDP estimate for algebraic lattices seems to occur rather due to both correlating with the discriminant than the two estimates correlating in general.

To summarize the brief overview of this subsection and [16], it seems that choosing Λ_e to be any full-diversity lattice with a reasonably good sphere-packing density will be a good choice of lattice. For example, taking Λ_b a full-diversity rotation of $\mathbb{Z}^n$ as conventional and $\Lambda_e = 2^k\Lambda_b$ will yield a good coset code. We point out that this seemingly simple result is not trivial at all since for the simpler but related AWGN channel model, it is known that a coset code $\Lambda_b = \mathbb{Z}^n$, $\Lambda_e = 2^k\mathbb{Z}^n$, or any other choice with orthogonal Λ_e, is *not* a good choice [15, 16].

Chapter 6

Conclusions

In this thesis, we studied the design of lattice codes based on algebraic number theory and geometry. The two main problems considered were the reliability problem of general lattice codes and the physical-layer security problem of lattice coset codes, both in noisy and Rayleigh fading channels. Good solutions to the former one are known based on algebraic number theory, whereas the latter one was an open question in the beginning of this work, although certain number-theoretic quantities had been suggested to yield possible design criteria.

We reviewed the information theory, lattice code theory, and the algebraic number theory relevant to the reliability and security problems. In the numerical computations, it however turned out that the probability estimate conventionally studied in the security problem, called the inverse norm sum, is too loose to order the interesting lattices. Nevertheless, one of the suggested number-theoretic design criteria for the coset code security, namely minimizing the discriminant of the underlying number field, seemed to prescribe very well the behaviour of a tighter probability estimate. We suggest that rather than due to its appearance in the inverse norm sum, this happend since the discriminant determines the sphere-packing density of the algebraic lattice, and a full diversity of the lattice guarantees the stability of vector-lengths under fading, hence together resulting in a dense sphere-packing after random fading. Regardless of the geometric interpretation of the results, they still pose a number-theoretic problem as all constructions of full-diversity lattices necessarily involve field extensions.

As the original hypothesis of this thesis turned inaccurate, we call the reader's attention to our suggested improvements. Since all of these improvements concern geometric properties number-theoretic code designs, we have chosen not to assimilate their content in the thesis that was almost complete at the time when the research projects were launched. In subsequent work based on the observations made in the numerical computations [16], we study the reliability of the different probability estimates and, using the tight ones, motivate geometric design criteria for full-diversity lattices that are supported by numerical computations. In [15], we study the security problem of coset codes as in this thesis and [16], but in the simpler AWGN channel model, and obtain a rigorous result on geometric lattice design. The ongoing project [17] concerns geometric lattice design based on the sphere-packing density after fading in channel models more complicated than the ones considered in this thesis. The sphere packing is related to both reliability and security problems. The approach of [17] can be seen as a generalization of the geometric design in Rayleigh fading channels considered in [16] and the locally diverse lattices addressed in [14].

Bibliography

[1] Belfiore, J.-C., and Oggier, F. Lattice Code Design for the Rayleigh Fading Wiretap Channel. In *Communications Workshops (ICC), 2011 IEEE International Conference on* (June 2011), pp. 1–5.

[2] Boutros, J., Viterbo, E., Rastello, C., and Belfiore, J.-C. Good Lattice Constellations for Both Rayleigh Fading and Gaussian Channels. *Information Theory, IEEE Transactions on 42*, 2 (1996), 502–518.

[3] Cohen, H. *A Course in Computational Algebraic Number Theory.* Graduate Texts in Mathematics. Springer, 1993.

[4] Cohen, H., Belabas, K., et al. PARI/GP Computer Algebra System. Available at *http://pari.math.u-bordeaux.fr/*. Referred to on 15 October 2015.

[5] Cover, T. M., and Thomas, J. A. *Elements of Information Theory.* Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, 2006.

[6] Ducoat, J., and Oggier, F. E. An analysis of small dimensional fading wiretap lattice codes. In *2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, June 29 - July 4, 2014* (2014), pp. 966–970.

[7] Ebeling, W. *Lattices and Codes: A Course Partially Based on Lectures by Friedrich Hirzebruch.* Advanced Lectures in Mathematics. Springer Fachmedien Wiesbaden, 2012.

[8] Everest, G., and Loxton, J. Counting Algebraic Units with Bounded Height . *Journal of Number Theory 44*, 2 (1993), 222 – 227.

[9] Fukshansky, L., Henshaw, G., Liao, P., Prince, M., Sun, X., and Whitehead, S. On Well-rounded Ideal Lattices II. *International Journal of Number Theory 09*, 01 (2013), 139–154.

[10] Fukshansky, L., and Petersen, K. On Well-rounded Ideal Lattices. *International Journal of Number Theory 08*, 01 (2012), 189–206.

[11] Hollanti, C., and Viterbo, E. Analysis on Wiretap Lattice Codes and Probability Bounds from Dedekind Zeta Functions. *International Congress on Ultra Modern Telecommunications and Control Systems*, 1-8 (2011).

[12] Jackson, J. *Classical Electrodynamics.* Wiley, 1975.

[13] Karpuk, D. A., Ernvall-Hytönen, A., Hollanti, C., and Viterbo, E. Probability Estimates for Fading and Wiretap Channels from Ideal Class Zeta Functions, 2014. Available at *http://arxiv.org/abs/1412.6946*. Referred to on 15 October 2015.

[14] Karpuk, D. A., and Hollanti, C. Locally Diverse Constellations from the Special Orthogonal Group, 2015. Available at *http://arxiv.org/abs/1505.02903*. Referred to on 15 October 2015.

[15] Karrila, A., and Hollanti, C. A Comparison of Skewed and Orthogonal Lattices in Gaussian Wiretap Channels. In *Information Theory Workshop (ITW) 2015 IEEE* (2015), pp. 1–5.

[16] Karrila, A., Karpuk, D., and Hollanti, C. Simple Geometric Lattice Design Criteria for Wiretap Coset Codes. In preparation.

[17] Karrila, A., Väisänen, N., Ó Catháin, P., Karpuk, D., and Hollanti, C. Sphere-packing Lattice Design for General Fading Channels Based on Hadamard Rotations. In preparation.

[18] Liang, Y., Poor, H., and Shamai, S. *Information Theoretic Security.* Foundations and trends in communications and information theory. Now Publishers, Incorporated, 2009.

[19] Ling, C., Luzzi, L., Belfiore, J., and Stehlé, D. Semantically Secure Lattice Codes for the Gaussian Wiretap Channel. *IEEE Transactions on Information Theory 60*, 10 (2014), 6399–6416.

[20] Lu, J., Harshan, J., and Oggier, F. A USRP Implementation of Wiretap Lattice Codes. In *Information Theory Workshop (ITW), 2014 IEEE* (Nov 2014), pp. 316–320.

[21] Moss, G. Introduction to Dirichlet series and the Dedekind zeta function. Available at

http://isites.harvard.edu/fs/docs/icb.topic482679.files/dirichlet0notes.pdf. Referred to on 15 October 2015.

[22] Näätänen, M., and Metsäkylä, T. Algebra. Available at *http://matematiikkalehtisolmu.fi/2010/algebra.pdf.* Referred to on 15 October 2015.

[23] Oggier, F., and Viterbo, E. *Algebraic Number Theory and Code Design for Rayleigh Fading Channels.* Foundations and Trends in Communications and Information Theory. Now, 2004.

[24] Oggier, F. E., Solé, P., and Belfiore, J. Lattice Codes for the Wiretap Gaussian Channel: Construction and Analysis, 2011. Available at *http://arxiv.org/abs/1103.4086v1.* Referred to on 15 October 2015.

[25] Ozarow, L., and Wyner, A. Wire-Tap Channel II. In *Advances in Cryptology*, T. Beth, N. Cot, and I. Ingemarsson, Eds., vol. 209 of *Lecture Notes in Computer Science.* Springer Berlin Heidelberg, 1985, pp. 33–50.

[26] Silverman, J. H. An Inequality Relating the Regulator and the Discriminant of a Number Field . *Journal of Number Theory 19*, 3 (1984), 437 – 442.

[27] Stein, W. Introduction to Algebraic Number Theory. Available at *http://modular.math.washington.edu/129-05/notes/129.pdf.* Referred to on 15 October 2015.

[28] Stewart, I. *Galois Theory, Third Edition.* Chapman Hall/CRC Mathematics Series. Taylor & Francis, 2003.

[29] Stewart, I., and Tall, D. *Algebraic Number Theory and Fermat's Last Theorem: Third Edition.* Ak Peters Series. Taylor & Francis, 2001.

[30] Svensson, P.-A. *Abstrakt algebra.* Studentlitteratur AB, 2001.

[31] Wyner, A. D. The Wire-Tap Channel. *Bell System Technical Journal 54*, 8 (1975), 1355–1387.

Appendix A

Probability bounds

In this appendix, we derive probability bounds for the eavesdropper's correct decision probability (ECDP) and the receiver's error probability (REP) for the gaussian, Rayleigh fast fading, and Rayleigh block fading channels. The bound for the ECDP is only valid for lattice coset codes, whereas the bound for the REP is valid for any coding strategy and any signaling constellation. The problem of minimizing Bob's error probability is addressed, *e.g.*, in [2, 23], and the ECDP can be found in [1, 24]. The probability bounds derived here constitute all the design criteria needed in the optimization of lattice coset codes. This is since once the dimension of the signaling lattice and the index $|\Lambda_b/\Lambda_e|$ have been fixed, then the transition rate of the lattice coset code is fixed, too.

One may wonder why we derive two bounds, as the two receivers have the same channel model. However, as dictated by the setup of the wiretap problem, there is a gap between their channel qualities. Hence, the two probability bounds should more precisely be considered as expansions of the probabilities of the same channel in two different limits, poor signal quality for the eavesdropper and good signal quality for the legitimate receiver. Of course, neither limit is completely realistic; truly, Λ_b should be approximately of Bob's best resolution of detection and Λ_e Eve's. However, an asymptotic expansion is the best we can do analytically, and an *expansion* is not an asymptotic *value*, so our results will describe the behaviour of the probability at the good and poor signal quality regimes.

A.1 On considering a single decoding event

All the derivations of this appendix are based on computing the probability of correct or wrong decoding for a single decoding event. It is appropriate to highlight why we can actually do so even in the case of block fading channels. Namely, in a realistic implementation of a block fading channel in a quasi-static physical environment, the subsequent fading coefficients are identically distributed but not independent, and the (approximative) independence of the components of a fading matrix $\operatorname{diag}(\mathbf{h})$ is only achieved by interleaving the fading coefficients of the different fading vectors $\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, ...$ Consequently, the subsequent fading vectors $\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, ...$ can be heavily correlated since their components with same index are the subsequent realizations of the fading coefficients.

Now, if Alice transmits lattice vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ...$ and a receiver (Eve or Bob) decodes the received vectors to $\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)}, ...$ the natural way to describe the probability of a decoding error in a quasi-static environment would be the *asymptotic error rate*

$$\lim_{m\to\infty} \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}_{\hat{\mathbf{x}}^{(j)} \neq \mathbf{x}^{(j)}}. \tag{A.1}$$

Correct decoding can be described analogously.

The question is whether the limit (A.1) exists (almost surely) and if so, is it a “number” in the sense of a constant almost sure limit, or a genuine random variable. The former will occur if the dependencies of $\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, ...$ are mild enough. What is “mild enough” is a nontrivial question in stochastics. To provide naive examples, if $\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, ...$ and hence also the decoding error indicators were i.i.d., then the Law of Large Numbers would guarantee that the constant $\mathbb{P}(\hat{\mathbf{x}}^{(1)} \neq \mathbf{x}^{(1)})$ is an almost sure limit in Eq. (A.1). If $\mathbf{h}^{(1)} = \mathbf{h}^{(2)} = ...$ and hence all the decoding error indicators are i.i.d. but with corresponding probability $\mathbb{P}(\hat{\mathbf{x}}^{(1)} \neq \mathbf{x}^{(1)} \text{ due to noise} \mid \text{fading } \mathbf{h}^{(1)}) = p$, then the limit again exists but is now given by the random variable p that depends on $\mathbf{h}^{(1)}$.

However, due to the identical distribution of $\mathbf{h}^{(j)}$ and hence $\mathbb{I}_{\hat{\mathbf{x}}^{(j)} \neq \mathbf{x}^{(j)}}$, the expectation of $\frac{1}{m}\sum_{j=1}^{m} \mathbb{I}_{\hat{\mathbf{x}}^{(j)} \neq \mathbf{x}^{(j)}}$ is given by the single decoding probability $\mathbb{P}(\hat{\mathbf{x}}^{(1)} \neq \mathbf{x}^{(1)})$ for any m. In particular, if a constant almost sure limit of Eq. (A.1) exists, as it should by the engineering setup, then the limit is given by the error probability of a single decoding event. By virtue of this discussion, we quantize probabilities also in quasi-static channels based on one decoding event, and call these quantizations simply decoding probabilities.

A.2 Bounds for the eavedropper's correct decision probability (ECDP)

In this subsection, we derive the ECDP bounds for the different channel models.

A.2.1 The ECDP in gaussian channels with lattice coset coding

Let us denote the vector transmitted by Alice by $\mathbf{x}$, its equivalence class in Λ_b/Λ_e by $[\mathbf{x}]$ and the sending region by $\mathcal{R} \subset \Lambda_b$. A possible received bit is denoted by $\mathbf{z}$. The probability $P_{c,e}$ of a correct guess for the transmitted vector, using the Voronoi cell criterion presented in Remark 19, is obtained as the probability of the received vector lying in the Voronoi cell of some $\mathbf{t} \in [\mathbf{x}]$, *i.e.*,

$$P_{c,e} = \sum_{\mathbf{t}\in[\mathbf{x}]\cap\mathcal{R}} \int_{\mathbf{z}\in\mathcal{V}(\mathbf{t})} g_n(\mathbf{z}-\mathbf{x})d^n z \le \sum_{\mathbf{t}\in[\mathbf{0}]=\Lambda_e} \int_{\mathbf{z}\in\mathcal{V}(\mathbf{t})} g_n(\mathbf{z})d^n z. \tag{A.2}$$

Here $\mathcal{V}$ denotes the Voronoi cells of Λ_b and $g_n(\mathbf{w}) = e^{-\|\mathbf{w}\|^2/(2\sigma_e^2)}/(2\pi\sigma_e^2)^{n/2}$ is the standard n-dimensional spherical zero-mean Gaussian distribution function with variance σ_e^2. Note that in lack of information about the lattice we implicitly assumed in the first step that if the noise makes the received message pop out of the sending region $\mathcal{R}$, the guess is automatically wrong. The contribution of such events, *i.e.*, the boundary effects, to the overall probability is hence assumed negligible.

Next, based on the Poisson formula for lattices, one can show that

$$\sum_{\mathbf{t}\in\Lambda_e} g_n(\mathbf{z}+\mathbf{t}) \le \sum_{\mathbf{t}\in\Lambda_e} g_n(\mathbf{t}),$$

with equality if and only if $\mathbf{z} \in \Lambda_e$ (see, *e.g.*, [15]). Hence, Eq. (A.2) yields

$$P_{c,e} \le \sum_{\mathbf{t}\in\Lambda_e} \mathrm{Vol}(\Lambda_b) g_n(\mathbf{t}). \tag{A.3}$$

The tightness of this bound has been discussed extensively in [16], concluding that it is an $\mathcal{O}(1/\sigma^2)$ approximation of $P_{c,e}$ as $\sigma \to 0^+$ and can be used for comparing lattices for any relevant σ. As $\sigma \to \infty$, the bound diverges. We also point out that the same bound can be derived by approximating in the last step of Eq. (A.2) the gaussian function by its middle-point value in each Voronoi cell. Analogously, truncating series (A.3) over the sending region is equivalent to this derivation for the first step of Eq. (A.2).

A.2.2 The ECDP in Rayleigh block fading channels with lattice coset coding

Let us next generalize the probability bound for the ECDP in AWGN channels to Rayleigh block fading channels. Using the standard generalization procedure described in Sec. 3.3.2, we start from the upper bound (A.3) for the ECDP:

$$\begin{aligned}
&\mathbb{P}\{\text{Eve decodes correctly}\}\\
\overset{(a)}{=}\;& \mathbb{E}_{\mathbf{h}_e}\{\mathbb{P}\{\text{Eve decodes correctly} \mid \text{fading } \mathbf{h}_e\}\}\\
\overset{(b)}{=}\;& \mathbb{E}_{\mathbf{h}_e}\left\{P_{c,e;\Lambda_{e,\mathbf{h}_e},\Lambda_{b,\mathbf{h}_e}}\right\}\\
\overset{(c)}{\leq}\;& \mathbb{E}_{\mathbf{h}_e}\left\{\sum_{\mathbf{t}\in\Lambda_{e,\mathbf{h}_e}} \mathrm{Vol}(\Lambda_{b,\mathbf{h}_e}) g_{Lm}(\mathbf{t})\right\}. \qquad \text{(A.4)}
\end{aligned}$$

Step (a) is basic conditioning, step (b) uses the fact that Eve performs a closest-point search on the faded lattice $\Lambda_{b,\mathbf{h}}$ — hence the subscripts in $P_{c,e;\Lambda_{e,\mathbf{h}_e},\Lambda_{b,\mathbf{h}_e}}$ which is otherwise as given by Eq. (A.2) — and step (c) is substituting the bound of Eq. (A.3). We have added a subscript $\mathbf{h}_e$ to emphasize that Eve is considered. It is worth noticing that if we truncate the series (A.3), then also the series (A.4) will be truncated.

It remains to compute the expectation, given the distribution of $|h_i|$. First, using the notation of Section 3.3.2,

$$\mathrm{Vol}(\Lambda_{b,\mathbf{h}_e}) = |\det(M_{h_e})| = \left(\prod_{i=1}^{m} |h_{i,e}|\right)^{L} \mathrm{Vol}(\Lambda_b).$$

Next, we change the summation of Eq. (A.4) back to the lattice Λ_e by expressing $\mathbf{t}\in\Lambda_{e,\mathbf{h}_e}$ in terms of $\mathrm{vec}(\mathbf{X}) := \mathbf{x}\in\Lambda_e$ as

$$\mathbf{t} = \mathrm{diag}(\mathrm{diag}(h_{i,e}), ..., \mathrm{diag}(h_{i,e}))\mathbf{x},$$

so $t_i = h_{i,e}x_i$ (the subscript i of $h_{i,e}$ interpreted mod L). Hence we can express the summand in terms of the non-random lattice,

$$g_{Lm}(\mathbf{t}) = e^{-\sum_{i=1}^{Lm}|h_{i,e}x_i|^2/(2\sigma_e^2)}/(2\pi\sigma_e^2)^{Lm/2}.$$

Finally, the expecation we want to find yields

$$\begin{aligned}
&\mathbb{P}\{\text{Eve decodes correctly}\}\\
\leq\;& \mathrm{Vol}(\Lambda_b)\sum_{\mathbf{x}\in\Lambda_e}\mathbb{E}_{\mathbf{h}_e}\left\{\left(\prod_{i=1}^{m}|h_{i,e}|\right)^{L} e^{-\sum_{i=1}^{Lm}|h_{i,e}x_i|^2/(2\sigma_e^2)}/(2\pi\sigma_e^2)^{Lm/2}\right\}.
\end{aligned}$$

Next, recall that the elements of $\mathbf{x}$ multiplied by the same fading coeffiecient $h_{i,e}$ are exactly the rows of $\mathbf{X}$. Using this fact and the independence of the variables $|h_{i,e}|$, we get

$$\begin{aligned}&\mathbb{P}\{\text{Eve decodes correctly}\}\\ \leq\ & \frac{\mathrm{Vol}(\Lambda_b)}{(2\pi\sigma_e^2)^{Lm/2}} \sum_{\mathrm{vec}(\mathbf{X})\in\Lambda_e} \prod_{i=1}^{m} \mathbb{E}_{\mathbf{h}_e}\{|h_{i,e}|^L e^{-\frac{|h_{i,e}|^2}{2\sigma_e^2}\sum_{k=1}^{L} X_{ik}^2}\}.\end{aligned}$$

Finally, modelling the absolute values $|h_{i,e}|$ as Rayleigh distributed, we use the Rayleigh distribution PDF (and some symbolic computation software). The expectation above can then be calculated by simple integration. The final result is

$$\begin{aligned}&\mathbb{P}\{\text{Eve decodes correctly}\}\\ \leq\ & \frac{\Gamma(L/2+1)^m \mathrm{Vol}(\Lambda_b)}{\pi^{Lm/2}} \left(\frac{\sigma_{h,e}}{\sigma_e}\right)^{Lm} \sum_{\mathrm{vec}(\mathbf{X})\in\Lambda_e} \prod_{i=1}^{m} \frac{1}{(1+\|\mathbf{X}_i\|^2 \frac{\sigma_{h,e}^2}{\sigma_e^2})^{L/2+1}}.\end{aligned} \tag{A.5}$$

It is worth noticing that the model for fading was only specified in this last step, except the fact that the fading coefficients were assumed independent. Thus, other fading models can be studied with minor modifications. In Eq. (A.5), Γ is the standard gamma function and $\mathbf{X}_i = (X_{i1}, ..., X_{iL})$ is the i^{th} row of $\mathbf{X}$. We point out that in the original (and so far only) reference [1] deriving this formula, there is a misprint in the constant of Eq. (A.5) — the same reference derives the fast fading case separately, and the formula given here coincides with their fast fading formula by setting $L = 1$. At this stage, it is reasonable to make some remarks considering the above formula.

Remark 115. A good sanity check for the formula is to see that the probability does not depend on the choice of electric-field units. The standard deviation σ_e and the components of $\mathbf{X}_i$ are in the units of the electric field, whereas $\mathrm{Vol}(\Lambda_b)$ is in these units to the power of Lm and all other numbers are dimensionless. The invariance follows immediately from these substitutions. It is also worth noticing that the expression rather depends on the quantity $\gamma_e := \sigma_{h,e}^2/\sigma_e^2$ rather that the parameters σ_e^2 or $\sigma_{h,e}^2$ alone. This was predicted when we studied the measure of channel quality in Section 3.2.

Remark 116. From this formula we can anticipate that to minimize the ECDP, $\mathbf{X}$ ought to have no zero rows (except in the zero-matrix, which is unavoidable in a lattice but not in a code).

A.2.3 The ECDP in Rayleigh fast fading channels with lattice coset coding

We recall that the Rayleigh fast fading channel is mathematically just the block fading channel with parameter $L = 1$. Since it is from the engineering point-of-view the most important model of wireless communications, we still state the result separately:

$$\mathbb{P}\{\text{Eve decodes correctly}\} \leq \frac{\text{Vol}(\Lambda_b)}{2^m}\gamma_e^{m/2} \sum_{\mathbf{x}\in\Lambda_e\cap\mathcal{R}} \prod_{i=1}^{m} \frac{1}{(1+\|x_i\|^2\gamma_e)^{3/2}}.$$

A.3 Bounds for the legitimate receiver's error probability (REP)

We derive an upper bound for the legitimate receiver's error probability (REP). The bound is valid for any signaling constellation and any code.

A.3.1 The pairwise error probability

In the following subsections, we find an upper bound for the REP in the considered channel models. This time, we will not do the generalization step as described in Section 3.3.2, but rather walk it backwards, *i.e.*, treat the gaussian channel as a special case of a block fading channel, where the fading coefficients are identically one.

Let us consider here the probability of Bob receiving a wrong lattice, which is a function of the realizations of the Rayleigh-distributed scaling coefficients $\mathbf{h}_b$. For notational simplicity in what follows, we denote the sending constellation by Λ_b and the skewed constellation as $\Lambda_{h_b,b}$, by which we otherwise denote lattices, even though the derivation does not use the fact that they are lattices. The bound is derived by computing the *pairwise error probability* (PEP) $P(\mathbf{x} \to \mathbf{w}|\mathbf{h}_b)$, *i.e.*, the probability that $\mathbf{x} \in \Lambda_b$ is transmitted, but due to the noise Bob interprets it as some $\Lambda_{h_b,b} \ni \mathbf{t} = \text{diag}(\text{diag}(|h_{i,b}|), ..., \text{diag}(|h_{i,b}|))\mathbf{w}$, given the fading coefficients.

Let us find an approximation for the PEP for a block fading channel. Denoting the block diagonal matrix related to block fading channels as $D := \text{diag}(\text{diag}(|h_{i,b}|), ..., \text{diag}(|h_{i,b}|))$, we find a bound

$$\begin{aligned} & P(\mathbf{x} \to \mathbf{w}|\mathbf{h}_b) \\ \leq\ & \mathbb{P}(\text{The received vector } D\mathbf{x} + \mathbf{v} \text{ is closer to } D\mathbf{w} \text{ than to } D\mathbf{x}) \\ =\ & \mathbb{P}(\|D\mathbf{x} + \mathbf{v} - D\mathbf{w}\|^2 \leq \|\mathbf{v}\|^2) \end{aligned}$$

Cancelling out $\|\mathbf{v}\|^2$ and denoting the diagonal elements of D as $h_{i,b}$, where the index i is again interpreted modulo L, computing the vector norms yields

$$\begin{aligned} & P(\mathbf{x} \to \mathbf{w}|\mathbf{h}_b) \\ \leq\ & \mathbb{P}(\sum_{i=1}^{Lm} h_{i,b}^2 (x_i - w_i)^2 + 2\sum_{i=1}^{Lm} h_{i,b}(x_i - w_i)v_i \leq 0). \end{aligned}$$

Now, v_i being i.i.d. one-dimensional zero-mean gaussian r.v. with a variance σ_b^2, given the realization of the fading coefficients the latter sum $\chi := \sum_{i=1}^{Lm} h_{i,b}(x_i - w_i)v_i$ is a one-dimensional zero-mean gaussian r.v. with a variance of $\sigma_\chi^2 = \sigma_b^2 \sum_{i=1}^{Lm} h_{i,b}^2 (x_i - w_i)^2$. Here we also used the channel model assumption that the fading coefficients $h_{i,b}$ are independent of the noise variables v_i. Denoting the constant first sum as $2A := \sum_{i=1}^{Lm} h_{i,b}^2 (x_i - w_i)^2$, we get

$$\begin{aligned} P(\mathbf{x} \to \mathbf{w}|\mathbf{h}_b) &\leq \Pr(2A + 2\chi \leq 0) \\ &= \operatorname{erfc}(A/\sigma_\chi), \end{aligned}$$

where erfc(x) is the integral giving the gaussian tail probability

$$\operatorname{erfc}(x) := \int_{t=x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

Finally, we use the approximation $\operatorname{erfc}(x) \leq e^{-x^2/2}/2$ for $x \geq 0$ (see below), yielding our desired PEP bound

$$P(\mathbf{x} \to \mathbf{w}|\mathbf{h}_b) \ \leq \ e^{-A^2/2\sigma_\chi^2}/2. \tag{A.6}$$

From our definitions of A and σ_χ^2, it follows that $A^2/2\sigma_\chi^2 = \sum_{i=1}^{Lm} h_{i,b}^2 (x_i - w_i)^2/(8\sigma_b^2)$. Finally,

$$\begin{aligned} \mathbb{P}\{\text{Bob decodes wrong}\} &\leq \sum_{\mathbf{w} \neq \mathbf{x}} P(\mathbf{x} \to \mathbf{w}|\mathbf{h}_b) \\ &= \sum_{\mathbf{w} \neq \mathbf{x}} \exp(-\sum_{i=1}^{Lm} h_{i,b}^2 (x_i - w_i)^2/(8\sigma_b^2))/2(\text{A.7}) \end{aligned}$$

Let us yet prove the bound $\mathrm{erfc}(x) \le e^{-x^2/2}/2$ for $x \ge 0$. For $x = 0$ the bound is trivial. By Markov's inequality, any random variable V satisfies for any $x > 0$

$$\mathbb{P}\{V \ge x\} = \mathbb{P}\{e^{xV} \ge e^{x^2}\} \le e^{-x^2}\mathbb{E}\{e^{xV}\},$$

and taking V normally distributed, $V \sim N(0,1)$, the above yields, after some integration

$$\mathrm{erfc}(x) = \mathbb{P}\{V \ge x\} \le e^{-x^2/2}/2.$$

A.3.2 The REP in gaussian channels

In the gaussian channel model, the fading coefficients used above are not random variables but constants 1. Then, Eq. (A.7) yields

$$\begin{aligned} & P(\text{Bob decodes wrong } \mid \mathbf{x} \text{ is transmitted}) \\ \le\ & \frac{1}{2} \sum_{\mathbf{x} \neq \mathbf{w} \in \Lambda_b \cap \mathcal{R}} e^{-\|\mathbf{x}-\mathbf{w}\|^2/8\sigma_b^2}. \end{aligned}$$

This bound can be made uniform in $\mathbf{x}$ (provided that Λ_b is a lattice; this is the only step where we need the assumption) by extending the summation to all of Λ_b,

$$P_{e,b} \le \frac{1}{2} \sum_{\mathbf{0} \neq \mathbf{w} \in \Lambda_b} e^{-\|\mathbf{w}\|^2/8\sigma_b^2}.$$

It is worth noticing that this expression largely resembles the ECDP of a gaussian channel. Hence, in gaussian channels, the probability bound design criteria of Bob and Eve coincide. This has been utilized in [15] to show that orthogonal lattices are always suboptimal in terms of these probabilities.

A.3.3 The REP in Rayleigh block fading channels

To investigate the block fading case, we just average the PEP (A.7) over the i.i.d. Rayleigh-distributed random variables $h_{i,b}$ since by indicator functions and conditioning, $\mathbb{P}\{\text{Bob interprets } \mathbf{x} \text{ as } \mathbf{w}\} = \mathbb{E}_{\mathbf{h}_b}\{P(\mathbf{x} \to \mathbf{w}|\mathbf{h}_b)\}$. Thus, compute

$$\begin{aligned} \mathbb{E}_{\mathbf{h}_b}\{P(\mathbf{x} \to \mathbf{w}|\mathbf{h}_b)\} &\le \mathbb{E}_{\mathbf{h}_b}\{e^{-A^2/2\sigma_X^2}/2\} \\ &= \mathbb{E}_{\mathbf{h}_b}\{e^{-(\sum_{i=1}^{Lm} h_{i,b}^2 (x_i - w_i)^2)/8\sigma_b^2}/2\}. \end{aligned}$$

Recall again that the indices of $h_{i,b}$ were interpreted modulo L, and each $h_{i,b}$ multiplies the rows of the original matrices $\mathbf{X}-\mathbf{W}$, so

$$\mathbb{E}_{\mathbf{h}_b}\{P(\mathbf{x}\to\mathbf{w}|\mathbf{h}_b)\} \;\leq\; \mathbb{E}_{\mathbf{h}_b}\{e^{-(\sum_{i=1}^{m}\sum_{k=1}^{L}h_{i,b}^2(X_{ik}-W_{ik})^2)/8\sigma_b^2}/2\}.$$

Finally, we express this in terms of $h_{i,b}$ and use their independence

$$\mathbb{E}_{\mathbf{h}_b}\{P(\mathbf{x}\to\mathbf{w}|\mathbf{h}_b)\} \;\leq\; \frac{1}{2}\prod_{i=1}^{m}\mathbb{E}_{h_{i,b}}\{e^{-h_{i,b}^2\sum_{k=1}^{L}(X_{ik}-W_{ik})^2/8\sigma_b^2}\}.$$

Now, calculating the expectations with the PDF of $h_{i,e}$ known is just a matter of integration. As in the ECDP bound, also here we only now specify the fading model, so the derivation can be generalized. We get

$$\mathbb{E}_{\mathbf{h}_b}\{P(\mathbf{x}\to\mathbf{w}|\mathbf{h}_b)\} \;\leq\; \frac{1}{2}\prod_{i=1}^{m}\frac{1}{\gamma_b\|(\mathbf{X}-\mathbf{W})_i\|^2/4+1},$$

where $\gamma_b=\sigma_{h,b}^2/\sigma_b^2$ is the measure of Bob's channel quality and $[\text{matrix}]_i$ denotes the i^{th} row vector of a matrix. Hence, finally

$$P_{e,b}\leq\frac{1}{2}\sum_{\mathbf{x}\neq\mathbf{w}\in\Lambda_b\cap\mathcal{R}}\prod_{i=1}^{m}\frac{1}{\gamma_b\|(\mathbf{X}-\mathbf{W})_i\|^2/4+1}.$$

As for AWGN channels and a lattice Λ_b, this bound can be made uniform by extending the summation to Λ_b and replacing $\mathbf{x}$ by $\mathbf{0}$.

In the limit of Bob's good signal quality $\gamma_b\to\infty$, the 1 in the denominator can be dropped, yielding the asymptotic approximation

$$P_{e,b}\leq\frac{1}{2}\sum_{\mathbf{x}\neq\mathbf{w}\in\Lambda_b\cap\mathcal{R}}\;\prod_{i=1,\mathbf{X}_i\neq\mathbf{W}_i}^{m}\frac{4}{\gamma_b\|(\mathbf{X}-\mathbf{W})_i\|^2}. \tag{A.8}$$

A.3.4 The REP in Rayleigh fast fading channels

For the important special case of Rayleigh fast fading channels, the upper bound for the REP becomes

$$P_{e,b}\leq\frac{1}{2}\sum_{\mathbf{x}\neq\mathbf{w}\in\Lambda_b\cap\mathcal{R}}\prod_{i=1}^{m}\frac{1}{\gamma_b\|x_i-w_i\|^2/4+1}.$$

In the limit of Bob's good signal quality $\gamma_b\to\infty$,

$$\mathbb{E}_{\mathbf{h}_b}\{P_{e,b}(\mathbf{h}_b)\} \;\leq\; \frac{1}{2}\sum_{\mathbf{x}\neq\mathbf{w}\in\Lambda_b}\prod_{i=1,x_i\neq w_i}^{m}\frac{4}{\gamma_b(x_i-w_i)^2}.$$

Let us yet connect this to what we know about lattice theory. Using the definitions of ℓ-product distance and the modulation diversity, this becomes

$$\mathbb{E}_{\mathbf{h}_b}\{P_{e,b}(\mathbf{h}_b)\} \leq \frac{1}{2} \sum_{\mathbf{x}\neq\mathbf{w}\in\Lambda_b} \left(\frac{4}{\gamma_b}\right)^{\ell(\mathbf{x},\mathbf{w})} d_p^{(\ell)}(\mathbf{x},\mathbf{w})^{-2},$$

which points out the two design criteria for a lattice:

i full diversity in order to minimize the factor $\left(\frac{4}{\gamma_b}\right)^{\ell(\mathbf{x},\mathbf{w})}$,

ii) minimization of the dominant term of the series, equivalently, maximization of the minimum product distance $d_{p,min}(\Lambda_b) := \min_{\mathbf{w}\in\Lambda_b} d_p^{(\ell)}(\mathbf{0},\mathbf{w})$.

This is the foundation of number-theoretic lattice design; as it is illustrated in Example 20, any construction of a lattice with diversity other than one necessarily uses field extensions. As a sort of a converse, Corollary 113 shows that full-diversity lattices can be constructed in any dimension using number-field extensions.

A.4 Geometric designs

Instead of using the analytical bounds derived in the preceding subsections, AWGN channels allow for a geometric design based on the spherical symmetry and fast decay of the PDF of the noise vector. For Bob's case, it is a part of the information theory folklore that the sphere-packing density of Λ_b measures the reliability of a lattice code. For Eve's probability and coset codes, we motivate in [15, 16] a heuristic stating that in this case, the sphere-packing density of Λ_e should be maximized. The heuristic is supported by semi-analytic computations.

Appendix B

Signal qualities relevant for coset coding

From the engineering setup it is clear that if Eve's signal is too good, no physical-layer security helps and, conversely, if Eve's signal is utterly poor, no physical-layer security is needed. We give here a motivation to obtain a rough approximation for this range of interest.

Proposition 117. *Optimization of the security of a coset code in an AWGN channel is relevant roughly in the noise range*

$$\frac{\Gamma(n/2+1)^{1/n}}{2\sqrt{n\pi}}\,Vol(\Lambda_b)^{1/n} \leq \sigma \leq \frac{\Gamma(n/2+1)^{1/n}}{\sqrt{\pi}}\,Vol(\Lambda_e)^{1/n}.$$

Proof sketch. For the lower bound, note that the ball $B(\mathbf{0}, 2\sqrt{n}\sigma)$ is the smallest one containing the axis-aligned cube with sides $[-2\sigma, 2\sigma]$. Eve will highly probably have a noise vector inside this ball, so she will "always decode correctly", when such a ball will fit inside $\mathcal{V}(\Lambda_b)$. Then, since Λ_b should have a good sphere packing, this will happen when the volumes of the ball and the lattice are approximately equal.

For the upper bound, complete secrecy is achieved when Eve's detection resolution σ is approximately the resolution of the lattice Λ_e, so we have required $\mathrm{Vol}(B(\mathbf{0}, \sigma)) = \mathrm{Vol}(\Lambda_e)$. Independently of this, we notice in [16] that also Λ_e should have a good sphere packing so this makes sense. □

For Rayleigh fast fading channels, using the range above, raising the inequality to power n, and the standard generalization with $\mathbb{E}\{\mathrm{Vol}(\Lambda_{b,\mathbf{h}})\} = \sigma_h^n$ yields

$$\frac{\Gamma(n/2+1)^{1/n}}{2\sqrt{n\pi}}\mathrm{Vol}(\Lambda_b)^{1/n} \leq \sigma/\sigma_h \leq \frac{\Gamma(n/2+1)^{1/n}}{\sqrt{\pi}}\mathrm{Vol}(\Lambda_e)^{1/n}.$$