Scuola Normale Superiore

Corso di Perfezionamento in Matematica
per la Tecnologia e l'Industria

A. A. 2006/07

Continued fractions, coding and wireless channels

Candidata
Laura Luzzi

Relatore
Prof. Stefano Marmi

Correlatore
Prof. Emanuele Viterbo

Acknowledgements

My first thanks go to prof. Stefano Marmi, without whom this work would not exist, for his energetic support and positive attitude during all these years. Most of the results in the first part of the thesis are based on his insights; I am also grateful to prof. Marcelo Viana for his precious suggestions that were essential for completing one of the proofs.

I am indebted to prof. Emanuele Viterbo for patiently explaining to me the basics of coding for wireless channels and for his insightful advice, to prof. Jean-Claude Belfiore for encouraging me to begin the study of the fascinating topic of non-commutative algebra, and to Ghaya Rekaya for letting me borrow her code for the simulation of the Golden Code transmission chain and for her competent advice on the interpretation of results.

I am grateful to prof. Da Prato, prof. Fagnani and prof. Profeti and to Carlo Carminati for accepting to be in the committee, and to prof. Nakada for refereeing my work and for several interesting discussions about possible developments. I also wish to acknowledge the financial support from Scuola Normale Superiore, which allowed me to spend long visits at Politecnico di Torino and Telecom Paris.

Finally, thanks to Rob for his moral *and* mathematical support and for always being at my side these past four years.

Contents

Introduction

Mathematicians have been studying continued fractions long before the modern theory of dynamical systems emerged. To this day, they remain one of the few models for which a comprehensive statistical analysis is available, including ergodicity, invariant measures and the decay of correlation functions.
The relevance of this model is not limited to the field of dynamical systems, but extends to number theory, information theory and the theory of algorithms.
As a tool for representing real numbers, continued fraction expansions are ideal to study diophantine approximation problems, they are more economical in terms of length than the decimal expansion, and aren't basis-dependent.
However, the major drawback of being hardly suited for computation (even simple operations like the sum and product become complex in this representation) is probably the reason why the literature describing the applications of continued fractions to engineering is so sparse.

Recently there has been an increasing interest in describing the behavior of families of dynamical systems at the boundary of chaoticity (a widely known example is the extensive study on the bifurcations of the logistic map). In this context interesting phenomena of phase transitions, self-similarity and fractal sets often arise.
The first part of my research concerns α-continued fractions for $\alpha \in [0, 1]$, a one-parameter family of interval maps giving rise to a whole class of continued fraction expansions. Just as the classical continued fractions can be viewed as an acceleration of the Euclidean division algorithm, α-continued fractions are obtained imposing the condition that the remainder in the Euclidean division should belong to the interval $[\alpha - 1, \alpha)$.
This allows to gain a wider perspective, bridging the gap between Gauss's classical continued fraction algorithm ($\alpha = 1$) and the expansion based on the *nearest integer* approximation ($\alpha = 1/2$), which has a faster convergence and a higher entropy; and more interestingly, between the latter and the *by-excess algorithm* ($\alpha = 0$), whose properties are markedly different: it is slower, and doesn't admit a finite invariant density, due to the presence of a parabolic fixed point.
It is then natural to investigate how this transition occurs, in particular by studying the *statistical stability* of the family of the invariant densities as a function of the parameter α. In §2.1 we prove that this family is in fact continuous in the L^1 norm.
Moreover, the explicit expression of the invariant densities can be used to compute the entropy $h(\alpha)$ of the system, which is related both to the complexity

of the corresponding algorithm[1] and to the rate of information creation of the system regarded as an information source [3] [14].
Unfortunately, in the general case there exists no purely mechanical algorithm for finding the invariant density; a general approach, introduced by Rohlin [19] and known as the *Natural Extension* method, involves finding a two-dimensional transformation $\overline{T}$ of which the initial map T is a factor, and a suitable domain where $\overline{T}$ is invertible. The density of T is then derived from the density of $\overline{T}$ simply by projecting on the first coordinate.
One of my main results is the expression of the natural extensions for all values of the parameter α in the sequence $\left\{\frac{1}{n}\right\}$. The shape of the domain of the natural extensions in this case is much more complex than expected, and the density is given by a long recursive formula.
Moreover, the result on L^1-continuity of the densities enables us to answer in the affirmative to a conjecture of Cassa [6] stating that the entropy vanishes when $\alpha \to 0$.
Our numerical study of the entropy map also reveals a surprisingly rich self-similar structure, resembling a devil's staircase, which is still unexplained. In particular, contrarily to our expectations the entropy doesn't seem to be monotonic in any neighborhood of the origin[2]. Numerical evidence also suggests the existence of countably many *phase transitions* or discontinuities of $h'(\alpha)$, in addition to the known discontinuity when α is equal to the Golden number.

The second part of the thesis was originally conceived in close relation to the first, and stemmed from the study of some recent applications of continued fractions to the design of space-time codes for wireless channels.
The wide diffusion of wireless communications has led to a growing demand for an increase in the capacity and reliability of digital transmission systems over fading channels.
The presence of fading effects, that is unpredictable perturbations and attenuations of the signal depending on the environment, causes a considerable loss in the capacity of these channels compared to the classical Additive White Gaussian Noise model. The use of coding together with multiple transmit and receive antennas can greatly reduce this loss without requiring any increase in the total transmitted power. Even though fading hinders transmission, its randomness can be seen as an advantage, and its negative effects can be reduced by increasing the number of independent transmit-receive paths or *diversity* of the system.
In a MIMO setting with M transmit antennas and N receive antennas, an information message u is encoded in an $M \times T$ matrix or *space-time block* $B(u) = (b_{ij})$, where b_{ij} is the signal emitted by antenna i at time $j \in \{1, \dots, T\}$, and T is the duration of the signal.
The maximum *rate* of transmission that can be achieved using space-time blocks is of $\min(M, N)$ symbols per channel use; the diversity is equal to MR, where R is the minimum rank of the matrices $B(u)$, and ought to be maximized. In the case of full diversity, the dominant term in the union bound estimate for the error probability is the *coding gain* $\Delta^{\frac{1}{M}}$, where $\Delta = \min_u \det(B(u)B(u)^H)$.

[1] More precisely, for $\alpha \in (0, 1]$ the average length of the continued fraction expansion of a rational number $\frac{p}{q}$ is $h(\alpha) \log q$; when $\alpha = 0$ the complexity is of the order of $\log^2 q$, see [23].

[2] This has been very recently proved by Nakada and Natsui (personal communication).

In [10] and [11], the problem of maximizing the coding gain for a class of full-rank MIMO codes called *Threaded Algebraic Space-Time Codes* or "TAST" is shown to be related to the diophantine approximation of complex numbers by algebraic numbers. Some bounds for the code performance are derived from a generalization of Liouville's Theorem. In particular, finding suitable algebraic numbers which have the worst order of approximation by rationals, that is, such that the elements in their continued fraction expansions are small, is the key to optimizing the code design.
This relation is not as surprising as it might appear at a first glance: in fact,

> "It is an interesting approach to see the design of space-time encoding as searching irrational numbers the "furthest" from rational approximations. On the other hand, the decoding process is equivalent to searching rational integers the closest to irrational numbers; and both, encoding and decoding, can be approached by the same algorithm (Sphere Decoder) of searching nonzero short vectors in a given lattice." [10]

Another application, described in [21], involves *differential diagonal space-time coding*, a design in which the information bits are encoded in the phase differences between one transmitted symbol and the next. In the 2-antenna case, code optimization turns out to be equivalent to finding an integer u such that the continued fraction expansion of $\frac{u}{L}$ has the smallest possible elements, where L is the cardinality of the signal set. In particular, quotients of Fibonacci numbers, which approximate the Golden number and have continued fraction elements all equal to 1, are a good choice.

Both TAST codes and diagonal space-time codes achieve full diversity; however, diagonal designs do not make full use of the antenna capacity; in fact, the transmit antennas are only used to ensure maximum diversity, while the rate of transmission is low, only one symbol per channel use.
TAST codes represent an improvement over diagonal codes, because they are full-rate; however, the major shortcoming of these codes is that the minimum determinant vanishes as the size of the signal set or "constellation" grows to infinity.
A new type of designs, based on suitable subsets of *division algebras*, solves this problem: in fact the minimum determinant, corresponding to the minimum of the reduced norm in a maximal order, is stable.
In the 2×2 case, one of the best schemes known up to date is Belfiore, Rekaya and Viterbo's *Golden Code* $\mathcal{G}$ (2005), a design based on a quaternion algebra containing the field $\mathbb{Q}(i, \theta)$, where θ is the Golden number. This code is full-rate and full-rank, and its *cubic shaping* is convenient for energy efficiency reasons and makes the decoding process faster.
It is possible to build longer, $2 \times 2L$ block codes using the Golden Code as the base alphabet; in particular, the structure of its ideals and quotients can be exploited to increase the minimum determinant, which can be written as a sum involving the determinants of the smaller blocks, and mixed terms of the form $\left\| \widetilde{X}_i X_j \right\|_F^2$, where $X \to \widetilde{X}$ is an involution, and $\| \ \|_F$ is the Frobenius norm. Thus the description of the lattice structure is not sufficient to obtain a good estimate of the coding gain, and the multiplicative structure plays an important role.

In §6.5, we consider block codes based on the cosets of a left ideal of $\mathcal{G}$ of index 4. In this simple case, the estimates of the mixed terms in the expression of the minimum determinant can be carried out in full detail, at least for short codes. When considering ideals of greater index, however, the approach based on direct computation of the codeword weights becomes impractical. Using two-sided ideals it is possible to obtain global estimates, as they are invariant with respect to involution and multiplication. Moreover, it is preferable to choose ideals whose index is a power of two, since binary partition schemes are simpler and better suited to digital data storage.
In §6.6.2, we describe the structure of the two-sided ideals of $\mathcal{G}$ whose index is a power of two and of the respective quotients, which turn out to be matrix rings over $\mathbb{F}_{2^n} + u\mathbb{F}_{2^n}$, where $u^2 = 0$. This structure can be exploited directly to build simple lifts of repetition codes on the quotient. The simulation results for the transmission chain using these codes show that they perform better than the uncoded case and confirm the expectations based on the estimates of the mixed terms.
In §6.8, we introduce some designs which improve the performance of the Golden Code in the *slow-fading* setting. When the channel changes so slowly that it can be considered constant for long time lapses, the ergodicity assumption must be dropped and the diversity of the system is reduced, leading to a performance loss.
To compensate for this loss, we combine a modulation scheme for the quotient ring $\mathcal{G}/2\mathcal{G}$ with an error-correcting code (a shortened Reed-Solomon code) to increase the minimum Hamming weight of the code. Performance simulations show that in the 4-QAM case, corresponding to a single signal point per coset, these codes achieve a remarkable gain with respect to the uncoded Golden Code at the same spectral efficiency, that is at the same bit-rate per channel use. These codes can be extended to the case of 16-QAM modulation with multiple points per coset, although the gain in this case is somewhat smaller, being limited by the minimum distance in the ideal.

Part I

α-continued fractions

Introduction

Let $\alpha \in [0,1]$. We will consider the one-parameter family of maps $T_\alpha : I_\alpha \to I_\alpha$, where $I_\alpha = [\alpha - 1, \alpha]$, defined by

$$T_\alpha(x) = \left|\frac{1}{x}\right| - \left[\left|\frac{1}{x}\right| + 1 - \alpha\right] \tag{1}$$

These dynamical systems generalize the Gauss map ($\alpha = 1$) and the nearest integer continued fraction map ($\alpha = \frac{1}{2}$); they were introduced by H. Nakada [16]. For all $\alpha \in (0,1]$ these maps are expanding, and even though in general they aren't Markovian nor have finite range structure, it can be shown that they admit a unique absolutely continuous invariant probability measure $d\mu_\alpha = \rho_\alpha(x)dx$ (for a detailed proof in this particular case see for example [3]). Nakada computed the invariant densities ρ_α for $\frac{1}{2} \leq \alpha \leq 1$ by finding an explicit representation of their natural extensions. The maps ρ_α turn out to be piecewise finite sums of linear fractional functions. The case $\sqrt{2} - 1 \leq \alpha \leq \frac{1}{2}$ was later studied by Moussa, Cassa and Marmi [15] for a slightly different version of the maps, that is $M_\alpha(x) : [0, \max(\alpha, 1-\alpha)] \to [0, \max(\alpha, 1-\alpha)]$ defined as follows:

$$M_\alpha(x) = \left|\frac{1}{x} - \left[\frac{1}{x} + 1 - \alpha\right]\right| \tag{2}$$

Notice that for a given α, M_α is a factor of T_α: in fact $T_\alpha \circ h = h \circ M_\alpha$, where $h : x \mapsto |x|$ is the absolute value. Since all the corresponding results for the maps M_α can be derived through this semiconjugacy, in the following paragraphs we will focus on the maps T_α.
Cassa found the invariant density for $\sqrt{2} - 1 \leq \alpha \leq \frac{1}{2}$ using an alternative method to the natural extension, which involves counting the poles of a meromorphic function [6]; like the natural extension, this method doesn't provide an algorithm to find the density, but only a means to verify that a certain candidate is valid. In §3.2, we include the natural extension for the maps T_α for this case.

It can be shown [8] that the Kolmogorov-Sinai entropy with respect to the unique absolutely continuous invariant measure μ_α of the T_α is given by Rohlin's formula:

$$h(T_\alpha) = \int_{\alpha-1}^{\alpha} \log|T'_\alpha(x)| d\mu_\alpha(x)$$

Actually, Rohlin's formula applies also to the M_α, and $h(T_\alpha) = h(M_\alpha)$. For $\sqrt{2} - 1 \leq \alpha \leq 1$, the entropy can be computed explicitly from the expression of the invariant densities [16], [15]:

$$h(T_\alpha) = \begin{cases} \frac{\pi^2}{6\log(1+\alpha)} & \text{for } g < \alpha \leq 1 \\ \frac{\pi^2}{6\log G} & \text{for } \sqrt{2} - 1 \leq \alpha \leq g \end{cases} \tag{3}$$

In particular, the entropy is constant when $\sqrt{2} - 1 \leq \alpha \leq g$ and its derivative has a discontinuity (*phase transition*) in $\alpha = g$.

The case $\alpha = 0$ requires a separate discussion; in fact, due to the presence of an indifferent fixed point, T_0 doesn't admit a finite invariant density, although

it is invariant with respect to the infinite measure $d\mu_0 = \frac{dx}{1+x}$. Therefore the entropy of T_0 can only be defined in *Krengel's sense*, that is up to multiplication by a constant (see Thaler [22] for a study of the general one-dimensional case). Following Thaler, for any subset A of $[0,1]$ with $0 < \mu_0(A) < \infty$ we can define

$$h(T_0, \mu_0) \doteqdot \mu_0(A) h((T_0)_A)$$

where $h((T_0)_A)$ is the entropy of the first return map of T_0 on A with respect to the normalized induced measure $\mu_A = \frac{\mu_0}{\mu_0(A)}$. This quantity is well-defined since the product $h(T_0, \mu_0)$ doesn't depend on the choice of A, and it has been computed exactly: $h(T_0, \mu_0) = \frac{\pi^2}{3 \log 2}$ [23]. Since this is a finite value, for a sequence A_k of subsets whose Lebesgue measure tends to 1 we would have $h((T_0)_{A_k}) = \frac{\pi^2}{(3 \log 2)\mu_0(A_k)} \to 0$. In this restricted sense we can say that "the entropy of T_0 is 0". Expression (3) suggests the notion that the dynamical systems T_α are somehow related and have a common origin; actually for $\frac{1}{2} \leq \alpha \leq g$ their natural extensions are all isomorphic. In fact, C. Kraaikamp proved that for these values of α the natural extensions are invertible Bernoulli shifts, and so having the same entropy is a sufficient condition for isomorphism [12]. Moreover, a recent result by R. Natsui [17] shows that the natural extensions of the Farey maps associated to the T_α are all isomorphic when $\frac{1}{2} \leq \alpha \leq 1$.
It is well-known that the maps T_1 and T_0 descend from the geodesic flow on the unit tangent bundle of the modular surface $PSL(2,\mathbb{Z})\backslash PSL(2,\mathbb{R})$ [21], [10]. Indeed we can represent this flow as a suspension flow over the natural extension of these maps and deduce in this way the invariant probability measures from the normalized Haar measure on $PSL(2,\mathbb{Z})\backslash PSL(2,\mathbb{R})$. It is natural to conjecture that the same happens for all the maps T_α, $\alpha \in [0,1]$. If this were true, one could (at least in principle) apply Abramov's formula to compute the entropies $h(\alpha)$ from the entropy of the geodesic flow.

We now summarize briefly the contents of the various sections of Part I.
In §1, we introduce α-continued fraction expansions and their basic properties, and remarking that for $\alpha \in \left[\frac{1}{2}, 1\right)$, the sequence of α-convergence can be seen as an acceleration of the sequence of standard (Gauss) convergents. We also recall how the exactness (and therefore ergodicity) of the system follows from the fact that the cylinder sets generate the Borel σ-algebra [16]. Finally, we remark how Rohlin's formula for the entropy holds in this case.
In §2.1 we prove that the entropy $h(\alpha)$ of T_α is continuous in α when $\alpha \in (0,1]$ and that $h(\alpha) \to 0$ as $\alpha \to 0$, as it had been conjectured by Cassa [6]. This result follows from the fact that the invariant densities are a continuous family in the L^1 norm with respect to α, and is based on a uniform version of the Lasota-Yorke inequality for the Perron-Frobenius operator of T_α, following M. Viana's approach [24]; in the uniform case, however, a further difficulty arises from the existence of arbitrarily small cylinders containing the endpoints, requiring *ad hoc* estimates.
In §2.2 we analyse the results of numerical simulations for the entropy obtained through Birkhoff sums, which suggest that the entropy function has a complex self-similar structure.
In §3.1, the notion of natural extension is introduced, following Schweiger [20]. Finally, in §3.3 we compute the natural extension and the invariant densities of the T_α for the sequence $\left\{\alpha = \frac{1}{r}\right\}_{r \in \mathbb{N}}$.

Chapter 1

α-continued fractions

In this chapter we introduce a family of piecewise monotonic maps of the interval which generalize the Gauss map, and give rise to a class of continued fraction approximations.

1.1 α-expansions

For $\alpha \in [0,1]$, let $I_\alpha = [\alpha - 1, \alpha)$. Consider the maps $T_\alpha : I_\alpha \to I_\alpha$ defined as follows [16]:

$$T_\alpha(x) = \left|\frac{1}{x}\right| - \left[\left|\frac{1}{x}\right|\right]_\alpha,$$

where $[x]_\alpha \doteqdot [x + 1 - \alpha]$. It is convenient to assume that $T_\alpha(0) = 0$.

Remark 1.1. When $\alpha = 1$, T_α is the Gauss map; for $\alpha = \frac{1}{2}$, it is the nearest integer continued fraction map.

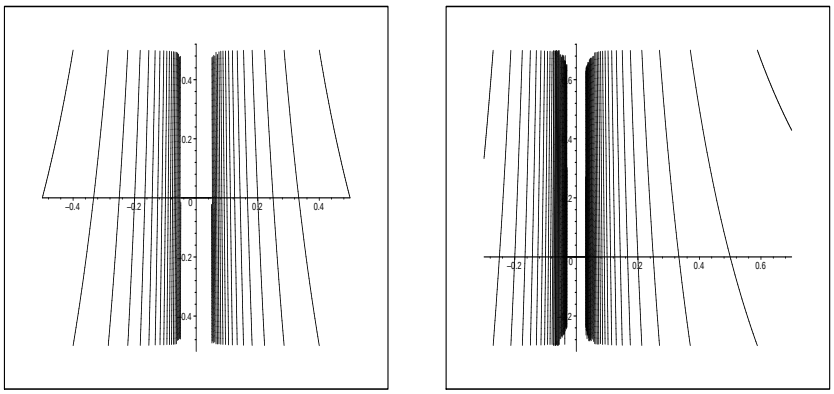

Figure 1.1: Graph of T_α when $\alpha = \frac{1}{2}$ and $\alpha = 0.7$ respectively.

The graph of T_α can also be obtained by intersecting the union of the sequence of hyperbolae $\left\{\left|\frac{1}{x}\right| - n\right\}, n \in \mathbb{N}$ the square $[\alpha - 1, \alpha - 1] \times [\alpha, \alpha]$ (see Figure

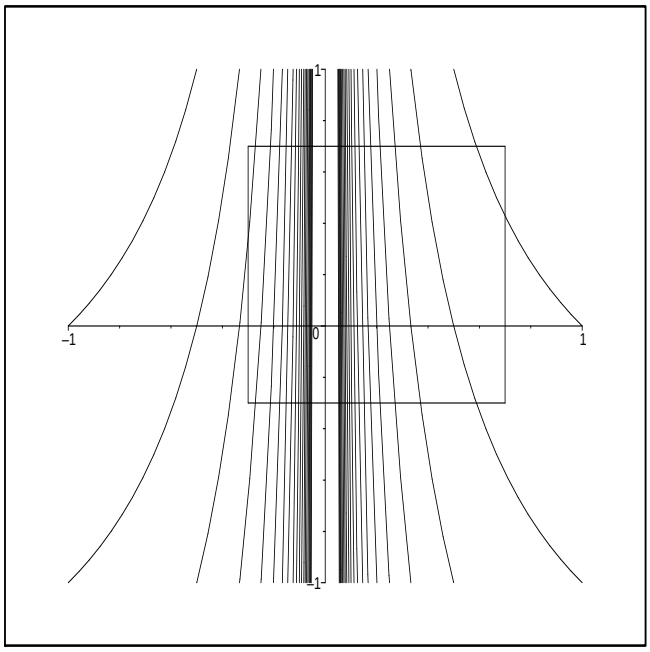

Figure 1.2: The graph of the map T_α is obtained by intersecting a family of hyperbolae with the square $[\alpha - 1, \alpha - 1] \times [\alpha, \alpha]$.

1.2). By moving the square along the diagonal, we obtain the whole family of α-continued fraction maps.
The maps T_α are related to the following *symbolic dynamics*: for α fixed, and $x \neq 0$, let

$$\begin{cases} a(x) = \left\lfloor \left| \dfrac{1}{x} \right| + 1 - \alpha \right\rfloor, \\ \varepsilon(x) = \operatorname{sign}(x), \end{cases}$$

and define $a(0) = \infty$, $\varepsilon(0) = 1$.
For any $x \in I_\alpha$, let $x_0 = x$, $x_n = T_\alpha^n(x)$, when $n \geq 1$, and

$$\begin{cases} a_n = a(x_{n-1}), \\ \varepsilon_n = \varepsilon(x_{n-1}) \end{cases}$$

Thus we obtain inductively a *continued fraction expansion* associated to T_α: $\forall n \geq 1$,

$$x = \cfrac{\varepsilon_1}{a_1 + \cfrac{\varepsilon_2}{a_2 + \cfrac{\varepsilon_3}{\ddots + \cfrac{\varepsilon_n}{a_n + x_n}}}}$$

For the sake of simplicity, we will denote this expression by

$$x = [(\varepsilon_1, a_1), (\varepsilon_2, a_2), \ldots, (\varepsilon_n, a_n); x_n]$$

The resulting expansion is infinite, of the form $[(\varepsilon_1, a_1), (\varepsilon_2, a_2), \ldots, (\varepsilon_n, a_n), \ldots]$ when x is irrational; when x is rational, the expansion is finite with length n, where n is the minimum index such that $x_n = 0$.
By truncating the expansion to the n-th step, we obtain the *α-convergents* of x, that is the reduced fraction

$$\frac{p_n}{q_n} = [(\varepsilon_1, a_1), (\varepsilon_2, a_2), \ldots, (\varepsilon_n, a_n)] = \cfrac{\varepsilon_1}{a_1 + \cfrac{\varepsilon_2}{\ddots + \cfrac{\varepsilon_n}{a_n}}},$$

with the convention that $p_{-1} = 1,\ q_{-1} = 0,\ p_0 = 0,\ q_0 = 1$.

Remark 1.2. We observe once and for all that the sequences $\{a_n\}$, $\{\varepsilon_n\}$, $\{x_n\}$, $\{p_n\}$, $\{q_n\}$ are a function of the parameter α and the starting point x. We will omit this dependence unless necessary, in order to simplify notation.

The following recursive relations among the convergents are easily proved by induction:

$$\begin{aligned} p_n &= a_n p_{n-1} + \varepsilon_n p_{n-2} \\ q_n &= a_n q_{n-1} + \varepsilon_n q_{n-2} \end{aligned} \tag{1.1}$$

Observe that

$$p_{n+1}q_n - q_{n+1}p_n = -\varepsilon_n(p_n q_{n-1} - q_n p_{n-1})$$

and so, since $p_0 q_1 - q_1 q_0 = -p_1 = -\varepsilon_1$,

$$p_n q_{n+1} - p_{n+1} q_n = \varepsilon_1 \varepsilon_2 \cdots \varepsilon_n (-1)^{n-1}, \quad |p_n q_{n+1} - p_{n+1} q_n| = 1 \tag{1.2}$$

Then, always by induction, we find

$$x = \frac{p_n + x_n p_{n-1}}{q_n + x_n q_{n-1}} \tag{1.3}$$

for $n \geq 0$. In fact, the basis of the induction is trivially $\frac{p_0 + x_0 p_{-1}}{q_0 + x_0 q_{-1}} = \frac{x_0}{1}$, and supposing that the relation (1.3) holds for some $n \geq 0$, using the recursive formulas (1.1) and the relation $x_{n+1} = \frac{\varepsilon_{n+1}}{x_n} - a_{n+1}$, we get

$$x = \frac{p_n(1 - \varepsilon_{n+1} a_{n+1} x_n) + \varepsilon_{n+1} x_n p_{n+1}}{q_n(1 - \varepsilon_{n+1} a_{n+1} x_n) + \varepsilon_{n+1} x_n q_{n+1}} = \frac{p_n x_{n+1} + p_{n+1}}{q_n x_{n+1} + q_{n+1}}$$

Now consider

$$\beta_n = |q_n x - p_n| \tag{1.4}$$

There are three useful alternative expression for this quantity: first, from the relations (1.3) and (1.2), we find

$$\beta_n = \left| \frac{x_n(q_n p_{n-1} - p_n q_{n-1})}{q_n + x_n q_{n-1}} \right| = \frac{|x_n|}{q_n + x_n q_{n-1}} \tag{1.5}$$

From equation (1.3), we can derive

$$x_n = -\left(\frac{p_n - x q_n}{p_{n-1} - x q_{n-1}} \right),$$

so we have

$$\beta_n = \left| \prod_{i=0}^{n} x_i \right| \tag{1.6}$$

But equations (1.5) and (1.6) also imply

$$\beta_n = \frac{\beta_{n+1}}{|x_{n+1}|} = \frac{1}{q_{n+1} + x_{n+1} q_n} \tag{1.7}$$

From (1.7), we obtain an estimate of the rate of convergence of the $\frac{p_n}{q_n}$ to x: if $x_n \geq 0$,

$$\frac{1}{q_n(q_{n+1} + \alpha q_n)} < \left| x - \frac{p_n}{q_n} \right| = \frac{\beta_n}{q_n} = \frac{1}{q_n(q_{n+1} + x_{n+1} q_n)} < \frac{1}{q_n q_{n+1}} \tag{1.8}$$

while for $x_n < 0$,

$$\frac{1}{q_n q_{n+1}} < \left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_n(q_{n+1} - (1-\alpha) q_n)} < \frac{1}{\alpha q_n q_{n+1}} \tag{1.9}$$

When $\alpha \in \left[\frac{1}{2}, 1\right)$, the α-convergents turn out to be a subsequence of the standard continued fraction convergents [4]. In this sense, α-continued fractions can be seen as an "acceleration" of 1-continued fractions:

Lemma 1.3. *Fix $\alpha \in \left[\frac{1}{2}, 1\right)$. Let $x \in \mathbb{R} \setminus \mathbb{Q}$, and denote by $\frac{P_n}{Q_n}$ the standard continued fraction convergents of x, and by $\frac{p_n}{q_n}$ its α-convergents. Then*

$$\frac{p_n}{q_n} = \frac{P_{k_\alpha(n)}}{Q_{k_\alpha(n)}},$$

where $k_\alpha : \mathbb{N} \to \mathbb{N}$ is defined inductively as follows:

$$k_\alpha(-1) = -1,$$
$$k_\alpha(n+1) = \begin{cases} k_\alpha(n) + 1 & \text{if } \varepsilon_{n+1} = 1 \\ k_\alpha(n) + 2 & \text{if } \varepsilon_{n+1} = -1. \end{cases}$$

Moreover, if $k_\alpha(n+1) = k_\alpha(n) + 2$ we have

$$q_{k_\alpha(n+1)} = q_{k_\alpha(n)+2} = q_{k_\alpha(n)+1} + q_{k_\alpha(n)}.$$

When $\alpha \in \left(0, \frac{1}{2}\right)$, this lemma doesn't hold any longer, and sequences of the form $\left\{ \frac{p_{n_j}}{q_{n_j}}, \ldots, \frac{p_{n_j+k}}{q_{n_j+k}} \right\}$, such that $\frac{p_{n_j+i}}{q_{n_j+i}}$ is not a standard convergent for $i = 0, \ldots, k$, appear. These correspond to sequences of length k of digits "$(2, -1)$", called *desingularization sequences*.

Now suppose that we know the standard continued fraction expansion $x = (w_1, w_2, w_3, \ldots)$ of an irrational number, and we want to derive its α-expansion $x = [(\varepsilon_1, a_1), (\varepsilon_2, a_2), (\varepsilon_3, a_3), \ldots]$. We do not know whether there exists a concise formula expressing this relation; however, it is not hard to define a step-by-step algorithm to pass from one expansion to the other.

The content of the following lemma is the same (although with different notations) of Theorem 7 in [15].

Lemma 1.4. *Fix $\alpha \in \left[0, \frac{1}{2}\right)$, and let $x \in [\alpha - 1, \alpha)$ be an irrational number with standard continued fraction expansion $x = w_0 + (w_1, w_2, \ldots)$, $w_0 \in \{0, -1\}$. Let $\frac{P_n}{Q_n}$ be the standard convergents of x, and $\frac{p_n}{q_n}$ its α-convergents.*
Then there exist two subsequences $\{n_j\}$ and $\{n_k\}$ such that

$$\frac{p_{n_j}}{q_{n_j}} = \frac{P_{n_k}}{Q_{n_k}}$$

More precisely, we define the following algorithm:

1.5 (One step of α-expansion). 1. First step:

- If $x \in (0, \alpha)$, then define $\varepsilon_1 = 1$ and $|x_0| = x = (w_1, w_2, \ldots)$. Obviously $\frac{p_0}{q_0} = \frac{P_0}{Q_0} = 0$.
- If $x \in [\alpha - 1, 0)$, define $\varepsilon_1 = -1$ and $|x_0| = -x$. We distinguish two cases:
  - If $w_1 = 1$, using the well-known identity

$$1 - \frac{1}{b + y} = \cfrac{1}{1 + \cfrac{1}{b - 1 + y}}$$

    for $b \geq 2$ and $y \in (0, 1)$, we find $-x = (w_2 + 1, w_3, \ldots)$.
  - If $w_1 > 1$, from the identity

$$1 - \cfrac{1}{n + \cfrac{1}{b}} = (\underbrace{(1, 2), (-1, 2), \ldots, (-1, 2)}_{n-1}; -\frac{1}{b+1}), \quad b \geq 1$$

    we get $a_1 = \ldots = a_{w_1 - 1} = 2, \varepsilon_2 = \ldots = \varepsilon_{w_1} = -1$, $|x_{w_1 - 1}| = (w_2 + 1, w_3, \ldots)$; $\frac{p_i}{q_i} = \frac{i}{i+1}$ for $i = 1, \ldots, w_1 - 1$, and

$$\frac{p_{w_1 - 1}}{q_{w_1 - 1}} = \frac{w_1}{w_1 + 1} = 1 - \frac{1}{w_1} = \frac{P_1}{Q_1}$$

2. Inductive step:
   Now suppose that we have found the first n digits of the α-expansion, such that $\frac{p_n}{q_n} = \frac{P_k}{Q_k}$ for some $k \geq 0$:

$$x = \cfrac{\varepsilon_1}{a_1 + \cfrac{\varepsilon_2}{a_2 + \ddots + \cfrac{\varepsilon_n}{a_n + \varepsilon_{n+1} |x_n|}}}, \quad \text{such that}$$

$$|x_n| = (w_{k+1}^{(n)}, w_{k+2}, w_{k+3}, \ldots) \in (0, 1 - \alpha), \qquad w_{k+1}^{(n)} \in \{w_{k+1}, w_{k+1} + 1\}$$

   Then

- If $T(|x_n|) < \alpha$, we have

$$\begin{aligned} &\varepsilon_{n+2} = 1, \ a_{n+1} = w_{k+1}^{(n)}, \\ &\frac{p_{n+1}}{q_{n+1}} = \frac{P_{k+1}}{Q_{k+1}}, \ |x_{n+1}| = (w_{k+2}, w_{k+3}, \ldots) \end{aligned} \tag{1.10}$$

- If $T(|x_n|) \geq \alpha$,

$$\begin{aligned}
&\varepsilon_{n+2} = -1,\ a_{n+1} = w_{k+1}^{(n)} + 1, \quad \text{if } w_{k+2} = 1\\
&\varepsilon_{n+2} = \cdots = \varepsilon_{n+w_{k+3}} = -1,\ a_{n+1} = w_{k+1}^{(n)} + 1,\\
&a_{n+2} = \cdots = a_{n+w_{k+2}} = 2 \ \text{ if } w_{k+2} \geq 2,\\
&\frac{p_{n+i}}{q_{n+i}} = \frac{iP_{k+1} + P_k}{iQ_{k+1} + Q_k} \quad \forall 1 \leq i \leq w_{k+2} - 1, \quad \frac{p_{n+w_{k+2}}}{q_{n+w_{k+2}}} = \frac{P_{k+2}}{Q_{k+2}},\\
&\left|x_{n+w_{k+2}}\right| = (w_{k+3} + 1, w_{k+4}, \ldots)
\end{aligned} \tag{1.11}$$

Proof. It is clear that equations (1.10) and (1.11) imply the existence of the two identical sequences $\frac{p_{n_j}}{q_{n_j}}$, $\frac{P_{n_k}}{Q_{n_k}}$ by induction, where the basis of the induction is given by the first step in the algorithm.

We have $a_{n+1} = \left[\left|\frac{1}{x_n}\right| + 1 - \alpha\right]$, and so

$$a_{n+1} = w_{k+1}^{(n)} \Leftrightarrow w_{k+1}^{(n)} + \frac{1}{w_{k+2} + \dfrac{1}{w_{k+3} + \cdots}} + 1 - \alpha < w_{k+1}^{(n)} + 1 \Leftrightarrow T(|x_n|) < \alpha$$

Clearly in this case $\varepsilon_{n+2} = 1$, and the remainder $|x_{n+1}|$ is equal to $T(|x_n|)$; otherwise we have $a_{n+1} = w_{k+1}^{(n)} + 1$ and $\varepsilon_{n+2} = -1$.
Observe that since the recursive relations defining the p_i and the q_i have the same form, it is sufficient to prove the statements above for the p_i.

When $T(|x_n|) < \alpha$, we distinguish two cases:

- If $\varepsilon_{n+1} = 1$, by inductive hypothesis $\frac{p_{n-1}}{q_{n-1}} = \frac{P_{k-1}}{Q_{k-1}}$, and $w_{k+1}^{(n)} = w_{k+1}$. Then
$$p_{n+1} = a_{n+1}p_n + \varepsilon_{n+1}p_{n-1} = w_{k+1}^{(n)}P_k + P_{k-1} = P_{k+1}$$

- If $\varepsilon_{n+1} = -1$, $w_{k+1}^{(n)} = w_{k+1} + 1$, again by inductive hypothesis

$$\begin{aligned}
&\frac{P_{k-1}}{Q_{k-1}} = \frac{p_{n-w_k^{(n)}}}{q_{n-w_k^{(n)}}},\\
&p_{n-1} = p_{(n-w_k^{(n)})+(w_k^{(n)}-1)} = (w_k^{(n)} - 1)P_{k-1} + P_{k-2} = P_k - P_{k-1},\\
&\Rightarrow p_{n+1} = (w_{k+1} + 1)P_k - P_k + P_{k-1} = w_{k+1}P_k + P_{k-1} = P_{k+1}
\end{aligned}$$

When $T(|x_n|) \geq \alpha$,

$$\begin{aligned}
\left|\frac{1}{x_n}\right| &= w_{k+1}^{(n)} + 1 - \left(1 - \frac{1}{w_{k+2} + \dfrac{1}{w_{k+3} + \cdots}}\right) =\\
&= a_{n+1} - \left(1 - \frac{1}{w_{k+2} + T^2(|x_n|)}\right) = a_{n+1} - |x_{n+1}|
\end{aligned}$$

Then if $w_{k+2} \geq 2$, we have $\left|\frac{1}{x_{n+1}}\right| = 1 + \frac{1}{w_{k+2}-1+T^2(|x_n|)}$, and since

$$1 \geq \frac{1}{w_{k+2} - 1 + T^2(x_n)} > \frac{1}{w_{k+2} + T^2(x_n)} \geq \alpha,$$

we find

$$a_{n+2} = \left\lceil \left| \frac{1}{x_{n+1}} \right| + 1 - \alpha \right\rceil = 2,\ \varepsilon_{n+3} = -1,$$

and so on: it is easy to prove by induction that for $1 \leq i \leq w_{k+2} - 1$,

$$\left|\frac{1}{x_{n+i}}\right| = 1 + \frac{1}{w_{k+2} - i + T^2(|x_n|)} \geq 1 + \alpha \Rightarrow a_{n+i+1} = 2,\ \varepsilon_{n+i+2} = -1,$$

up to

$$\left|\frac{1}{x_{n+w_{k+2}}}\right| = 1 + \frac{1}{T^2(|x_n|)} = 1 + w_{k+3} + \frac{1}{w_{k+4} + \cdots}$$

which is true also when $w_{k+2} = 1$.
In conclusion,

$$|x_n| = [(w_{k+1}^{(n)} + 1, -), \underbrace{(2,-)(2,-), \ldots, (2,-)}_{w_{k+2}-1}; \left|x_{n+w_{k+2}}\right|] \tag{1.12}$$

Again we distinguish two cases:

- If $\varepsilon_{n+1} = 1$, then by inductive hypothesis $w_{k+1}^{(n)} = w_{k+1}$, $\frac{p_{n-1}}{q_{n-1}} = \frac{P_{k-1}}{Q_{k-1}}$.

$$p_{n+1} = a_{n+1}p_n + \varepsilon_{n+1}p_{n-1} = (w_{k+1} + 1)P_k + P_{k+1} = P_{k+1} + P_k$$

- If $\varepsilon_{n+1} = -1$, then $w_{k+1}^{(n)} = w_{k+1} + 1$,

$$p_{n-1} = p_{(n-w_k^{(n)})+(w_k^{(n)}-1)} = (w_k^{(n)} - 1)P_{k-1} + P_{k-2} = P_k - P_{k-1},$$
$$p_{n+1} = (w_{k+1} + 2)P_k + P_k - P_{k-1} = (w_{k+1} + 1)P_k + P_{k-1} = P_{k+1} + P_k$$

So in both cases we have $p_{n+1} = P_{k+1} + P_k$, and

$$p_{n+2} = a_{n+2}p_n + \varepsilon_{n+2}p_n = 2(P_{k+1} + P_k) - P_k = 2P_{k+1} + P_k$$

By induction we can prove that for $1 \leq i \leq w_{k+2} - 1$,

$$p_{n+i} = iP_{k+1} + P_k,$$

up to

$$p_{n+w_{k+2}} = w_{k+2}P_{k+1} + P_k = P_{k+2},$$

which completes the proof. □

1.2 Symbolic dynamics

1.6 (Cylinders of rank 1). Let $\alpha \in (0,1]$. The map T_α is piecewise monotonic and piecewise analytic on the countable partition $\mathcal{P} = \{I_j^+\}_{j \geq j_{\min}} \cup \{I_j^-\}_{j \geq j'_{\min}}$, where $j_{\min} = \left[\left|\frac{1}{\alpha}\right| + 1 - \alpha\right], j'_{\min} = \left[\left|\frac{1}{1-\alpha}\right| + 1 - \alpha\right]$ and the elements of $\mathcal{P}$ are called *cylinders of rank* 1:

$$I_j^+ \doteq \left(\frac{1}{j+\alpha}, \frac{1}{j-1+\alpha}\right], \ j \in [j_{\min}+1, \infty), \quad I_{j_{\min}}^+ \doteq \left(\frac{1}{j_{\min}+\alpha}, \alpha\right],$$
$$I_j^- \doteq \left[-\frac{1}{j-1+\alpha}, -\frac{1}{j+\alpha}\right), \ j \in [j'_{\min}+1, \infty), \quad I_{j'_{\min}}^- \doteq \left[\alpha - 1, -\frac{1}{j'_{\min}+\alpha}\right)$$

T_α is monotone on each cylinder and we have

$$\begin{cases} T_\alpha(x) = \frac{1}{x} - j, & x \in I_j^+, \ j \in \mathbb{N} \cap [j_{\min}, \infty) \\ T_\alpha(x) = -\frac{1}{x} - j, & x \in I_j^-, \ j \in \mathbb{N} \cap [j'_{\min}, \infty) \end{cases}$$

We also find that for $\alpha \in (0,1)$, T_α is *expanding*, that is $|T'_\alpha(x)| > 1$ almost everywhere[1]: in fact for all $x \in I_\alpha$,

$$\frac{1}{|T'_\alpha(x)|} \leq \lambda = (1-\alpha) < 1 \tag{1.13}$$

1.7 (Cylinders of rank n; full cylinders). Let $\mathcal{P}^{(n)} = \bigvee_{i=0}^{n-1} T_\alpha^{-i}(\mathcal{P})$ be the induced partition in monotonicity intervals of T_α^n. Each cylinder $I_\eta^{(n)} \in \mathcal{P}^{(n)}$ is uniquely determined by the sequence

$$((j_0(\eta), \varepsilon_0(\eta)), \ldots, (j_{n-1}(\eta), \varepsilon_{n-1}(\eta)))$$

such that for all $x \in I_\eta^{(n)}$, $T_\alpha^i(x) \in I_{j_i(\eta)}^{\varepsilon_i(\eta)}$. On each cylinder T_α^n is a Möbius map $T_\alpha^n(x) = \frac{ax+b}{cx+d}$, where $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL(2,\mathbb{Z})$. We will say that a cylinder $I_\eta^{(n)} \in \mathcal{P}$ is *full* if $T_\alpha^n(I_\eta^{(n)}) = I_\alpha$.

1.3 The Perron-Frobenius operator

1.8 (Perron-Frobenius operator). Let $V_\eta : T_\alpha^n(I_\eta^{(n)}) \to I_\eta^{(n)}$ be the inverse branches of T_α^n, and P_{T_α} the Perron-Frobenius operator associated with T_α. Then for every $\varphi \in L^1(I_\alpha)$,

$$(P_{T_\alpha}^n \varphi)(x) = \sum_{I_\eta^{(n)} \in \mathcal{P}^n} \frac{\varphi(V_\eta(x))}{\left|(T_\alpha^n)'(V_\eta(x))\right|} \chi_{T_\alpha^n(I_\eta^{(n)})}(x) \tag{1.14}$$

On $I_\eta^{(n)}$ we have the following bound:

$$\sup_{I_\eta^{(n)}} \frac{1}{\left|(T_\alpha^n)'(x)\right|} = \sup_{I_\eta^{(n)}} \frac{1}{\left|T'_\alpha\left(T_\alpha^{n-1}(x)\right) \cdots T'_\alpha(x)\right|} \leq \lambda_\eta^{(n)} \leq \lambda^n, \tag{1.15}$$

[1] The only value of α in which $|T'_\alpha(x)| = 1$ for any point is actually the Gauss map T_1, with its fixed point $x = 1$. The by-excess map T_0, which we are not taking into account here, also has a parabolic point, and is *not* expanding.

where $\lambda_\eta^{(n)} \doteqdot \lambda_{j_0(\eta)} \cdots \lambda_{j_{n-1}(\eta)}$. Recall that for $f_1, \ldots, f_n \in BV$,

$$\mathrm{Var}(f_1 \cdots f_n) \leq \sum_{k=1}^{n} \mathrm{Var}(f_k) \prod_{i \neq k} \sup |f_i| \tag{i}$$

and consequently

$$\operatorname*{Var}_{I_\eta^{(n)}} \frac{1}{\left|(T_\alpha^n)'(x)\right|} = \operatorname*{Var}_{I_\eta^{(n)}} \frac{1}{\left|T'_\alpha\left(T_\alpha^{n-1}(x)\right) \cdots T'_\alpha(T_\alpha(x)) \cdot T'_\alpha(x)\right|} \leq n\lambda_\eta^{(n)}$$

Finally, we state the following *bounded distortion property*, that we are going to use several times in the sequel:

Proposition 1.9 (Bounded distortion). $\forall \alpha > 0$, $\exists C_1$ *such that* $\forall n \geq 1$, $\forall I_\eta^{(n)} \in \mathcal{P}^{(n)}, \forall x, y \in I_\eta^{(n)}$,

$$\left|\frac{(T_\alpha^n)'(y)}{(T_\alpha^n)'(x)}\right| \leq C_1$$

Moreover, for all measurable set $B \subseteq I_\alpha$, for all full cylinders $I_\eta^{(n)} \in \mathcal{P}^{(n)}$,

$$m(V_\eta(B)) \geq \frac{m(B)m(I_\eta^{(n)})}{C_1},$$

where m denotes the Lebesgue measure.

The proof of this statement follows a standard argument:

Proof. Observe that $\exists k > 0$ such that $\forall I_j^\varepsilon \in \mathcal{P}, \forall x, y \in I_j^\varepsilon$,

$$\left|\frac{T'_\alpha(x)}{T'_\alpha(y)} - 1\right| \leq k\,|T_\alpha(x) - T_\alpha(y)|$$

In fact, if $x, y \in I_{j,\alpha}^\varepsilon$, then

$$\left|\frac{T'_\alpha(x)}{T'_\alpha(y)} - 1\right| \frac{1}{|T_\alpha(x) - T_\alpha(y)|} = \left|\frac{y^2}{x^2} - 1\right| \frac{|xy|}{|x-y|} \leq \left|\frac{y}{x}\right| |x+y| \leq 4$$

Let $n \geq 1$, $I_\eta^{(n)} \in \mathcal{P}^{(n)}$, $x, y \in I_\eta^{(n)}$. Define $\lambda = \sup\left|\frac{1}{T'_\alpha}\right| = (1-\alpha)^2$: then

$$\begin{aligned}
\log\left|\frac{(T_\alpha^n)'(y)}{(T_\alpha^n)'(x)}\right| &= \sum_{i=0}^{n-1} \log\left|\frac{T'_\alpha(T_\alpha^i(y))}{T'_\alpha(T_\alpha^i(x)}\right| \leq \sum_{i=0}^{n-1} \left|\frac{T'_\alpha(T_\alpha^i(y))}{T'_\alpha(T_\alpha^i(x))} - 1\right| \leq \\
&\leq 4\sum_{i=0}^{n-1} \left|T_\alpha^{i+1}(y) - T_\alpha^{i+1}(x)\right| = 4\sum_{i=1}^{n} \left|T_\alpha^i(y) - T_\alpha^i(x)\right| \leq \\
&\leq 4\sum_{i=1}^{n} \lambda^{n-i} \left|T_\alpha^n(y) - T_\alpha^n(x)\right| \leq 4\sum_{i=0}^{\infty} \lambda^i = \frac{4}{1-(1-\alpha)^2} = C_2
\end{aligned} \tag{1.16}$$

Then $\left|\frac{(T_\alpha^n)'(y)}{(T_\alpha^n)'(x)}\right| \leq e^{C_2} = C_1$. Let $I_\eta^{(n)}$ be a full cylinder: $T_\alpha^n(I_\eta^{(n)}) = I_\alpha$. Now consider any measurable set B:

$$\frac{m(B)}{m(I_\alpha)} = \frac{\int\limits_{V_\eta(B)} |(T_\alpha^n)'(y)|\,dy}{\int\limits_{I_\eta^{(n)}} |(T_\alpha^n)'(x)|\,dx} \leq \frac{m(V_\eta(B)) \sup\limits_{y\in I_\eta^{(n)}} |(T_\alpha^n)'(y)|}{m(I_\eta^{(n)}) \inf\limits_{x\in I_\eta^{(n)}} |(T_\alpha^n)'(x)|} \leq C_1 \frac{m(V_\eta(B))}{m(I_\eta^{(n)})}$$

$$\Rightarrow m(V_\eta(B)) \geq m(B)\frac{m(I_\eta^{(n)})}{C_1} \quad (1.17)$$

which concludes the proof. □

1.4 Invariant measures

Sufficient conditions for the existence of absolutely continuous invariant measures (a.c.i.m.) for expanding maps have been extensively studied in the literature. A desirable property in most cases is the *Markov property*:

1.10 (Markov map). Let I be an interval, $\mathcal{P}$ a countable partition of I, $T : I \to I$ such that the restriction of T to each interval of the partition is monotonic and C^2. T is called a *Markov map* if the set $\mathcal{I}^* \doteq \bigcup_{n\geq 1} T^n(\mathcal{P})$ is finite.

In fact, a folklore theorem states that

Theorem 1.11. *If $T : I \to I$ is Markov and expanding, then there exists a unique invariant probability measure for f absolutely continuous with respect to the Lebesgue measure.*

Unfortunately, the Markov condition is not satisfied by the maps T_α except for a set of measure 0 in the parameter α. In fact

$$T_\alpha(\mathcal{P}) = \{[\alpha - 1, \alpha], [T_\alpha(1-\alpha), \alpha], [T_\alpha(\alpha), \alpha]\}$$

The union $\bigcup_{n\geq 1} T_\alpha^n(\mathcal{P})$ is finite only in the following cases:

a) $\exists n, m \in \mathbb{N}$ such that $T_\alpha^n(\alpha) = 0, T_\alpha^m(1-\alpha) = 0$, which happens if and only if α is *rational*;

b) the sequences $\{T_\alpha^i(\alpha)\}_{i\in\mathbb{N}}$ e $\{T_\alpha^i(1-\alpha)\}_{i\in\mathbb{N}}$ are periodic, that is α is *algebraic of degree* 2.

However, it can be proved [3] that for all $\alpha \in (0, 1]$ the maps T_α admit a unique absolutely continuous invariant probability measure μ_α, whose density ρ_α is of bounded variation (and therefore bounded). The proof follows a more general framework, see the study by A. Broise [5]:

Theorem 1.12 (Bourdon, Daireaux, Vallée). *Consider an interval map $T : I \to I$ which is monotone and C^2 on each interval of a countable partition $\mathcal{P}$ of I. Let $I_\eta^{(n)}$ denote the open cylinders of rank n, and V_η the local inverse of T on $I_\eta^{(n)}$. Suppose that T satisfies the following properties:*

a) (Expansivity) $\sup_j \sup_{x \in T(I_j)} \left|V_j'(x)\right| \leq 1$

b) (Strong expansivity) $\exists n_0, \exists \gamma < 1$ *such that* $\sup_{I_\eta^{(n_0)}} \sup_{x \in T^{n_0}(I_\eta^{(n_0)})} \left|V_\eta'(x)\right| < \gamma$

c) $\exists c > 0$ *such that* $\forall I_j \in \mathcal{P}$, $\forall x \in T(I_j)$, $\left|V_j''(x)\right| \leq c \left|V_j'(x)\right|$

d) (Quasi-Markov) $\forall n$, $\inf_{I_\eta^{(n)} \in \mathcal{P}^n} m\left(T^n\left(I_\eta^{(n)}\right)\right) > 0$, *where* m *denotes the Lebesgue measure.*

Then T *admits an invariant density of bounded variation.*

We have already seen in equation (1.13) that the maps T_α are expanding for all $\alpha \in [0,1]$, and strongly expanding for $\alpha \in (0,1]$ (actually, we can take $n_0 = 1$ for $\alpha \in (0,1)$, and $n_0 = 2$ for $\alpha = 1$, see also equation (1.15)). Condition (c) holds for all α, and can be checked directly.
Condition (d) follows trivially from the fact that since $T_\alpha(\mathcal{P})$ is finite (actually it has at most four elements), $T_\alpha^n(\mathcal{P}^{(n)})$ is also finite for each n, and the length of its intervals must be bounded from below. However, there is no uniform bound in α or n for these measures, as we will see in the sequel.
We also remark that a priori the invariant density might be discontinuous in every point of the partition $\bigcup_n T_\alpha^n(\mathcal{P})$ [3].

The uniqueness of the a.c.i.m. is a consequence of the ergodicity of the system:

1.13 (Ergodic system). A measure-preserving dynamical system $(X, \mathcal{A}, T, \mu)$ is *ergodic* if for every measurable set $A \in \mathcal{A}$ such that $T^{-1}(A) = A$, $\mu(A) = 0$ or 1.

1.14 (Exact endomorphism). A surjective measure-preserving dynamical system $(X, \mathcal{A}, T, \mu)$ is said to be *exact* if

$$\bigcap_{n=0}^{\infty} T^{-n}(\mathcal{A}) = \{X, \emptyset\} \qquad (\text{mod } 0) \tag{1.18}$$

In particular, every invariant set is trivial and so the system is also ergodic.
For a proof of the following classical theorem, see for example [7]:

Theorem 1.15. *Consider a dynamical system* $(X, T, \mathcal{A})$ *and two measures* μ_1, μ_2 *on* $(X, \mathcal{A})$ *which are invariant for* T. *If both* $(X, T, \mathcal{A}, \mu_1)$ *and* $(X, T, \mathcal{A}, \mu_2)$ *are ergodic, then either* $\mu_1 = \mu_2$, *or* μ_1 *and* μ_2 *are singular with respect to each other.*

Because of the previous theorem, if T_α is exact it admits at most one invariant density absolutely continuous with respect to the Lebesgue measure. We remark however that there is an infinite number of singular invariant measures for T_α (consider for example any linear combination of Dirac deltas in the fixed points of T_α).

Lemma 1.16 (Exactness). *For all* $\alpha \in [0,1]$, *the dynamical system* (T_α, μ_α) *is exact (and therefore ergodic).*

The proof of this Lemma for $\alpha \in \left[\frac{1}{2}, 1\right]$ was given by H. Nakada ([16], Theorem 2) and his argument can be adapted to our case with slight changes.[2]
Let $\omega_i = (j_i, \varepsilon_i)$ for brevity. The crucial property that we need in order to prove Lemma 1.16 is the following:

Proposition 1.17. *The family of the cylinder sets $I_\eta^{(n)} = (\omega_1, \dots, \omega_n) \in \mathcal{P}^{(n)}$ such that $T_\alpha^n(I_\eta^{(n)}) = I_\alpha$ generates the Borel sets.*

Proof of Proposition 1.17. Consider the sets

$$E_n = \{(\omega_1, \dots, \omega_n) \mid T_\alpha(\omega_1) \neq I_\alpha, T_\alpha^2(\omega_1, \omega_2) \neq I_\alpha, \dots, T_\alpha^n(\omega_1, \dots, \omega_n) \neq I_\alpha\}$$

and let $M_n = m\left(\bigcup_{I_\eta^{(n)} \in E_n} I_\eta^{(n)}\right)$. Consider the orbits of the endpoints with respect to T_α:

$$\alpha = (a_1, a_2, a_3, \dots), \qquad \alpha - 1 = (b_1, b_2, b_3, \dots)$$

Then $E_1 = \{(a_1), (b_1)\}$, and

$$E_n = \{(\omega_1, \dots, \omega_n) \in \mathcal{P}^{(n)} \mid (\omega_2, \dots, \omega_n) \in E_{n-1} \text{ and } \omega_1 = a_1 \text{ or } b_1\} \cup \{(a_1, a_2, \dots, a_n), (b_1, b_2, \dots, b_n)\}$$

In fact if $\omega_1 \notin \{a_1, b_1\}$, we would have $T_\alpha(\omega_1) = I_\alpha$; moreover, if $(\omega_2, \dots, \omega_n) \neq (a_2, \dots, a_n)$, the monotonicity of T_α on (a_1) implies that either $(\omega_2, \dots, \omega_n) \cap T_\alpha(a_1) = \varnothing$, or $(\omega_2, \dots, \omega_n) \subseteq T_\alpha(a_1)$. In this last case $T_\alpha^n(a_1, \omega_2, \dots, \omega_n) = T_\alpha^{n-1}(\omega_2, \dots, \omega_n)$. So we get

$$M_n \leq ((1-\alpha)^2 + \alpha^2) M_{n-1} + m((a_1, \dots, a_n) \cup (b_1, \dots, b_n)),$$

and since $(1-\alpha)^2 + \alpha^2 < 1$ and $m(w_1, \dots, w_n)$ vanishes as $n \to \infty$, we have $M_n \to 0$ as $n \to \infty$, that is, $E = \{x \mid \forall n,\ T_\alpha^n(I_\eta^{(n)}(x)) \neq I_\alpha\}$ has Lebesgue measure 0, where $I_\eta^{(n)}(x)$ is the cylinder in $\mathcal{P}^{(n)}$ containing x. Then, recalling that T_α is non-singular, $m(T_\alpha^{-n}(E))$ is also 0 for all $n \geq 0$, and so $m\left(\bigcup_n T_\alpha^{-n}(E)\right) = 0$. That is, for almost all x there is a subsequence $\{n_i\}$ such that $T^{n_i}(I_\eta^{n_i}(x)) = I_\alpha$ for all $i \in \mathbb{N}$. Then for almost all x, $\forall U$ open neighborhood of x we can find n and a full cylinder $x \in I_\eta^{(n)} \subset U$. □

Proof of Lemma 1.16. We have just proved that the full cylinders generate the Borel sets. Then a sufficient and necessary condition for exactness, due to Rohlin [19], is the following: $\exists C > 0$ such that $\forall n$, $\forall I_\eta^{(n)}$ full cylinder of rank n, $\forall X \subset I_\eta^{(n)}$,

$$\mu_\alpha(T_\alpha^n(X)) \leq C \frac{\mu_\alpha(X)}{\mu_\alpha(I_\eta^{(n)})} \tag{1.19}$$

We recall that the T_α satisfy the *bounded distortion* property: Then, recalling that the density of μ_α with respect to the Lebesgue measure is bounded from above and from below by constants, we get for some constant C,

$$\mu_\alpha(V_\eta(B)) \geq \frac{1}{C} \mu_\alpha(B) \mu_\alpha(I_\eta^{(n)}),$$

that is, Rohlin's characterization (1.19). □

[2] The fact that T_0 is exact follows from a result of M. Thaler [22].

1.5 Entropy

Knowing the invariant densities allows to compute the *entropy* of the system. We briefly recall the relevant definitions:

1.18 (Entropy of a partition). Let $(X, \mathcal{A}, \mu)$ be a probability space, $\xi = \{X_1, \dots, X_n\}$ a finite measurable partition of X, that is $X_i \in \mathcal{A}\ \forall i \in \mathbb{N}$ and $X = \bigsqcup_{i=1}^{n} X_i (\mathrm{mod}\, 0)$. $H(\xi) = -\sum_{i=1}^{n} \mu(X_i) + \log \mu(X_i)$ is called the *entropy* of the partition ξ.

1.19. Given $\xi_1, \dots, \xi_k$ partitions of X, and $X^1 \in \xi_1, \dots, X^k \in \xi_k$, we denote by $\bigvee_{i=1}^{k} \xi_i$ the partition $\{X^1 \cap \dots \cap X^k\}$; given $T : X \to X$ measurable, we denote by $T^{-1}(\xi)$ the partition $\{T^{-1}(X_i), X_i \in \xi\}$.

1.20 (Kolmogorov-Sinai entropy). Let $(X, \mathcal{A}, \mu, T)$ be a measurable dynamical system, μ an invariant probability measure for T, ξ a finite partition of X. The quantity

$$H_\xi(T) = \lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right)$$

is called the *entropy of T with respect to ξ*. $H(T) = \sup_\xi H_\xi(T)$, where the sup is taken over all finite partitions ξ of X, is called the *Kolmogorov-Sinai entropy* of T.

A dynamical system's entropy and information are deeply related, as can be seen from the following

Theorem 1.21 (Shannon, Breiman, McMillan). *Let $(X, \mathcal{A}, \mu, T)$ be a measurable and ergodic dynamical system, ξ a finite partition of X. Given $x \in X$, let $\xi^n(x)$ be the element of $\bigvee_{i=1}^{n-1} T^{-i}\xi$ which contains x. Then for μ-almost every $x \in X$,*

$$H_\xi(T) = \lim_{n\to\infty} -\frac{1}{n} \log \mu(\xi^n(x))$$

Supposing the initial point x to be unknown to us, we may be interested in the quantity of information provided to us by some initial segment of the symbolic dynamics of x. If $\xi = \{X_1, \cdots X_k\}$, knowing the set $\xi^n(x)$ is equivalent to knowing (for μ-almost every x) the indices $j_0, \dots, j_{n-1} \in \{1, \dots, k\}$ such that $T^i(x) \in X_{j_i}$. Intuitively, the smaller $\mu(\xi^n(x))$ is, the better we have "located" the point x in our space, and the more information we have obtained. This corresponds to the system having high entropy. The Shannon-Breiman-McMillan Theorem then states that the entropy with respect to a partition represents the "average information production rate" of the input obtained with the partition ξ.

The α-continued fraction maps belong to a class of interval transformations for which an explicit formula for the entropy is available:

1.22 (AFU map). Let I be an interval, $T : I \to I$ which is piecewise C^2 with respect to a countable partition $\mathcal{P}$ of I in subintervals $\{I_j\}_{j\geq 1}$, and such that the following conditions hold:

a) *Adler's condition:* $\exists K > 0,\ \frac{|T''(x)|}{(T'(x))^2} < K\ \forall x \in I$;

b) *Finite range:* $\#\{T(I_j),\ j \geq 1\} < \infty$;

c) *Uniform expansivity:* $\exists \lambda :\ |T'(x)| \geq \lambda > 1\ \forall x \in I$

The maps T_α are AFU: we have seen that they are uniformly expanding in (2.1); moreover they clearly have finite range, and $\frac{|T''_\alpha(x)|}{(T'_\alpha(x))^2} = \frac{2}{|x|^3}x^4 = 2\,|x| < 2$.

Theorem 1.23 (Rohlin's formula). *Let I be an interval, $T : I \to I$ an AFU map, $d\mu = \rho(x)dx$ the a.c.i.m. for T. Then the Kolmogorov-Sinai entropy of T is given by*

$$h_\mu(T) = \int_I \log |T'(x)|\, d\mu(x)$$

For Rohlin's original proof we refer the reader to [19], while the proof in the case of AFU maps can be found in [8].

Remark 1.24. We remark that the algorithm introduced in the proof of Lemma 1.4 to extract two identical subsequences $\frac{p_{n_j}}{q_{n_j}} = \frac{P_{n_k}}{Q_{n_k}}$ from the 1-convergents and α-convergents respectively could be used to compute the entropy $h(\alpha)$ of T_α. In fact Birkhoff's ergodic theorem implies that

$$h(\alpha) = -\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} \log \left|T^i_\alpha(x)\right| = -\lim_{n\to\infty} \frac{1}{n} \log \beta_{n-1},$$

where β_n is the product defined in equation (1.4). The estimates (1.8) and (1.9) imply that

$$\frac{1}{(1+\alpha)q_n} < \beta_{n-1} < \frac{1}{\alpha q_n}$$

But since $\lim_{n\to\infty} \frac{1}{n}\log(cq_n) = \lim_{n\to\infty} \frac{1}{n}\log(q_n)$ for every constant $c > 0$,

$$h(\alpha) = \lim_{n\to\infty} \frac{1}{n} \log q_n$$

In particular, even without knowing all the values of the α-convergents $\frac{p_n}{q_n}$, the entropy $h(\alpha)$ could be approximated simply with the limit $\frac{1}{n_k}\log Q_{n_k}$, requiring only the knowledge of the 1-convergents.

Chapter 2

Statistical stability for α-continued fractions

In the previous chapter we have described the dynamical properties of the system (I_α, T_α) for a fixed value of the parameter α. We now wish to understand to what extent these properties remain stable when α varies; in a sense, we wish to study the behavior of the system under *deterministic perturbations*.
Several notions for the stability of a dynamical system have been introduced. In the case of smooth systems, *structural stability* requires that the orbits should be preserved up to homeomorphism; however, this notion is too strong for our case.
We will adopt the point of view of Alves and Viana [2], and we will call a family of interval maps $\{(I, \phi_t)\}_{t\in\mathbb{R}}$ *statistically stable* if the SRB measures μ_t of the maps ϕ_t are continuous in t with respect to the L^1 norm.
It may be convenient to assume that the maps $\{\phi_t\}$ are all defined on the same interval; up to conjugation, at least locally, the maps T_α can always be rescaled to a suitable fixed interval.

2.1 Continuity of the entropy

We will denote the entropy of T_α by $h(\alpha)$. The main goal of the present section is the following

Theorem 2.1. *The function $\alpha \to h(\alpha)$ is continuous in $(0, 1]$, and*

$$\lim_{\alpha\to 0^+} h(\alpha) = 0$$

Since in the case $\alpha \geq \sqrt{2} - 1$ the entropy has been computed exactly by Nakada [16] and Marmi, Moussa, Cassa [15], we can restrict our study to the case $0 < \alpha \leq \sqrt{2} - 1$.
To prove continuity we adopt the following approach: by means of a uniform Lasota-Yorke-type inequality for the Perron-Frobenius operator, we prove that the variations of the invariant densities are equibounded as α varies in some neighborhood of any fixed $\bar{\alpha} > 0$ (see Proposition 2.2 below). Our argument follows quite closely [24], except that we have to deal with a further difficulty

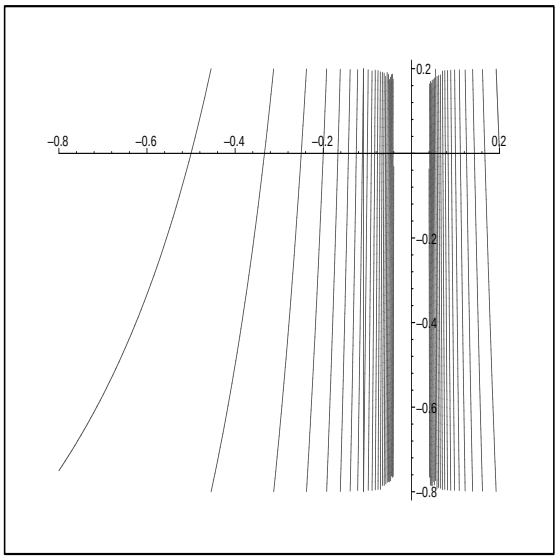

Figure 2.1: Graph of the map T_α when $\alpha = 0.2$

arising from the fact that the cylinders containing the endpoints α and $\alpha - 1$ can be arbitrarily small. After translating the maps so that their interval of definition does not depend on α around $\bar{\alpha}$, we prove the L^1-continuity of the invariant densities ρ_α using Helly's Theorem (Lemma 2.5). Then the continuity of the entropy follows from Rohlin's formula.

2.1.1 Uniformly bounded variation of the invariant densities

Let $\bar{\alpha} \in (0, \sqrt{2} - 1]$ and $\varepsilon < \bar{\alpha}$ be fixed, and choose $\alpha \in [\bar{\alpha} - \varepsilon, \bar{\alpha} + \varepsilon]$. In this case, recalling the definitions in §1.2, we have $j'_{\min} = 2$, and for $x \in I_j^{\pm}$,

$$\frac{1}{|T'_\alpha(x)|} \leq \lambda_j \leq \lambda < 1, \tag{2.1}$$

where

$$\lambda = (1 - \bar{\alpha} + \varepsilon)^2, \quad \lambda_j = \frac{1}{(j - 1 + \bar{\alpha} - \varepsilon)^2},\ j > 2, \quad \lambda_2 = \lambda \tag{2.2}$$

depend only on $\bar{\alpha}$ and ε. Moreover, we have that $\mathrm{Var}_{I_j^{\pm}} \left| \frac{1}{T'_\alpha(x)} \right| \leq \lambda_j \ \forall \alpha \in [\bar{\alpha} - \varepsilon, \bar{\alpha} + \varepsilon]$.

As we have seen in §1.4, for all $\alpha \in (0, 1]$ the maps T_α admit a unique absolutely continuous invariant probability measure μ_α, whose density ρ_α is of bounded variation (and therefore bounded). In addition, a result of R. Zweimüller entails that ρ_α is bounded from below (see [25], Lemma 7):

$$\forall \alpha \in (0, \sqrt{2} - 1],\ \exists C > 0 \text{ s.t } \forall x \in I_\alpha, \rho_\alpha(x) \geq C \tag{2.3}$$

Proposition 2.2. *$\forall \bar{\alpha} \in (0, \sqrt{2}-1]$, $\rho_{\bar{\alpha}}$ is of bounded variation, and $\exists \varepsilon$, $\exists K > 0$ such that for all $\alpha \in [\bar{\alpha} - \varepsilon, \bar{\alpha} + \varepsilon]$, $\operatorname{Var}(\rho_\alpha) < K$.*

The main result we need in order to prove Proposition 2.2 is the following

Lemma 2.3 (Uniform version of the Lasota-Yorke inequality). *Let $\bar{\alpha}$ be fixed. Then there exist $\lambda_0 < 1$, $C, K_0 > 0$ such that $\forall n$, $\forall \varphi \in BV(I)$, $\forall \alpha \in [\bar{\alpha} - \varepsilon, \bar{\alpha} + \varepsilon]$,*

$$\operatorname*{Var}_{I_\alpha} \left(P^n_{T_\alpha} \varphi\right) \leq C(\lambda_0)^n \operatorname{Var} \varphi + K_0 \int\limits_{I_\alpha} |\varphi| \, dx \tag{2.4}$$

Assuming Lemma 2.3 the Proposition then follows easily. Indeed it is enough to recall that the Cesaro sums

$$\rho_n = \frac{1}{n} \sum_{j=0}^{n-1} P^j_{T_\alpha} 1$$

of the sequence $\{P^j_{T_\alpha} 1\}_{j \in \mathbb{N}}$ converge almost everywhere to the invariant density ρ_α of T_α. Both the variations and the L^∞ norms of the $\{\rho_n\}$ are uniformly bounded:

$$\operatorname{Var} \rho_n \leq \frac{1}{n} \sum_{j=0}^{n-1} \operatorname{Var} \left(P^j_{T_\alpha} 1\right) \leq \frac{1}{n} \sum_{j=0}^{n-1} K_0 m(I_\alpha) = K_0 \quad \forall n$$

$$\int\limits_{I_\alpha} \rho_n dx = \frac{1}{n} \sum_{j=0}^{n-1} \int\limits_{I_\alpha} P^j_{T_\alpha} 1 dx = m(I_{\bar{\alpha}}) = 1 \quad \forall n \Rightarrow$$

$$\sup_{I_\alpha} |\rho_n| \leq \operatorname*{Var}_{I_\alpha} \rho_n + \frac{1}{m(I_\alpha)} \leq K_0 + 1 \quad \forall n,$$

where K_0 is the constant we found in Lemma 2.3. Then we also have $\operatorname{Var} \rho_\alpha \leq K_0$, $\sup |\rho_\alpha| \leq K_0 + 1$, which concludes the proof of Proposition 2.2.

Proof of Lemma 2.3. We have

$$\operatorname{Var} \left(P^n_{T_\alpha} \varphi\right) \leq \sum_\eta \left(\operatorname*{Var}_{T^n_\alpha(I^{(n)}_\eta)} \frac{\varphi\left(V_\eta(x)\right)}{\left|(T^n_\alpha)'\left(V_\eta(x)\right)\right|} + 2 \sup_{T^n_\alpha(I^{(n)}_\eta)} \left| \frac{\varphi\left(V_\eta(x)\right)}{(T^n_\alpha)'\left(V_\eta(x)\right)} \right| \right) =$$

$$= \sum_\eta \left(\operatorname*{Var}_{I^{(n)}_\eta} \frac{\varphi(y)}{|(T^n_\alpha)'(y)|} + 2 \sup_{I^{(n)}_\eta} \left| \frac{\varphi(y)}{(T^n_\alpha)'(y)} \right| \right) \tag{2.5}$$

For the last equality, observe that since $V_\eta : T^n_\alpha(I^{(n)}_\eta) \to I^{(n)}_\eta$ is a homeomorphism, $\operatorname{Var}_{T^n_\alpha(I^{(n)}_\eta)} \left(\frac{\varphi}{|(T^n_\alpha)'|} \circ V_\eta \right) = \operatorname{Var}_{I^{(n)}_\eta} \frac{\varphi}{|(T^n_\alpha)'|}$. The first term in expression (2.5) can be estimated using (i):

$$\sum_\eta \operatorname*{Var}_{I^{(n)}_\eta} \frac{\varphi(y)}{|(T^n_\alpha)'(y)|} \leq \sum_\eta \left(\operatorname*{Var}_{I^{(n)}_\eta} \varphi \sup_{I^{(n)}_\eta} \frac{1}{|(T^n_\alpha)'(y)|} + \operatorname*{Var}_{I^{(n)}_\eta} \frac{1}{|(T^n_\alpha)'(y)|} \sup_{I^{(n)}_\eta} |\varphi| \right) \leq$$

$$\leq \sum_\eta \left(\lambda^{(n)}_\eta \operatorname*{Var}_{I^{(n)}_\eta} \varphi + n \lambda^{(n)}_\eta \sup_{I^{(n)}_\eta} |\varphi| \right)$$

For the second term, we have $2\sum_\eta \sup_{I_\eta^{(n)}} \left|\frac{\varphi(y)}{(T_\alpha^n)'(y)}\right| \le 2\sum_\eta \lambda_\eta^{(n)} \sup_{I_\eta^{(n)}} |\varphi(y)|$. In conclusion, from equation (2.5) we get

$$\operatorname{Var}\left(P_{T_\alpha}^n \varphi\right)(x) \le \lambda^n \operatorname*{Var}_{I_\alpha} \varphi + \sum_\eta (n+2)\lambda_\eta^{(n)} \sup_{I_\eta^{(n)}} |\varphi| \tag{2.6}$$

We want to give an estimate of the sum in equation (2.6); recall that for $\varphi \in BV$,

$$\sup_{I_\eta^{(n)}} |\varphi| \le \operatorname*{Var}_{I_\eta^{(n)}} \varphi + \frac{1}{m(I_\eta^{(n)})} \int_{I_\eta^{(n)}} |\varphi|\, dx \tag{ii}$$

However, equation (ii) doesn't provide a global bound independent from η for two reasons. In the first place, the lengths of the intervals $I_\eta^{(n)}$ are not bounded from below when the indices $j_i(\eta)$ grow to infinity. Furthermore, a difficulty that arises only in the case of uniform continuity and that was not dealt with in reference [24] is that the measures of the cylinders of rank n containing the endpoints $\bar\alpha$ and $\bar\alpha - 1$ are *not* uniformly bounded from below in α, and require a careful handling.
To overcome the first difficulty, following [24], we split the sum into two parts: for n fixed, let k be such that

$$\sum_{j>k} \lambda_j \le \frac{\lambda^n}{2^{2n-1}} \tag{2.7}$$

Since λ doesn't depend on α, neither does k. Define the set of "intervals with bounded itineraries"

$$G(n) = \{I_\eta^{(n)} \in \mathcal{P}^n \mid \max(j_0(\eta), \dots, j_{n-1}(\eta)) \le k\} \tag{2.8}$$

To get rid of the measures of the cylinders containing the endpoints, we combine them with full cylinders; the measures of the latter can be estimated using Lagrange's Theorem, since the derivatives are bounded under the hypothesis of bounded itineraries. When combining intervals, we have to consider the sum of the corresponding $\lambda_v^{(n)}$ and make sure that it is smaller than 1. This requires additional care when $I_\eta^{(n)} \ni \alpha - 1$.

Remark 2.4. Let $r = r(\alpha)$ be such that

$$v_{r+1} \le \alpha < v_r, \qquad \text{where } v_r = -\frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{4}{r}} \tag{2.9}$$

(clearly r is bounded by $r(\bar\alpha)+1$ in a small neighborhood of $\bar\alpha$). Then $T_\alpha^i(\alpha-1) = \frac{(i+1)\alpha-1}{1-i\alpha} \in I_2^-$ for $i = 0, \dots, r-1$ and $T_\alpha^r(\alpha-1) \notin I_2^-$. Thus any cylinder with more than r consecutive digits "$(2,-)$" is empty, and the cylinder $((2,-),\dots,(2,-))$ of rank r may be arbitrarily small when α varies. The cylinder $(j_{\min}, +)$ can be arbitrarily small too.

Consider the function $\sigma : G(n) \to G(n)$ which maps every nonempty cylinder $I_\eta^{(n)}$ in $I_\xi^{(n)}$ in the following way:

a. If $(j_i(\eta), \varepsilon_i(\eta)) = (j_{\min}, +)$ for some i, then $(j_i(\xi), \epsilon_i(\xi)) = (j_{\min}+1, +)$;

b. If $\exists i$ such that

$$((j_i(\eta),\varepsilon_i(\eta)),\dots,(j_{i+r}(\eta),\varepsilon_{i+r}(\eta))) = ((2,-),(2,-),\dots,(2,-)),$$

then $((j_i(\xi),\varepsilon_i(\xi)),\dots,(j_{i+r}(\xi),\varepsilon_{i+r}(\xi))) = ((2,-),\dots,(2,-),(3,-))$;

c. Otherwise, $(j_i(\xi),\varepsilon_i(\xi)) = (j_i(\eta),\varepsilon_i(\eta))$.

We want to show that there exists $\delta_n > 0$, depending only on $\bar{\alpha}$, such that for all $\xi \in \sigma(G(n))$, $m(\xi) \geq \delta_n$. For this purpose, we group together the sequences of consecutive digits $(2,-)$, and obtain a new alphabet $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$, where

$$\begin{aligned}\mathcal{A}_1 &= \{(3,-),\dots,(k,-)\} \cup \{(j_{\min}+1,+),\dots,(k,+)\}\\ \mathcal{A}_2 &= \{(2,-),((2,-),(2,-)),\dots,(\underbrace{(2,-),\dots,(2,-)}_{r-1})\}\end{aligned}$$

Then each $\xi \in \sigma(G(n))$ can be seen as a sequence in $\mathcal{A}'^s = \{(a_1,\dots,a_s) \in \mathcal{A}^s \,|\, a_i \in \mathcal{A}_2 \Rightarrow a_{i+1} \in \mathcal{A}_1\}$ for some $n \geq s \geq \frac{n}{r}$. Let $\widetilde{T_\alpha}$ be the first return map on $\mathcal{A}_1$ restricted to $\sigma(G(n))$:

$$\begin{aligned}&\widetilde{T}(x) = T_\alpha(x) \quad \text{for } x \in (a) \in \mathcal{A}_1;\\ &\widetilde{T}(x) = T_\alpha^i(x) \quad \text{if } \exists i : x \in \underbrace{((2,-),\dots,(2,-))}_{i}, x \notin \underbrace{((2,-),\dots,(2,-))}_{i+1}\end{aligned}$$

Let $\widetilde{V}_a$ be the inverse branch of $\widetilde{T}$ relative to the cylinder (a). Observe that

$$\forall (a_1,\dots,a_s) \in \mathcal{A}'^s, \widetilde{T}^s(a_1,\dots,a_s) = \widetilde{T}(a_s) \tag{2.10}$$

This can be proved by induction on s: when $s = 1$ it is trivial; supposing that the property (2.10) holds for all sequences of length s, we have

$$\begin{aligned}\widetilde{T}^{s+1}(a_1,\dots,a_{s+1}) = \widetilde{T}^{s+1}\left((a_1,\dots,a_s) \cap \widetilde{V}_{a_1}\cdots\widetilde{V}_{a_s}(a_{s+1})\right) =\\ = \widetilde{T}(\widetilde{T}^s(a_1,\dots,a_s) \cap (a_{s+1}))\end{aligned}$$

since $\widetilde{T}^s$ is injective on $(a_1,\dots,a_s)$; this is equal to $\widetilde{T}(\widetilde{T}(a_s)\cap(a_{s+1}))$ by inductive hypothesis.

- If $a_{s+1} \in \mathcal{A}_2$, we have $a_s \in \mathcal{A}_1$ and $\widetilde{T}(a_s) = I$: then $\widetilde{T}^{s+1}(a_1,\dots,a_{s+1}) = \widetilde{T}(a_{s+1})$.

- If $a_{s+1} \in \mathcal{A}_1$, $\widetilde{T}(a_s) \supseteq (a_{s+1})$. In fact for all $i = 0,\dots,r-1$,

$$\begin{aligned}T_\alpha^i(\underbrace{(2,-),\dots,(2,-)}_{i}) = T_\alpha^i\left(\left[\alpha-1, V^i_{(2,-)}(\alpha)\right)\right) =\\ = [T_\alpha^i(\alpha-1),\alpha) \supseteq \left[-\frac{1}{2+\alpha},\alpha\right) \supseteq \bigcup_{a\in\mathcal{A}_1}(a) \supseteq \left[-\frac{1}{3},0\right]\end{aligned} \tag{2.11}$$

Equation (2.10) provides a lower bound on the measures of the intervals in $\sigma(G(n))$:

$$\frac{1}{3} \le m(\widetilde{T}(a_s)) = m(\widetilde{T}^s(a_1,\dots,a_s)) \le m(a_1,\dots,a_s)\sup\left|(\widetilde{T}^s)'\right|, \quad \text{and}$$
$$M(\bar\alpha) \doteqdot \left(\max\left((k+\alpha+\varepsilon)^2, (2+\alpha+\varepsilon)^{2r(\bar\alpha)}\right)\right) \ge \sup\left|\widetilde{T}'\right|$$

in a neighborhood of $\bar\alpha$. Thus for all $I_\xi^{(n)} \in \sigma(G(n))$,

$$\delta_n \doteqdot \frac{1}{3M(\bar\alpha)^n} \le m(I_\xi^{(n)})$$

Returning to the sum in equation (2.6), and defining $\overline{I}_\xi^{(n)} = \bigcup\{I_\eta^{(n)} \mid \sigma(I_\eta^{(n)}) = I_\xi^{(n)}\}$, we find:

$$\sum_{I_\eta^{(n)}\in G(n)} \left(\lambda_\eta^{(n)} \sup_{I_\eta^{(n)}} |\varphi|\right) \le \sum_{I_\xi^{(n)}\in\sigma(G(n))} \sup_{\overline{I}_\xi^{(n)}} |\varphi| \left(\sum_{\sigma(I_\eta^{(n)})=I_\xi^{(n)}} \lambda_\eta^{(n)}\right)$$

We want to estimate $\lambda' = \sup_{\sigma(G(n))} \sum_{\sigma(I_\eta^{(n)})=I_\xi^{(n)}} \lambda_\eta^{(n)}$: each sum can be computed distributively as a product of at most n factors λ_i', each of which corresponds to one of the cases a), b), c) that we have listed in the definition of σ:

- In the case a), we have $\lambda_i' = \lambda_{j_{\min}} + \lambda_{j_{\min}+1} \le 2(\bar\alpha+\varepsilon)^2 < \frac{1}{2}$ (remark that $j_{\min} \ge 3$ when $\alpha \le \sqrt{2}-1$).
- In the case b), $\lambda_i' = \lambda_2^r + \lambda_2^{r-1}\lambda_3 = (1-\alpha)^{2(r-1)}\left((1-\alpha)^2 + \frac{1}{(2+\alpha)^2}\right) < 0.9$. In fact, when $\alpha > \frac{1}{5}$ we have $(1-\alpha)^2 + \frac{1}{(2+\alpha)^2} < \frac{9}{10}$; otherwise, $(1-\alpha)^2 + \frac{1}{(2+\alpha)^2} < \frac{5}{4}$, and for $\alpha \ge \eta_{r+1}$, we have $r-1 \ge \frac{1}{\alpha^2+\alpha} - 2$, and
$$(1-\alpha)^{2(r-1)} = \left(\frac{1-\alpha^2}{1+\alpha}\right)^{2(r-1)} \le \frac{1}{(1+\alpha)^{2(r-1)}} \le \frac{1}{1+2\alpha(r-1)} \le$$
$$\le \frac{1+\alpha}{3-3\alpha-4\alpha^2} < \frac{3}{5}$$
- In the case c), $\lambda_i' = \lambda_{j_i}$.

(The constants in the previous discussion are far from optimal, but they are sufficient for our purposes.)
Then $\lambda' \le \max\left(\lambda^n, \left(\frac{9}{10}\right)^{\frac{n}{r(\bar\alpha)+1}}\right) = \tilde\lambda^n < 1$. Note that $\tilde\lambda$ only depends on $\bar\alpha$ and not on α.
We can finally complete our estimate for the sum over $I_\eta^{(n)} \in G(n)$:

$$\lambda' \sum_{I_\xi^{(n)}\in\sigma(G(n))} \sup_{\overline{I}_\xi^{(n)}} |\varphi| \le \tilde\lambda^n \sum_{I_\xi^{(n)}\in\sigma(G(n))} \left(\operatorname*{Var}_{\overline{I}_\xi^{(n)}} |\varphi| + \frac{1}{m(\overline{I}_\xi^{(n)})}\int_{\overline{I}_\xi^{(n)}} \varphi\right) \le$$
$$\le \tilde\lambda^n \left(\operatorname{Var}\varphi + \sum_{I_\xi^{(n)}\in\sigma(G(n))} \frac{1}{m(I_\xi^{(n)})}\int_{\overline{I}_\xi^{(n)}} \varphi\right) \le \tilde\lambda^n \operatorname{Var}\varphi + \frac{\tilde\lambda^n}{\delta_n}\|\varphi\|_1 \qquad (2.12)$$

On the other hand, for the sum over $I_\eta^{(n)} \notin G(n)$ we have the following estimate:

$$\sum_{I_\eta^{(n)} \notin G(n)} \left((n+2)\lambda_\eta^{(n)} \sup_{I_\eta^{(n)}} |\varphi| \right) \leq$$

$$\leq \sup_{I_\alpha} |\varphi| \sum_{j>k} \sum_{l=0}^{n-1} \sum_{j_l(\eta)=\max\{j_0(\eta),\ldots,j_{n-1}(\eta)\}=j} (n+2)\lambda_{j_0(\eta)} \cdots \lambda_{j_{n-1}(\eta)} \tag{2.13}$$

where in the third sum of expression (2.13) we take l to be the smallest integer that realizes the maximum, to avoid counting the same sequences twice. Observe that when we take the sum over $j_0(\eta), \ldots, j_{n-1}(\eta)$, since we are not taking into account the signs $\varepsilon_i(\eta)$, we are actually counting at most 2^n distinct sequences.

$$\sum_{\substack{(j_0(\eta),\ldots,j_{n-1}(\eta)) \\ j_l(\eta)=j}} \lambda_{j_0(\eta)} \cdots \lambda_{j_{n-1}(\eta)} \leq$$

$$\leq \lambda_j \left(2 \sum_{(j_0(\eta),\ldots,j_{l-1}(\eta),j_{l+1}(\eta),\ldots,j_{n-1}(\eta))} \lambda_{j_0(\eta)} \cdots \lambda_{j_{n-1}(\eta)} \right) \leq$$

$$\leq \lambda_j \left(2 \prod_{\substack{i=0 \\ i\neq l}}^{n-1} \sum_{j_i=2}^{j} 4\lambda_{j_i} \right) \leq \lambda_j 2^{2n-1}$$

since $\sum_2^\infty \lambda_j \leq \frac{\pi^2}{6} \leq 2$. Therefore

$$\sum_{I_\eta^{(n)} \notin G(n)} \left((n+2)\lambda_\eta^{(n)} \sup_{I_\eta^{(n)}} |\varphi| \right) \leq \sup_{I_\alpha} |\varphi| \sum_{j>k} \sum_{l=0}^{n-1} (n+2)\lambda_j 2^{2n-1} \leq$$

$$\leq \sup_{I_\alpha} |\varphi|\, n(n+2)2^{2n-1} \sum_{j>k} \lambda_j \leq \sup_{I_\alpha} |\varphi|\, n(n+2)\lambda^n$$

where in the last inequality we have used the hypothesis (2.7) on k.
In conclusion, $\operatorname*{Var}_{I_\alpha}(P_{T_\alpha}^n \varphi)$ is bounded by

$$\tilde{\lambda}^n \left((n^2+3n+3) \operatorname*{Var}_{I_\alpha} \varphi + (n+2) \left(\frac{1}{\delta_n} + n \right) \|\varphi\|_1 \right)$$

and we recall that we have chosen δ_n and $\tilde{\lambda}$ so that they do not depend on α. Choose any $\bar{\lambda} \in (\tilde{\lambda}, 1)$, and let $\overline{K} > 0, N \in \mathbb{N}$ be such that

$$\forall n \geq 1,\ (n^2+3n+3)\tilde{\lambda}^n \leq \overline{K}\bar{\lambda}^n \text{ and } \forall n \geq N,\ \overline{K}\bar{\lambda}^n \leq \frac{1}{2}$$

Let $L(n) = (n+2)\left(\frac{1}{\delta_n} + n\right)\bar{\lambda}^n$, $\hat{K} = \max_{1\leq n\leq N} L(n)$. For any n, we can perform the Euclidean division $n = qN + r$ for some $q \geq 0$ and $0 \leq r < N$. Then

$$\operatorname*{Var}_{I_\alpha} \left(P_{T_\alpha}^N \varphi \right) \leq \overline{K}\bar{\lambda}^N \operatorname*{Var}_{I_\alpha} \varphi + \hat{K} \|\varphi\|_1 \tag{2.14}$$

More generally, we can show by induction on q that

$$\operatorname*{Var}_{I_\alpha}\left(P_{T_\alpha}^{qN}\varphi\right) \le (\overline{K}\lambda^N)^q \operatorname*{Var}_{I_\alpha}\varphi + C(q)\hat{K}\left\|\varphi\right\|_1 \tag{2.15}$$

where $C(q) = 1 + \frac{1}{2} + \cdots + \frac{1}{2^{q-1}} < 2$ for all q. In fact if (2.15) is true for some q, recalling that the Perron-Frobenius operator P_{T_α} preserves the L^1 norm, we get

$$\begin{aligned}
&\operatorname*{Var}_{I_\alpha}\left(P_{T_\alpha}^{(q+1)N}\varphi\right) \le (\overline{K}\bar{\lambda}^N)^q \operatorname*{Var}_{I_\alpha}\left(P_{T_\alpha}^N\varphi\right) + C(q)\hat{K}\left\|P_{T_\alpha}^N\varphi\right\|_1 \le \\
&\le (\overline{K}\bar{\lambda}^N)^{q+1} \operatorname*{Var}_{I_\alpha}\varphi + \left(C(q) + \frac{1}{2^q}\right)\hat{K}\left\|\varphi\right\|_1 dx \le \\
&\le (\overline{K}\bar{\lambda}^N)^{q+1} \operatorname*{Var}_{I_\alpha}\varphi + C(q+1)\hat{K}\left\|\varphi\right\|_1 dx
\end{aligned}$$

For $0 \le r < N$, $\operatorname{Var}\left(P_{T_\alpha}^r\varphi\right) \le \overline{K}\bar{\lambda}^r \operatorname{Var}\varphi + \hat{K}\left\|\varphi\right\|_1$. In general, for $n = qN + r$, we obtain

$$\begin{aligned}
&\operatorname*{Var}_{I_\alpha}\left(P_{T_\alpha}^n\varphi\right) \le (\overline{K}\bar{\lambda}^N)^q \operatorname*{Var}_{I_\alpha}\left(P_{T_\alpha}^r\varphi\right) + C(q)\hat{K}\left\|\varphi\right\|_1 \le \\
&\le (\overline{K}\bar{\lambda}^N)^q\overline{K}\bar{\lambda}^r \operatorname*{Var}_{I_\alpha}\varphi + \hat{K}\left((\overline{K}\bar{\lambda}^N)^q + C(q)\right)\left\|\varphi\right\|_1 \le \frac{\overline{K}}{2^q}\bar{\lambda}^r \operatorname*{Var}_{I_\alpha}\varphi + 3\hat{K}\left\|\varphi\right\|_1
\end{aligned}$$

Now take $\lambda_0 \ge \max\left(\frac{1}{2^{\frac{1}{N}}}, \bar{\lambda}\right)$, so that $\frac{\bar{\lambda}^r}{2^q} \le (\lambda_0)^r(\lambda_0)^{Nq} = (\lambda_0)^n$. This concludes the proof of Lemma 2.3. □

2.1.2 L^1 continuity of the densities ρ_α and continuity of the entropy

Let $\bar{\alpha} \in (0, \sqrt{2} - 1]$ be fixed. To study the L^1-continuity property of the densities ρ_α (and the continuity of the entropy $h(\alpha)$) it is convenient to work with measures supported on the same interval. Thus we rescale the maps T_α with α in a neighborhood of $\bar{\alpha}$ to the interval $[\bar{\alpha} - 1, \bar{\alpha}]$ by applying the translation $\tau_{\bar{\alpha}-\alpha}$. Let $A_{\alpha,\bar{\alpha}} = \tau_{\bar{\alpha}-\alpha} \circ T_\alpha \circ \tau_{\bar{\alpha}-\alpha}^{-1}$ be the new maps:

$$A_{\alpha,\bar{\alpha}}(x) = \left|\frac{1}{x - \bar{\alpha} + \alpha}\right| - \left[\left|\frac{1}{x - \bar{\alpha} + \alpha}\right| + 1 - \alpha\right] + \bar{\alpha} - \alpha \tag{2.16}$$

Let $J_j^\pm = I_j^\pm + \bar{\alpha} - \alpha$ be the translated versions of the intervals of the original partition, and $\tilde{\rho}_\alpha(x) = \rho \circ \tau_{\bar{\alpha}-\alpha}^{-1}(x) = \rho(x - \bar{\alpha} + \alpha)$ the invariant densities for $A_{\alpha,\bar{\alpha}}$. Clearly the bounds for the sup and the variation of ρ_α are still valid for $\tilde{\rho}_\alpha$.

Lemma 2.5. *Let $\bar{\alpha} \in (0, \sqrt{2} - 1]$ be fixed, and let ε be given by Proposition 2.2. Then if $\{\alpha_n\} \subset [\bar{\alpha} - \varepsilon, \bar{\alpha} + \varepsilon]$ is a monotone sequence converging to $\bar{\alpha}$, we have $\tilde{\rho}_{\alpha_n} \xrightarrow{L^1} \tilde{\rho}_{\bar{\alpha}}$.*

Proof. Since $\sup|\tilde{\rho}_{\alpha_n}| \le K$, $\operatorname{Var}\tilde{\rho}_{\alpha_n} \le K\ \forall n$, we can apply the following theorem:

Theorem 2.6 (Helly's Theorem). *Let $\{\rho_n\}$ be a sequence in $BV(I)$ such that:*

1. $\sup|\rho_n| \le K_1 \quad \forall n,$
2. $\mathrm{Var}\,\rho_n \le K_2 \quad \forall n$

Then there exists a subsequence ρ_{n_k} *and a function* $\rho \in BV(I)$ *such that* $\rho_{n_k} \xrightarrow{L^1} \rho$, $\rho_{n_k} \to \rho$ *almost everywhere, and*

$$\sup|\rho| \le K_1, \qquad \mathrm{Var}\,\rho \le K_2$$

Thus we can find a subsequence $\{\tilde\rho_{\alpha_{n_k}}\}$ converging in the L^1 norm and almost everywhere to some function ρ_∞ such that $\sup|\rho_\infty| \le K$, $\mathrm{Var}\,\rho_\infty \le K$. We want to show that $\rho_\infty = \tilde\rho_{\bar\alpha}$: we observe that it is sufficient to show that ρ_∞ is an invariant density for $A_{\bar\alpha} = A_{\bar\alpha,\bar\alpha} = T_{\bar\alpha}$, and then use the uniqueness of the invariant density. To simplify notations, we will write α_k for α_{n_k}, $\tilde\rho_k$ for $\tilde\rho_{\alpha_{n_k}}$, and A_k for $A_{\alpha_{n_k},\bar\alpha}$.
Our goal is to show that $\forall B \subseteq I_{\bar\alpha}$, $\int \chi_B(A_{\bar\alpha}(x))\rho_\infty(x)dx = \int \chi_B(x)\rho_\infty(x)dx$. Observe that every $\chi_B(x)$ belongs to $L^1(I_{\bar\alpha})$ and so can be approximated arbitrarily well by compactly supported C^1 functions with respect to the L^1 norm. Then it will be sufficient to prove that $\forall\varphi \in C^1$ with compact support contained in $I_{\bar\alpha}$,

$$\left|\int \varphi\left(A_{\bar\alpha}(x)\right)\rho_\infty(x)dx - \int \varphi(x)\rho_\infty(x)dx\right| = 0 \tag{2.17}$$

Observe that $\left|\int \varphi(A_{\bar\alpha}(x))\rho_\infty(x)dx - \int \varphi(x)\rho_\infty(x)dx\right| \le I_1+I_2+I_3$, with I_1,I_2,I_3 given below:

$$I_1 = \left|\int \varphi(A_{\bar\alpha}(x))\rho_\infty(x)dx - \int \varphi(A_{\bar\alpha}(x))\tilde\rho_k(x)dx\right| \le \|\varphi\|_\infty \|\tilde\rho_k - \rho_\infty\|_{L^1}$$

$$I_3 = \left|\int \varphi(A_k(x))\tilde\rho_k(x)dx - \int \varphi(x)\rho_\infty(x)dx\right| =$$

$$= \left|\int \varphi(x)\tilde\rho_k(x)dx - \int \varphi(x)\rho_\infty(x)dx\right| \le \|\varphi\|_\infty \|\tilde\rho_k - \rho_\infty\|_{L^1}$$

which vanish as $k \to \infty$. Finally, $I_2 = \int |\varphi(A_{\bar\alpha}(x)) - \varphi(A_k(x))|\,\tilde\rho_k(x)dx$ is bounded by $K\int |\varphi(A_k(x)) - \varphi(A_{\bar\alpha}(x))|\,dx$, and we need to show that

$$\int |\varphi(A_k(x)) - \varphi(A_{\bar\alpha}(x))|\,dx \to 0 \text{ when } k \to \infty \tag{2.18}$$

Recall that for $x \in J^-_{j,\alpha_k} = \left[\frac{-1}{j-1+\alpha_k} + \bar\alpha - \alpha_k, -\frac{1}{j+\alpha_k} + \bar\alpha - \alpha_k\right)$, $A_k(x) = -\frac{1}{x-\bar\alpha+\alpha_k} - j + \bar\alpha - \alpha_k$, and for $x \in J^-_{j,\bar\alpha} = \left[-\frac{1}{j-1+\alpha}, -\frac{1}{j+\alpha}\right)$, $A_{\bar\alpha}(x) = -\frac{1}{x} - j$. We will examine in detail the case $\alpha_k < \bar\alpha\ \forall k$, $x < \bar\alpha - \alpha_k$; the other cases can be dealt with in a similar way. In this case, $0 < \frac{1}{j+\alpha_k} - \frac{1}{j+\bar\alpha} < \bar\alpha - \alpha_k$, and if $j < \frac{1}{\sqrt{\bar\alpha-\alpha_k}} = N(k)$, then $-\frac{1}{j-1+\alpha_k} + \frac{1}{j+\bar\alpha} < \frac{\alpha_k-\bar\alpha-1}{j^2} < \alpha_k - \bar\alpha$ and so

$$-\frac{1}{j-1+\alpha_k} + \bar\alpha - \alpha_k < -\frac{1}{j+\bar\alpha} < -\frac{1}{j+\alpha_k} + \bar\alpha - \alpha_k \tag{2.19}$$

$I_{N(k)} = \bigcup_{j\ge N(k)} J^-_{j,\alpha_k}$ contains the set in which condition (2.19) isn't satisfied, and its measure $m(I_{N(k)}) = \sum_{j=N(k)}^\infty \left|\frac{1}{(j-1+\alpha_k)(j+\alpha_k)}\right|$ vanishes when $k \to \infty$.

Given $\varepsilon'' > 0$, choose $\bar{k}$ such that $m(I_{N(\bar{k})}) < \varepsilon''$, and let $k \geq \bar{k}$. Define

$$\xi_j^- = \left[-\frac{1}{j-1+\alpha_k} + \bar{\alpha} - \alpha_k, -\frac{1}{j+\bar{\alpha}}\right), \ \eta_j^- = \left[-\frac{1}{j+\bar{\alpha}}, -\frac{1}{j+\alpha_k} + \bar{\alpha} - \alpha_k\right)$$
$$\text{when } 2 < j \leq N(\bar{k}), \text{ and } \quad \xi_2^- = \left[\bar{\alpha} - 1, -\frac{1}{2+\alpha_k}\right)$$

Then we can split the integral (2.18) in three parts in the following way:

$$\int_{\bar{\alpha}-1}^{\bar{\alpha}} |\varphi(A_k(x)) - \varphi(A_{\bar{\alpha}}(x))|\, dx \leq \int\limits_{I_{N(\bar{k})}} |\varphi(A_k(x)) - \varphi(A_{\bar{\alpha}}(x))|\, dx +$$
$$+ \sum_{j=3}^{N(\bar{k})} \left(\int_{\eta_j^-} |\varphi(A_k(x)) - \varphi(A_{\bar{\alpha}}(x))|\, dx \right) + \sum_{j=2}^{N(\bar{k})} \left(\int_{\xi_j^-} |\varphi(A_k(x)) - \varphi(A_{\bar{\alpha}}(x))|\, dx \right)$$

The first integral in this expression is bounded by $2\varepsilon'' \|\varphi\|_\infty$. Moreover, the measures of the sets η_j^- tend uniformly to 0 when $k \to \infty$:

$$m(\eta_j^-) \leq |\bar{\alpha} - \alpha_k| + \frac{|\bar{\alpha} - \alpha_k|}{(j+\bar{\alpha})(j+\alpha_k)} \leq C_1(\bar{\alpha} - \alpha_k)$$
$$\Rightarrow \sum_{j=3}^{N(\bar{k})} \int_{\eta_j^-} |\varphi(A_k(x)) - \varphi(A_{\bar{\alpha}}(x))|\, dx \leq N(\bar{k}) 2 \|\varphi\|_\infty C_1(\bar{\alpha} - \alpha_k)$$

Finally, $m(\xi_j^-) \leq \frac{C_2}{j^2} + |\bar{\alpha} - \alpha_k| \leq \frac{C_3}{j^2}$ when $k \geq \bar{k}$, $j < N(\bar{k})$, and for $x \in \xi_j^-$, $x \leq -\frac{1}{j+\bar{\alpha}}$ and $x - \bar{\alpha} + \alpha_k < -\frac{1}{j+\alpha_k}$, therefore

$$|A_k(x) - A_{\bar{\alpha}}(x)| = \left| -\frac{1}{x} + \frac{1}{x - \bar{\alpha} + \alpha_k} - \bar{\alpha} + \alpha_k \right| \leq$$
$$\leq |\bar{\alpha} - \alpha_k| + \frac{|\bar{\alpha} - \alpha_k|}{|x(x - \bar{\alpha} + \alpha_k)|} \leq |\bar{\alpha} - \alpha_k| \left(1 + (j+1)^2\right) \quad (2.20)$$

Since φ is C^1 on a compact interval, it is also lipschitzian for some Lipschitz constant L_φ, and

$$\sum_{j=2}^{N(\bar{k})} \int_{\xi_j^-} |\varphi(A_k(x)) - \varphi(A_{\bar{\alpha}}(x))|\, dx \leq \sum_{j=2}^{N(\bar{k})} m(\xi_j^-) L_\varphi |A_k(x) - A_{\bar{\alpha}}(x)| \leq$$
$$\leq \sum_{j=2}^{N(\bar{k})} \frac{C_3(1 + (j+1)^2)}{j^2} L_\varphi |\bar{\alpha} - \alpha_k| \leq C_4 N(\bar{k}) |\bar{\alpha} - \alpha_k| \leq C_4 \sqrt{|\bar{\alpha} - \alpha_k|}$$

when k is large. This establishes the claim that the third integral vanishes when $x < \bar{\alpha} - \alpha_k$. In the case $x > \bar{\alpha} - \alpha_k$ we have similar estimates: for $j < \frac{1}{\sqrt{|\bar{\alpha} - \alpha_k|}}$, we have

$$\frac{1}{j+\bar{\alpha}} < \frac{1}{j+\alpha_k} + \bar{\alpha} - \alpha_k < \frac{1}{j-1+\bar{\alpha}}$$

and we can define the intervals

$$\gamma_j^+ = \left(\frac{1}{j+\alpha_k} + \bar{\alpha} - \alpha_k, \frac{1}{j-1+\bar{\alpha}}\right], \ \delta_j^+ = \left(\frac{1}{j-1+\bar{\alpha}}, \frac{1}{j-1+\alpha_k} + \bar{\alpha} - \alpha_k\right]$$

We have $m(\delta_j^+) \leq C_5 \, |\bar{\alpha} - \alpha_k|$, $m(\gamma_j^+) \leq \frac{C_5}{j^2}$, and

$$|A_k(x) - A_{\bar{\alpha}}(x)| \leq C_7 j^2 \, |\alpha_k - \bar{\alpha}| \text{ for } x \in \gamma_j^+$$

Finally, we leave it to the reader to check that the case $\bar{\alpha} < \alpha_k$ can be treated in same way. Thus we can conclude that (2.18) holds.
Therefore we have shown that $\rho_\infty = \tilde{\rho}_{\bar{\alpha}}$. This is also true if we extract a converging sub-subsequence from any subsequence of $\tilde{\rho}_{\alpha_n}$, and so $\tilde{\rho}_{\alpha_n} \to \tilde{\rho}_{\bar{\alpha}}$ both in L^1 and almost everywhere for $n \to \infty$. This completes the proof of Lemma 2.5. □

The L^1-continuity of the map $\alpha \mapsto \rho_\alpha$ is sufficient to prove that the entropy map $\alpha \mapsto h(\alpha)$ is also continuous. This is achieved by applying the following lemma (for a proof see for example [1]) to Rohlin's formula.

Lemma 2.7. *Let $\{\rho_n\}$ be a sequence of functions in $L^1(I)$ such that*

1. $\|\rho_n\|_\infty \leq K \quad \forall n,$
2. $\rho_n \xrightarrow{L^1} \rho$ *for some* $\rho \in L^1(I)$

Then for any $\psi \in L^1(I)$,

$$\int \psi(\rho_n - \rho) \to 0$$

Applying Rohlin's Formula for the entropy, we get for any $\alpha \in [\bar{\alpha} - \varepsilon, \bar{\alpha} + \varepsilon]$

$$h(\alpha) = \int_{\alpha-1}^{\alpha} \log \frac{1}{(x - \bar{\alpha} + \alpha)^2} \tilde{\rho}_\alpha(x) dx = 2 \int_{\alpha-1}^{\alpha} |\log |x - \bar{\alpha} + \alpha|| \, \tilde{\rho}_\alpha(x) dx$$

Consider a sequence $\{\alpha_n\} \to \bar{\alpha}$. Then

$$\begin{aligned}
&|h(\bar{\alpha}) - h(\alpha_n)| \leq 2 \int_{\bar{\alpha}-1}^{\bar{\alpha}} \left| \log |x - \bar{\alpha} + \alpha_n| \, \tilde{\rho}_{\alpha_n}(x) - \log |x| \, \tilde{\rho}_{\bar{\alpha}}(x) \right| dx \leq \\
&\leq 2 \left(\int_{\bar{\alpha}-1}^{\bar{\alpha}} \left| \log |x - \bar{\alpha} + \alpha_n| \, (\tilde{\rho}_{\alpha_n}(x) - \tilde{\rho}_{\bar{\alpha}}(x)) \right| dx + \right. \\
&\left. + \int_{\bar{\alpha}-1}^{\bar{\alpha}} \left| (\log |x - \bar{\alpha} + \alpha_n| - \log |x|) \, \tilde{\rho}_{\bar{\alpha}}(x) \right| dx \right)
\end{aligned}$$

The second integral is bounded by $2(K_0 + 1) \int_{\bar{\alpha}-1}^{\bar{\alpha}} |\log |x - \bar{\alpha} + \alpha_n| - \log |x|| \, dx$ and vanishes when $n \to \infty$ because of the continuity of translation in L^1. If we take $\tilde{\rho}_n = \tilde{\rho}_{\alpha_n}$, $\tilde{\rho} = \tilde{\rho}_{\bar{\alpha}}$, $\psi(x) = |\log |x||$ in Lemma 2.7, we find that the first integral also tends to 0. □

2.1.3 Behavior of the density and entropy when $\alpha \to 0$

In this section we will prove that the entropy has a limit as $\alpha \to 0^+$ and that $\lim_{\alpha \to 0^+} h(\alpha) = 0$.
The continuity of the entropy on the interval $(0, \sqrt{2} - 1]$ followed from the L^1-continuity of the densities. The vanishing of the entropy as $\alpha \to 0$ is a consequence of the fact that the densities converge to the Dirac delta at the parabolic fixed point of T_0 as $\alpha \to 0$.

Proposition 2.8. *When $\alpha \to 0$, the invariant measures $\tilde{\mu}_\alpha$ of the translated maps $A_{\alpha,0} : [-1,0] \to [-1,0]$ converge in the sense of distributions to the Dirac delta in -1.*

From the previous Proposition the vanishing of the entropy follows easily:

Corollary 2.9. *Let $h(\alpha)$ be the metric entropy of the map T_α with respect to the absolutely continuous invariant probability measure μ_α. Then $h(\alpha) \to 0$ as $\alpha \to 0$.*

Proof of the Corollary. We compute the entropy of the T_α through Rohlin's formula:

$$h(\alpha) = 2 \int_{\alpha-1}^{\alpha} |\log|x|| \, d\mu_\alpha \tag{2.21}$$

Observe that $\forall E \subseteq (c_1, 0]$, $\mu_\alpha(E) = \frac{1}{C(\alpha)} \nu_\alpha(E) \leq \frac{C_0}{C(\alpha)} m(E)$. Therefore if ρ_α is the density of μ_α, $\rho_\alpha < \frac{C_0}{C(\alpha)}$ in $(c_1, 0]$. Given ε, let c_k be such that $|\log|x|| < \varepsilon$ for $x \in [-1, c_k]$, and choose α small such that $\alpha - 1 < c_k$, $\mu_\alpha([c_k, \alpha]) = \tilde{\mu}_\alpha([\tilde{c}_k, 0]) < \varepsilon$ and $\frac{C_0}{C(\alpha)} < \varepsilon$. Then

$$\begin{aligned} h(\alpha) &\leq \int_{\alpha-1}^{c_k} |\log|x|| \, d\mu_\alpha + \int_{c_k}^{c_1} |\log|x|| \, d\mu_\alpha + \int_{c_1}^{\alpha} |\log|x|| \, \rho_\alpha dx \leq \\ &\leq |\log|c_k|| + \left|\log \frac{1}{3}\right| \mu_\alpha([c_k, c_1]) + \frac{C_0}{C(\alpha)} \left\| \log|x| \right\|_1 \to 0 \end{aligned}$$

which concludes the proof. □

To prove Proposition 2.8 we adopt the following strategy: we introduce the jump transformations G_α of the maps T_α over the cylinder $(2, -)$, whose derivatives are strictly bounded away from 1 even when $\alpha \to 0$; we can then prove that their densities $\frac{d\nu_\alpha}{dx}$ are bounded from above and from below by uniform constants. Using the relation between μ_α and the induced measure ν_α, we conclude that for any measurable set B such that $-1 \notin B$, $\tilde{\mu}_\alpha(B) = \mu_\alpha(B + \alpha) \to 0$ when $\alpha \to 0$.

Proof of Proposition 2.8. Given $v_{r+1} \leq \alpha < v_r$ as in equation (2.9), and $0 \leq j \leq r$, let

$$L_0 = I_\alpha \setminus (2, -), \quad L_j = [c_{j+1}, c_j) = \underbrace{((2,-), \ldots, (2,-))}_{j} \setminus \underbrace{((2,-), \ldots, (2,-))}_{j+1}$$

for $1 \leq j \leq r$. Thus $I_\alpha = \bigcup_{0 \leq j \leq r} L_j \pmod 0$. It is easy to prove by induction that for $r \geq j \geq 1$, $c_j = V_{(2,-)}^{j-1}\left(-\frac{1}{2+\alpha}\right) = -1 + \frac{1}{j + \frac{1}{1+\alpha}}$, that is, $-\frac{j}{j+1} < c_j \leq -\frac{j-1}{j}$, while $c_0 = \alpha$, $c_{r+1} = \alpha - 1$. Let

$$G_\alpha|_{L_j} = T_\alpha^{j+1}|_{L_j}$$

be the jump transformation associated to the return time $\tau(x) = j + 1 \iff x \in L_j$. Observe that τ is bounded and therefore integrable with respect to μ_α.

Then a result of R. Zweimuller ([26], Theorem 1.1) guarantees that G_α admits an invariant measure $\nu_\alpha \ll \mu_\alpha$ such that for all measurable E,

$$\mu_\alpha(E) = \frac{1}{C(\alpha)} \left(\sum_{n\geq 0} \nu_\alpha \left(\{\tau > n\} \cap T_\alpha^{-n}(E) \right) \right) \tag{2.22}$$

where $C(\alpha)$ is a suitable normalization constant. Actually from equation (2.22) it follows that $\nu_\alpha(I_\alpha) = \nu_\alpha(\{\tau > 0\}) \leq C(\alpha)\mu_\alpha(I_\alpha)$ is finite, and so by choosing a suitable $C(\alpha)$ we can take $\nu_\alpha(I_\alpha) = 1$. We will prove the following:

Lemma 2.10. There exists $\tilde{\alpha} > 0$ such that for $0 < \alpha < \tilde{\alpha}$, the densities ψ_α of ν_α are bounded from above and from below by constants that do not depend on α: $\exists C_0$ *s.t.* $C_0^{-1} \leq \psi_\alpha \leq C_0$.

Proof of Lemma 2.10. In order to prove that ψ_α is bounded from above, we can proceed as in Lemma 2.3, and show that $\exists C'$ such that for all α, $\forall \varphi \in L^1(I_\alpha)$, $\mathrm{Var}_{I_\alpha} P_{G_\alpha}^n \varphi < C'$. Since the outline of the proof is very similar to that of Lemma 2.3, we will only list the passages where the estimates are different, and emphasize how in this case all the constants can be chosen uniform in α. The cylinders of rank 1 for G_α are of the form

$$I_j^{k,\varepsilon} = (j,k,\varepsilon) \doteq (\underbrace{(2,-),\ldots,(2,-)}_{j},(k,\varepsilon)), \quad 0 \leq j \leq r,$$

so they are also cylinders for T_α, although of different rank. On $I_j^{k,\varepsilon}$, $j \geq 1$ we have

$$\left|\frac{1}{G'_\alpha(x)}\right| = (T_\alpha^j(x)\cdots T_\alpha(x)x)^2 \leq \lambda_j^k = \frac{4}{(j+2)^2(k-1)^2} \leq \frac{1}{(k-1)^2} \leq \frac{1}{4},$$
$$\left|\frac{1}{G'_\alpha(x)}\right| \geq \frac{1}{9j^2(k+1)^2},$$

while on $I_0^{k,\varepsilon}$, $\left|\frac{1}{G'_\alpha(x)}\right| = x^2 < \frac{1}{(k-1)^2} \leq \frac{1}{4}$, and so $\lambda = \sup\left|\frac{1}{G'_\alpha}\right| < \frac{1}{4}$ for all α. Letting $\mathcal{Q} = \bigcup_{j=0}^r \{I_j^{k,\varepsilon}\}$, and $\lambda_\eta^{(n)} = \sup_{I_\eta^{(n)}} \left|\frac{1}{(G_\alpha^n)'}\right|$, we can obtain the analogue of equation (2.6) for the maps G_α:

$$\mathrm{Var}\left(P_{G_\alpha}^n \varphi\right)(x) \leq \lambda^n \operatorname*{Var}_{I_\alpha} \varphi + \sum_{I_\eta^{(n)} \in \mathcal{Q}^{(n)}} (n+2)\lambda_\eta^{(n)} \sup_{I_\eta^{(n)}} |\varphi|,$$

and similarly to (2.7), we can choose h such that $\sum_{i\geq h} \frac{1}{i^2} \leq \frac{\lambda^n}{2^{4n-2}}$, and the set of intervals with bounded itineraries

$$G(n) = \{I_\eta^{(n)} = ((j_0,k_0,\varepsilon_0),\ldots,(j_{n-1},k_{n-1},\varepsilon_{n-1})) \in \mathcal{Q}^n \mid \max(j_0,\ldots,j_{n-1}) \leq h,\ \max(k_0,\ldots,k_{n-1}) \leq h\}$$

Again we can define a function $\sigma : G(n) \to G(n)$ that maps every cylinder $I_\eta^{(n)} = ((j_0,k_0,\varepsilon_0),\ldots,(j_{n-1},k_{n-1},\varepsilon_{n-1}))$ to $I_\xi^{(n)} = ((j'_0,k'_0,\varepsilon'_0),\ldots,(j'_{n-1},k'_{n-1},\varepsilon'_{n-1}))$ as follows:

a. If $(j_i, k_i, \varepsilon_i) = (j, j_{\min}, +)$, $j < r-1$ for some i, then $(j_i', k_i', \varepsilon_i') = (j, j_{\min} + 1, +)$;

b. If for some i, $(j_i, k_i, \varepsilon_i) = (r, k, \varepsilon)$ with $(k, \varepsilon) \neq (j_{\min}, +), (j_{\min}+1, +)$, then $(j_i', k_i', \varepsilon_i') = (r-1, k, \varepsilon)$;

c. If $(j_i, k_i, \varepsilon_i) \in \{(r, j_{\min}, +), (r, j_{\min}+1, +), (r-1, j_{\min}, +)\}$, then $(j_i', k_i', \varepsilon_i') = (r-1, j_{\min}+1, +)$;

d. Otherwise, $(j_i', k_i', \varepsilon_i') = (j_i, k_i, \varepsilon_i)$.

With this definition, the cylinders in $\sigma(G(n))$ are all full, because as we have seen in equation (2.11), for $0 \leq i \leq r-1$,

$$T_\alpha^i(\underbrace{(2,-),\dots,(2,-)}_{i}) \supseteq \left[-\frac{1}{2+\alpha}, \alpha\right) \supseteq \bigcup_{(k,\varepsilon), k\geq 3} I_k^\varepsilon$$

Then $\forall I_\xi^{(n)} \in \sigma(G(n))$,

$$1 \leq m(I_\xi^{(n)}) \sup_{\sigma(G(n))} |(G_\alpha^n)'| = m(I_\xi^{(n)})(9h^4)^n \Rightarrow m(I_\xi^{(n)}) \geq \frac{1}{\delta_n} = \frac{1}{(9h^4)^n},$$

which doesn't depend on α. Again we need to estimate the supremum over $I_\xi^{(n)} \in \sigma(G(n))$ of the sums $\sum_{\sigma(I_\eta^{(n)}) = I_\xi^{(n)}} \lambda_\eta^{(n)}$, each of which is the product of n terms λ_i', that correspond to one of the cases a), b), c) d) we listed previously:

- In the case a), $\lambda_i' = \lambda_j^{j_{\min}} + \lambda_j^{j_{\min}+1} \leq \frac{1}{(j_{\min}-1)^2} + \frac{1}{j_{\min}^2} \leq \frac{1}{2}$ (observe that for $\alpha < \sqrt{2}-1$, $j_{\min} \geq 3$).
- In the case b), $\lambda_i' = \lambda_r^k + \lambda_{r-1}^k \leq \frac{4}{(k-1)^2}\left(\frac{1}{(r+1)^2} + \frac{1}{(r+2)^2}\right) \leq \frac{1}{2}$ when $\alpha < v_2$;
- In the case c), $\lambda_i' = \lambda_r^{j_{\min}} + \lambda_{r-1}^{j_{\min}} + \lambda_r^{j_{\min}+1} + \lambda_{r-1}^{j_{\min}+1}$ $< 4\left(\frac{1}{(j_{\min}-1)^2} + \frac{1}{j_{\min}^2}\right)\left(\frac{1}{(r+1)^2} + \frac{1}{(r+2)^2}\right) < \frac{1}{2}$ for $\alpha < v_2$;
- In the case d), $\lambda_i' < \lambda = \frac{1}{4}$.

Then $\lambda' \leq \tilde{\lambda} = \frac{1}{2}$, and as in equation (2.12), we find for $\alpha < v_2$,

$$\sum_{I_\eta^{(n)} \in G(n)} \left(\lambda_\eta^{(n)} \sup_{I_\eta^{(n)}} |\varphi|\right) \leq \lambda' \sum_{I_\xi^{(n)} \in \sigma(G(n))} \sup_{\overline{I}_\xi^{(n)}} |\varphi| \leq \tilde{\lambda}^n \operatorname{Var} \varphi + \frac{\tilde{\lambda}^n}{\delta_n} \|\varphi\|_1$$

For the sum over intervals with unbounded itineraries we proceed in a similar way to (2.13):

$$\sum_{I_\eta^{(n)} \notin G(n)} \lambda_\eta^{(n)} \leq \sum_{i\geq h} \sum_{l=0}^{n-1} \left(\sum_{\substack{j_l(\eta) = i+1 = \\ \max(j_0(\eta),\dots,j_{n-1}(\eta))}} \lambda_{j_0(\eta)}^{k_0(\eta)} \cdots \lambda_{j_{n-1}(\eta)}^{k_{n-1}(\eta)} + \right.$$

$$\left. + \sum_{\substack{k_l(\eta) = i = \\ \max(k_0(\eta),\dots,k_{n-1}(\eta))}} \lambda_{j_0(\eta)}^{k_0(\eta)} \cdots \lambda_{j_{n-1}(\eta)}^{k_{n-1}(\eta)} \right) \leq \sum_{i\geq h} \sum_{l=0}^{n-1} \frac{4}{i^2} \left(2 \sum_{I_j^{k,\varepsilon} \in \mathcal{Q}} \lambda_\eta^{(n)}\right)^{n-1}$$

(This expression is redundant, but sufficient for our purpose.) Observe that

$$\sum_{I_j^{k,\varepsilon}\in\mathcal{Q}} \lambda_\eta^{(n)} \le 2\sum_{j=0}^{\infty}\sum_{k=3}^{\infty}\frac{4}{(j+2)^2(k-1)^2} \le 8\left(\sum_{2}^{\infty}\frac{1}{k^2}\right)^2 \le 8, \text{ and so}$$

$$\sum_{I_\eta^{(n)}\notin G(n)} \lambda_\eta^{(n)} \le \sum_{i\ge h}\frac{4}{i^2}n2^{4n-4} \le n\lambda^n$$

Then we can prove relation (2.14) and complete our argument exactly like in Lemma 2.3. Notice that all the constants involved are uniform in α.

To prove that the densities ψ_α of ν_α are uniformly bounded from below, we use a bounded distortion argument. We follow the same outline as in §1.9, but with the advantage that in this case the derivatives are uniformly bounded from above.
Since T_α satisfies *Adler's condition* (see definition 1.22), $\left|\frac{T_\alpha''}{(T_\alpha')^2}\right| < K$ (here $K = 2$), then there exists K' independent of n and of α such that $\forall n > 0$, $\left|\frac{(T_\alpha^n)''}{((T_\alpha^n)')^2}\right| < K'$ (see [25], Lemma 10). Then $\forall x, y$ belonging to the same cylinder $I_j^{k,\varepsilon}$ of rank 1 of G_α,

$$\left|\frac{G_\alpha'(x)}{G_\alpha'(y)} - 1\right| = |G_\alpha''(\xi)|\left|\frac{x-y}{G_\alpha'(y)}\right| = \left|\frac{G_\alpha''(\xi)(G_\alpha(x)-G_\alpha(y))}{G_\alpha'(y)G_\alpha'(\eta)}\right| \le$$
$$\le 36^2\left|\frac{G_\alpha''(\xi)}{(G_\alpha'(\xi))^2}\right||G_\alpha(x)-G_\alpha(y)| \le K''|G_\alpha(x)-G_\alpha(y)|, \quad \text{and}$$

$$\log\left|\frac{(G_\alpha^n)'(y)}{(G_\alpha^n)'(x)}\right| \le \sum_{i=0}^{n-1}\left|\frac{G_\alpha'(G_\alpha^i(y))}{G_\alpha'(G_\alpha^i(x))} - 1\right| \le K''\sum_{i=0}^{n-1}\left|G_\alpha^{i+1}(y) - G_\alpha^{i+1}(x)\right| \le$$
$$\le K''\sum_{i=1}^{n}\left(\frac{1}{4}\right)^{n-i}|G_\alpha^n(y) - G_\alpha^n(x)| \le K''\sum_{i=0}^{\infty}\left(\frac{1}{4}\right)^i \Rightarrow \left|\frac{(G_\alpha^n)'(y)}{(G_\alpha^n)'(x)}\right| \le C_1$$

where C_1 does not depend on α. Letting $W_\eta : G_\alpha^n(I_\eta^{(n)}) \to I_\eta^{(n)}$ be the local inverses of G_α^n, for every full cylinder $I_\eta^{(n)} \in \mathcal{P}^{(n)}$ and for every measurable set B,

$$\frac{m(B)}{m(I_\alpha)} = \frac{\int_{W_\eta(B)}|(G_\alpha^n)'(y)|\,dy}{\int_{I_\eta^{(n)}}|(G_\alpha^n)'(x)|\,dx} \le \frac{m(W_\eta(B))\ \sup_{y\in I_\eta^{(n)}}|(G_\alpha^n)'(y)|}{m(I_\eta^{(n)})\ \inf_{x\in I_\eta^{(n)}}|(G_\alpha^n)'(x)|} \le C_1\frac{m(W_\eta(B))}{m(I_\eta^{(n)})}$$

$$\Rightarrow m(W_\eta(B)) \ge m(B)\frac{m(I_\eta^{(n)})}{C_1} \quad (2.23)$$

Finally, we can show that the measure of the union S_n of all full cylinders of rank n is strictly greater than 0. In fact we have the following characterization: $I_\eta^{(n)} \in \mathcal{Q}^{(n)}$ is not full $\Rightarrow$ it has an initial segment of the orbit (with respect to G_α) of one of the endpoints α and $\alpha - 1$ as its final segment. That is, if $\alpha =$

$(a_1, a_2, a_3, \ldots)$ and $\alpha - 1 = (b_1, b_2, b_3, \ldots)$, then there exists $1 \leq k \leq n$ such that $I_\eta^{(n)} = (\omega_1, \ldots, \omega_{n-k}, a_1, \ldots, a_k)$ or $I_\eta^{(n)} = (\omega_1, \ldots, \omega_{n-k}, b_1, \ldots, b_k)$. To prove this, observe that if $I_\eta^{(n)}$ doesn't contain any initial segment of $(a_1, a_2, a_3, \ldots)$ or $(b_1, b_2, b_3, \ldots)$, it is clearly full, and if every such segment $(a_1, \ldots, a_k)$ or $(b_1, \ldots, b_k)$ is followed by $\omega_{k+1} \neq a_{k+1}$ or b_{k+1} respectively, then it is either full or empty because G_α^n is monotone on each cylinder. Then

$$\nu_\alpha(S_n^{\mathcal{C}}) \leq \nu_\alpha\left(\bigcup_{k=1}^{n} G_\alpha^{-(n-k)}(a_1, \ldots, a_k)\right) + \nu_\alpha\left(\bigcup_{k=1}^{n} G_\alpha^{-(n-k)}(b_1, \ldots, b_k)\right) \leq$$

$$\leq \sum_{k=1}^{n} \left(\nu_\alpha(a_1, \ldots, a_k) + \nu_\alpha(b_1, \ldots, b_k)\right)$$

since ν_α is G_α-invariant. We have already shown that ν_α is bounded from below, and so $\nu_\alpha(a_1, \ldots, a_k) + \nu_\alpha(b_1, \ldots, b_k) < C'(m(a_1, \ldots, a_k) + m(b_1, \ldots, b_k))$.
In order to prove that $\nu_\alpha(S_n^C) < 1$, we take advantage of the fact that the cylinders containing the endpoints become arbitrarily small when α approaches 0. Recall that $(a_1) = (j_{\min}) = \left(\left\lfloor \frac{1}{\alpha} + 1 - \alpha \right\rfloor\right)$, and consequently $\inf_{(a_1)} |G'_\alpha(x)| \geq (j_{\min} - 1)^2$, and since $\inf_{I_\alpha} |G'_\alpha(x)| \geq 4$, from Lagrange's theorem we get

$$m(a_1, \ldots, a_k) \leq \frac{1}{4^{k-1}(j_{\min} - 1)^2}$$

Since $j_{\min} \to \infty$ as $\alpha \to 0$, we can choose $\tilde{\alpha}$ such that $\forall \alpha < \tilde{\alpha}$, $m(a_1, \ldots, a_k) \leq \frac{1}{4^k C'}$. Similarly, recall from Remark 2.4 that for $v_{r+1} \leq \alpha < v_r$, where $v_r = \frac{-1+\sqrt{1+4/r}}{2}$, we have $(b_1) = (\underbrace{(2, -), \ldots, (2, -)}_{r}, (k, \varepsilon))$ for some $k \geq 3$, and recalling that $T_\alpha^i(\alpha - 1) = \frac{(i+1)\alpha - 1}{1 - i\alpha}$, we find

$$\inf_{(b_1)} |G'_\alpha(x)| = \inf_{(b_1)} \prod_{i=0}^{r} \frac{1}{(T_\alpha^i(x))^2} \geq \frac{1}{(k-1)^2} \prod_{i=0}^{r-1} \frac{(1 - i\alpha)^2}{(1 - (i+1)\alpha)^2} \geq \frac{4}{(1 - r\alpha)^2}$$

But $\alpha \geq v_{r+1} \Rightarrow \frac{1}{r+1} < \alpha^2 + \alpha \Rightarrow 1 - r\alpha < \frac{\alpha(2+\alpha)}{1+\alpha} < 3\alpha$, and so by taking $\tilde{\alpha}$ small enough we can ensure that $\forall \alpha < \tilde{\alpha}$, $\inf_{(b_1)} |G'_\alpha(x)| \geq 4C'$ and consequently $m(b_1, \ldots, b_k) \leq \frac{1}{4^k C'}$ $\forall k$.
Then for α small enough, $m(S_n^{\mathcal{C}}) \leq \frac{2}{C'} \sum_{k=1}^{n} \frac{1}{4^k} \leq \frac{2}{3C'} \Rightarrow \nu_\alpha(S_n) \leq \frac{2}{3} \Rightarrow \nu_\alpha(S_n) \geq \frac{1}{3} \Rightarrow m(S_n) > \frac{\nu_\alpha(S_n)}{C'} \geq \frac{1}{3C'}$. Taking the sum over all full cylinders $I_\eta^{(n)}$ in (2.23), we find that for all measurable $B \subseteq I_\alpha$,

$$m(G_\alpha^{-n}(B)) \geq m(G_\alpha^{-n}(B) \cap S_n) \geq \frac{m(B)m(S_n)}{C_1} \geq \frac{m(B)}{3C_1C'}$$

Now recall that the density of ν_α is equal almost everywhere to the limit of the Cesaro sums $\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P_{G_\alpha}^i 1$, and so for $\alpha < \tilde{\alpha}$,

$$\nu_\alpha(B) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \int_B P_{G_\alpha}^i 1 dx = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} m(G_\alpha^{-n}(B))$$

and consequently we have $\nu_\alpha(B) \geq \frac{m(B)}{3C_1C'}$ $\forall B$. □

We can finally conclude the proof of Proposition 2.8. The following properties hold:

- $C(\alpha) \to \infty$ *when* $\alpha \to 0$. In fact when α is small, $C_0^{-1} \leq \frac{d\nu_\alpha}{dm} \leq C_0$ for some C_0, and

$$1 = \sum_{k=0}^{r} \mu_\alpha(L_k) = \sum_{k=0}^{r} \frac{1}{C(\alpha)} \sum_{n \geq k} \nu_\alpha(L_n) = \frac{1}{C(\alpha)} \sum_{k=0}^{r} \nu_\alpha([\alpha - 1, c_k]) \geq$$
$$\geq \frac{1}{C_0 C(\alpha)} \sum_{k=0}^{r} m([\alpha - 1, c_k]) \geq \frac{1}{C_0 C(\alpha)} \left(\sum_{k=0}^{r} \frac{1}{k+1} - (r+1)\alpha \right) \geq$$
$$\geq \frac{1}{C_0 C(\alpha)} \left(\log\left(\frac{1}{\alpha^2 + \alpha} \right) - 1 \right) \quad (2.24)$$

  since $r \leq \frac{1}{\alpha^2+\alpha} \leq r+1$. Therefore the normalization constant $C(\alpha) \geq \frac{1}{C_0} \left(\log\left(\frac{1}{\alpha^2+\alpha} \right) - 1 \right) \to \infty$ when $\alpha \to 0$.

- Finally, $\forall L_k,\ k \geq 0$ *finite,* $\mu_\alpha(L_k) \to 0$ *when* $\alpha \to 0$. In fact we have

$$\mu_\alpha(L_k) = \sum_{j \geq k} \frac{\nu_\alpha(L_j)}{C(\alpha)} \leq \frac{C_0}{C(\alpha)} \sum_{j \geq k} m(L_j) \leq \frac{C_0}{C(\alpha)} |1 - c_k| \leq \frac{C_0}{C(\alpha)} \frac{1}{k} \to 0^1$$

Consider now the translated versions $A_{\alpha,0}$ of the T_α with respect to $\bar{\alpha} = 0$, and let $\tilde{c}_j = c_j - \alpha$ be the translated versions of the c_j (we omit the dependence on α for simplicity). Then we have $\tilde{\mu}_\alpha((\tilde{c}_k, 0]) \to 0$ for all finite k. Let $f \in C^\infty([-1, 0])$ be a test function: we want to show that $\forall \varepsilon > 0$, $\exists \alpha'$ such that $\forall \alpha \leq \alpha'$, $\left| \int_{-1}^{0} f(x) d\tilde{\mu}_\alpha - f(-1) \right| < \varepsilon$. Since f is uniformly continuous, $\exists \delta$ such that $\forall\, |x - 1| < \delta$, $|f(x) - f(-1)| < \varepsilon$. Choose k so that $\tilde{c}_k < -1 + \delta$. Then for all α such that $\tilde{\mu}_\alpha((\tilde{c}_k, 0]) < \varepsilon$,

$$\left| \int_{-1}^{0} (f(x) - f(-1)) d\tilde{\mu}_\alpha \right| \leq \int_{-1}^{\tilde{c}_k} |f(x) - f(-1)|\, d\tilde{\mu}_\alpha + \int_{\tilde{c}_k}^{0} |f(x)|\, d\tilde{\mu}_\alpha +$$
$$+ \int_{\tilde{c}_k}^{0} |f(-1)|\, d\tilde{\mu}_\alpha \leq \varepsilon + \varepsilon(\|f\|_\infty + |f(-1)|) \qquad \square$$

2.2 Numerical results

In this section we collect our numerical results on the entropy of Japanese continued fractions. We already know that the function $\alpha \to h(T_\alpha)$ is continuous in $(0, 1]$ and that in the case $\alpha \geq \sqrt{2} - 1$ the entropy has been computed exactly by Nakada [16] and Marmi, Moussa, Cassa [15]. For values

[1] The reader might be wondering whether the estimates of the densities of T_α in Paragraph 2.1.1 and the continuity of the entropy might be derived directly from Lemma 2.10. This would follow from equation (2.24) if we possessed a suitable lower bound for $\frac{d\nu_\alpha}{dm}$ when α varies; but we haven't been able to provide such a bound except for small α. As for the continuity of the entropy, the fact that $h(T_\alpha)$ and $h(G_\alpha)$ are related by the Generalized Abramov Formula [26] suggests that proving the continuity of $h(G_\alpha)$ might be a valid alternative approach; however, taking expansivity into account, we believe that the estimates necessary to prove L^1-continuity of the invariant densities of G_α as in §2.1.2 would be far more taxing than for T_α.

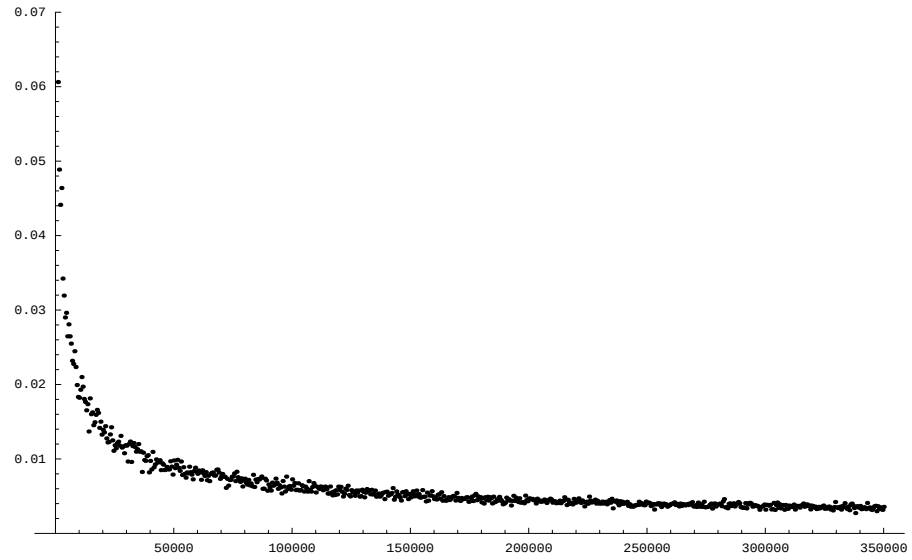

Figure 2.2: The dependence on n of the standard deviation of the normally distributed $h(\frac{1}{2}, n, x_k)$ where n ranges from 500 to 350000 and $N = 100$.

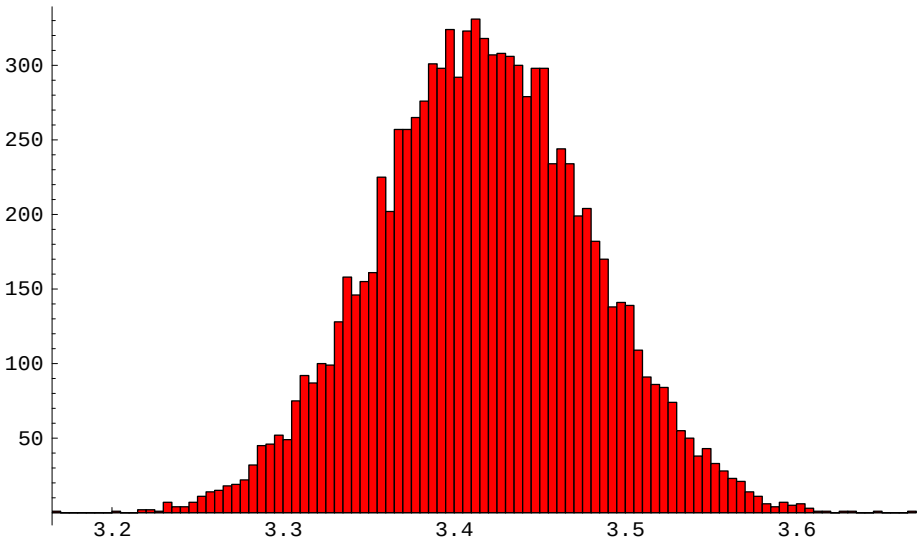

Figure 2.3: The distribution of $h(\frac{1}{2}, 1000, x_k)$ for 10000 random initial conditions. The average $h(\frac{1}{2}, n, N) = 3.41711$ must be compared to the exact value $h(T_{\frac{1}{2}}) = \frac{\pi^2}{6 \log G} = 3.418315971\ldots$.

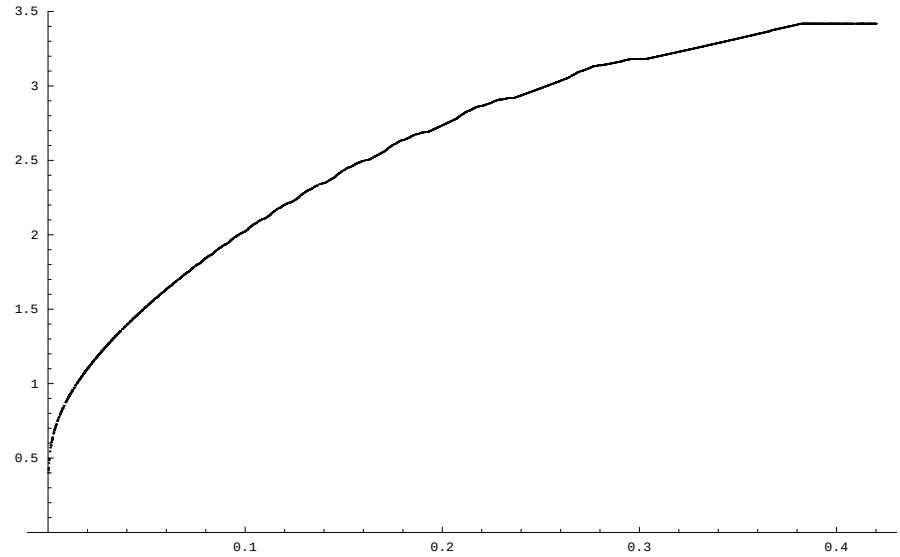

Figure 2.4: The entropy of the map T_α at 4080 uniformly distributed values of α from 0 to 0.42. The estimated error is less than $2 \cdot 10^{-4}$.

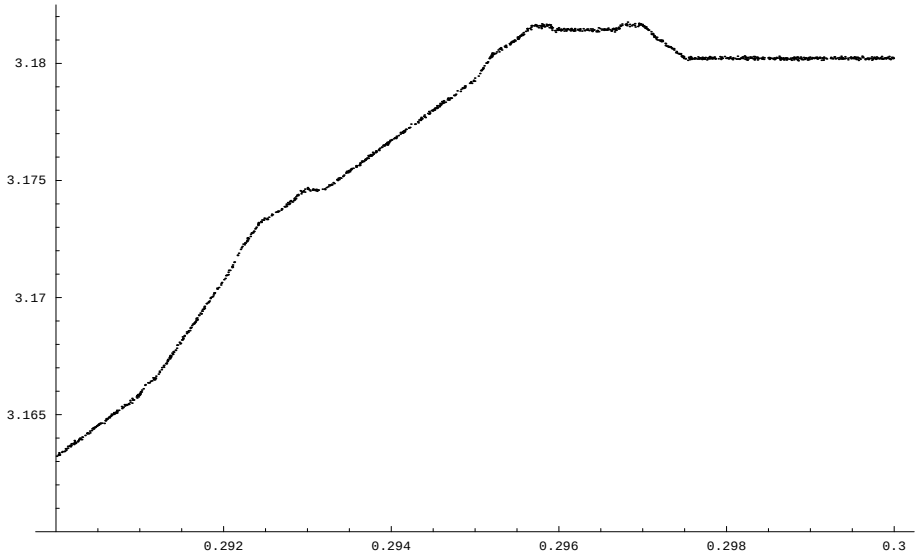

Figure 2.5: The entropy of the map T_α at 1314 uniformly distributed values of α from 0.29 to 0.30. The estimated error is less than $1 \cdot 10^{-4}$.

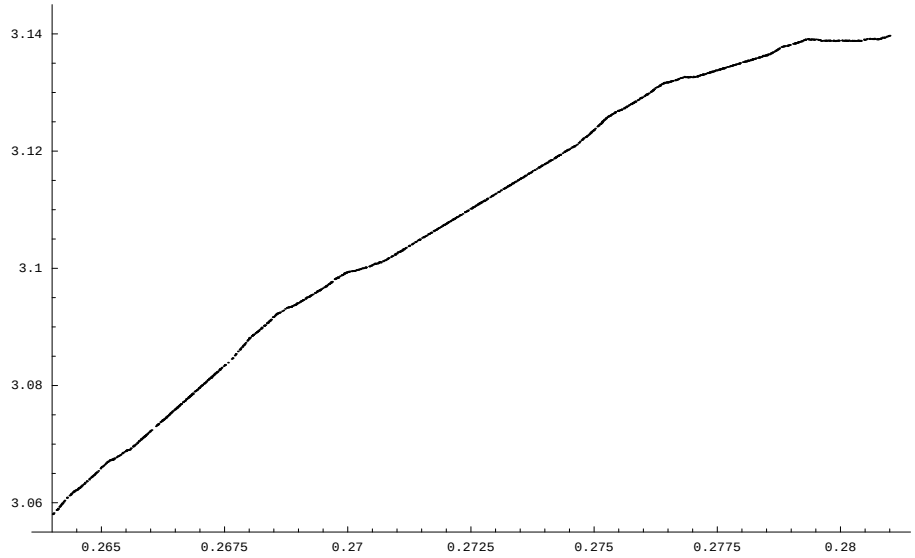

Figure 2.6: The entropy of the map T_α at 1600 uniformly distributed values of α from 0.265 to 0.281. The estimated error is less than $1.5 \cdot 10^{-4}$.

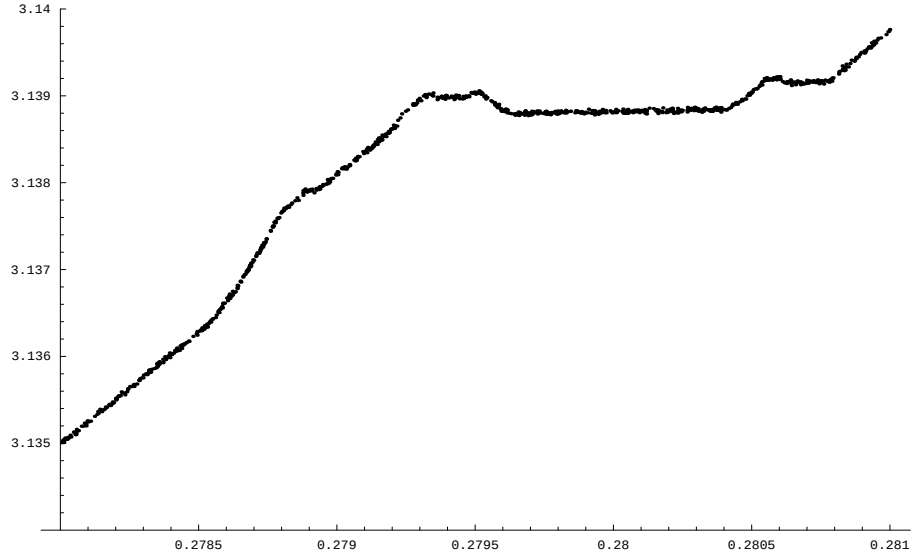

Figure 2.7: The entropy of the map T_α at 989 uniformly distributed values of α from 0.278 to 0.281. The estimated error is less than $4 \cdot 10^{-5}$.

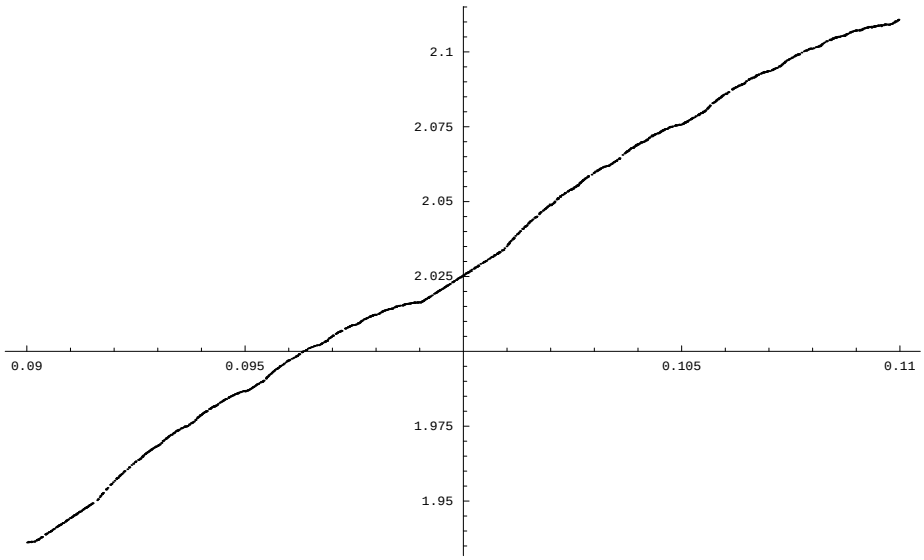

Figure 2.8: The entropy of the map T_α at 1799 uniformly distributed values of α from 0.09 to 0.11. The estimated error is less than $2.5 \cdot 10^{-4}$.

of α in the interval $(0, \sqrt{2} - 1]$ we have numerically computed the entropy of the maps applying Birkhoff's ergodic theorem and replacing the integral $h(T_\alpha) = -2 \int_{\alpha-1}^{\alpha} \log|x| \, \rho_\alpha(x) dx$ in Rohlin's formula with the Birkhoff averages

$$h(\alpha, n, x) = -\frac{2}{n} \sum_{j=0}^{n-1} \log \left| T_\alpha^j(x) \right|$$

which converge to $h(T_\alpha)$ for almost all choices of $x \in (\alpha - 1, \alpha)$. In order to get rid of the dependence on the choice of an initial condition we have computed $h(\alpha, n, x_k)$ for a large number N of uniformly distributed values of $x_k \in (\alpha - 1, \alpha)$, $k = 1, \ldots N$, and we have taken the average on all the results:

$$h(\alpha, n, N) = \frac{1}{N} \sum_{k=1}^{N} h(\alpha, n, x_k) \, .$$

Unsurprisingly, it turns out that the values $h(\alpha, n, x_k)$ are normally distributed around their average $h(\alpha, n, N)$ (see Figure 2.2). We have also computed the standard deviations of the normal distributions for values of n from 500 to 350000 (see Figure 2.3): a least squares fit suggests that they decay as $1/\sqrt{n}$ (we refer to A. Broise [5] for a general treatment of Central Limit Theorems that may apply also to our maps).
In Figure 2.4 we see a graph of $h(\alpha, 10^4, N)$ at 4080 uniformly distributed random values of α in the interval $(0, \sqrt{2} - 1)$: the values of N range from 10^5 to $4 \cdot 10^5$ increasing as α decreases so as to keep the standard deviation approximately constant. The estimated error for the entropy is less than $2 \cdot 10^{-4}$.
As $\alpha \to 0$ the entropy decreases (although non monotonically, see below) and the graph exhibits a quite rich self–similar structure that we have just started

to investigate: for example the entropy seems to be independent of α as α varies in the intervals whose endpoints have Gauss continued fraction expansions of the form $[0, n, n-1, 1, \bar{n}]$ and $[0, \bar{n}]$ respectively, and to depend linearly on α in the intervals $([\bar{n}], [0, \overline{n, 1}])$. Compare with Figure 2.6, where $h(\alpha, 10^4, 4 \cdot 10^5)$ is computed at 1600 values of $\alpha \in (0.264, 0.281)$ and with Figure 2.8 where $h(\alpha, 10^4, 2 \cdot 10^5)$ is computed at 1799 values of $\alpha \in (0.09, 0.11)$.
Figure 2.5 is a graph of $h(\alpha, 10^4, 4 \cdot 10^5)$ at 1314 uniformly distributed random values of α in the interval $(0.29, 0.3)$: here the non-monotone character of the function $\alpha \mapsto h(T_\alpha)$ is quite evident. A magnification of Figure 2.6 corresponding to values $\alpha \in (0.278, 0.281)$, showed in Figure 2.7, suggests that the same phenomenon occurs at the end of each of the plateaux exhibited in Figure 2.4.

Chapter 3

Natural Extensions

3.1 Fibred systems

Natural extensions were introduced by Rohlin [19] as a general method to pass from a given endomorphism to an automorphism of which the first is a factor. Nakada, Ito e Tanaka applied it to continued fractions [18] [16].
The main reference for this section is Schweiger's book [20], whose approach is based on the notion of fibred systems.

3.1 (Fibred system). Consider a set B, and an application $T : B \to B$. (B, T) is called a *fibred system* if there exists a finite or countable alphabet A, and a surjective function $k : B \to A$ such that $T_{|k^{-1}\{a\}}$ is injective $\forall a \in A$.

Let $\Sigma = A^{\mathbb{N}}$, $\sigma : \Sigma \to \Sigma$ the shift map. We define a *representation function* $\phi : B \to \Sigma$ as follows:

$$(\phi(x))_i = k(T^{i-1}x)$$

Then we have the commuting diagram

$$\begin{array}{ccc} B & \xrightarrow{T} & B \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\phi} \\ \Sigma & \xrightarrow[\sigma]{} & \Sigma \end{array}$$

The elements of $\phi(B) \subseteq \Sigma$ are called *admissible sequences.*

3.2 (Cylinders). We call *cylinders of rank* 1 of the system (B, T) the sets

$$B_a = k^{-1}\{\bar{k}\},\ \bar{k} \in A;$$

The *cylinder of rank* n associated to the block $(k_1, \ldots, k_n)$, $k_i \in A$ is defined as follows:

$$B(k_1, \ldots, k_n) = B(k_1) \cap T^{-1}B(k_2) \cap \ldots \cap T^{-n+1}B(k_n)$$

We will say that $(k_1, \ldots, k_n)$ is an *admissible block* if $B(k_1, \ldots, k_n) \neq \varnothing$.

Proposition 3.3. *$\forall n \geq 1$, the following properties hold:*

a) $\bigcup_{a\in A} B(k_1,\dots,k_n,a) = B(k_1,\dots,k_n)$

b) $T^{-1}B(k_1,\dots,k_n) = \bigcup_{a\in A} B(a,k_1,\dots,k_n)$

In the sequel we will assume that the following additional conditions are satisfied:

1. $B \subseteq \mathbb{R}$ and the σ-algebra generated by cylinders is the σ-algebra $\mathcal{B}(B)$ of Borel sets of B;
2. T is measurable with respect to $\mathcal{B}(B)$;
3. $\forall (k_1,\dots,k_n)$, $T_{|\mathring{B}(k_1,\dots,k_n)}$ is differentiable.

We will denote by $V_{(k_1,\dots,k_n)}$ the local inverse of $T_{|\mathring{B}(k_1,\dots,k_n)}$, and by $w_{(k_1,\dots,k_n)}$ its Jacobian with respect to the Lebesgue measure.

3.4 (Dual system). Let (B,T) be a fibred system with respect to the alphabet A, $B^\#$ a set, $T^\# : B\times B^\# \to B^\#$ such that:

a) $\forall x\in B$ fixed, $(B^\#, T^\#(x,\cdot))$ is a fibred system, whose alphabet is still A;

b) Let $B^\#_k(x)$, $k\in A$ the cylinders of $(B^\#, T^\#(x,\cdot))$. Then $\forall k\in A$, $B^\#_k(x)\neq \varnothing$ and

$$V_{k_1}x \text{ is well-defined} \;\Rightarrow\; V^\#_{k(x)}(Tx)B^\#_{k_1}(x) \text{ is well-defined} \tag{3.1}$$

$(B^\#, T^\#(x,\cdot))$ is called a *dual system for* (B,T) *in* x.[1]

Let $k^\#(x,y)$, $y\in B^\#$ be the representation function on $(B^\#, T^\#(x,\cdot))$, with $V^\#_{(k_1,\dots,k_n)}(x)$ the local inverse of $T^\#(x,\cdot)_{|\mathring{B}^\#_{(k_1,\dots,k_n)}(x)}$, and with $w^\#_{(k_1,\dots,k_n)}(x)$ its Jacobian with respect to the Lebesgue measure. Moreover we denote by $\mathcal{B}^\#$ the σ-algebra of Borel sets of $B^\#$.

3.5. We define the cylinders of $B^\#$ with respect to x as follows:

$$\begin{aligned} B^\#_{(k_1,\dots,k_n)}(x) \doteq B^\#_{k_1}(x) \cap \Big(T^\#(x,\cdot)^{-1}B^\#_{k_2}(V_{k_1}x)\Big) \cap \\ \cap \Big(T^\#(x,\cdot)^{-1}T^\#(V_{k_1}x,\cdot)^{-1}B^\#_{k_3}(V_{(k_2,k_1)}x)\Big) \cap \cdots \\ \cdots \cap \Big(T^\#(x,\cdot)^{-1}\cdots T^\#(V_{(k_{n-2},\dots,k_1)}x,\cdot)^{-1}B^\#_{k_n}(V_{(k_{n-1},\dots,k_1)}x)\Big) \end{aligned}$$

Condition (3.1) ensures that if $V_{(k_n,\dots,k_1)}x$ is well-defined, $B^\#_{(k_1,\dots,k_n)}(x)$ is non-empty. Also remark that if $y\in B^\#_{(k_1,\dots,k_n)}(x)$,

$$\begin{aligned} \Big(T^\#(x,y)\in B^\#_{k_2}(V_{k_1}x)\Big) \cap \Big(T^\#(V_{k_1}x,\cdot)^{-1}B^\#_{k_3}(V_{(k_2,k_1)}x)\Big) \cap \cdots \\ \cdots \cap \Big(T^\#(V_{k_1}x,\cdot)^{-1}\cdots T^\#(V_{(k_{n-2},\dots,k_1)}x,\cdot)^{-1}B^\#_{k_n}(V_{(k_{n-1},\dots,k_1)}x)\Big) = \\ = B^\#_{(k_2,\dots,k_n)}(V_{k_1}x) \end{aligned}$$

[1] We remark here that for the sake of our proofs we need a more general definition of dual system than the one adopted in Schweiger [20], that is our dual map $T^\#$ is dependent on x. All the duality properties still hold, as we will show in the following pages.

3.6 (Kernel). $K : B \times B^{\#} \to \mathbb{R}$ is called a *kernel* if it is non-negative, measurable and $\forall (x,y) \in B \times B^{\#}$ such that $T^{-1}\{x\} \cap B(k) \neq \varnothing$, $(T^{\#}(x,\cdot))^{-1}\{y\} \cap B_k^{\#}(x) \neq \varnothing$, we have

$$K(x, V_k^{\#}(x,y)) w_k^{\#}(x,y) = K(V_k(x), y) w_k(x) \tag{3.2}$$

Proposition 3.7. *Define*

$$D(x) \doteqdot \left\{ y \in B^{\#} \mid \forall n > 0,\ y \in B^{\#}_{(k_1,\dots,k_n)}(x) \Leftrightarrow T^{-n}\{x\} \cap B(k_n,\dots,k_1) \neq \varnothing \right\}$$

Then the following properties hold:

a) $y \in D(x) \Rightarrow V^{\#}_{k(x)}(Tx, y) \in D(Tx)$

b) $y \in D(x) \Rightarrow T^{\#}(x,y) \in D(V_{k(x,y)}(x))$

c) $D(x) = \bigcup_k \left(B_k^{\#}(x) \cap V_k^{\#}(x,\cdot) D(V_k(x)) \right)$

Proof. a) If $y \in B^{\#}_{(k_1,\dots,k_n)}(x) \cap D(x)$, $V^{\#}_{k(x)}(Tx,y) \in B^{\#}_{(k(x),k_1,\dots,k_n)}(Tx)$,

$$V^{\#}_{k(x)}(Tx,y) \in D(Tx) \ \Leftrightarrow\ T^{-n-1}\{Tx\} \cap B(k_n,\dots,k_1,k(x)) \neq \varnothing$$

b) If $y \in B^{\#}_{(k_1,\dots,k_n)}(x) \cap D(x)$, $T^{\#}(x,y) \in B^{\#}_{(k_2,\dots,k_n)}(V_{k_1}x)$,

$$T^{\#}(x,y) \in D(V_{k(x,y)}(x)) \ \Leftrightarrow\ T^{-n-1}\{V_{k(x,y)}(x)\} \cap B_{(k_n,\dots,k_2)} \neq \varnothing \ \Leftrightarrow$$
$$T^{-n}\{x\} \cap B_{(k_n,\dots,k_2,k_1)}\} \neq \varnothing \ \Leftrightarrow\ y \in D(x)$$

c) Let $y \in D(x)$: then point (b) implies that $T^{\#}(x,y) \in D(V_{k(x,y)}(x))$, and so $D(x) \subseteq \bigcup_k B_k^{\#}(x) \cap T^{\#}(x,\cdot)^{-1} D(V_k(x))$. But point (a) entails that $T^{\#}(x,y) \in D(V_{k(x,y)}(x))$, and so

$$V^{\#}_{k(V_{k(x,y)}(x))}(x,\cdot) T^{\#}(x,y) = V^{\#}_{k(x,y)}(x,\cdot) T^{\#}(x,y) \in D(T V_{k(x,y)}(x))$$
$$\Rightarrow\ y \in D(x) \Rightarrow D(x) \supseteq \bigcup_k B_k^{\#}(x) \cap V_k^{\#}(x,\cdot) D(V_k(x)) \qquad \square$$

Remark 3.8. The statement (c) in Proposition 3.7 is equivalent to the definition of $D(x)$. In fact, suppose that (c) holds: we can prove by induction on n that

$$y \in \left(B^{\#}_{(k_1,\dots,k_n)}(x) \cap D(x) \right) \Rightarrow \left(T^{-n}\{x\} \cap B_{(k_n,\dots,k_1)} \right) \neq \varnothing \tag{3.3}$$

For $n = 1$, $y \in B^{\#}k_1(x) \Rightarrow y \in B^{\#}k_1(x) \cap V_{k_1}^{\#}(x,\cdot) D(V_{k_1}(x))$. Therefore $D(V_{k_1}(x)) \neq \varnothing$, and $V_{k_1}(x)$ is well-defined.
Now suppose that (3.3) holds for $n-1$, and consider $y \in B^{\#}_{(k_1,\dots,k_n)}(x) \cap D(x)$. Then $T^{\#}(x,y) \in B^{\#}_{(k_2,\dots,k_n)}(V_{k_1}(x)), T^{\#}(x,y) \in D(V_{k_1}(x))$. For inductive hypothesis, $T^{-n+1}\{V_{k_1}(x)\} \cap B_{(k_n,\dots,k_2)} \neq \varnothing \ \Rightarrow\ T^{-n}\{x\} \cap B_{(k_n,\dots,k_1)} \neq \varnothing$. On the other side, if $T^{-n}\{x\} \cap B_{(k_n,\dots,k_1)} = \varnothing$, if there existed $y \in B^{\#}_{(k_1,\dots,k_n)}(x) \cap D(x)$, we would have $T^{\#}(x,y) \in D(V_{k_1}(x))$ and, recursively, $(T^{\#})^i(V_{(k_1,\dots,k_{i-1})}, y) \in D(V_{(k_1,\dots,k_i)}(x)), \forall i = 1,\dots,n$, that is $D(V_{(k_1,\dots,k_i)}(x)) \neq \varnothing\ \forall i = 1,\dots,n$, a contradiction.

Theorem 3.9. *Let (B,T) be a fibred system, $B^{\#}$ a set, $T^{\#}: B\times B^{\#}$ such that $(B^{\#},T^{\#}(x,\cdot))$ is a dual fibred system $\forall x\in B$, and $D(x)\neq\varnothing\ \forall x\in B$. Let*

$$\begin{aligned}&\overline{B}=\{(x,y)\mid x\in B,\ y\in D(x)\}\\&\overline{T}:\overline{B}\to\overline{B},\qquad \overline{T}:(x,y)\mapsto(Tx,V^{\#}_{k(x)}(Tx,y))\\&\overline{\mathcal{B}}=\left\{A\in\mathcal{B}\times\mathcal{B}^{\#}\mid A\subseteq\overline{B}\right\}\end{aligned}$$

Then $(\overline{B},\overline{T})$ is a fibred system and $\overline{T}$ is invertibile. Moreover if $\overline{T}$ is measurable with respect to $\overline{\mathcal{B}}$ and $K:B\times B^{\#}\to\mathbb{R}$ is a kernel, $K_{|\overline{B}}$ is an invariant density for $\overline{T}$.

3.10 (Natural extension). $(\overline{B},\overline{T})$ is called a *natural extension* of (B,T).

Proof of Theorem 3.9. The map $\widetilde{T}:(x,y)\mapsto(V_{k(y)}(x),T^{\#}(x,y))$ is well defined thanks to Proposition 3.7 (b), and it is an inverse of $\overline{T}$:

$$\begin{aligned}&\widetilde{T}\overline{T}(x,y)=\widetilde{T}(Tx,V^{\#}_{k(x)}(Tx,y))=(V_{k(x)}Tx,T^{\#}(Tx,\cdot)V^{\#}_{k(x)}(Tx,y)=(x,y)\\&\overline{T}\widetilde{T}(x,y)=\overline{T}(V_{k(y)}(x),T^{\#}(x,y))=(x,V^{\#}_{k(x,y)}(x,\cdot)T^{\#}(x,y))=(x,y)\end{aligned}$$

Remark that

$$w_k(x)=\frac{1}{|T'(V_k(x))|}\Rightarrow w_k(Tx)=\frac{1}{|T'(x)|}$$

From equation (3.2) it follows that

$$K(Tx,V^{\#}_{k(x)}(Tx,y))w^{\#}_{k(x)}(Tx,y)=K(V_{k(x)}T(x),y)w_{k(x)}(T(x))$$

Then from the change of variable formula we find that $\forall C\subseteq\bar{B}$,

$$\begin{aligned}&\int_{\bar{T}(C)}K(x,y)dxdy=\int_C K(Tu,V^{\#}_{k(u)}(Tu,v))\,|T'(u)|\,w^{\#}_{k(u)}(Tu,v)dudv=\\&=\int_C K(V_{k(u)}T(u),v)w_{k(u)}(T(u))\,|T'(u)|\,dudv=\int_C K(u,v)dudv\end{aligned}$$ □

Corollary 3.11. *$h(x)=\int_{D(x)}K(x,y)dy$ is an invariant density for T.*

Proof. It is sufficient to consider the commuting diagram

$$\begin{array}{ccc}\overline{B} & \xrightarrow{\overline{T}} & \overline{B}\\ {\scriptstyle p^+}\downarrow & & \downarrow{\scriptstyle p^+}\\ B & \xrightarrow[T]{} & B\end{array}$$ □

3.2 Natural extensions for $\alpha\in[\sqrt{2}-1,1)$

In this section we summarize the known results on the natural extensions and invariant densities for α-continued fraction algorithms. For $\alpha\in\left[\frac{1}{2},1\right)$ these constructions are due to Nakada [16], while for $\alpha\in\left[\sqrt{2}-1,\frac{1}{2}\right)$ the invariant densities have been found by Cassa [6], who did not employ the natural extension method, but another technique that involves counting the poles of a meromorphic function. Here we translate his result into the language of natural extensions, providing an independent proof.

Theorem 3.12. *Let $\alpha \in [\sqrt{2}-1, 1]$, and define the domain $D_\alpha \subset \mathbb{R}^2$ as follows:*

1. *For $\alpha \in (g, 1)$,*

$$D_\alpha \doteq \left(\left[\alpha - 1, \frac{1-\alpha}{\alpha} \right] \times \left[0, \frac{1}{2} \right] \right) \cup \left(\left(\frac{1-\alpha}{\alpha}, \alpha \right) \times [0,1] \right),$$

2. *For $\alpha \in \left[\frac{1}{2}, g\right]$,*

$$D_\alpha \doteq \left(\left[\alpha - 1, \frac{1-2\alpha}{\alpha} \right] \times [0, 1-g] \right) \cup \left(\left(\frac{1-2\alpha}{\alpha}, \frac{2\alpha-1}{1-\alpha} \right) \times \left[0, \frac{1}{2} \right] \right) \cup$$
$$\cup \left(\left[\frac{2\alpha-1}{1-\alpha}, \alpha \right) \times [0, g] \right)$$

3. *For $\alpha \in \left[\sqrt{2}-1, \frac{1}{2}\right)$,*

$$D_\alpha \doteq \left(\left[\alpha - 1, \frac{2\alpha-1}{1-\alpha} \right] \times [0, 1-g] \right) \cup$$
$$\cup \left(\left(\frac{2\alpha-1}{1-\alpha}, \frac{1-2\alpha}{\alpha} \right) \times ([0, 1-g]) \cup \left[\frac{1}{2}, g \right] \right) \cup \left(\left[\frac{1-2\alpha}{\alpha}, \alpha \right) \times [0, g] \right)$$

Define $T_\alpha : D_\alpha \to D_\alpha$ as follows:

$$\overline{T}_\alpha(x, y) = \left(T_\alpha(x), \frac{1}{k(x) + \operatorname{sign}(x) y} \right)$$

where $k(x) = \left[\left| \frac{1}{x} \right| + 1 - \alpha \right]$.
Then $\overline{T}_\alpha : D_\alpha \to D_\alpha$ is well-defined and bijective, and is a representation of the natural extension of T_α. Moreover, $\overline{T}_\alpha$ preserves the density $\frac{C_\alpha}{(1+xy)^2}$, where C_α is a suitable normalizing constant.

The vertical sections of the domain D_α for $x \in I_\alpha$ correspond to the sets $D(x)$ defined in Proposition 3.7.

Corollary 3.13. *Let g and G denote the Golden numbers $\frac{\sqrt{5}-1}{2}$ and $\frac{\sqrt{5}+1}{2}$ respectively. Then the unique invariant density ρ_α for T_α is given by the following expressions:*

- *For $g < \alpha \leq 1$,*

$$\rho_\alpha(x) = \frac{1}{\log(1+\alpha)} \left(\chi_{\left[\alpha-1, \frac{1-\alpha}{\alpha}\right]}(x) \frac{1}{x+2} + \chi_{\left(\frac{1-\alpha}{\alpha}, \alpha\right)}(x) \frac{1}{x+1} \right)$$

- *For $\frac{1}{2} < \alpha \leq g$,*

$$\rho_\alpha(x) = \frac{1}{\log G} \left(\chi_{\left[\alpha-1, \frac{1-2\alpha}{\alpha}\right]}(x) \frac{1}{x+G+1} + \right.$$
$$\left. + \chi_{\left(\frac{1-2\alpha}{\alpha}, \frac{2\alpha-1}{1-\alpha}\right)}(x) \frac{1}{x+2} + \chi_{\left[\frac{2\alpha-1}{1-\alpha}, \alpha\right)}(x) \frac{1}{x+G} \right)$$

- *For* $\sqrt{2}-1 \leq \alpha \leq \frac{1}{2}$,

$$\rho_\alpha(x) = \frac{1}{\log G}\left(\chi_{[\alpha-1,\frac{2\alpha-1}{1-\alpha})}(x)\frac{1}{x+G+1}+\right.$$
$$\left.+\chi_{[\frac{2\alpha-1}{1-\alpha},\frac{1-2\alpha}{\alpha})}\left(\frac{1}{x+G+1}+\frac{1}{x+G}-\frac{1}{x+2}\right)+\chi_{[\frac{1-2\alpha}{\alpha},\alpha)}(x)\frac{1}{x+G}\right)$$

This corollary follows easily from Theorem 3.12: since T_α is a factor of $\overline{T}_\alpha$, its invariant measure $\rho_\alpha dx$ is simply the image measure of $K(x,y)dxdy$ with respect to the projection on the first coordinate, that is $\rho_\alpha(x) = \int_{D_\alpha(x)} \frac{dy}{(1+xy)^2}$.

Proof in the case $\alpha \in (g,1)$. For the sake of simplicity, we will write T instead of T_α, and $\overline{T}$ instead of $\overline{T}_\alpha$.
In this case, we have $k(\alpha) = 1$, and $T(\alpha) = \frac{1-\alpha}{\alpha}$. Let $r = k(1-\alpha)$: then $k(T(\alpha)) = r-1$:

$$k(T(\alpha)) = \left[\frac{\alpha}{1-\alpha}+1-\alpha\right] = \left[\frac{1}{1-\alpha}-\alpha\right] = \left[\frac{1}{1-\alpha}+1-\alpha\right]-1 = r-1$$

Moreover, we also have $T^2(\alpha) = T(\alpha-1)$:

$$T(\alpha-1) = \frac{1}{1-\alpha} - r = \frac{\alpha}{1-\alpha} - (r-1) = T^2(\alpha)$$

In order to show that $\overline{T}$ is bijective on the domain D_α, we consider a partition of D_α into suitable rectangles, as shown in Figure 3.1. Their images are shown in Figure 3.2. If we show that we can divide the domain D_α into rectangles where the components of $\overline{T}$ are monotonic, and that these are mapped into distinct rectangles which cover the whole domain, we will prove at once that $\overline{T}$ is well-defined, one-to-one and onto.
For the sake of simplicity, we will not make any distinction between open and closed intervals, since the natural extension and invariant density are defined only modulo negligible sets. To be more precise, we should remark that if $D_\alpha(x)$ is the vertical section corresponding to x, the union of the boundaries of the rectangles $\bigcup_{x\in I_j}\{x\}\times D_\alpha(x)$ is still a set of measure 0.
Let $\xi = T^2(\alpha) = T(\alpha-1)$. Then

$$\overline{T}\left(\left[\alpha-1,\frac{-1}{r+\alpha}\right]\times\left[0,\frac{1}{2}\right]\right) = [\xi,\alpha]\times\left[\frac{1}{r},\frac{1}{r-\frac{1}{2}}\right] \tag{a}$$

For all $h > r$, we have

$$\overline{T}\left(\left[\frac{-1}{h-1+\alpha},\frac{-1}{h+\alpha}\right]\times\left[0,\frac{1}{2}\right]\right) = [\alpha-1,\alpha]\times\left[\frac{1}{h},\frac{1}{h-\frac{1}{2}}\right] \tag{b}$$

Similarly, for all $h \geq r$ we find

$$\overline{T}\left(\left[\frac{1}{h+\alpha},\frac{1}{h-1+\alpha}\right]\times\left[0,\frac{1}{2}\right]\right) = [\alpha-1,\alpha]\times\left[\frac{1}{h+\frac{1}{2}},\frac{1}{h}\right] \tag{c}$$

Moreover,

$$\overline{T}\left(\left[\frac{1}{r-1+\alpha},\frac{1-\alpha}{\alpha}\right]\times\left[0,\frac{1}{2}\right]\right) = [\xi,\alpha]\times\left[\frac{1}{r-\frac{1}{2}},\frac{1}{r-1}\right] \tag{d}$$

$$\overline{T}\left(\left[\frac{1-\alpha}{\alpha},\frac{1}{r-2\alpha}\right]\times[0,1]\right) = [\alpha-1,\xi]\times\left[\frac{1}{r},\frac{1}{r-1}\right] \tag{e}$$

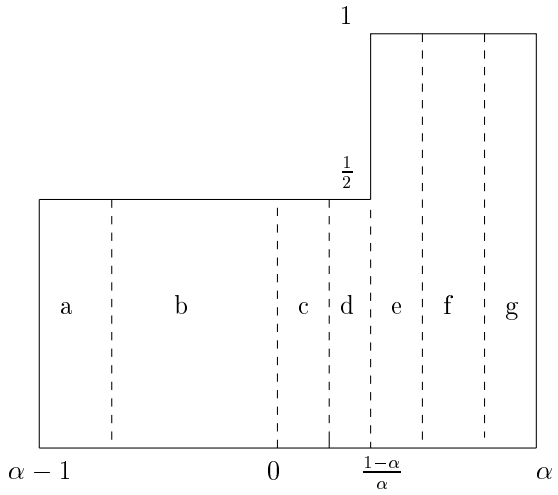

Figure 3.1: A simplified diagram showing the blocks (a)-(g) in the domain D_α when $\alpha \in (g, 1)$.

For $2 \leq h \leq r-2$,

$$\overline{T}\left(\left[\frac{1}{h+\alpha}, \frac{1}{h-1+\alpha}\right] \times [0,1]\right) = [\alpha - 1, \alpha] \times \left[\frac{1}{h+1}, \frac{1}{h}\right] \tag{f}$$

Finally,

$$\overline{T}\left(\left[\frac{1}{2+\alpha}, \alpha\right] \times [0,1]\right) = \left[\frac{1-\alpha}{\alpha}, \alpha\right] \times \left[\frac{1}{2}, 1\right] \tag{g}$$

Thus $\overline{T}$ is bijective. Finally, the fact that $K(x,y)$ is invariant for $\overline{T}$ can be easily checked through the change of variables formula: the determinant of the Jacobian for $\overline{T}$ is respectively $\frac{1}{x^2(k(x)+y)^2}$ when $x > 0$ and $\frac{1}{x^2(k(x)-y)^2}$ when $x < 0$. Then for any $A \subset D_\alpha$, if we put $A^+ = A \cap \{x > 0\}$ and $A^- = A \cap \{x < 0\}$, we have

$$\begin{aligned}\int_{\overline{T}(A)} K(x,y)dxdy &= \frac{1}{C_\alpha}\left(\int_{A^+} \frac{1}{u^2(1/u+v)^2}dudv + \right. \\ \left. + \int_{A^-} \frac{1}{u^2(-1/u-v)^2}dudv\right) &= \frac{1}{C_\alpha}\left(\int_{A^+} \frac{dudv}{(1+uv)^2} + \int_{A^-} \frac{dudv}{(1+uv)^2}\right)\end{aligned} \tag{3.4}$$

This concludes the proof. □

Proof in the case $\alpha \in \left[\frac{1}{2}, g\right]$. This proof is identical to the previous one, except for the shape of the domain D_α; we need to check again that $\overline{T}$ is well-defined, one-to-one and onto.
In this case, we have $\alpha^2 + \alpha - 1 \leq 0$, so $k(\alpha) = 2$ and $k(\alpha - 1) = 2$. Then $T(\alpha) = \frac{1-2\alpha}{\alpha}$, $T(\alpha - 1) = \frac{2\alpha - 1}{1-\alpha}$. It is easy to check that $k(T(\alpha)) = 3$, $k(T(\alpha - 1)) = 2$, and $T^2(\alpha) = T^2(1-\alpha) = \frac{3-5\alpha}{2\alpha-1}$. Let $\xi = T^2(\alpha)$.
As in the previous case, we compute the images through $\overline{T}$ of a suitable partition of the domain D_α into rectangles, as shown in Figures 3.3 and 3.4:

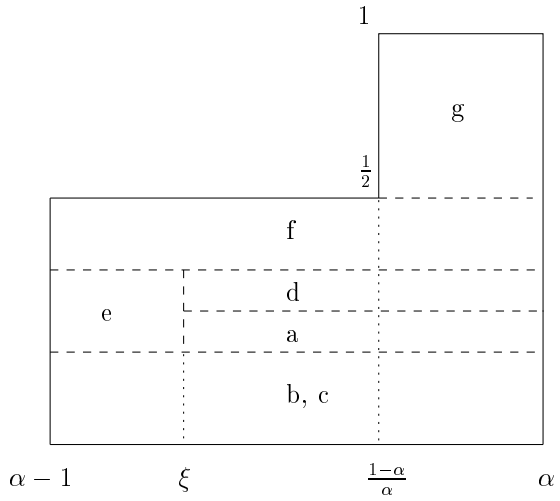

Figure 3.2: A simplified diagram showing the images with respect to $\overline{T}$ of the blocks (a)-(g) in the domain D_α when $\alpha \in (g, 1)$.

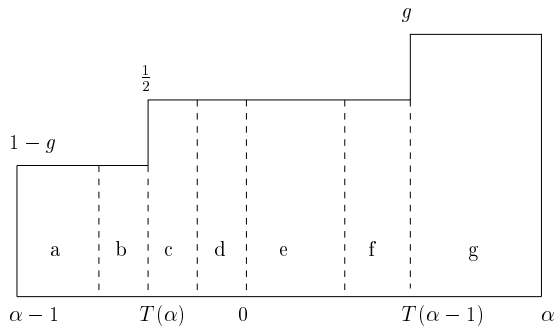

Figure 3.3: A simplified diagram showing the blocks (a)-(g) in the domain D_α when $\alpha \in \left[\frac{1}{2}, g\right]$.

$$\overline{T}\left(\left[\alpha - 1, \frac{-1}{2+\alpha}\right] \times [0, 1-g]\right) = [T(\alpha - 1), \alpha] \times \left[\frac{1}{2}, g\right] \quad \text{(a)}$$

$$\overline{T}\left(\left[\frac{-1}{2+\alpha}, T(\alpha)\right] \times [0, 1-g]\right) = [\alpha - 1, \xi] \times \left[\frac{1}{3}, 1-g\right] \quad \text{(b)}$$

$$\overline{T}\left(\left[T(\alpha), \frac{-1}{3+\alpha}\right] \times \left[0, \frac{1}{2}\right]\right) = [\xi, \alpha] \times \left[\frac{1}{3}, \frac{2}{5}\right] \quad \text{(c)}$$

Observe that in equations (a) and (b), we have used the fact that $\frac{1}{1+g} = g$ and $\frac{1}{2+g} = 1 - g$ respectively.
For $h \geq 4$, we have

$$\overline{T}\left(\left[\frac{-1}{h-1+\alpha}, \frac{-1}{h+\alpha}\right] \times \left[0, \frac{1}{2}\right]\right) = [\alpha - 1, \alpha] \times \left[\frac{1}{h+\frac{1}{2}}, \frac{1}{h}\right] \quad \text{(d)}$$

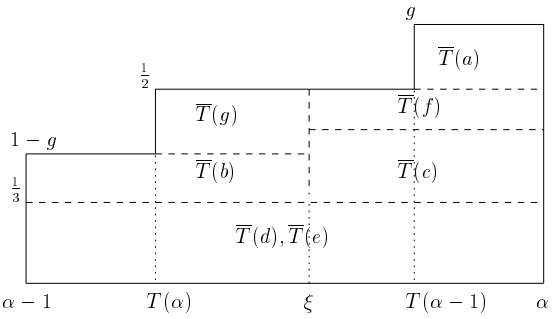

Figure 3.4: A simplified diagram showing the images with respect to $\overline{T}$ of the blocks (a)-(g) in the domain D_α when $\alpha \in \left[\frac{1}{2}, g\right]$.

Similarly, for $h \geq 3$ we have

$$\overline{T}\left(\left[\frac{1}{h+\alpha}, \frac{1}{h-1+\alpha}\right] \times \left[0, \frac{1}{2}\right]\right) = [\alpha, 1-\alpha] \times \left[\frac{1}{h+\frac{1}{2}}, \frac{1}{h}\right] \tag{e}$$

Finally,

$$\overline{T}\left(\left[\frac{1}{2+\alpha}, T(1-\alpha)\right]\right) \times \left[0, \frac{1}{2}\right] = [\xi, \alpha] \times \left[\frac{2}{5}, \frac{1}{2}\right] \tag{f}$$

$$\overline{T}\left([T(1-\alpha), \alpha] \times [0, g]\right) = [T(\alpha), \xi] \times \left[1-g, \frac{1}{2}\right] \tag{g}$$

This completes the proof. □

Proof in the case $\alpha \in \left[\sqrt{2}-1, \frac{1}{2}\right)$. Also in this case we have $k(\alpha) = 2$, $k(\alpha - 1) = 2$. Then $T(\alpha) = \frac{1-2\alpha}{\alpha}$, $T(\alpha - 1) = \frac{2\alpha-1}{1-\alpha}$. It is easy to check that if $k(T(\alpha - 1)) = r$, then $k(T(\alpha)) = r - 1$, and $r \geq 4$. Once again, we have $T^2(\alpha) = T^2(\alpha - 1) = \xi$. Recalling that $\frac{1}{1+g} = g$, we have

$$\overline{T}\left(\left[\alpha - 1, \frac{-1}{2+\alpha}\right] \times [0, 1-g]\right) = [T(\alpha - 1), \alpha] \times \left[\frac{1}{2}, g\right] \tag{a}$$

For $3 \leq h \leq r - 1$,

$$\overline{T}\left(\left[\frac{-1}{h-1+\alpha}, \frac{-1}{h+\alpha}\right] \times [0, 1-g]\right) = [\alpha - 1, \alpha] \times \left[\frac{1}{h}, \frac{1}{h-1+g}\right] \tag{b}$$

$$\overline{T}\left(\left[\frac{-1}{r-1+\alpha}, T(\alpha - 1)\right] \times [0, 1-g]\right) = [\alpha - 1, \xi] \times \left[\frac{1}{r}, \frac{1}{r-1+g}\right] \tag{c}$$

$$\overline{T}\left(\left[T(\alpha - 1), \frac{-1}{r+\alpha}\right] \times \left([0, 1-g] \cup \left[\frac{1}{2}, g\right]\right)\right) = \\ = [\xi, \alpha] \times \left(\left[\frac{1}{r}, \frac{1}{r-1+g}\right] \cup \left[\frac{1}{r-\frac{1}{2}}, \frac{1}{r-g}\right]\right) \tag{d}$$

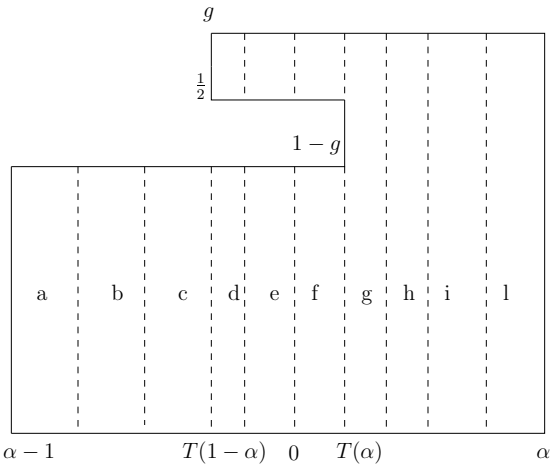

Figure 3.5: A simplified diagram showing the blocks (a)-(l) in the domain D_α when $\alpha \in \left[\sqrt{2}-1, \frac{1}{2}\right)$.

For $h \geq r+1$,

$$\overline{T}\left(\left[\frac{-1}{h-1+\alpha}, \frac{-1}{h+\alpha}\right] \times \left([0,1-g] \cup \left[\frac{1}{2}, g\right]\right)\right) = \\ = [\alpha-1, \alpha] \times \left(\left[\frac{1}{h-\frac{1}{2}}, \frac{1}{h-g}\right] \cup \left[\frac{1}{h+1-g}, \frac{1}{h}\right]\right) \quad \text{(e)}$$

For $h \geq r$,

$$\overline{T}\left(\left[\frac{1}{h+\alpha}, \frac{1}{h-1+\alpha}\right] \times \left([0,1-g] \cup \left[\frac{1}{2}, g\right]\right)\right) = \\ = [\alpha-1, \alpha] \times \left(\left[\frac{1}{h+1-g}, \frac{1}{h}\right] \cup \left[\frac{1}{h+g}, \frac{1}{h+\frac{1}{2}}\right]\right) \quad \text{(f)}$$

Similarly, we have

$$\overline{T}\left(\left[\frac{1}{r-1+\alpha}, T(\alpha)\right]\right) \times \left([0,1-g] \cup \left[\frac{1}{2}, g\right]\right) = \\ = [\xi, \alpha] \times \left(\left[\frac{1}{r-1+g}, \frac{1}{r-\frac{1}{2}}\right] \cup \left[\frac{1}{r-g}, \frac{1}{r-1}\right]\right) \quad \text{(g)}$$

$$\overline{T}\left(\left[T(\alpha), \frac{1}{r-2+\alpha}\right] \times [0,g]\right) = [\alpha-1, \xi] \times \left[\frac{1}{r-1+g}, \frac{1}{r-1}\right] \quad \text{(h)}$$

For $3 \leq h \leq r-2$,

$$\overline{T}\left(\left[\frac{1}{h+\alpha}, \frac{1}{h-1+\alpha}\right] \times [0,g]\right) = [\alpha-1, \alpha] \times \left[\frac{1}{h+g}, \frac{1}{h}\right] \quad \text{(i)}$$

Finally,

$$\overline{T}\left(\left[\frac{1}{2+\alpha}, \alpha\right] \times [0,g]\right) = \left[\frac{1-2\alpha}{\alpha}, \alpha\right] \times \left[\frac{1}{2+g}, \frac{1}{2}\right] \quad \text{(l)}$$

This completes the proof. □

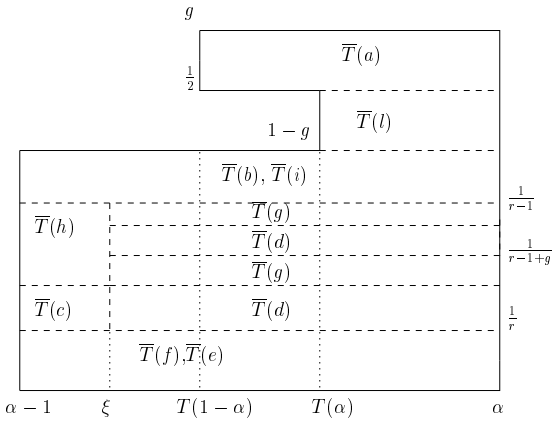

Figure 3.6: A simplified diagram showing the images with respect to $\overline{T}$ of the blocks (a)-(l) in the domain D_α when $\alpha \in \left[\sqrt{2}-1, \frac{1}{2}\right)$.

3.3 Natural extension for $\alpha = \frac{1}{r}$

In the case $\alpha \in (0, \sqrt{2}-1]$, the structure of the domain D_α of the natural extension for T_α seems to be much more intricate than for $\alpha > \sqrt{2}-1$. Here we find the exact expression for D_α and the invariant density of T_α when $\alpha \in \left\{\frac{1}{r}, r \in \mathbb{N}\right\}$.

3.3.1 The by-excess continued fraction map

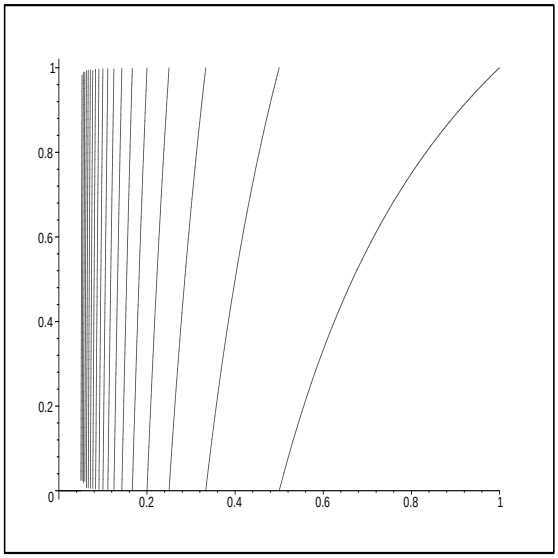

Figure 3.7: Graph of the map M_0.

Before stating our main theorem, we introduce some notations. In the following paragraphs we will often refer to the *by-excess* continued fraction expansion of a number, that is the expansion related to the map $M_0(x) = -\frac{1}{x} + \left[\frac{1}{x} + 1\right]$, $M_0 : [0,1] \to [0,1]$. To simplify notations, we will omit the minus signs and use brackets:

$$\langle a_0, a_1, a_2, \ldots \rangle \doteqdot \cfrac{1}{a_0 - \cfrac{1}{a_1 - \cfrac{1}{a_2 - \cdots}}}, \qquad a_i \in \{2, 3, 4, \ldots\}$$

We will denote a non-integer remainder $x > 1$ by a semicolon:

$$\langle a_0, a_1, \ldots, a_n; x \rangle \doteqdot \cfrac{1}{a_0 - \cfrac{1}{a_1 - \cfrac{\quad}{\ddots - \cfrac{1}{a_n - \cfrac{1}{x}}}}}, \qquad a_i \in \{2, 3, 4, \ldots\}$$

We also recall that the by-excess expansion of any real number $y \in (0,1)$ is infinite, and that

$$y = \langle a_1, a_2, a_3, \ldots \rangle \in \mathbb{Q} \Rightarrow \exists i \text{ s. t. } \forall j \geq i,\ a_j = 2$$

3.3.2 Reflection rules

We begin by making some preliminary observations on the relation between the symbolic dynamics of the map M_0 and the reflection map $x \mapsto 1 - x$ on $[0,1]$, which reveal a sort of "duality"between the digit 2 and the digits greater than 2, and will prove very useful to construct a "dual"fibred system for T_α in the sense of Schweiger [20].

Let $x = \langle a_1, a_2, a_3, \ldots \rangle \in [0,1]$. We would like to determine the by-excess continued fraction expansion of $1 - x$. Since the general solution to this problem turns out to be quite complicated, we will only describe a single step of the algorithm, that is, we will suppose to have computed the first i digits of the expansion and the remainder:

$$1 - x = \cfrac{1}{a'_1 - \cfrac{1}{a'_2 - \cfrac{1}{\ddots - \cfrac{1}{a'_i - (1 - z)}}}}, \qquad z \in [0,1)$$

$$z = \langle h_1, h_2, h_3, \ldots \rangle, \qquad h_i \geq 2$$

We want to determine the first digit of the remainder $1 - z$. For reasons that will become clear later, we will treat any sequence of the kind

$$\underbrace{2, 2, \ldots, 2}_{n}$$

as a single digit.
We will make use of the following well-known identity (see for example [15]) that can be easily proved by induction on n:

$$1 - \cfrac{1}{n + \cfrac{1}{y-1}} = \langle \underbrace{2, 2, \ldots, 2}_{n-1}; y \rangle \qquad \forall y \in \mathbb{R} \tag{3.5}$$

There are three separate cases to consider:

- If $h_1, h_2 \geq 3$, then from the identity (3.5) with $n = h_1$ and

$$y - 1 = -\left(h_2 - \cfrac{1}{h_3 - \ldots}\right), \qquad \frac{1}{y} = -\cfrac{1}{h_2 - 1 - \cfrac{1}{h_3 - \ldots}}$$

we get

$$1 - \langle h_1, h_2, h_3, \ldots \rangle = \big\langle \underbrace{2, \ldots, 2}_{h_1 - 2}; 2 + \langle h_2 - 1, h_3, \ldots \rangle \big\rangle =$$
$$= \big\langle \underbrace{2, \ldots, 2}_{h_1 - 2}; 3 - \big(1 - \langle h_2 - 1, h_3, \ldots \rangle\big) \big\rangle$$

We sum up our observations in the following

Rule 1. If $h_1, h_2 \geq 3$,

$$1 - \langle h_1, h_2, h_3, \ldots \rangle = \left\langle \underbrace{2, \ldots, 2}_{h_1 - 2}, 3; \frac{1}{(1 - \langle h_2 - 1, h_3, \ldots \rangle)} \right\rangle$$

- If $z = \langle h_1, \underbrace{2, \ldots, 2}_{n}, h_3, \ldots \rangle$ with $h_1, h_3 \geq 3$, then

$$1 - z = \left\langle \underbrace{2, \ldots, 2}_{h_1 - 2}; 2 + \frac{1}{1 - \langle \underbrace{2, \ldots, 2}_{n-1}, h_3, \ldots \rangle} \right\rangle$$

We want to use the identity (3.5), with

$$y - 1 = -\frac{1}{\langle \underbrace{2, \ldots, 2}_{n}, h_3, \ldots \rangle} = -\left(2 - \langle \underbrace{2, \ldots, 2}_{n-1}, h_3, \ldots \rangle\right),$$
$$-\frac{1}{y} = \frac{1}{1 - \langle \underbrace{2, \ldots, 2}_{n-1}, h_3, \ldots \rangle}$$

Observe that

$$1-\langle\underbrace{2,\ldots,2}_{n-1},h_3,\ldots\rangle=\frac{1}{n+\langle h_3-1,h_4,\ldots\rangle}$$
$$\Rightarrow\frac{1}{1-\langle\underbrace{2,\ldots,2}_{n-1},h_3,\ldots\rangle}=n+1-(1-\langle h_3-1,h_4,\ldots\rangle)$$

In conclusion, we find

Rule 2. If $h_1,h_3\geq 3$,

$$1-\langle h_1,\underbrace{2,\ldots,2}_{n},h_3,h_4,\ldots\rangle=\left\langle\underbrace{2,\ldots,2}_{h_1-2},n+3;\frac{1}{1-\langle h_3-1,h_4,\ldots\rangle}\right\rangle$$

- If $z=\langle\underbrace{2,\ldots,2}_{n},h_2,\ldots\rangle$, $h_2\geq 3$, then using again the identity (3.5) with

$$y=\frac{1}{\langle h_2,h_3,\ldots\rangle},\qquad\frac{1}{y-1}=\langle h_2-1,h_3,\ldots\rangle$$

we find

$$1-\langle\underbrace{2,\ldots,2}_{n},h_2,h_3,\ldots\rangle=\frac{1}{n+1+\langle h_2-1,h_3,\ldots\rangle}=$$
$$=\frac{1}{n+2-(1-\langle h_2-1,h_3,\ldots\rangle)}$$

Rule 3. If $h_2\geq 3$,

$$1-\langle\underbrace{2,\ldots,2}_{n},h_2,h_3,\ldots\rangle=\left\langle n+2;\frac{1}{1-\langle h_2-1,h_3,\ldots\rangle}\right\rangle$$

Notice that we have taken into consideration all the possible cases. Also remark that Rule 1 and Rule 2 guarantee that in the new digits h_i' a sequence of twos is never followed by another.
Let $\alpha=\frac{1}{r}$, for a fixed $r\geq 3$. Observe that $T_\alpha(\alpha)=0$, and

$$T_\alpha^i(\alpha-1)=\frac{-(r-i-1)}{r-i}\leq 0\qquad\text{for } i=0,\ldots,r-2 \tag{3.6}$$

Let β be the fixed point for M_0 corresponding to the branch $r+1$, and $\xi=\frac{1}{r-\beta}$:

$$\begin{aligned}\beta&=\frac{r+1-\sqrt{(r+1)^2-4}}{2}=\langle r+1,r+1,r+1,r+1,\ldots\rangle\\ \xi&=\frac{2}{r-1+\sqrt{(r+1)^2-4}}=\langle r,r+1,r+1,r+1,\ldots\rangle\end{aligned} \tag{3.7}$$

Then

$$1-\beta=\langle\underbrace{2,\ldots,2}_{r-1},3,\overline{\underbrace{2,\ldots,2}_{r-2},3}\rangle,\qquad 1-\xi=\langle\overline{\underbrace{2,\ldots,2}_{r-2},3}\rangle \tag{3.8}$$

3.3.3 Domain of the natural extension

Let $n \geq 1$, and define

$$H_n^+ = \Bigg\{ (h_1, h_2, \ldots, h_n) \Bigg| h_1 \in \{2, (2,2), \ldots, \underbrace{(2,2,\ldots,2)}_{r-1}\} \cup \{3,4,\ldots,r\},$$
$$h_2, \ldots, h_n \in \{2, (2,2), \ldots, \underbrace{(2,2,\ldots,2)}_{r-2}\} \cup \{3,4,\ldots,r,r+1\},$$
$$\text{and such that } h_i = \underbrace{(2,\ldots,2)}_{s} \Rightarrow h_{i+1} \geq 3 \Bigg\}$$

$$H_n^- = \Bigg\{ (h_1, h_2, \ldots, h_{n-1}) \Bigg| h_1, h_2, \ldots, h_n \in \{2, (2,2), \ldots, \underbrace{(2,2,\ldots,2)}_{r-2}\} \cup$$
$$\cup \{3,4,\ldots,r,r+1\}, \text{ and such that } h_i = \underbrace{(2,\ldots,2)}_{s} \Rightarrow h_{i+1} \geq 3 \Bigg\}$$

Moreover, for $i = 2, 3, \ldots, r-1$ define

$$H_n^i = \Bigg\{ (h_1, h_2, \ldots, h_n) \Bigg| h_1 \in \{2, (2,2), \ldots, \underbrace{(2,2,\ldots,2)}_{r-1-i}\} \cup \{3,4,\ldots,r+1\},$$
$$h_2, \ldots, h_n \in \{2, (2,2), \ldots, \underbrace{(2,2,\ldots,2)}_{r-2}\} \cup \{3,4,\ldots,r,r+1\},$$
$$\text{and such that } h_i = \underbrace{(2,\ldots,2)}_{s} \Rightarrow h_{i+1} \geq 3 \Bigg\}$$

Also define

$$\hat{H}_n^+ = \left\{ (h_1, h_2, \ldots, h_n) \in H_n^+ \mid h_n \geq 3 \right\},$$
$$\hat{H}_n^- = \left\{ (h_1, h_2, \ldots, h_n) \in H_n^- \mid h_n \geq 3 \right\},$$
$$\hat{H}_n^i = \left\{ (h_1, h_2, \ldots, h_n) \in H_n^i \mid h_n \geq 3 \right\}, \quad i = 2, 3, \ldots, r-1$$

Let $V_i(x) = \frac{1}{i-x}$ denote the inverse branches of M_0, and

$$V_{\underbrace{(2,\ldots,2)}_{s}}(x) \doteqdot \underbrace{(V_2 \circ V_2 \circ \cdots \circ V_2)}_{s}(x)$$

Define

$$B^+ = \bigcup_{n=1}^{\infty} \bigcup_{(h_1,h_2,\ldots,h_n) \in \hat{H}_n^+} (V_{h_1} \circ V_{h_2} \circ \cdots \circ V_{h_n})((1-\xi, 1)),$$

and similarly

$$B^- = \bigcup_{n=1}^{\infty} \bigcup_{(h_1,h_2,\ldots,h_n)\in\hat{H}_n^-} (V_{h_1}\circ V_{h_2}\circ\cdots\circ V_{h_n})((1-\xi,1)),$$

$$B^i = \bigcup_{n=1}^{\infty} \bigcup_{(h_1,h_2,\ldots,h_n)\in\hat{H}_n^i} (V_{h_1}\circ V_{h_2}\circ\cdots\circ V_{h_n})((1-\xi,1)), \qquad i=2,\ldots,r-1$$

Finally, let $E, B, D \subset \mathbb{R}^2$ be defined as follows:

$$E = \bigcup_{i=1}^{r-1} \left(\left[-\frac{i}{i+1}, -\frac{(i-1)}{i} \right] \times \left[0, M_0^{i-1}(1-\xi)\right] \right) \cup \left(\left[0, \frac{1}{r}\right] \times [0, 1-\beta] \right),$$

$$B = \bigcup_{i=2}^{r-1} \left(\left[-\frac{i}{i+1}, -\frac{(i-1)}{i} \right] \times B^i \right) \cup \left(\left[-\frac{1}{2}, 0 \right] \times B^- \right) \cup \left(\left[0, \frac{1}{r}\right] \times B^+ \right),$$

$$D = E \setminus B$$

Remark that we have omitted the dependence on r of the sets B^+, B^-, B^i, E, D for simplicity of notation.

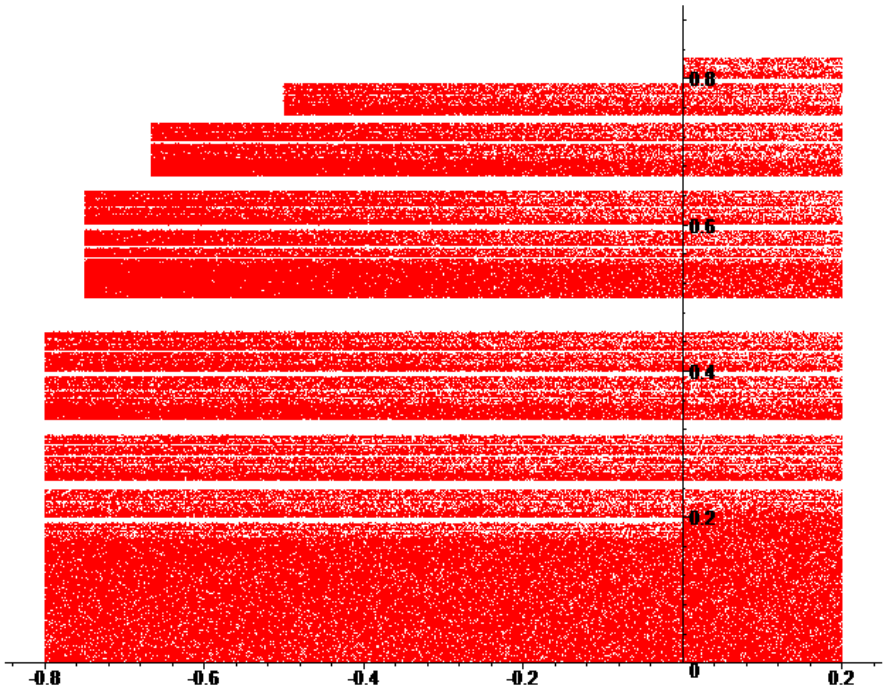

Figure 3.8: A computer simulation for the domain D when $r = 5$.

Theorem 3.14 (Natural extension for $\alpha = \frac{1}{r}$). *Let $\alpha = \frac{1}{r}$, $r \geq 3$ be fixed,*

and let $D \subset \mathbb{R}^2$ be defined as in §3.3.3. Let $k(x) = \left[\left|\frac{1}{x}\right| + 1 - \alpha\right]$, and

$$\overline{T}_\alpha(x,y) = \left(T_\alpha(x), \frac{1}{k(x) + \mathrm{sign}(x)y}\right) \tag{3.9}$$

Then $\overline{T}_\alpha : D \to D$ is well defined, one-to-one and onto, and it preserves the density $K_\alpha(x,y) = \frac{1}{C_\alpha}\frac{1}{(xy+1)^2}$, where $C_\alpha = \int_D \frac{1}{(xy+1)^2}dxdy$. In other words, $\overline{T}_\alpha : D \to D$ is a natural extension for T_α.

Here the reader should remark that the domain D and the function k also depend on α. In the following paragraphs, however, we will write $\overline{T}$ instead of $\overline{T}_\alpha$ for the sake of simplicity.
To prove Theorem 3.14 we shall need the following two lemmas:

Lemma 3.15. *Let $z = \langle h_1, h_2, \ldots, h_n; y\rangle$, where $y > 2$ is a real number and $n \geq 1$. Then $1 - z$ is of the form $\left\langle h'_1, h'_2, \ldots, h'_m; \frac{1}{1-1/(y-1)}\right\rangle$, and*

$$(h_1, h_2, \ldots, h_n) \in H_n^- \Rightarrow (h'_1, h'_2, \ldots, h'_m) \in H_m^+,$$
$$(h_1, h_2, \ldots, h_n) \in H_n^+ \Rightarrow (h'_1, h'_2, \ldots, h'_m) \in H_m^-$$

Lemma 3.16.

$$B^+ \cup \left(1 - B^-\right) = [\xi, 1 - \beta] \pmod 0 \tag{3.10}$$

and their union is disjoint.

Proof of Lemma 3.15. Suppose that $(h_1, h_2, \ldots, h_n) \in H_n^-$. From the application of the Rules 1-3, it is straightforward to check that after a suitable number of steps in the algorithm we will obtain a remainder of the form $\frac{1}{1-1/(y-1)} > 1$. We need to verify that at each step of the reflection algorithm described in Paragraph 3.3.2 the newly introduced digits in the by-excess expansion are in accordance with the definition of H_n^+. We will consider separately the first step and the ensuing ones. In the first step, we will have $h_1 \in \{3, \ldots, r, r+1\}$ or $h_1 \in \{2, (2,2), \ldots, (\underbrace{2, 2, \ldots, 2}_{r-2})\}$.

If $h_1 \geq 3, h_2 \geq 3$, applying Rule 1 we get

$$1 - \langle h_1, h_2, h_3, \ldots\rangle = \left\langle \underbrace{2, 2, \ldots, 2}_{h_1 - 2}, 3; \frac{1}{1 - \langle h_2 - 1, h_3, \ldots\rangle}\right\rangle \tag{a}$$

where $h_1 - 2 \leq r - 1$.
If $h_1 \geq 3, h_2 = (\underbrace{2, 2, \ldots, 2}_{n})$, $n \leq r - 2$, using Rule 2 we find

$$1 - \langle h_1, h_2, h_3, \ldots\rangle = \left\langle \underbrace{2, \ldots, 2}_{h_1 - 2}, n + 3; \frac{1}{1 - \langle h_3 - 1, h_4, \ldots\rangle}\right\rangle \tag{b}$$

where $0 \leq h_1 - 2 \leq r - 1, n + 3 \leq r + 1$.
Lastly, for $h_1 \in (\underbrace{2, 2, \ldots, 2}_{n})$, $n \leq r - 2, h_2 \geq 3$, we have

$$1 - \langle \underbrace{2, \ldots, 2}_{n}, h_2, h_3, \ldots\rangle = \left\langle n + 2; \frac{1}{1 - \langle h_2 - 1, h_3, \ldots\rangle}\right\rangle \tag{c}$$

where $n+2 \leq r$ as needed. In all three cases we found an admissible initial segment for H_m^+.
The subsequent steps can be treated in a similar way, although we have to take into account the ways in which the remainder $\langle h_{i+1}, h_{i+2}, \ldots\rangle$ from the original sequence has been modified by the reflection rules. More precisely: if $h_{i+1} \geq 3$ it will be replaced by $h_{i+1}-1 \in \{2,3,\ldots,r\}$; thus when $h_{i+1}-1 \geq 3$, applying Rules 1 and 2, we will find

$$\underbrace{(2,\ldots,2)}_{h_{i+1}-3}$$

as the next digit, with $1 \leq h_{i+1}-3 \leq r-2$, which is admissible for H_n^+. Moreover, when $h_{i+1}-1=2$ and h_{i+2} is a sequence of twos they will be considered as a single digit, and it is possible to obtain the sequence

$$\underbrace{(2,\ldots,2)}_{r-1}$$

which gives the new digit $r+1$ when we apply Rule 3. We have thus completed the proof for $(h_1,h_2,\ldots,h_n) \in H_n^-$.
When switching the roles of H_n^+ and H_n^-, we can follow the same basic outline. We briefly list the few differences that the reader can easily check for himself: if $(h_1,h_2,\ldots,h_n) \in H_n^+$,

- in (a) and (b), we find $h_1 \leq r \Rightarrow h_1 - 2 \leq r-2$
- in (c), $n \leq r-1 \Rightarrow n+2 \leq r+1$

and so the reflected sequence is in accordance with the definition of H_n^-. □

Before moving on to the next Lemma, we make a few observations.
First of all, notice that since the inverse branches $V_i : x \mapsto \frac{1}{i-x}$ of M_0 are all non-decreasing functions, from the by-excess expansions of a sequence of reals we can obtain full knowledge of their ordering. In fact,

$$\begin{gathered}\langle h_1,h_2,\ldots,h_n,\ldots\rangle < \langle h_1',h_2',\ldots,h_n',\ldots\rangle \\ \Updownarrow \\ \exists i \geq 1 \text{ s. t. } \forall j<i,\ h_j=h_j' \text{ and } h_i > h_i'\end{gathered} \tag{3.11}$$

Recalling the expansions of $\beta, 1-\beta, \xi, 1-\xi$ from equations (3.7) and (3.8), it follows that $B^- \subset [\beta, 1-\xi]$ and $B^+ \subset [\xi, 1-\beta]$, and moreover these are the minimal intervals containing B^+ and B^-: for example, the sequence

$$(V_r \circ V_{r+1} \circ V_{r+1} \circ \cdots \circ V_{r+1})(x),\ x \in (1-\xi,1),$$

goes arbitrarily close to ξ as the number of pre-images grows.
We also observe that if $x \in (1-\xi,1)$, its by-excess expansion must be of the form

$$x = \langle \underbrace{2,2,\ldots,2}_{r-1}, h_r, h_{r+1}, \ldots\rangle, \quad h_r \geq 2$$

Proof of Lemma 3.16. We first want to prove that B^+ and $1 - B^-$ are disjoint. Let $x \in B^-$; then there exists $l \geq 1$ such that $M_0^l(x) \in (1-\xi, 1)$, $M_0^j(x) \in [\beta, 1-\xi]$ $\forall j < l$. Observe that

$$z \in (1-\xi,1) \Rightarrow z = \Big\langle \underbrace{\underbrace{2,\dots,2}_{r-2},3,\dots,\underbrace{2,\dots,2}_{r-2},3}_{k},\underbrace{2,\dots,2}_{r-1},\dots \Big\rangle, \quad k \geq 0$$

Equivalently, for some $i \geq 1$ we have $x = \langle h_1, h_2, \dots, h_i, \underbrace{2,\dots,2}_{n}, \dots\rangle$, where

$$n \geq r-1,\ h_i \geq 3,\ (h_1,h_2,\dots,h_i) \in \hat{H}_i^-,\ \ (h_1,h_2,\dots,h_{i-1}) \in H_{i-1}^-$$

Then from Lemma 3.15 we get

$$1 - x = \Big\langle h_1', \dots, h_m'; \frac{1}{1 - \langle h_i - 1, \underbrace{2,\dots,2}_{n}, \dots\rangle}\Big\rangle,$$

with $(h_1', \dots, h_m') \in H_m^+$, and applying Rule 2 (or Rule 3 if $h_i = 3$), we find

$$1 - x = \Big\langle h_1', \dots, h_m', \underbrace{2,\dots,2}_{h_i-3}, n+3; z \Big\rangle,\ n+3 \geq r+2,\ \ 0 \leq h_i - 3 \leq r-2$$

Observe that $(h_1', \dots, h_m', \underbrace{2,\dots,2}_{h_i-3}) \in H_{m+1}^+$ (or to H_m^+ if $h_i - 3 = 0$), but clearly $(h_1', \dots, h_m', \underbrace{2,\dots,2}_{h_i-3}, n+3)$ does not belong to B^+ because it contains the forbidden digit $n+3$. Since none of the iterates of $1-x$ up to that point belongs to $(1-\xi, 1)$, we find that $1 - x \notin B^+$.

Next we want to show that $B^+ \cup (1 - B^-) = (\xi, 1-\beta)$.
Let $x = \langle h_1, h_2, \dots\rangle \in (\xi, 1-\beta) \setminus B^+$. We must prove that for almost every such x we have $1 - x \in B^-$. We have to consider two cases:

- $\forall n \geq 1, (h_1, \dots, h_n) \in H_n^+$, and so none of the iterates $M_0^{n-1}(x)$ belongs to $(1-\xi, 1)$
- For some i, the by-excess expansion of x contains a forbidden digit h_i: either $h_i = (\underbrace{2,\dots,2}_{n}), n \geq r$ or $h_i \geq r+1$ when $i = 1$, or $h_i = (\underbrace{2,\dots,2}_{n}), n \geq r-1$ or $h_i \geq r+2$ when $i > 1$.

However, observe that since the first condition entails in particular that all the elements h_i in the by-excess expansion of x should be bounded, it is satisfied only for a set of Lebesgue measure 0, and therefore it is negligible for our purposes (equivalently, recall that M_0 is ergodic).
Next, observe that $x < 1-\beta$ implies that the digit 2 cannot appear r consecutive times in the initial segment of the by-excess expansion of x, and $x > \xi$ implies $h_1 \leq r$. Let i be the minimum integer such that $\forall j < i$, $(h_1, \dots, h_j) \in H_j^+$ and

$(h_1,\dots,h_i)\notin H_i^+$ (we have just seen that $i>1$). Then h_i cannot be of the form $(\underbrace{2,\dots,2}_{n})$, $n\geq r-1$ because then $\langle h_i,h_{i+1},\dots\rangle>1-\xi$ and x would belong to B^+. The only case left to consider is then $h_i\geq r+2$. Equivalently, one of the iterates M_0^{i-k}, $k\geq 0$ of x is of the form $\langle\underbrace{r+1,\dots,r+1}_{k},r+2,\dots\rangle<\beta$. Applying Lemma 1 with $n=i-k-1>1$, $\frac{1}{y}=\langle\underbrace{r+1,\dots,r+1}_{k},r+2,\dots\rangle$, we get $1-x=\left\langle h_1',\dots,h_m';\frac{1}{1-1/(y-1)}\right\rangle$, $(h_1',\dots,h_m')\in H_m^-$. Now observe that $\beta=\frac{1}{r+1-\beta}\Rightarrow\frac{1}{\beta}-1=r-\beta=\frac{1}{\xi}$. Then $y-1>\frac{1}{\beta}-1>\frac{1}{\xi}\Rightarrow 1-\frac{1}{y-1}>1-\xi$. Now if $h_m'\geq 3$, we have $(h_1',\dots,h_m')\in\hat{H}_m^-$ and $1-x\in B^-$. But if $h_m'=(\underbrace{2,\dots,2}_{s})$, we have $h_{m-1}'\geq 3$ and $\left\langle h_m';\frac{1}{1-1/(y-1)}\right\rangle$ is still greater than $1-\xi$, and again $1-x\in B^-$ (observe that $m>1$, otherwise $1-x=\left\langle h_m';\frac{1}{1-1/(y-1)}\right\rangle>1-\beta$). □

3.3.4 Proof of Theorem 3.14

First of all, we observe that Lemma 2 implies that $\overline{T}$ is one-to-one on D. In fact, suppose that $\overline{T}(x_1,y_1)=\overline{T}(x_2,y_2)$. Since $y_2\in[0,1]$, we must have $k(x_2)\in\{k(x_1)-1,k(x_1),k(x_1)+1\}$.

- If $k(x_1)=k(x_2)$ and $\operatorname{sign}(x_1)=\operatorname{sign}(x_2)$, then obviously $x_1=x_2,y_1=y_2$.
- If $k(x_1)=k(x_2)$ and $\operatorname{sign}(x_1)=-\operatorname{sign}(x_2)$, we find $y_1=-y_2$, which is possible only for $\{y_1=y_2=0\}$, a negligible set.
- Lastly, if $x_1>0,x_2<0$ and $k(x_2)=k(x_1)+1$, we get $y_2=1-y_1$. But $(x_1,y_1)\in D\Rightarrow y_1\in[0,\xi]\cup([\xi,1-\beta]\setminus B^+)\Rightarrow y_2\in B^-\cup(1-\xi,1)\Rightarrow(x_2,y_2)\notin D$. Thus $\overline{T}$ is one-to-one (mod 0).

Then we can write $\overline{T}(D\setminus B)=\overline{T}(D)\setminus\overline{T}(B)$. Now it is quite straightforward to check that $\overline{T}(D)=D$. In fact, recalling that

$$\begin{aligned}
\{x>0\mid k(x)=n\}&=\left(\frac{1}{n+\alpha},\frac{1}{n-1+\alpha}\right],\quad n>r\\
\{x>0\mid k(x)=r\}&=\left(\frac{1}{r+\alpha},\alpha\right],\quad n>r\\
\{x<0\mid k(x)=n\}&=\left[-\frac{1}{n-1+\alpha},-\frac{1}{n+\alpha}\right),\quad n>2\\
\{x<0\mid k(x)=2\}&=\left[\alpha-1,-\frac{1}{2+\alpha}\right)
\end{aligned}$$

we find

$$\begin{aligned}
&\overline{T}\left(\left[-\frac{i}{i+1},-\frac{(i-1)}{i}\right]\times\left(\left[0,M_0^{i-1}(1-\xi)\right]\setminus B^i\right)\right)=\\
&=\left[-\frac{(i-1)}{i},-\frac{(i-2)}{i-1}\right]\times\left(\left[\frac{1}{2},M_0^{i-2}(1-\xi)\right]\setminus B^{i-1}\right),\ i=2,\dots,r-1,\quad\text{(a)}
\end{aligned}$$

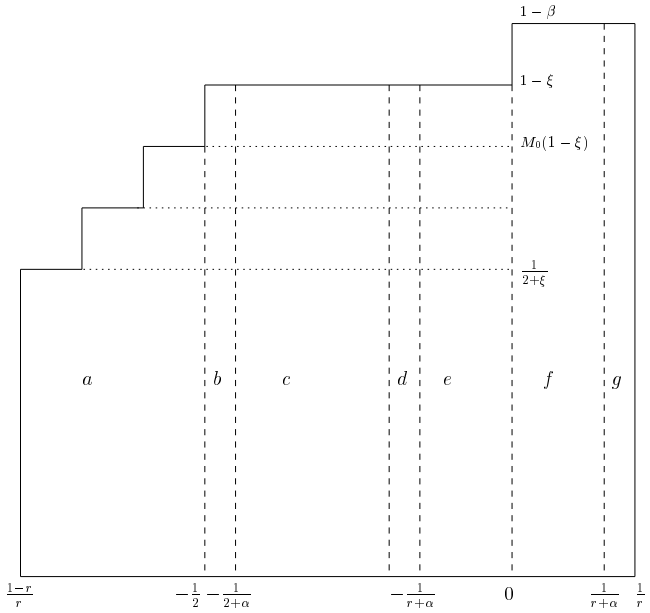

Figure 3.9: A simplified diagram showing the blocks (a)-(g) in the domain.

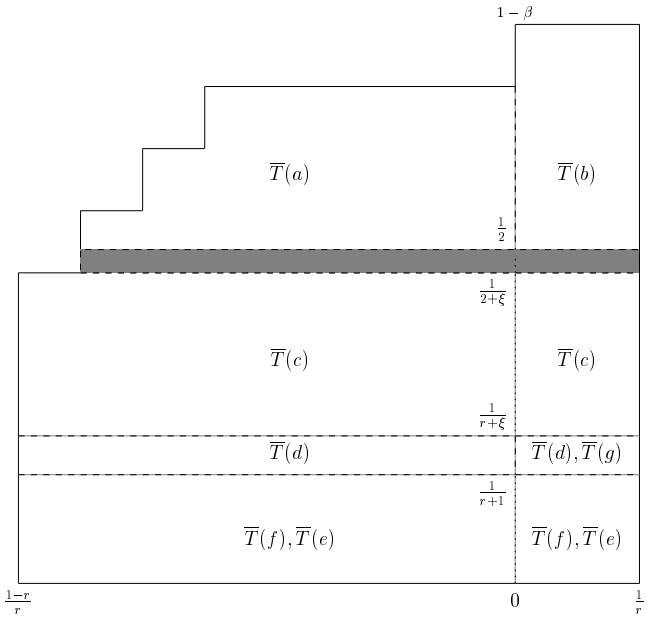

Figure 3.10: A simplified diagram showing the images with respect to $\overline{T}$ of the blocks (a)-(g).

$$\overline{T}\left(\left[-\frac{1}{2},-\frac{1}{2+\alpha}\right]\times\left([0,1-\xi]\setminus B^-\right)\right) = [0,\alpha]\times\left(\left[\frac{1}{2},1-\beta\right]\setminus B^+\right), \quad \text{(b)}$$

$$\overline{T}\left(\left[-\frac{1}{n-1+\alpha},-\frac{1}{n+\alpha}\right]\times\left([0,1-\xi]\setminus B^-\right)\right) = \\ = [\alpha-1,\alpha]\times\left(\left[\frac{1}{n},\frac{1}{n-1+\xi}\right]\setminus B^-\right), \quad n=3,\dots,r \quad \text{(c)}$$

Here we observe that $B^+\cup\left[\frac{1}{r},\frac{1}{2+\xi}\right] = B^-\cup\left[\frac{1}{r},\frac{1}{2+\xi}\right] = B^i\cup\left[\frac{1}{r},\frac{1}{2+\xi}\right]$ for $i=2,3,\dots,r-1$. Also remark that the rectangles $\left[-\frac{(r-2)}{r-1},\alpha\right]\times\left[\frac{1}{2+\xi},\frac{1}{2}\right]$ and $[\alpha-1,\alpha]\times\left[\frac{1}{n+\xi},\frac{1}{n}\right]$ for $n=3,\dots,r$ both belong to B.

$$\overline{T}\left(\left[-\frac{1}{r+\alpha},-\frac{1}{r+1+\alpha}\right]\times\left([0,1-\xi]\setminus B^-\right)\right) = \\ = \left([\alpha-1,0]\times\left(\left[\frac{1}{r+1},\frac{1}{r+\xi}\right]\cap D\right)\right)\cup \\ \cup\left([0,\alpha]\times\left(\left[\frac{1}{r+1},\frac{1}{r+\xi}\right]\setminus V_{r+1}(B^-)\right)\right) \quad \text{(d)}$$

(Here we wanted to highlight the fact that $B^+\cap\left[\frac{1}{r+1},\frac{1}{r+\xi}\right]=\varnothing$.)

$$\overline{T}\left(\left[-\frac{1}{n-1+\alpha},-\frac{1}{n+\alpha}\right]\times\left([0,1-\xi]\setminus B^-\right)\right) = \\ = [\alpha-1,\alpha]\times\left(\left[\frac{1}{n},\frac{1}{n-1+\xi}\right]\setminus V_n(B^-)\right), \quad n\geq r+2 \quad \text{(e)}$$

$$\overline{T}\left(\left[\frac{1}{n-1+\alpha},\frac{1}{n-2+\alpha}\right]\times\left([0,1-\beta]\setminus B^+\right)\right) = \\ = [\alpha-1,\alpha]\times\left(\left[\frac{1}{n-\beta},\frac{1}{n-1}\right]\setminus V^+_{n-1}(B^+)\right), \quad n\geq r+1 \quad \text{(f)}$$

where we set $V_n^+(x)=\frac{1}{n+x}$.

$$\overline{T}\left(\left[\frac{1}{r+\alpha},\alpha\right]\times\left([0,1-\beta]\setminus B^+\right)\right) = \\ = [0,\alpha]\times\left(\left[\frac{1}{r+1-\beta},\frac{1}{r}\right]\setminus V_r^+(B^+)\right) \quad \text{(g)}$$

To conclude the proof observe that

$$B^+\cup(1-B^-)=[\xi,1-\beta]\Rightarrow V_r^+(B^+)=\left[\frac{1}{r+1-\beta},\frac{1}{r+\xi}\right]\setminus V_{r+1}(B^-)$$

which together with (d) proves that $\overline{T}$ is onto.
The fact that $K(x,y)$ is invariant for $\overline{T}$ can be easily checked through the change of variables formula, as we have already seen in equation 3.4.

Remark 3.17. It is important to remark that it is still true that the boundary of D has measure 0. In fact, as we have already seen, except for the points y whose 0-continued fraction digits are all less or equal to $r+1$, which are a set of measure 0, all the other points in the vertical sections $D(x)$ belong to some cylinder of M_0 which is fully contained in $D(x)$.
Roughly speaking, for almost all points in $D(x)$ the "cutting process" described in §3.3.3 ends after a finite time.
Therefore the boundary of $D(x)$ is the union of a countable number of points, and has measure 0.

3.3.5 Invariant densities and entropy for $\alpha = \frac{1}{r}$

Since $\pi_1 \circ \overline{T}_\alpha = T_\alpha \circ \pi_1$, where π_1 is simply the projection on the first coordinate, the invariant density for T_α is obtained by integrating $K_\alpha(x,y)$ with respect to the second coordinate. Given a sequence $(h_1, h_2, \dots, h_n)$, define

$$a(h_1, h_2, \dots, h_n) = \frac{1}{\left\langle h_1, h_2, \dots, h_n; \frac{1}{1-\xi} \right\rangle} > 1$$
$$b(h_1, h_2, \dots, h_n) = \frac{1}{\langle h_1, h_2, \dots, h_n - 1 \rangle} > 1$$

Let

$$\psi_+(x) = \sum_{n=1}^{\infty} \sum_{(h_1,\dots,h_n)\in H^+} \left(\frac{1}{x + b(h_1,\dots,h_n)} - \frac{1}{x + a(h_1,\dots,h_n)} \right),$$
$$\psi_-(x) = \sum_{n=1}^{\infty} \sum_{(h_1,\dots,h_n)\in H^-} \left(\frac{1}{x + b(h_1,\dots,h_n)} - \frac{1}{x + a(h_1,\dots,h_n)} \right),$$
$$\psi_i(x) = \sum_{n=1}^{\infty} \sum_{(h_1,\dots,h_n)\in H^i} \left(\frac{1}{x + b(h_1,\dots,h_n)} - \frac{1}{x + a(h_1,\dots,h_n)} \right)$$

for $i = 2, \dots, r-1$, and observe that

$$\int_a^b \frac{1}{(1+xy)^2} dy = \frac{1}{x + \frac{1}{b}} - \frac{1}{x + \frac{1}{a}}$$

It follows that, for a suitable normalization constant c_α,

$$\rho_\alpha = \frac{\psi_\alpha(x)}{c_\alpha} = \frac{1}{c_\alpha} \left(\sum_{i=2}^{r-1} \left(\chi_{\left[-\frac{i}{i+1}, -\frac{(i-1)}{i}\right]}(x) \left(\frac{1}{x + \frac{1}{M_0^{i-1}(1-\xi)}} - \psi_i(x) \right) \right) + \right.$$
$$\left. + \chi_{\left[-\frac{1}{2}, 0\right]}(x) \left(\frac{1}{x + \frac{1}{1-\xi}} - \psi_-(x) \right) + \chi_{\left[0, \frac{1}{r}\right]}(x) \left(\frac{1}{x + \frac{1}{1-\beta}} - \psi_+(x) \right) \right)$$

is an invariant density for T_α.

Remark 3.18. Even in the case $\alpha = \frac{1}{r}$ the domain of the natural extension seems too complicated to allow for a direct computation of the entropy. However, as far as Corollary 2.9 is concerned, it is probably possible to prove a much

stronger result. In fact Nakada [16] showed that in the case of $\frac{1}{2} \leq \alpha \leq 1$, the integral $-2\int_{\alpha-1}^{\alpha} \log|x| \int_{D_\alpha(x)} K(x,y)dydx$ (where $D_\alpha(x)$ are the vertical sections of the domain of the natural extension) is constant[2] and equal to $\frac{\pi^2}{6}$. We conjecture that the same should be true for $0 < \alpha < \frac{1}{2}$.

Remark 3.19. One may ask whether the proof of Theorem 3.14 could be adapted to the general case of $\alpha \in (0, \sqrt{2}-1)$ with relatively small changes. We observe that our proof makes use of the fact that $T_{\frac{1}{r}}\left(\frac{1}{r}\right) = 0 = T_{\frac{1}{r}}^{r-1}\left(\frac{1}{r}-1\right)$. Also in the case $\alpha \in [\sqrt{2}-1, 1]$, as shown in [16] and [6], the construction of the Natural Extension depends on the fact that the tails of the α-expansions of α and $\alpha-1$ coincide after one or two iterations (more precisely, $T_\alpha^2(\alpha) = T_\alpha(\alpha-1)$ when $\alpha \in [g,1]$, and $T_\alpha^2(\alpha) = T_\alpha^2(\alpha-1)$ when $\alpha \in [\sqrt{2}-1, g]$).
In the general case, one would need an explicit relation between the α-expansions of α and of $\alpha-1$, which at present is not known.

[2] The theory of *S-expansions* provides an explanation of this surprising fact, see [11].

Part II

Coding for wireless channels

Introduction

Recently, the diffusion of wireless networks has led to the development of new coding schemes in order to improve performances on fading channels; algebraic number theory has proven to be an effective tool for their design.
Wireless transmission introduces new problems with respect to the classical model of the Additive White Gaussian Noise (AWGN) channel: in fact the electromagnetic signal, propagating along multiple paths, is affected by attenuations, delays and frequency shifts, collectively called "fading", that make this channel much less reliable than the AWGN. The most effective strategy to counterbalance fading is to introduce *diversity* in the transmission, that is to send the same information through multiple independent channels. There are several ways to increase diversity:

- *in space*, by receiving the same signal through multiple antennas, that must be sufficiently spaced to ensure that the fadings on the different paths are uncorrelated;
- *in time*, by receiving the same signal at sufficiently long time delays;
- *in frequency*, by transmitting the same signal over different frequencies.

However, the second method has the drawback of introducing heavy delays in the communication, while the third entails a waste of the available bandwidth. The use of multiple antennas both at the transmitter and at the receiver (*Multiple Input, Multiple Output* or MIMO) allows for a potential diversity of MN, where M is the number of transmit antennas and N is the number of receive antennas.
In general, the implementation of coding for wireless channels must take into account the actual availability of resources (bandwidth, power, cost of the appliances) and answer three basic and often conflicting needs:

- increasing the *rate* of transmission,
- increasing *diversity*,
- keeping a low *decoding complexity*.

In the MIMO setting, the information vector u, belonging to a finite signal subset or "constellation" S, is encoded in a *space-time block*, that is an $M \times T$ matrix $B(u)$, where M is the number of transmit antennas and T is the duration of the signal.
In this context the fundamental parameters to assess the system performance are the *diversity gain* $\min_{u \neq u'} (\mathrm{rk}\, A(u,u'))$, and the *coding gain*

$$\frac{1}{E_S} \min_{u \neq u'} (\det A(u,u'))^{\frac{1}{M}},$$

where $A(u,u') = (B(u) - B(u'))(B(u) - B(u'))^H$, $u, u' \in S$, and E_S is the average energy of the constellation.

In 2002 Belfiore, Damen and Tewfik [10] proposed a 2×2 *full rate* code which guarantees maximum diversity, and their method can be extended to the higher-dimensional case. Its major drawback, however, is that the coding gain vanishes when the size of the constellation grows to infinity.

In the 2-dimensional case, this problem has been solved by Belfiore, Rekaya and Viterbo [5] with the *Golden Code* $\mathcal{G}$, a *full-rate*, *full-rank* code whose minimum determinant does not vanish when the size of the constellation tends to infinity. This design is based on a principal ideal $A\mathcal{O}$ of a maximal order $\mathcal{O}$ in a quaternion algebra $\mathcal{A}$ of matrices over $\mathbb{Q}(i)$, containing as a maximal subfield the number field $K = \mathbb{Q}(i, \theta)$, where θ is the Golden number.[3]
Each information vector $u \in \mathbb{Z}[i]^4$ can be mapped to a matrix $AB(u)$ in $A\mathcal{O}$. $\mathcal{A}$ turns out to be a division ring, and the determinant of $AB(u)$ is nothing but its reduced norm, so that it is a nonzero Gaussian integer, modulo a normalization constant. Therefore the minimum determinant is bounded from below by a fixed constant $\delta_{\min} = \frac{1}{5}$, for any size of the constellation $S \subset \mathbb{Z}[i]^4$.
It can be shown that this is the best performance one can obtain from this kind of construction using the field $\mathbb{Q}(i, \sqrt{p})$, where p is a prime number, $p \equiv 5 \pmod 8$.
Moreover, the design guarantees *cubic shaping*, which is convenient both for energy efficiency and fast decoding: when vectorized, $A\mathcal{O}$ is a rotated version of the lattice $\mathbb{Z}[i]^4$, and allows for effective decoding using the *Sphere Decoder* and the *Viterbo-Boutros algorithm*.

It is possible to obtain a further increase in the coding gain by using a suitable subcode of the Golden Code, in the choice of which decoding complexity must be taken into account.
Belfiore, Hong and Viterbo [4] have recently described a chain of nested left ideals of the form $\mathcal{G}_k = \mathcal{G}B^k$, $1 \leq k \leq 4$, such that the index $[\mathcal{G}_k : \mathcal{G}_{k+1}]$ between two successive subcodes is 4, and the minimum determinant in $\mathcal{G}_k$ is $2^k\delta_{\min}$.
Moreover, each ideal $\mathcal{G}_k$ is isometric to a well-known lattice; in particular, $\mathcal{G}_2$ is isometric to the *Gosset lattice* $\mathbb{E}_8$.
A more general problem consists in building a block code $X = (X_1, \ldots, X_L)$, where each component X_i is a Golden codeword. Choosing the $X_i \in \mathcal{G}_k$ independently, we obtain a very simple block code. For small sizes of the signal constellation these subcodes already yield a performance gain with respect to the "uncoded" Golden Code (that is, with respect to choosing $X_i \in \mathcal{G}$ independently). However, this gain is cancelled out asymptotically by the loss of rate as the size of the signal set grows to infinity [4], since an energy increase is required to mantain the same spectral efficiency, or bit-rate per channel use.
A better performance is achieved when the X_i are not chosen in an independent fashion. For example it is possible to exploit the hierarchic structure of the partition chain $\{\mathcal{G}_k\}$ previously described by combining two encoders: a *trellis encoder* which outputs the cosets of $\mathcal{G}_k/\mathcal{G}_l$, $1 \leq k < l \leq 4$, and a lattice encoder for $\mathcal{G}_l$ (*Trellis Coded Modulation*). The Viterbi algorithm (*soft decoding*) can be employed for the trellis decoding, in association with a Sphere Decoder in each coset.

In the case of block codes, the coding gain is a power of

$$\Delta = \min_{X \neq 0} \det \left(\sum_{i=1}^{L} X_i X_i^H \right) \geq \min_{X \neq 0} \sum_{i=1}^{L} \det \left(X_i X_i^H \right) = \Delta'$$

[3] Following the notation in [5], from now on we will denote the Golden number $\frac{1+\sqrt{5}}{2}$ by the letter θ.

The expression Δ is difficult to handle because it contains mixed terms of the form $\left\| \widetilde{X}_i X_j \right\|_F^2$, where $X \mapsto \widetilde{X}$ is an involution of $\mathcal{G}$, and $\|\cdot\|_F$ is the Frobenius norm. The codes in [4] are designed to maximize the approximate parameter Δ' and so a priori they might be suboptimal; in §6.4.1 we treat the mixed terms.
In §6.5 we describe some simple block codes designs for $L = 2, 3$ that are lifts of linear codes over the quotient group $\mathcal{G}/\mathcal{G}_1$. We describe the lattice structure of $\mathcal{G}_1$ and compute the minima of the Frobenius norms of the products over all pairs of cosets; we use this information to select the codes with the best weight enumerator polynomials in dimension 2 and 3.

When using ideals of $\mathcal{G}$ to build block codes, it is preferable to:

- choose ideals whose index is a power of two, in order to have a variety of simple binary *set partitioning* schemes available;
- choose *two-sided* ideals, so that the quotient has a nice ring structure.

The ideals in [4] satisfied the first condition but not the second; in §6.6.2, we show that (unfortunately) the only two-sided ideals of $\mathcal{G}$ whose index is a power of two are the trivial ones, that is, the ideals of the form $c\mathcal{G}$ with $c \in \mathbb{Z}[i]$, and $|c|^2$ a power of two.
In particular, we study the quotient rings $\mathcal{G}/(1+i)\mathcal{G}$ and $\mathcal{G}/2\mathcal{G}$ which turn out to be isomorphic to the rings of 2×2 matrices over $\mathbb{F}_2$ and $\mathbb{F}_2[i]$ respectively. Unfortunately, only a very sparse literature is available on the subject of codes over non-commutative rings, especially as far as efficient decoding algorithms are concerned, and for the time being we have been unable to exploit the ring structure directly for code construction, except in the simple case of the repetition code over the cosets of $(1+i)\mathcal{G}$. This basic construction provides a first application of the criteria based on the estimate of the mixed terms, and our performance simulations show that it can lead to up to 2.9 dB of gain with respect to the "uncoded" case.
However, it is still possible to take advantage of the structure of $\mathbb{F}_2$-module of the quotient; from the additive point of view, the quotient $\mathcal{G}/2\mathcal{G}$ is indistinguishable from $\mathbb{F}_{256}$, for which a wide variety of error-correcting codes are available. In §6.8, we combine a shortened Reed-Solomon code with the encoder of the quotient ring to increase the minimum Hamming distance of the code. The main advantage with respect to trellis codes is the relative ease of decoding.
Simulation results show that using 4-QAM constellations, that is using only one lattice point per coset, we obtain a gain of up to 6 dB with respect to the uncoded Golden Code at the same spectral efficiency, under the hypothesis that the channel remains constant for the entire length of the block. This assumption corresponds to the *slow fading* case, and may be considered realistic when the block length does not exceed one hundred, leaving space for further improvements. This construction can be extended to the 16-QAM case, yielding a gain of up to 3.8 dB.

Chapter 4

Coding for wireless channels

4.1 The wireless channel model

4.1.1 The transmitter

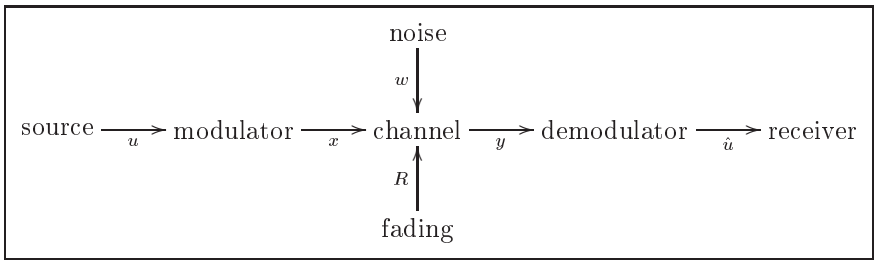

Figure 4.1: The various stages of transmission.

In the typical model, the *source* emits a binary vector $u \in \{0,1\}^m$. This information is coded in such a way as to be optimized for transmission:

- first, the *source encoder* compresses the original message, removing any redundancies;
- afterwards, the *channel encoder* embeds some redundancy in the message in order to protect the information from the errors that may arise during the transmission.

In the following paragraphs, we will not take the source encoder into account.

4.1.2 The modulator

The *digital modulator* associates an analog waveform x to the coded information vector. In *amplitude modulation*, it is the amplitude of the wave which carries the information, while in *frequency modulation*, information is stored in the instantaneous frequencies of the wave.
First, the binary message is mapped into a point of a real or complex lattice Λ.

The set of all possible emitted waveforms is a finite subset $\mathcal{S} \subset \Lambda$, of cardinality 2^m, called a *signal constellation*. The components of the lattice vector are used to modulate a linear combination of basis waveforms.
In most cases it is best to choose the mapping $\{0,1\}^m \to \mathcal{S}$ in such a way that nearest binary neighbors (with respect to the Hamming distance) correspond to nearest neighbors in the lattice (with respect to the Euclidean distance). In particular, the mapping is called a *Gray mapping* if the labellings of every pair of adjacent lattice points have Hamming distance 1.
In the following chapters we will always consider *Quadrature Amplitude Modulation* or QAM; with this type of modulation, data is conveyed by changing the amplitude of two orthogonal carrier waves. The amplitudes of the first and second wave are respectively called the "in phase" component and the "quadrature" component.
Figure 4.2 shows the representation of some widely used QAM constellations in the in phase-quadrature plane. We can identify these constellations with subsets of a $\mathbb{Z}[i]$-lattice shifted by $\frac{1}{2} + i\frac{1}{2}$. The minimum Euclidean distance between two points will always be 1. We also report the average energy for each case, which will be useful in the sequel:

$$\begin{array}{ll} E_{4-QAM} = 0.5, & E_{8-QAM} = 1.5, \\ E_{16-QAM} = 2.5, & E_{32-QAM} = 5 \end{array} \tag{4.1}$$

The number of information bits transmitted for each channel use is called the *spectral efficiency* of the modulation scheme, and is measured in bpcu (bits per channel use).
Clearly, increasing the cardinality or "size" of the constellation allows for a growth in spectral efficiency, but corresponds to a greater average energy (or equivalently, to a decrease in the minimum Euclidean distance).

4.1.3 The channel

In the AWGN case, across the channel the signal x is perturbed by a random noise w, so that the received signal is $y = x + w$. With wireless transmission, further complications appear compared to the AWGN model: the signal (an electromagnetic wave) propagates along multiple paths, with reflection, refraction and scattering effects due to the presence of massive obstacles like houses or mountains. In the case of mobile telephony, moreover, other relevant perturbations include the Doppler effect due to the relative motion between the transmitter and the receiver, and the decrease in signal power when the distance from the transmitter antenna increases.
Thus the received signal is distorted not only by the noise, but also by attenuations, delays and frequency shifts, collectively denoted by the term *fading*.
Since the fading effects depend on the particular environment in which transmission takes place, it is virtually impossible to determine their sum. In practice, a good working approximation consists in assuming that the very great number of propagation parameters can be modelled as independent random variables so that the Central Limit Theorem holds.
In the case of a modulated signal, in general the different frequencies composing the signal are subject to independent fadings and phase shifts (*frequency selective fading*); we only consider the case where the bandwidth range is narrow

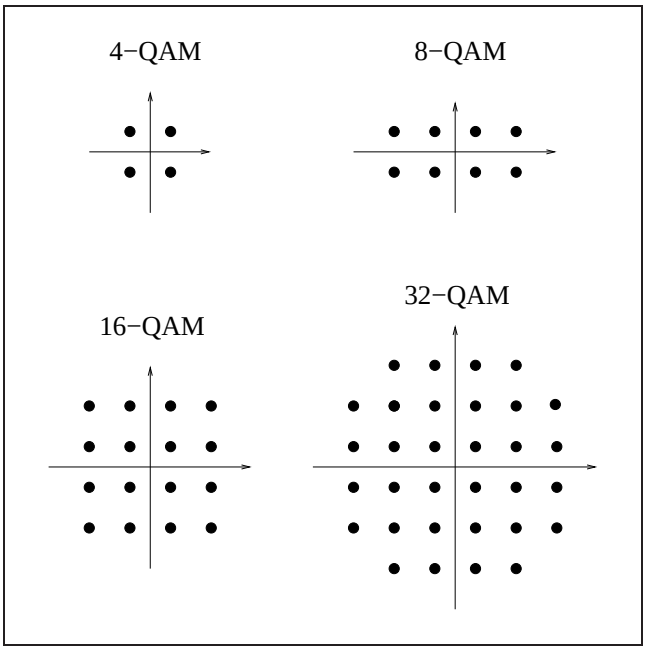

Figure 4.2: Some QAM (Quadrature Amplitude Modulation) constellations.

so that we can assume that the fading acts multiplicatively. In this case the received signal y, seen as a point in $\mathbb{R}^n$, is given by

$$y = Hx + w \tag{4.2}$$

where w is a complex Gaussian random variable of zero mean and variance $\frac{N_0}{2}$, and H is a random diagonal matrix whose entries α_i are independent random variables with *Rayleigh* density[1]:

$$p(r) = \begin{cases} \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} & 0 \leq r < \infty \\ 0 & r < 0 \end{cases} \tag{4.3}$$

We also assume $\mathbb{E}[\alpha_i^2] = 1 \; \forall i$, so that $\sigma_i^2 = \frac{1}{2}$, and $p(\alpha_i) = 2\alpha_i e^{-\alpha_i^2}$.
The assumption that the α_i are independent is reasonable when using an *interleaver*, which permutes the codeword components so that fadings on adjacent components can be considered uncorrelated.
Since we have supposed the fading matrix to be normalized, we can define the *signal-to-noise ratio* or SNR ζ as the ratio between the average signal power (see §4.1.2) and the average noise power: $\zeta = 2\frac{E_S}{N_0}$.

4.1.4 Decoding

Now we suppose that the receiver knows the fading coefficients α_i (*perfect Channel State Information* or CSI). In fact, when the fading varies slowly compared

[1] The Rayleigh density is the distribution of the modulus of a complex Gaussian random variable.

to the duration of the signal, the information signal can be preceded by a sequence of *pilot symbols*, allowing the receiver to estimate the fading parameters. The *demodulator* must recover from y an estimate $\hat{x} \in \mathcal{S}$ of the original signal. Figure 4.3 provides an intuitive explanation of the fact that rotated lattices

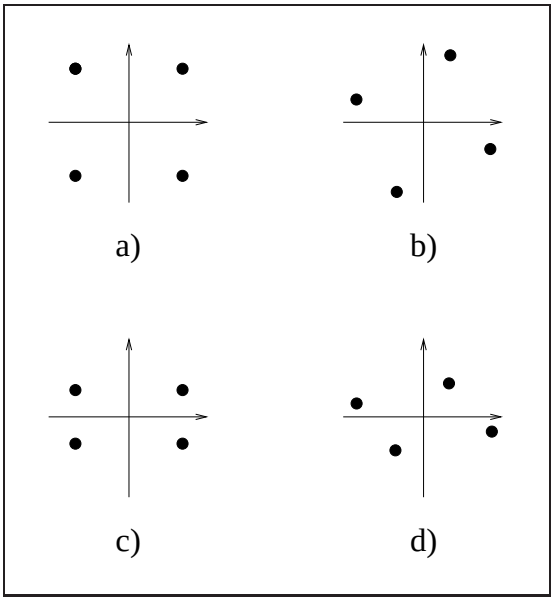

Figure 4.3: An example of the effect of rotation on the resistance to fading: figures a)-c) show a 4-QAM constellation before and after fading, figures b)-d) show the rotated case. In the first situation, with strong fading on the second coordinate, lattice points can be easily confused. In the rotated case, the lattice points are easier to tell apart.

work better to contrast fading in two dimensions. In fact, rotating corresponds to increasing the "algebraic dimension" of the lattice but entails no extra cost. When the dimension of the lattice increases, the performance can approach that of a Gaussian channel without fading; at the same time, however, decoding becomes more complex.
The points of $\mathcal{S}$ must be well-spaced in order to guarantee a low error probability in the decoding; on the other side, the square norm $\|x\|^2$ represents the power of the transmitted signal, and so the diameter of the constellation $\mathcal{S}$ should be kept as small as possible. In theory, the most advantageous shape for $\mathcal{S}$ would be a sphere; however, labelling the lattice points contained in a sphere turns out to be too costly in terms of computations, and usually a *cubic shape* is preferred. Thus, constellations carved from rotated $\mathbb{Z}^n$ lattices are a more desirable choice than constellations coming from skew lattices.
Since the coefficients α_i are known *a posteriori*, the received signal components $y_i = \alpha_i x_i + w_i$ can be modelled as Gaussian random variables $\mathcal{N}\left(\alpha_i x_i, \left(\frac{N_0}{2}\right)^2\right)$. The Maximum Likelihood criterion (*ML decoding*) coincides with the minimum

Euclidean distance:

$$\hat{x}(y) = \underset{x\in\mathcal{S}}{\operatorname{argmax}}\, P(y|x) = \underset{x\in\mathcal{S}}{\operatorname{argmax}} \frac{1}{\left(\sqrt{2\pi}\frac{N_0}{2}\right)^n} e^{-\sum_{j=1}^{n} \frac{2(y_j-\alpha_j x_j)^2}{N_0^2}} =$$

$$= \underset{x\in\mathcal{S}}{\operatorname{argmin}} \sum_{j=1}^{n} (y_j - \alpha_j x_j)^2 = \underset{z\in H(\mathcal{S})}{\operatorname{argmin}} \sum_{j=1}^{n} (y_j - z_j)^2 \quad (4.4)$$

4.1.5 The Sphere Decoder

The search for the minimum in equation (4.4), that is the search for the closest lattice point to a given received point in the deformed lattice $H(\mathcal{S})$, becomes too costly in terms of computation time when $\#\mathcal{S}$ increases. For an efficient and fast decoding, the *Sphere Decoder* may be used [18].
This algorithm exploits the fact that, when the dimension of the space grows, the number of lattice points contained in a sphere becomes much smaller than the number of points inside a cube of the same radius.
We consider here the case when $\Lambda = M\mathbb{Z}^n$ is a real lattice with generator matrix M. In the complex case $\Lambda = M\mathbb{Z}[i]^n$, we can simply separate the real and imaginary parts, obtaining a real lattice of dimension $2n$.
The decoder examines only the points found inside a sphere of radius $\sqrt{C}$ and centered in the received point y. Let $\Lambda' = y - H\Lambda$ be the deformed lattice with the origin translated to y:

$$\min_{z\in H\Lambda} \|y - z\|^2 = \min_{w\in\Lambda'} \|w\|^2$$

We have $x = Mv \in \Lambda$ for some $v \in \mathbb{Z}^n$, $y = HM\rho$ for some $\rho \in \mathbb{R}^n$, $w = HM(\rho - v) = HM\xi$. Then

$$\|w\|^2 = w^H w = \xi^H G \xi = \sum_{i=1}^{n} g_{ij}\xi_i\xi_j \le C$$

where G is the Gram matrix associated to the lattice generator matrix HM. The set $\{\sum_{i=1}^{n} g_{ij}\xi_i\xi_j = C\}$ is an ellipsoid in the ξ coordinates.
Applying Cholesky's factorization, we can write $G = R^H R$ with R upper triangular. Then

$$\xi^H G \xi = \|R\xi\|^2 = \sum_{i=1}^{n} \left(r_{ii} + \sum_{j=i+1}^{n} r_{ij}\xi_j \right)^2 = r_{ii}^2 \left(\xi_i + \sum_{i=1}^{n} \frac{r_{ij}\xi_j}{r_{ii}} \right)^2 \le C$$

Choosing new coordinates $\nu_i = \xi_i + \sum_{j=i+1}^{n} \frac{r_{ij}}{r_{ii}}$, we find the equation of an n-dimensional ellipsoid centered in the origin whose axes are the carthesian axes. The search is then conducted inside the boundaries of this ellipsoid [18].

When a lattice point is found inside, its distance to the center is compared to $\sqrt{C}$; if it is smaller, the search radius is updated. If C is too small and no lattice points are found inside the ellipsoid, an *erasure* is declared and the search is renewed with a bigger radius. Thus it is very important to get a good estimate of C from the beginning. In general the choice of the initial radius $\sqrt{C}$ is based

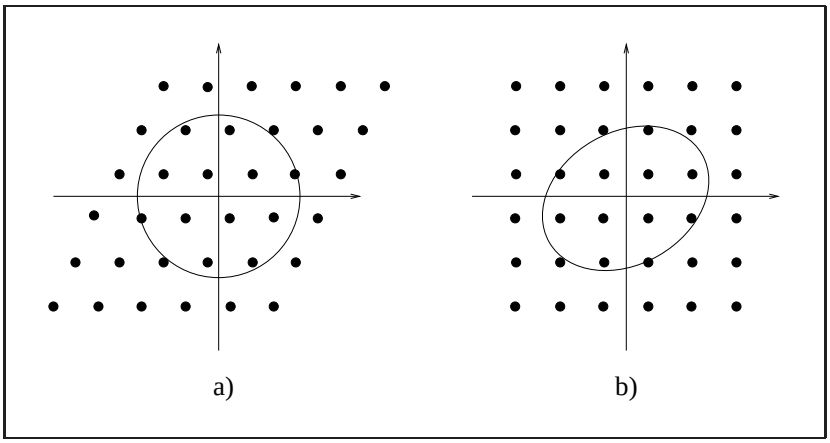

Figure 4.4: The problem of enumerating the points of a skew lattice that fall inside a sphere of given radius can be reduced to the problem of finding the points of a $\mathbb{Z}^n$ lattice inside an ellipsoid.

on the SNR. In practice, a good choice for C might be the smallest of the fading coefficients α_i, which corresponds to the shortest of the axes of the ellipsoid. One of the drawbacks of the Sphere Decoder algorithm in the presence of fading is that H, and thus the lattice, vary with each received symbol, so that a new Cholesky factorization has to be performed each time.

4.1.6 The union bound estimate

The problem of choosing a constellation S that minimizes the error probability with respect to ML decoding is extremely complex. In general, the classical *union bound* estimate holds, since S is a finite subset of the linear lattice Λ:

$$P_e(S) \leq P_e(\Lambda) \leq \sum_{x' \neq x} P(x \to x') \tag{4.5}$$

In the previous equation, the *pairwise error probability* (PEP) $P(x \to x')$ denotes the probability that y is closer to x' than to x with respect to the Euclidean metrics. The conditioned probability with respect to the channel yields [6]:

$$\begin{aligned} P(x \to x' \,|\, H) &= P\left(\sum_{i=1}^{n} \left|y_i - \alpha_i (x_i')^2\right| \leq \sum_{i=1}^{n} \left|y_i - \alpha_i x_i^2\right|\right) = \\ &= P\left(\sum_{i=1}^{n} \left|\alpha_i (x_i - x_i') + w_i\right|^2 \leq \sum_{i=1}^{n} |w_i|^2\right) = \\ &= P\left(\sum_{i=1}^{n} \alpha_i^2 (x_i - x_i')^2 + 2\sum_{i=1}^{n} \alpha_i (x_i - x_i') w_i \leq 0\right) \end{aligned}$$

Consider the random variable $\chi = \sum_{i=1}^{n} \alpha_i (x_i - x_i') w_i$: χ is a linear combination of the Gaussian random variables $w_i \sim \mathcal{N}(0, \frac{N_0}{2})$, so it is also Gaussian with 0 mean and variance $\sigma^2 = \frac{N_0}{2} \sum_{i=1}^{n} \alpha_i^2 (x_i - x_i')^2$. The PEP can thus be written

as

$$P(x \to x' \mid H) = P\left(\chi \geq \frac{1}{2}\sum_{i=1}^{n} \alpha_i^2 (x_i - x_i')^2\right)$$

Let A be the known quantity $\frac{1}{2}\sum_{i=1}^{n} \alpha_i^2 (x_i - x_i')^2$. Then

$$P(\chi \geq A) = \int_A^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}} dt = \int_{\frac{A}{\sigma^2}}^\infty \frac{1}{\sqrt{2\pi}} e^{\frac{-s^2}{2}} ds$$

Recalling the Gaussian tail function estimate

$$\frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-\frac{t^2}{2}} dt \leq \frac{1}{2} e^{-\frac{a^2}{2}}$$

we obtain

$$P(x \to x' \mid H) \leq \frac{1}{2} e^{-\frac{A}{2\sigma^2}} = \frac{1}{2} e^{-\frac{\left(\sum_{i=1}^n \alpha_i^2 (x_i - x_i')^2\right)^2}{4N_0 \sum_{i=1}^n \alpha_i^2 (x_i - x_i')^2}} = \frac{1}{2} e^{-\frac{1}{4N_0}\sum_{i=1}^n \alpha_i (x_i - x_i')^2}$$

Notice that in the last expression we can omit the sum over the indices i such that $x_i = x_i'$. In order to obtain the unconditioned PEP, we average over the fading coefficients α_i:

$$P(x \to x') = \int P(x \to x' \mid H) p(\alpha_1) \cdots p(\alpha_n) d\alpha_1 \cdots d\alpha_n \leq$$
$$\leq \frac{1}{2} e^{-\frac{1}{4N_0}\sum_{i=1}^n \alpha_i^2 (x_i - x_i')^2} p(\alpha_1) \cdots p(\alpha_n) d\alpha_1 \cdots d\alpha_n$$

Recalling that the α_i are assumed to be independent and Rayleigh distributed (see equation (4.3)) with $\mathbb{E}[\alpha_i^2] = 1$, so that $p(\alpha_i) = 2\alpha_i e^{-\alpha_i^2}$, we obtain

$$P(x \to x') \leq \prod_{i=1}^n \left(\int_0^\infty e^{-\frac{\alpha_i^2 (x_i - x_i')^2}{4N_0}} \alpha_i e^{-\alpha_i^2} d\alpha_i\right) =$$
$$= \prod_{i=1}^n \left(\int_0^\infty \alpha_i e^{-\alpha_i^2\left(1 + \frac{(x_i - x_i')^2}{4N_0}\right)} d\alpha_i\right) = \prod_{i=1}^n \frac{1}{2\left(1 + \frac{(x_i - x_i')}{4N_0}\right)}$$

If the signal-to-noise ratio is big enough, we can write

$$P(x \to x') \leq \frac{1}{2} \prod_{x_i \neq x_i'} \frac{4N_0}{(x_i - x_i')^2}$$

Introducing the auxiliary functions

$$l(x, x') = \#\{i \mid x_i \neq x_i'\}, \qquad d_P(x, x') = \prod_{x_i \neq x'_i} |x_i - x_i'|,$$

we obtain

$$P(x \to x') \leq \frac{1}{2} \prod_{x_i = x_i'} \frac{4N_0}{(x_i - x_i')^2} = \frac{1}{2} \frac{(4N_0)^{l(x,x')}}{(d_P(x, x'))^2} \tag{4.6}$$

The dominant term in the union bound sum (4.5) is the one which achieves the minimum L of $l(x, x')$, and is called the *diversity* of the system. The goal is then to maximize L and, for maximum L, the *product distance* d_P.
Clearly in general the error probability $P_e(S)$ increases when the size of the constellation S increases.

4.2 Multiple antenna systems

Thanks to the increased diversity, the use of multiple antennas both at the transmitter and at the receiver can make for a substantial gain in the reliability of wireless communications. The main objective when designing MIMO schemes is to achieve a tradeoff between transmission rate and diversity. At the opposite ends of the spectrum, one can place the following behaviors:

- sending independent signals through each antenna simultaneously, getting the maximum rate gain and no diversity gain;
- sending the same signal from all antennas, obtaining the maximum diversity gain and no rate gain.

The maximum rate that can be attained using M transmit antennas and N receive antennas is $\min\{M, N\}$, while the maximum diversity is MN (corresponding to the number of independent paths between transmit and receive antennas).
As we will see, in some special cases, depending on the number of transmit and receive antennas, it is possible to achieve both ends, full rate and full diversity, without any loss, using tools from number theory.

4.2.1 Error probability and determinant criterion

We consider a system with M transmit antennas and N receive antennas. The information vector $u = (u_1, \dots, u_k)$ belonging to a constellation $\mathcal{C} \subset \mathbb{Z}[i]^k$ is coded into an $M \times T$ matrix, $B(u) = (b_{mt})$, whose entries are in another constellation $\mathcal{S}$. b_{mt} is the symbol transmitted by the m-th antenna at the time $t \in \{1, \dots, T\}$. $B(u)$ is called a *space-time block*. The signal received by the antenna n at the time t is

$$y_{nt} = \sum_{m=1}^{M} h_{nm} b_{mt} + w_{nt}, \tag{4.7}$$

Here we suppose that the fading coefficients h_{nm} between transmit antenna m and receive antenna n stay constant for a time T (*quasi-static fading*).
This assumption is realistic if the duration T of the block is smaller than the *coherence time* of the channel, that is the time length for which the effects of the Doppler shift on the channel can be ignored. If T is large, we say that the channel is *slow fading*; this behavior may be caused by large obstructions between transmitter and receiver.
For most practical applications, it has been estimated [2] that the coherence time is greater than 0.01 seconds, so that $T < 200$ is a legitimate assumption.
Moreover, we suppose that there is perfect CSI at the receiver (*coherent case*).
As in the single antenna case, it is possible to estimate the error probability for a pair of space-time blocks $P_e(B(u) \to B(u'))$, and it turns out that

$$P_e(B(u) \to B(u')) \leq \frac{1}{\left((\prod_{i=1}^{r} \lambda_i)^{\frac{1}{r}} \frac{E_{\mathcal{S}}}{N_0} \right)^{rN}}, \tag{4.8}$$

where r is the rank of the matrix $B(u, u') = (B(u) - B(u'))(B(u) - B(u'))^H$, λ_i are its nonzero eigenvalues, and $E_{\mathcal{S}}$ is the average energy of the constellation

used [24]. In order to minimize the error probability, then, the following criteria must be adopted:

a) Maximize $\min_{u \neq u'} r = \min_{u \neq u'} \mathrm{rk}(B(u) - B(u'))$, called *diversity gain* ;

b) Maximize $\Delta = \min_{u \neq u'} (\prod_{i=1}^{n} \lambda_i)^{\frac{1}{r}}$, called *coding gain*.

Using the Gaussian tail estimate (4.8), we can compare two different coding schemes at the same spectral efficiency by considering the ratio

$$\gamma_{\mathrm{as}} = \frac{(\Delta_1)^{\frac{1}{r}} / E_{\mathcal{S},1}}{(\Delta_2)^{\frac{1}{r}} / E_{\mathcal{S},2}} \tag{4.9}$$

This parameter, called *asymptotic coding gain*, yields a good estimate for high SNR. When this ratio is favourable, the same word error probability can be achieved at a lower SNR; this gain in SNR, measured in decibels, is equal to $10 \log_{10} \gamma_{\mathrm{as}}$.

4.2.2 Channel capacity

In this section we briefly recall a few notions from information theory.
According to Shannon's Channel Coding Theorem, each transmission channel H with noise admits a *cutoff rate* or *capacity* $C(H)$, that is a limit rate under which reliable communication is possible. By *reliable communication* for a certain rate R, we mean that the word error probability for a sequence of random codes of length n and rate R in the Shannon ensemble goes to 0 as the code length n tends to infinity.
Conversely, if the rate is greater than the cutoff rate, Shannon's Theorem tells us that *any* code will have a positive error probability.
Shannon's result also implies that, roughly speaking, a code chosen at random is very likely to be good; however, in practice random codes are not a good solution because they would be too difficult to decode, and it is preferable to focus on the problem of designing deterministic codes that come as close as possible to the cutoff rate.
For a memoryless channel, the capacity is given by

$$C(H) = \max_{X} I(X, Y),$$

where $I(X, Y)$ is the mutual information between the input X and the output Y, and the maximum is taken over all the probability distributions of X.
When evaluating the capacity of multiple-antenna systems over fading channels, it is important to distinguish between the *high SNR* regime and the *low SNR regime*.
In the low SNR case, the most important parameter to consider is the *diversity advantage*; we say that the diversity advantage of the system is d if the average error probability decays like $\frac{1}{\mathrm{SNR}^d}$.
If on the contrary the signal-to-noise ratio is high, the "number of degrees of freedom" or independent fading paths available for transmission plays a more important role. We say that a scheme has *spatial multiplexing gain* r if the capacity of the channel is approximately $r \log(\mathrm{SNR})$. It has been shown by Foschini [13] that the maximum spatial multiplexing gain attainable is $\min(M, N)$,

where M is the number of transmit antennas and N is the number of receive antennas.
Zheng and Tse [26] proved that there is a *fundamental tradeoff* between diversity and spatial multiplexing; in fact for a block length $L \geq M + N - 1$, the optimal diversity $d(r)$ is given by $d(r) = (M - r)(N - r)$. Intuitively, if r transmit and receive antennas are used to increase multiplexing, only the remaining $M - r$ transmit and $N - r$ receive antennas can provide diversity.
Achieving the diversity-multiplexing tradeoff is the key for optimizing transmission for both high and low SNR regimes.

Remark 4.1. The above discussion concerns channels for which the *ergodicity* assumption holds, that is channels for which the fading components can be regarded as uncorrelated. In the non-ergodic case, which includes the slow fading model, the definition of the capacity of the channel is more problematic.
Since there can be no time averaging for fading over long codewords, for any given rate it might happen that the capacity of the channel at a certain time doesn't support that rate, and Shannon's theorem is no longer valid. A more fruitful approach consists in considering the capacity itself as a random variable, depending on the instantaneous mutual information. We say that an *information outage* occurs if the transmission rate exceeds the instantaneous capacity. Thus instead of the capacity it is more useful to consider the *outage probability* for a given spectral efficiency.

Chapter 5

Space-time codes and continued fractions

5.1 Diagonal Space-Time Codes (DAST)

One coding scheme whose focus is on achieving maximum diversity from multiple antenna systems, without increasing at all the capacity, is *diagonal space-time* coding (DAST for short).
In this case, regardless of the number of receive antennas, the capacity is that of an $M \times 1$ system; the role of the transmit antennas is that of providing independent fading paths.

5.1.1 Diagonal space-time codes and continued fractions

An interesting relation between MIMO codes and continued fractions is presented in [21].
We consider a system with 2 transmit and receive antennas. Each codeword is a block of two 2×2 diagonal matrices belonging to the finite set

$$\mathcal{V} = \left\{ V_l = \begin{pmatrix} \eta^l & 0 \\ 0 & \eta^{ul} \end{pmatrix}, \quad l \in 1, \ldots, L \right\}$$

where $\eta = e^{\frac{2\pi i}{L}}$ is a primitive L-th root of unity, and u is a suitable integer. If the data stream to transmit is the sequence $\{l_1, l_2, \ldots\}$, $l_i \in \{1, \ldots, L\} \forall i$, the transmitted blocks $B_1, B_2, B_3, \ldots$ are defined as follows:

$$B_1 = \begin{pmatrix} I \\ V_{l_1} \end{pmatrix}, \quad B_i = \begin{pmatrix} \prod_{k=1}^{i-1} V_{l_k} \\ \prod_{k=1}^{i} V_{l_k} \end{pmatrix} \quad \forall i \geq 2$$

We consider the problem of obtaining the best tradeoff between maximizing the diversity product

$$d_P(\mathcal{V}) = \min_{1 \leq l < k \leq L} |\det(V_l - V_k)| \geq \min_{1 \leq l \leq L} \left|1 - \eta^l\right| \left|1 - \eta^{ul}\right| = \zeta(u, L)$$

and maximizing the size L of $\mathcal{V}$. In particular we want to determine in an efficient way the optimal parameter u that maximizes $\zeta(u,L)$ for a given L. Observe that since

$$\left|1-e^{ix}\right| = |1-\cos x - i\sin x| = \sqrt{2-2\cos x} = 2\sin\left(\frac{x}{2}\right),$$

we have

$$\zeta(u,L) = \min_{1\le l\le L}\left|1-e^{\frac{2\pi i l}{L}}\right|\left|1-e^{\frac{2\pi i u l}{L}}\right| = 4\min_{1\le l\le L}\left|\sin\left(\frac{\pi l}{L}\right)\sin\left(\frac{\pi l u}{L}\right)\right| =$$
$$= 4\min_{(x,y)}\left|\sin\left(\frac{\pi x}{L}\right)\sin\left(\frac{\pi y}{L}\right)\right|,$$

where the last expression ranges over the pairs (x,y) with $x\in\mathbb{Z}$, $y\equiv xu \pmod L$, $|x|,|y|<L$, that is, (x,y) belongs to the subset $\Delta_{u,L}=\Lambda_{u,L}\cap((0,L)\times(0,L))$ of the two-dimensional lattice

$$\Lambda_{u,L} = \{(x, ux-zL) \mid x,z\in\mathbb{Z}\} = M\mathbb{Z}^2,$$

where

$$M = \begin{pmatrix} 1 & 0 \\ u & L \end{pmatrix}$$

It can be shown [21] that

$$\left(1-\frac{\pi^2}{24}\right)\frac{\pi^2}{2L^2}\min_{(x,y)\in\Delta_{u,L}}|xy| \le \zeta(u,L) \le \frac{\pi^2}{2L^2}\min_{(x,y)\in\Delta_{u,L}}|xy|,$$

that is, the behavior of the function $\zeta(u,L)$ is roughly similar to that of

$$\mu(u,L) = \min_{(x,y)\in\Delta_{u,L}}|xy|$$

It is natural to ask whether the value u which maximizes $\mu(u,L)$ also maximizes $\zeta(u,L)$. Unfortunately this is false in general.
The function $\mu(u,L)$ turns out to be related to the approximations of $\frac{u}{L}$ by the convergents of its continued fraction expansion:

Proposition 5.1. *Let $\frac{p_1}{q_1},\frac{p_2}{q_2},\dots,\frac{p_t}{q_t}=\frac{u}{L}$ be the convergents of $\frac{u}{L}$. Then*

$$\mu(u,L) = \min_{1\le l\le t-1} q_l\,|q_l u - p_l L|$$

Proof. Observe that

$$|xy| = |x(ux-zL)|,$$

For $0<x<L$, let l be such that $q_l\le x<q_{l+1}$. Then the thesis is an easy consequence of the fact that continued fraction convergents are "best approximations": $\forall(x,z)\in\mathbb{Z}^2$ with $0<x\le q_l$, $\left|\frac{u}{L}-\frac{z}{x}\right|\ge\left|\frac{u}{L}-\frac{p_l}{q_l}\right|$, and so

$$\left|u-\frac{zL}{x}\right|\ge\left|u-\frac{p_l L}{q_l}\right| \Rightarrow x\,|xu-zL| \ge q_l^2\left|u-\frac{zL}{x}\right| \ge q_l\,|q_l u - p_l L|$$

Remark that for x negative, the same argument works using $-x$, since there is no condition on the sign of z. □

The following result [21] shows that, roughly speaking, *the smaller the elements of the continued fraction expansion of* $\frac{u}{L}$ *are, the better*:

Proposition 5.2. *Let* $\frac{u}{L} = [0; a_1, a_2, \ldots, a_t]$, $\frac{w}{L} = [0; b_1, b_2, \ldots, b_s]$ *with*

$$\max_{1 \leq j \leq t} a_j + 1 < \max_{1 \leq i \leq s} b_i$$

Then $\mu(u, L) > \mu(w, L)$.

It is then natural to look to the sequence $\{F_n\}$ of the Fibonacci numbers as a way to build "good rational numbers" $\frac{u}{L}$, since, as is well-known,

$$\frac{F_n}{F_{n+1}} = [0; \underbrace{1, 1, \ldots, 1}_{n}]$$

Indeed, it turns out that it is the sequence

$$\frac{F_n}{F_{n+2}} = [0; \underbrace{1, \ldots, 1, 2}_{n-1}]$$

which realizes the maximum:

Proposition 5.3. *We have*

$$\max_{1 \leq u < F_n} \mu(u, F_n) = \mu(F_{n-2}, F_n) = F_{n-2}$$

Moreover, Shokrollahi conjectures that $\zeta(F_n) = \zeta(F_{n-2}, F_n)$ also holds.

5.2 Threaded-Algebraic Space-Time Codes

Of course, the main drawback of using diagonal space-time codes consists in the fact that the actual transmission rate is only one symbol per channel use.
Recently Damen and El Gamal [11] introduced a multi-antenna code design (*Threaded Algebraic Space-Time Codes*, TAST for short) which allows for full diversity, *full rate* of transmission and a polynomial complexity of decoding. The problem of choosing the parameters in order to optimize performance of these codes is related to diophantine approximations, and is still open.
Before describing these codes, we will introduce some useful algebraic tools that will guarantee the full rate condition.

5.2.1 Algebraic lattices from totally real number fields

Let $\mathbb{Q}(\theta)$ be a number field of degree n over $\mathbb{Q}$, $\mathcal{O}$ its ring of integers, and $\{w_1, \ldots, w_n\}$ a basis of integers.
It is well-known that there exist n embeddings $\sigma_i : \mathbb{Q}(\theta) \to \mathbb{C}$, leaving $\mathbb{Q}$ fixed, defined by $\sigma_i(\theta) = \theta_i$, where $\theta = \theta_1, \theta_2, \ldots, \theta_n$ are the conjugates of θ.
Let r_1 be the number of embeddings of $\mathbb{Q}(\theta)$ whose image is contained in $\mathbb{R}$, and $2r_2$ the number of embeddings whose image contains some complex number (it is clearly an even number since the complex roots of the minimal polynomial of θ come in pairs of conjugates.) The pair (r_1, r_2) is called the *signature* of $\mathbb{Q}(\theta)$. $\mathbb{Q}(\theta)$ is called *totally real* if $r_2 = 0$.
The embeddings σ_i provide a geometrical interpretation of $\mathbb{Q}(\theta)$ as a sublattice of $\mathbb{C}^n$: let $\sigma : \mathbb{Q}(\theta) \to \mathbb{R}^n$ be the *canonical embedding*

$$\sigma(x) = (\sigma_1(x), \ldots, \sigma_{r_1}(x), \Re(\sigma_{r_1+1}(x)), \Im(\sigma_{r_1+1}(x)), \ldots, \Re(\sigma_{r_1+r_2}(x)), \Im(\sigma_{r_1+r_2}(x))$$

where we take only one embedding from each pair of complex conjugates.
Consider the matrix $A = (a_{ij}) \in M_n(\mathbb{R})$, with $a_{ij} = \sigma_i(w_j)$: its columns $\sigma(w_j)$ are linearly independent [23] and generate an algebraic lattice $\Lambda(\mathcal{O}) = A\mathbb{Z}^n$. The volume of the fundamental parallelotope of $\Lambda(\mathcal{O})$ is

$$\mathrm{vol}(\Lambda(\mathcal{O})) = 2^{-r_2}\sqrt{|d_k|},$$

where $d_k = (\det A)^2$ is the *discriminant* of the lattice.
It can be shown [18] that algebraic lattices from totally real number fields achieve maximum diversity:

Theorem 5.4. *The diversity of an algebraic lattice is $L = r_1 + r_2$.*

Thus in the totally real case, $L = r_1 = n$.
If $I = (\alpha)$ is a principal ideal of $\mathcal{O}$, we define $N(I) = |N(\alpha)|$. I has an integer basis $\{\alpha w_1, \ldots, \alpha w_n\}$, and we can again consider the algebraic lattice $\Lambda(I)$ generated by $A' = (\sigma_i(\alpha w_j))$. In this case,

$$\mathrm{vol}(\Lambda(I)) = 2^{-r_2} N(I)\sqrt{|d_k|}$$

In order to compute the minimum of the product distance when $\mathcal{S} = \Lambda(\mathcal{O})$, it is enough to consider the case $u \in \Lambda(\mathcal{O}) \setminus \{0\}$, $u' = 0$, since $\Lambda(\mathcal{O})$ is linear. If $u = (u_1, \ldots, u_n)^t \in \mathbb{Z}^n$, we have

$$Au = \left(\sigma_1\left(\sum_{i=1}^{n} u_i w_i\right), \ldots, \sigma_n\left(\sum_{i=1}^{n} u_i w_i\right)\right)^t = (\sigma_1(s), \ldots, \sigma_n(s))^t$$

for some algebraic integer $s \in \mathcal{O}$. In this case the product distance coincides with the algebraic norm of s over $\mathbb{Q}$:

$$d_P(u, 0) = \prod_{i=1}^{n} |u_i| = \prod_{i=1}^{n} |\sigma_i(s)| = |N(s)|$$

But for any algebraic integer $s \neq 0$, we have $N(s) \in \mathbb{Z}\setminus\{0\}$; therefore $|N(s)| \geq 1$, and this construction ensures that for $\mathcal{S} \subset \Lambda$ the product distance doesn't vanish [18].
Clearly d_P can be increased by choosing a principal ideal I instead of $\mathcal{O}$; however, in this case the number of lattice points available for a certain constellation radius is smaller, since the volume of the fundamental parallelotope is bigger.
As we will see in §6, the same basic ideas we introduced here can be extended to the non-commutative case, by considering a quaternion ring over $\mathbb{Q}(i)$ and its maximal order instead of a number field with its ring of integers.

5.2.2 Threaded-Algebraic Space-Time Codes

Consider a system with M transmit antennas and N receive antennas, with block length $T = M$. Given $L \leq M$, let $l_j \subset \{1, \ldots, M\} \times \{0, \ldots, M-1\}$ be defined as follows:

$$l_j = \{([t + j - 1]_M + 1, t),\ 0 \leq t < M\}, \quad 1 \leq j \leq L$$

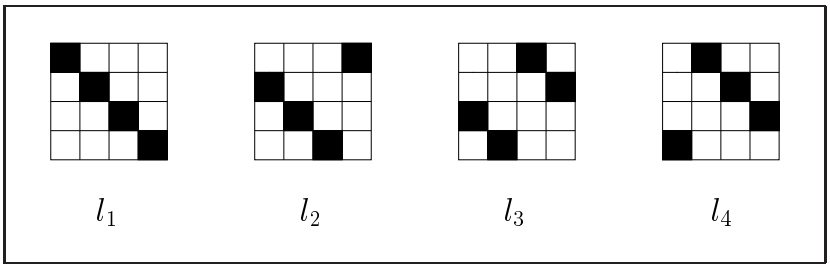

Figure 5.1: A subdivision of a matrix into "threads" when $M = 4$.

where $[\,]_M$ denotes the remainder $(\mathrm{mod}\, M)$.
Given $K = kL$, we consider a partition of the information vector u in intervals of length k: $u = (u^{(1)}, \ldots, u^{(L)})$. Each $u^{(j)}$ is separately mapped to a vector $\gamma_j(u^{(j)}) \in \mathcal{S}^M$ and its components are assigned to the elements of $B(u)$ corresponding to the *thread* l_j. If there are any entries in $B(u)$ whose indices do not belong to any thread, these are assigned the value 0.
The matrix $B(u)$ is thus decomposed into threads, which behave like independent codes, transparent to one another, each of which exploits all the space-time diversity: at the time t, the thread l_j transmits a symbol using the antenna $[t + j - 1]_M + 1$; the number of simultaneously active antennas at a given time is L.
Observe that the diagonal space-time codes described in §5 are a special case of TAST, corresponding to a single thread. The problem of coding is now reduced to choosing the encoders γ_j in such a way that the resulting code B achieves maximum diversity.
Define

$$\gamma_j(u^{(j)}) = \phi_j A u^{(j)},$$

where $\{\phi_1, \ldots, \phi_L\}$ are complex numbers to be determined, and A is an $M \times M$ matrix that maximizes

$$\min_{\substack{u \neq u', \\ s = A(u-u')}} \prod_{i=1}^{M} |s_j|$$

It can be shown that a good choice for A is the matrix defined in equation (5.2.1) in Paragraph §5.2.1, constructed from a suitable number field $\mathbb{Q}(\theta)$ that is a Galois extension of $\mathbb{Q}$ of degree M.
We can take $\phi_1 = 1, \phi_2 = \phi^{\frac{1}{M}}, \ldots, \phi_L = \phi^{\frac{L-1}{M}}$, where ϕ is an algebraic integer such that $\mathbb{Q}(\phi) \supset \mathbb{Q}(\theta)$ and $\{1, \phi, \ldots, \phi^{L-1}\}$ are linearly independent over $\mathbb{Q}(\theta)$.

Theorem 5.5. *The TAST code built in this way has maximum diversity; moreover, its coding gain $\delta_{\mathcal{C}}$ is greater or equal to*

$$(1 + |\phi| + L\nu)^{-\frac{m(L-1)}{M}}, \tag{5.1}$$

where m is the degree of $\mathbb{Q}(\theta)$ and ν depends only on the initial constellation $\mathcal{C}$.

The proof of Theorem 5.5 relies on the fact that $\det(B(u) - B(u'))$ is a polynomial of degree $L - 1$ in ϕ, whose coefficients are algebraic integers in $\mathbb{Q}(\theta)$; then we can apply the following theorem on the *simultaneous approximation of algebraic numbers*:

Theorem 5.6 (Generalization of Liouville's Theorem). *Let $\alpha_1, \ldots, \alpha_m$ be algebraic numbers, and let n be the degree of the smallest number field $\mathbb{Q}(\theta)$ which contains them. Let $p(X_1, \ldots, X_m) \in \mathbb{Z}[X_1, \ldots, X_m]$ be a polynomial of degree less or equal to k, such that the maximum modulus of its coefficients is H. Then if $p(\alpha_1, \ldots, \alpha_m) \neq 0$, we have*

$$|p(\alpha_1, \ldots, \alpha_m)| \geq \frac{1}{r^{nk}(1 + \sum_{i=0}^{m} |\overline{\alpha_i}|)^{(n-1)k} H^{n-1}},$$

where r is such that $r\alpha_i$ is an algebraic integer for $i = 1, \ldots, m$, and $|\overline{\alpha_i}|$ is the maximum modulus of the conjugates of α_i.

Unfortunately, a higher number of lattice points available means a better approximation and a smaller gain: the parameter ν in equation (5.1) grows when $\#\mathcal{C}$ increases.
A slightly different problem from the one we have been addressing is how to determine ϕ in such a way that $\delta_{\mathcal{C}}(\phi)$ decreases as slowly as possible when the size of the constellation grows to infinity.
Theorem 5.6 suggests that $\{\phi_1, \ldots, \phi_L\}$ ought to be chosen in such a way as to be "badly approximated" by algebraic numbers. A different strategy consists in choosing ϕ trascendental, for example $\phi = e^{i\lambda}$ with λ algebraic. It can be shown [11] that even in this case maximum diversity is achieved, thanks to the following well-known theorem:

Theorem 5.7 (Lindemann). *Let $\alpha_1, \ldots, \alpha_m$ be distinct algebraic numbers, and $c_1, \ldots, c_m$ algebraic coefficients not all equal to zero. Then $\sum_{i=1}^{m} c_i e^{\alpha_i} \neq 0$.*

For this choice of ϕ, however, we have no bounds for the coding gain.

5.2.3 The 2-dimensional case

This special case of TAST codes was introduced by Damen, Tewfik and Belfiore [10]. Again, there is a surprising relation between these codes and continued fractions; in fact the coding gain turns out to be greater when the fundamental parameter is "badly approximable" by continued fractions.
Consider the TAST code with $M = T = 2$, built from the matrix

$$A_\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & \phi \\ 1 & -\phi \end{pmatrix}$$

Let $\phi = \theta = e^{i\lambda}$, with λ to be determined. Then for $u = (u_1, u_2, u_3, u_4) \in \mathbb{Z}[i]^4$,

$$B_\phi(u) = \frac{1}{\sqrt{2}} \begin{pmatrix} u_1 + \phi u_2 & \sqrt{\phi}(u_3 + \phi u_4) \\ \sqrt{\phi}(u_3 - \phi u_4) & u_1 - \phi u_2 \end{pmatrix}$$

If we take as a constellation the whole lattice $\mathbb{Z}[i]^4$, the coding gain is

$$\delta_{\mathcal{C}}(\phi) = \inf_{u \in \mathbb{Z}[i]^4 \setminus \{0\}} \sqrt{\left|\det(B_\phi B_\phi^H)\right|} = \frac{1}{\sqrt{2}} \sqrt{\inf_{u \in \mathbb{Z}[i]^4 \setminus \{0\}} |u_1^2 - u_3^2\phi - u_2^2\phi^2 + u_4^2\phi^3|}$$

Clearly if $\mathcal{C} \subset \mathbb{Z}[i]^4$, the coding gain over $\mathcal{C}$ will in general be greater. In this case a stronger result than Theorem 5.5 holds:

Proposition 5.8. *Suppose that one of the following is true:*

a) ϕ *is algebraic of degree greater or equal to* 4 *over* $\mathbb{Q}(i)$*;*

b) ϕ *is algebraic of degree* 2 *over* $\mathbb{Q}(i)$ *and* $\phi^2 \in \mathbb{Q}(i)$*;*

c) ϕ *is trascendental*

Then the previous design achieves maximum diversity over all the constellations $\mathcal{C}$ *carved from* $Z[i]^4$*.*

When the condition (a) or (c) hold, the determinant of $B(u)B(u)^H$ is a polynomial of degree 3 in ϕ, so the thesis follows from the fact that $\{1, \phi, \phi^2, \phi^3\}$ are linearly independent over $\mathbb{Q}(i)$.
As for condition (b), it is enough to remark that the determinant can be written in the form $\frac{1}{2}((u_1^2 - u_2^2\phi^2) - \phi(u_3^2 - u_4^2\phi^2))$. If it were 0, since $\{1, \phi\}$ is a free set over $\mathbb{Q}(i)$ and $\phi^2 \in \mathbb{Q}(i)$, we would have $u_1^2 - u_2^2\phi^2 = u_3^2 - u_4^2\phi^2 = 0$; but ϕ is not a square in $\mathbb{Q}(i)$, being algebraic of degree 2.
In general, given $\mathcal{C}$, no criterion to maximize $\delta_{\mathcal{C}}$ is known; in the case of 4-QAM and 16-QAM constellations, computer simulations seem to suggest that the values of ϕ that maximize the coding gain are $e^{\frac{i}{2}}$ and $e^{0.521i}$ respectively. However, the computer search for the optimal parameters becomes extremely complex when the size of the constellation increases; moreover, the parameters found by computer search which give an optimal coding gain for one constellation might work very poorly with another, so that a theoretical approach would be preferable.
As a heuristic approach, one can observe that

$$\delta_{\mathcal{C}} \leq \min_{u \neq 0} \min(\left|u_1^2 - u_3^2\phi\right|, \left|u_1^2 - u_2^2\phi^2\right|, \left|u_1^2 + u_4^2\phi^3\right|),$$

and try to choose ϕ badly approximable by rational numbers, that is:

- ϕ algebraic of small degree and such that the moduli of the coefficients of its minimum polynomial are small,
- ϕ trascendental such that the digits of its continued fraction expansion are small.

In the case of ϕ trascendental, however, there is no lower bound available on the decrease of the coding gain when the spectral efficiency grows.

Chapter 6

Algebraic space-time block coded modulation

As we have seen, the main inconvenience of the 2×2 TAST code is the fact that the minimum determinant vanishes when the size of the constellation tends to infinity. This problem can be solved in an elegant manner by ensuring that the block codewords belong to a division algebra. Moreover, in the two-dimensional case the algebra can be chosen in such a way that the code will be isometric to a rotated cubic lattice.

6.1 Quaternion Algebras

This section summarizes some basic facts about quaternion algebras that will be useful later. Our main references are the books of Vignéras [25] and Reiner [19].

6.1 (Quaternion algebras). Let K be a field. A *quaternion algebra* $\mathcal{H}$ of center K is a central simple algebra of dimension 4 over K, such that there exists a separable quadratic extension L of K, and an element $\gamma \in K^*$, such that

$$\mathcal{H} = L \oplus Le, \qquad e^2 = \gamma, \qquad ex = \sigma(x)e \quad \forall x \in L$$

where σ is the non-trivial K-automorphism of L. L is called a *maximal subfield* of $\mathcal{H}$. $\mathcal{H}$ will be denoted by the triple $(L/K, \sigma, \gamma)$.

Quaternion algebras are a special case of *cyclic algebras*.
To obtain a representation of $\mathcal{H}$ as a K-module, consider a primitive element i such that $L = K(i)$, and let $j = e$, $k = ij = j\sigma(i)$. Then

$$\mathcal{H} = \{a + bi + cj + dk \mid a, b, c, d \in K\} \tag{6.1}$$

The following theorem gives a sufficient condition for a quaternion algebra to be a division ring:

Theorem 6.2. *Let $\mathcal{H} = (L/K, \sigma, \gamma)$ be a quaternion algebra. If γ is not a reduced norm of any element of L, then $\mathcal{H}$ is a skewfield.*

6.3 (Splitting fields). Let $\mathcal{H}$ be a central simple K-algebra. An extension field E of K *splits* $\mathcal{H}$, or is a *splitting field* for $\mathcal{H}$, if

$$E \otimes_K \mathcal{H} \cong M_r(E)$$

In the case of division algebras, every maximal subfield is a splitting field:

Theorem 6.4. *Let $\mathcal{D}$ be a skewfield with center K, with finite degree over K. Then every maximal subfield E of $\mathcal{D}$ contains K, and is a splitting field for $\mathcal{D}$.*

In the following paragraphs we will always consider a Dedekind domain R, its quotient field K, and a quaternion algebra $\mathcal{H}$ over K.

6.5 (Lattices and orders). A *full R-lattice* or *ideal* in $\mathcal{H}$ is a finitely generated R-submodule I in $\mathcal{H}$ such that $KI = \mathcal{H}$, where

$$KI = \left\{ \sum_{i=1}^{n} k_i x_i \;\middle|\; k_i \in K,\, x_i \in I,\, n \in N \right\}$$

An *R-order* Θ in $\mathcal{H}$ is a full R-lattice which is also a subring of $\mathcal{H}$ with the same unity element. A *maximal R-order* is an order which is not properly contained in any other order of $\mathcal{H}$.

For the following proposition see for example Reiner [19]:

Proposition 6.6. *A subring of $\mathcal{H}$ containing a basis for $\mathcal{H}$ over K is an order if and only if all its elements are integral over R.*

Remark 6.7. The notion of order is a generalization of the notion of the ring of integers for commutative extensions. However, in the non-commutative case the set of elements which are integral over the base field might not be a ring.

6.8 (Properties of ideals). Given an ideal I of $\mathcal{H}$, we can define the *left order* and the *right order* of I as follows:

$$\Theta_l(I) = \{x \in \mathcal{H} \mid Ix \subset I\},$$
$$\Theta_r(I) = \{x \in \mathcal{H} \mid xI \subset I\}$$

$\Theta_l(I)$ and $\Theta_r(I)$ are orders. I is called

- *two-sided* if $\Theta_l(I) = \Theta_r(I)$,
- *normal* if $\Theta_l(I)$ and $\Theta_r(I)$ are maximal,
- *integral* if $I \subset \Theta_l(I)$, $I \subset \Theta_r(I)$,
- *principal* if $I = \Theta_l(I)x = x\Theta_r(I)$ for some $x \in \mathcal{H}$

The *inverse* of I is the fractional ideal $I^{-1} = \{x \in \mathcal{H} \mid IxI \subset I\}$.
The *norm* $N(I)$ of an ideal I is the set of reduced norms of its elements, and it is an ideal of R. If $I = \Theta x$ is principal, $N(I) = RN(x)$.

6.2 Space-time codes from quaternion algebras

Theorem 6.2 provides a sufficient condition (albeit one that is not simple to check) for building division algebras. For the applications to coding, in order to ensure a uniform distribution of the average energy among the different antennas, it is preferable to choose the generator γ of the quaternion algebra $(L/K, \sigma, \gamma)$ such that $|\gamma| = 1$.

6.2.1 The Alamouti Code

The first example of space-time code that can be interpreted in the framework of quaternion algebras is the famous *Alamouti Code* [1]. This code is optimal for the case of 2 transmit antennas and one receive antenna. In addition to being full rate and full rank, its *orthogonal* structure allows for linear decoding. Thanks to its simplicity of implementation and good performance, this code is already integrated in the UMTS standard.

The Alamouti code can be derived from the skewfield $\mathcal{H}$ of *Hamilton quaternions*, which corresponds to choosing, following the notation in Definition 6.1, the field of real numbers $\mathbb{R}$ as the base field K, the field of complex numbers $\mathbb{C}$ as its quadratic extension L, and the element $\gamma = -1$.
Since $\mathbb{C}$ is a maximal subfield of $\mathcal{H}$, it follows from Theorem 6.4 that $\mathcal{H}$ admits a matrix representation as a subset of $M_2(\mathbb{C})$. More precisely, $\mathcal{H} = \{a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \mid a, b, c, d \in \mathbb{R}\}$, where

$$\mathbf{i} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad \mathbf{j} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \mathbf{k} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$$

Given an information vector $u = (x_1, x_2) = (a+bi, c+di) \in \mathbb{C}^2$, the transpose of the corresponding Alamouti codeword for block length $T = 2$ is the quaternion

$$X^t = x_1 + x_2\mathbf{j} = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} = \begin{pmatrix} a+bi & c+di \\ -(c-di) & a-ib \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \\ -\overline{x}_2 & \overline{x}_1 \end{pmatrix},$$

where the overline denotes the complex conjugate. Remark that the columns of each codeword

$$X = \begin{pmatrix} x_1 & -\overline{x}_2 \\ x_2 & \overline{x}_1 \end{pmatrix}$$

are orthogonal with respect to the Hermitian product.
Moreover, if X is nonzero its determinant $\det(X) = |x_1|^2 + |x_2|^2 > 0$, so that the full rank condition is always satisfied. Moreover, if the symbols x_1,x_2 belong to a QAM constellation carved from $\mathbb{Z}[i]$, the minimum determinant will be in $\mathbb{Z}$ and therefore $|\det(X)| \geq 1$.
In the setting of §4.2.1, the received codeword is $Y = HX + W$:

$$\begin{pmatrix} y_1 & y_2 \end{pmatrix} = \begin{pmatrix} h_1 & h_2 \end{pmatrix} \begin{pmatrix} x_1 & -\overline{x}_2 \\ x_2 & \overline{x}_1 \end{pmatrix} + \begin{pmatrix} w_1 & w_2 \end{pmatrix} =$$
$$= \begin{pmatrix} h_1x_1 + h_2x_2 + w_1 & -h_1\overline{x}_2 + h_2\overline{x}_1 + w_2 \end{pmatrix}$$

In order to recover X, it is convenient to consider the vector

$$Z = \begin{pmatrix} y_1 \\ \overline{y}_2 \end{pmatrix} = \begin{pmatrix} h_1 & h_2 \\ \overline{h}_2 & -\overline{h}_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} w_1 \\ \overline{w}_2 \end{pmatrix}$$

That is, our system is equivalent to a transmission scheme with 2 receive antennas and 2 transmit antennas, but with block length $T = 1$, and such that the columns of the channel matrix

$$H' = \begin{pmatrix} h_1 & h_2 \\ \overline{h}_2 & -\overline{h}_1 \end{pmatrix}$$

are orthogonal. By multiplying on the left by $(H')^H$, we obtain

$$(H')^H Z = (|h_1|^2 + |h_2|^2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

where the components of the new noise vector (v_1, v_2) are still independent random variables.
Thus the Alamouti Code admits a very simple decoding process, which consists in applying a linear transformation followed by a single symbol ML detection.

6.3 The Golden Code

This code, introduced by Belfiore, Rekaya and Viterbo [5], is optimal for the case of 2 transmit and 2 receive antennas, and belongs to a class of $n \times n$ MIMO codes called *perfect codes* [17]. These codes have been shown to exist only for $n = 2, 3, 4$ and 6.

6.9 (Perfect code). A code is *perfect* if:

1. it is full rate using constellations carved from either $\mathbb{Z}[i]$ or $\mathbb{Z}[e^{\frac{2}{3}\pi i}]$;
2. its minimum determinant is nonzero, so that it is full rank;
3. the real lattice generated by the vectorized codewords is $\mathbb{Z}^{2n^2}$ or $A_2^{n^2}$, where A_2 is the hexagonal lattice;
4. the symbols in the code matrix have the same average energy.

Consider the number field $\mathbb{Q}(i, \theta)$, where $\theta = \frac{\sqrt{5}+1}{2}$ is the Golden number, and its ring of integers $\mathbb{Z}[i, \theta]$.
If $x = a + b\theta \in \mathbb{Q}(i, \theta)$, let $\sigma(x) = \bar{x}$ be its canonical conjugate $a + b\bar{\theta}$, where

$$\bar{\theta} = 1 - \theta, \quad \theta\bar{\theta} = -1$$

$N(x) = x\bar{x}$ is the norm of x, and $x \in \mathbb{Z}[i, \theta]$ implies that $N(x) \in \mathbb{Z}[i]$.
Consider the quaternion algebra $\mathcal{A}$, with center $\mathbb{Q}(i)$, and maximal subfield $\mathbb{Q}(i, \theta)$:

$$\mathcal{A} = \mathbb{Q}(i, \theta) \oplus \mathbb{Q}(i, \theta)j = \mathcal{A}(\mathbb{Q}(i, \theta)/\mathbb{Q}(i), \sigma, i)$$

where $j^2 = i$, $xj = j\bar{x}\ \forall x \in \mathbb{Q}(i, \theta)$.
It can be shown [5] that i is not a norm in $\mathbb{Q}(i, \theta)/\mathbb{Q}(i)$, so Theorem 6.2 implies that $\mathcal{A}$ is a division algebra.
As a $\mathbb{Q}(i)$-module, we have

$$\mathcal{A} = \mathbb{Q}(i) \oplus \mathbb{Q}(i)\theta \oplus \mathbb{Q}(i)j \oplus \mathbb{Q}(i)\theta j,$$

From Theorem 6.4, it follows that $\mathbb{Q}(i,\theta)$ is a splitting field for $\mathcal{A}$, and so

$$\mathbb{Q}(i,\theta) \otimes_{\mathbb{Q}(i)} \mathcal{A} \cong M_2(\mathbb{Q}(i,\theta))$$

Consequently, $\mathcal{A}$ admits a matrix representation as a subset of $M_2(\mathbb{Q}(i,\theta))$, where the inclusion is given by

$$x \mapsto \begin{pmatrix} x & 0 \\ 0 & \bar{x} \end{pmatrix}, \ \forall x \in \mathbb{Q}(i,\theta), \quad j \mapsto \begin{pmatrix} 0 & 1 \\ i & 0 \end{pmatrix} \tag{6.2}$$

Then $\mathcal{A}$ can be written in the form

$$\mathcal{A} = \left\{ \begin{bmatrix} x_1 & x_2 \\ i\bar{x}_2 & \overline{x}_1 \end{bmatrix}, \ x_1, x_2 \in \mathbb{Q}(i,\theta) \right\} \tag{6.3}$$

and every element of $\mathcal{A}$ has nonzero determinant $N(x_1) - iN(x_2)$.

Remark 6.10. An alternative representation of $\mathcal{A}$ as a cyclic algebra can be obtained by considering the field extension $\mathbb{Q}(\sqrt{i})/\mathbb{Q}(i)$ and the quaternion algebra $\mathcal{A}(Q(\sqrt{i})/\mathbb{Q}(i), \tau, 5)$, where τ is the canonical conjugacy in $\mathbb{Q}(\sqrt{i})/\mathbb{Q}(i)$. Thus

$$\mathcal{A}' = \mathbb{Q}(i) \oplus \mathbb{Q}(i)\sqrt{i} \oplus \mathbb{Q}(i)f \oplus \mathbb{Q}(i)\sqrt{i}f, \qquad f^2 = 5, \qquad xf = f\tau(x) \quad \forall x \in \mathbb{Q}(\sqrt{i})$$

The isomorphism $\phi : \mathcal{A} \mapsto \mathcal{A}'$ of $Q(i)$-algebras between the two representations is given by $\phi(\sqrt{5}) = f$, $\phi(j) = \sqrt{i}$. It is sufficient to check that it is a ring isomorphism, namely that

$$\phi(j)\phi(\sqrt{5}) = \sqrt{i}f = -f\sqrt{i} = -\phi(\sqrt{5})\phi(j),$$

the other products between the generators being automatically preserved, since their characteristic polynomials are the same.

The construction of $\mathcal{A}$ ensures that the first two conditions in Definition 6.9 are satisfied. In order to have *cubic shaping*, we will take a suitable ideal of $\mathcal{A}$. Consider the order of $\mathcal{A}$

$$\mathcal{O} = \left\{ \begin{bmatrix} x_1 & x_2 \\ i\bar{x}_2 & \overline{x}_1 \end{bmatrix}, \ x_1, x_2 \in \mathbb{Z}(i,\theta) \right\} \tag{6.4}$$

and let

$$\alpha = 1 + i\bar{\theta}, \quad A = \begin{bmatrix} \alpha & 0 \\ 0 & \bar{\alpha} \end{bmatrix}, \tag{6.5}$$

The *Golden Code* $\mathcal{G} = \frac{1}{\sqrt{5}} A\mathcal{O}$ is a rescaled version of the right principal ideal $A\mathcal{O}$ of $\mathcal{O}$. Every codeword $X \in \mathcal{G}$ is of the form

$$X = \frac{1}{\sqrt{5}} \begin{bmatrix} \alpha(a + b\theta) & \alpha(c + d\theta) \\ \bar{\alpha}i(c + d\bar{\theta}) & \bar{\alpha}(a + b\bar{\theta}) \end{bmatrix} \tag{6.6}$$

It is easy to verify that $\mathcal{G}$ is a *two-sided ideal*: in fact if $w = w_1 + w_2 j \in \mathcal{O}$, $w_1, w_2 \in \mathbb{Z}[i,\theta]$,

$$\alpha(w_1 + w_2 j) = w_1\alpha + w_2 j\bar{\alpha} = (w_1 + i\theta w_2 j)\alpha,$$

observing that $\alpha i\theta = i\theta + 1 = \bar{\alpha}$. But

$$\xi : w_1 + w_2 j \mapsto w_1 + i\theta w_2 j \tag{6.7}$$

is an homomorphism of $\mathbb{Z}[i]$-modules that maps $\mathcal{O}$ into itself bijectively, therefore $\alpha\mathcal{O} = \mathcal{O}\alpha$. Moreover, if we neglect the normalization constant $\frac{1}{\sqrt{5}}$, $\mathcal{G}$ is an *integral ideal* because it is contained in $\mathcal{O}$.

Remark 6.11. $\forall W \in \mathcal{O} \setminus \{0\}$, $|\det(W)| \geq 1$. Consequently, $\forall X \in \mathcal{G} \setminus \{0\}$, $|\det(X)| \geq \frac{1}{\sqrt{5}}$.

Proof. Since $\mathcal{A}$ is a division algebra, $N(x_1) - iN(x_2) \neq 0$ if $(x_1, x_2) \neq (0,0)$. Moreover, when $x_1, x_2 \in \mathbb{Z}[i][\theta]$, $N(x_1) - iN(x_2) \in \mathbb{Z}[i]$ and so its absolute value is at least 1. If $X = \frac{A}{\sqrt{5}}W$, $|\det(X)| = \frac{|N(\alpha)|}{5}|\det(W)| = \left|\frac{\det W}{\sqrt{5}}\right|$, since $|N(\alpha)| = |2 + i| = \sqrt{5}$. □

Remark 6.11 implies that the Golden Code $\mathcal{G}$ is *full-rank* and has non-vanishing determinant. By construction, it is also *full-rate*, that is, each codeword transmits four information symbols.[1]

6.3.1 Lattice representation of $\mathcal{G}$

We follow the column convention for vectors, so that lattices have the form $\Lambda = \{M\mathbf{u} \,|\, \mathbf{u} \in \mathbb{Z}[i]^n\}$. Two lattices $\Lambda = \{M\mathbf{u}\}$ and $\Lambda' = \{M'\mathbf{u}\}$ are *equivalent* if there exist U unimodular with Gaussian integer entries and T unitary such that $M' = TMU$.
Let $a, b, c, d \in \mathbb{C}$, and consider the linear mapping $\phi : \mathcal{A} \to \mathbb{C}^4$ that vectorizes matrices:

$$\phi\left(\begin{bmatrix} a & c \\ b & d \end{bmatrix}\right) = (a, b, c, d) \in \mathbb{C}^4$$

Obviously, the mapping ϕ preserves the norm: $\forall A \in \mathcal{A}$, $\|A\|_F^2 = \|\phi(A)\|^2$.
The left multiplication function $l_Y : \mathcal{A} \to \mathcal{A}$ that maps W to YW induces a linear mapping $Y_l = \phi \circ l_Y \circ \phi^{-1} : \phi(\mathcal{A}) \to \phi(\mathcal{A})$ that can be seen as a 4×4 complex matrix. Similarly, we can define the linear function $Y_r = \phi \circ r_Y \circ \phi^{-1}$, where $r_Y : W \mapsto WY$ is the multiplication on the right by Y.

Remark 6.12. If $Y = \begin{bmatrix} x & z \\ y & w \end{bmatrix}$, then

$$Y_l = \begin{bmatrix} x & z & 0 & 0 \\ y & w & 0 & 0 \\ 0 & 0 & x & z \\ 0 & 0 & y & w \end{bmatrix}, \qquad Y_r = \begin{bmatrix} x & 0 & y & 0 \\ 0 & x & 0 & y \\ z & 0 & w & 0 \\ 0 & z & 0 & w \end{bmatrix}$$

Another useful mapping to consider is $\Phi_{\mathcal{O}} : \mathbb{Z}[i]^4 \to \phi(\mathcal{O})$,

$$\Phi_{\mathcal{O}} : \mathbf{u} = (a, b, c, d) \mapsto \phi\left(\begin{bmatrix} a + b\theta & c + d\theta \\ i(c + d\bar{\theta}) & a + b\bar{\theta} \end{bmatrix}\right)$$

[1] It has also been shown [12] that the Golden Code achieves the diversity-multiplexing gain tradeoff (see §4.2.2).

It is easy to check that

$$\Phi_{\mathcal{O}} = \begin{bmatrix} 1 & \theta & 0 & 0 \\ 0 & 0 & i & i\bar{\theta} \\ 0 & 0 & 1 & \theta \\ 1 & \bar{\theta} & 0 & 0 \end{bmatrix}$$

Remark 6.13. The map $R = \frac{1}{\sqrt{5}} A_l \Phi_{\mathcal{O}}$ sends $\mathbb{Z}[i]^4$ to $\phi(\mathcal{G})$. R is the unitary matrix

$$R = \frac{1}{\sqrt{5}} \begin{bmatrix} 1+i\bar{\theta} & \theta - i & 0 & 0 \\ 0 & 0 & -\theta + i & 1 + i\bar{\theta} \\ 0 & 0 & 1 + i\bar{\theta} & \theta - i \\ 1 + i\theta & \bar{\theta} - i & 0 & 0 \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} \alpha & -\bar{\alpha}i & 0 & 0 \\ 0 & 0 & \bar{\alpha}i & \alpha \\ 0 & 0 & \alpha & -\bar{\alpha}i \\ \bar{\alpha} & -\alpha i & 0 & 0 \end{bmatrix}$$

Thus we have shown that the Golden Code has *cubic shaping*, since it is isometric to $\mathbb{Z}[i]^4$ with respect to the Frobenius norm.

Remark that the pre-images of the rotated canonical basis are

$$\begin{aligned} \phi^{-1}(R\mathbf{e}_1) &= \frac{1}{\sqrt{5}} \begin{bmatrix} \alpha & 0 \\ 0 & \bar{\alpha} \end{bmatrix} = \frac{A}{\sqrt{5}} \\ \phi^{-1}(R\mathbf{e}_2) &= \frac{1}{\sqrt{5}} \begin{bmatrix} -\bar{\alpha}i & 0 \\ 0 & -\alpha i \end{bmatrix} = \frac{\Theta A}{\sqrt{5}} \\ \phi^{-1}(R\mathbf{e}_3) &= \frac{1}{\sqrt{5}} \begin{bmatrix} 0 & \alpha \\ \bar{\alpha}i & 0 \end{bmatrix} = \frac{iJ\Theta A}{\sqrt{5}} \\ \phi^{-1}(R\mathbf{e}_4) &= \frac{1}{\sqrt{5}} \begin{bmatrix} 0 & -\bar{\alpha}i \\ \alpha & 0 \end{bmatrix} = \frac{-iJA}{\sqrt{5}}, \end{aligned} \tag{6.8}$$

where

$$J = \begin{bmatrix} 0 & 1 \\ i & 0 \end{bmatrix}, \quad \Theta = \begin{bmatrix} \theta & 0 \\ 0 & \bar{\theta} \end{bmatrix}$$

are the matrix representations of j and θ respectively (see equation (6.2)).

6.4 Golden Block Codes

We now consider the case of a *slow fading* channel, meaning that the channel coefficients remain constant for a certain time frame L. We want to define a block code of length L using the Golden Code as the "alphabet", in order to improve even further its performance.
As usual, we assume that the fading coefficients are known at the receiver. The received signal is given by

$$\mathbf{Y} = H\mathbf{X} + \mathbf{W}, \qquad \mathbf{X}, \mathbf{Y}, \mathbf{W} \in \mathbb{C}^{2\times 2L}, \tag{6.9}$$

where the entries of $H \in \mathbb{C}^{2\times 2}$ are i.i.d. complex Gaussian random variables, $\mathbf{W}$ is the Gaussian noise with i.i.d. entries of zero mean and variance N_0, and the transmitted signal $\mathbf{X} = (X_1, \ldots, X_L)$ belongs to a suitable subset $\mathcal{S} \subset \mathcal{G}^L$. As we have seen in equation (4.8), the pairwise error probability is given by

$$P(X \mapsto X') \leq \frac{1}{\left(\sqrt{\Delta_{\min}} \frac{N_0}{E_S}\right)^4},$$

where $E_{\mathcal{S}}$ is the average energy of $\mathcal{S}$ and

$$\Delta_{\min} = \min_{\mathbf{X}\in\mathcal{S}\setminus\{0\}} \left|\det(\mathbf{X}\mathbf{X}^H)\right|$$

In order to minimize the PEP for a given SNR and average energy, we should maximize $\Delta_{\min}$.
First of all, we would like to find an explicit formula for $\det(\mathbf{X}\mathbf{X}^H)$. Consider the following involution of the cyclic algebra $\mathcal{A}$ (corresponding to the quaternionic conjugate):

$$X = \begin{bmatrix} x_1 & x_2 \\ i\overline{x_2} & \overline{x_1} \end{bmatrix} \quad \mapsto \quad \widetilde{X} = \begin{bmatrix} \overline{x_1} & -x_2 \\ -i\overline{x_2} & x_1 \end{bmatrix}$$

Remark 6.14. $\forall X \in \mathcal{A}$,

$$\widetilde{X}X = \det(X)\mathbb{1} \tag{6.10}$$

$$\widetilde{X} + X = (x_1 + \overline{x_1})\mathbb{1} = \mathrm{tr}(X)\mathbb{1} \tag{6.11}$$

$$\det(X) = \det(\widetilde{X}) \tag{6.12}$$

Recall the definition of the *Frobenius norm* of a matrix:

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad a,b,c,d \in \mathbb{C} \quad \Rightarrow \quad \|M\|_F = \sqrt{|a|^2 + |b|^2 + |c|^2 + |d|^2}$$

Lemma 6.15. $\forall \mathbf{X} = (X_1, \ldots, X_L) \in \mathcal{A}^L$,

$$\det(\mathbf{X}\mathbf{X}^H) = \det(X_1X_1^H + \ldots + X_LX_L^H) = \\ |\det(X_1)|^2 + \ldots + |\det(X_L)|^2 + \sum_{j>i} \left\|\widetilde{X}_jX_i\right\|_F^2$$

Proof. For all $i = 1, \ldots, L$, let $Q_i = X_iX_i^H$: then

$$\det(X_1X_1^H + \ldots + X_LX_L^H)\mathbb{1} = \det(Q_1 + \ldots + Q_L)\mathbb{1} = \\ = (\widetilde{Q}_1 + \ldots + \widetilde{Q}_L)(Q_1 + \ldots + Q_L)\mathbb{1} = \sum_{i,j=1}^{L} \widetilde{Q}_iQ_j = \sum_{i=1}^{L} \det(Q_i)\mathbb{1} + \sum_{i\neq j} \widetilde{Q}_iQ_j$$

We need to show that $\widetilde{Q}_iQ_j + \widetilde{Q}_jQ_i = \left\|\widetilde{X}_jX_i\right\|_F^2 \mathbb{1}$.
But $\|X\|_F^2 = \mathrm{tr}(XX^H)$, and therefore $\left\|\widetilde{X}_jX_i\right\|_F^2 = \mathrm{tr}(\widetilde{X}_jX_iX_i^H\widetilde{X}_j^H)$, and

$$\widetilde{Q}_jQ_i = \widetilde{X}_j^H\widetilde{X}_jX_iX_i^H, \quad \widetilde{Q}_iQ_j = \widetilde{\widetilde{Q}_jQ_i} \\ \Rightarrow \widetilde{Q}_iQ_j + \widetilde{Q}_jQ_i = \mathrm{tr}(\widetilde{Q}_iQ_j)\mathbb{1} = \mathrm{tr}(\widetilde{X}_jX_iX_i^H\widetilde{X}_j^H)\mathbb{1},$$

recalling that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. □

6.4.1 Estimates of the Frobenius norm

Remark 6.16. If $W \in \mathcal{O}$, $\|W\|_F^2 \in \mathbb{Z}$.

Proof. Let

$$W = \begin{bmatrix} w_1 & w_2 \\ i\overline{w}_2 & \overline{w}_1 \end{bmatrix}, \quad w_1 = t_1 + is_1, w_2 = t_2 + is_2, \quad t_1, t_2, s_1, s_2 \in \mathbb{Z}[\theta]$$

Then $\|W\|_F^2 = |w_1|^2 + |\overline{w}_1|^2 + |w_2|^2 + \overline{w}_2^2$. But $w_1 = a + b\theta + i(c + d\theta)$ for some $a, b, c, d \in \mathbb{Z}$, and

$$|w_1|^2 + |\overline{w}_1|^2 = (a + b\theta)^2 + (c + d\theta)^2 + (a + b\bar{\theta})^2 + (c + d\bar{\theta})^2 = \\ = 2a^2 + 3b^2 + 2ab + 2c^2 + 3d^2 + 2cd \in \mathbb{Z}$$

The same is true for $|w_2|^2 + |\overline{w}_2|^2$. □

Remark 6.17. Let X, Y be two 2×2 complex-valued matrices. Then

$$\|X\|_F^2 \geq 2\,|\det(X)|\,, \tag{6.13}$$

$$\left\|\widetilde{X}Y\right\|_F^2 \geq 2\,|\det(X)|\,|\det(Y)| \tag{6.14}$$

Proof. If $X = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, then

$$\|X\|_F^2 = |a|^2 + |b|^2 + |c|^2 + |d|^2 \geq 2(|ad| + |bc|) \geq 2\,|ad - bc| = 2\,|\det(X)|$$

and $\left\|\widetilde{X}Y\right\|_F^2 \geq 2\left|\det(\widetilde{X}Y)\right| = 2\,|\det(X)\det(Y)|$. □

Remark 6.18. $\forall W \in \mathcal{O} \setminus \{0\}$, $\|W\|_F^2 \geq 2\,|\det(W)| \geq 2$. Moreover, the minimum is actually achieved when $W_1 = W_2 = \mathbb{1}$.

Remark 6.19. If $X_1, X_2 \in \mathcal{G} \setminus \{0\}$, $\left\|\widetilde{X}_2 X_1\right\|_F^2 \geq \frac{2}{5}$.

Proof. Let $X_1 = \frac{1}{\sqrt{5}} A W_1, X_2 = \frac{1}{\sqrt{5}} A W_2$, $W_1, W_2 \in \mathcal{O}$. Then

$$\left\|\widetilde{X}_2 X_1\right\|_F^2 = \frac{1}{25}\left\|\widetilde{W}_2 \widetilde{A} A W_1\right\|_F^2 = \frac{|N(\alpha)|^2}{25}\left\|\widetilde{W}_2 W_1\right\|_F^2 = \frac{1}{5}\left\|\widetilde{W}_2 W_1\right\|_F^2 \geq \frac{2}{5},$$

since $W = \widetilde{W}_2 W_1$ belongs to $\mathcal{O}$. □

From Remark 6.17, it follows that:

Lemma 6.20. *Let* $\mathbf{X} = (X_1, \ldots, X_L) \in \mathcal{G}^L$. *Then*

$$\det(\mathbf{X}\mathbf{X}^H) \geq \left(\sum_{i=1}^{L} |\det(X_i)|\right)^2 \geq \frac{(w_H(\mathbf{X}))^2}{5},$$

where $w_H(\mathbf{X}) = \#\{i \in \{1, \ldots, L\} \mid X_i \neq 0\}$ *is the Hamming weight of the block* $\mathbf{X}$.

6.4.2 The product lattice

We can study the structures induced on $\mathbb{Z}[i]^4$ by the map $X \mapsto \widetilde{X}$ and by the product $(X, Y) \mapsto XY$.

$$\tilde{\mathbf{x}} \doteqdot (R^{-1}\phi)(\widetilde{\phi^{-1}R\mathbf{x}}), \tag{6.15}$$

$$\mathbf{x} * \mathbf{y} \doteqdot (R^{-1}\phi)(\phi^{-1}R\mathbf{x} \cdot \phi^{-1}R\mathbf{y}) \tag{6.16}$$

Then recalling expression (6.8) for the lifts of the vectors of the canonical basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}$, we get

$$\widetilde{\mathbf{e}}_1 = i\mathbf{e}_2, \widetilde{\mathbf{e}}_2 = -i\mathbf{e}_1, \widetilde{\mathbf{e}}_3 = -\mathbf{e}_3, \widetilde{\mathbf{e}}_4 = -\mathbf{e}_4$$

Table 6.1 lists all the products of the vectors of the canonical basis. For example we have:

$$\mathbf{e}_1 * \mathbf{e}_1 = (R^{-1}\phi)\left(\frac{A^2}{5}\right) = \frac{1}{5}(R^{-1}\phi)((1+i)A - \widetilde{A}) = \frac{1}{\sqrt{5}}((1+i)\mathbf{e}_1 - i\mathbf{e}_2)$$

$\mathbf{e}_1 * \mathbf{e}_1 = \mathbf{e}_3 * \mathbf{e}_4 = \frac{1}{\sqrt{5}}((1+i)\mathbf{e}_1 - i\mathbf{e}_2)$
$\mathbf{e}_2 * \mathbf{e}_2 = -\mathbf{e}_4 * \mathbf{e}_3 = \frac{1}{\sqrt{5}}(\mathbf{e}_1 + (1-i)\mathbf{e}_2)$
$-\mathbf{e}_3 * \mathbf{e}_3 = \mathbf{e}_4 * \mathbf{e}_4 = \mathbf{e}_1 * \mathbf{e}_2 = \mathbf{e}_2 * \mathbf{e}_1 = \frac{1}{\sqrt{5}}(i\mathbf{e}_1 - \mathbf{e}_2)$
$\mathbf{e}_1 * \mathbf{e}_3 = i\mathbf{e}_3 * \mathbf{e}_2 = \frac{1}{\sqrt{5}}((1+i)\mathbf{e}_3 - i\mathbf{e}_4)$
$\mathbf{e}_1 * \mathbf{e}_4 = \mathbf{e}_2 * \mathbf{e}_3 = i\mathbf{e}_4 * \mathbf{e}_2 = -i\mathbf{e}_3 * \mathbf{e}_1 = \frac{1}{\sqrt{5}}(-i\mathbf{e}_3 + \mathbf{e}_4)$
$\mathbf{e}_4 * \mathbf{e}_1 = i\mathbf{e}_2 * \mathbf{e}_4 = \frac{1}{\sqrt{5}}(i\mathbf{e}_3 + (i+1)\mathbf{e}_4)$

Table 6.1: The products $\{\mathbf{e_i} * \mathbf{e_j}\}$, $i, j \in \{1, 2, 3, 4\}$.

Observe that with the definition (6.16), $\mathbb{Z}[i]^4 * \mathbb{Z}[i]^4 \subset \frac{1}{\sqrt{5}}\mathbb{Z}[i]^4$.

In order to design a code with good minima of the Frobenius norms $\left\|\widetilde{X}_j X_i\right\|_F^2$ in Lemma 6.15, we need more information on the set of products $\mathcal{P} = \{\widetilde{X}Y \mid X, Y \in \mathcal{G}\}$. Since $\widetilde{\mathcal{G}} = \mathcal{G}$, this set is simply $\mathcal{G}^2$. More precisely, if $\widetilde{X} = \frac{1}{\sqrt{5}}AW$, $\widetilde{Y} = \frac{1}{\sqrt{5}}AW'$ with $W, W' \in \mathcal{O}$, recalling that $\mathcal{G}$ is a two-sided ideal we have

$$\widetilde{X}Y = \frac{1}{5}AWAW' = \frac{1}{5}A^2\xi^{-1}(W)W' = \frac{1}{5}A^2W''W',$$

where ξ is the bijection defined in equation (6.7). The last expression ranges over all $W', W'' \in \mathcal{O}$, so that $\mathcal{P} = \frac{1}{5}A^2\mathcal{O}^2$. But since $1 \in \mathcal{O}$, we have $\mathcal{O}^2 = \mathcal{O}$ and $\mathcal{P} = \frac{1}{\sqrt{5}}A\mathcal{G}$. With the notation of §6.3.1, we can say that $\phi(\mathcal{P}) = \frac{1}{\sqrt{5}}A_lR\mathbb{Z}[i]^4 = \frac{1}{\sqrt{5}}RA'_l\mathbb{Z}[i]^4$, where $A'_l = R^{-1}A_lR = R^HA_lR$. A simple calculation yields

$$A'_l = \begin{bmatrix} 1+i & -i & 0 & 0 \\ -i & 1 & 0 & 0 \\ 0 & 0 & 1+i & -i \\ 0 & 0 & -i & 1 \end{bmatrix}$$

The columns of A'_l are a reduced basis for the lattice $A'_l\mathbb{Z}[i]^4$; we already know from Remark 6.19 that the minimal norm in this lattice is 2.

6.4.3 A Golden partition chain

For $k = 1, \ldots, 4$, we consider the left principal ideals of the Golden Code [4]

$$\mathcal{G}_k = \left\{ XB^k \mid X \in \mathcal{G} \right\}, \quad B = \begin{bmatrix} i\bar{\theta} & \bar{\theta} \\ i\theta & i\theta \end{bmatrix}, \tag{6.17}$$

The property that $\det(B) = 1 + i$ makes these subcodes an ideal choice for a binary set partitioning.
We remark that the codes $\mathcal{G}_k$ are obtained from the right principal ideal $\mathcal{O}B^k$ of $\mathcal{O}$. Let

$$W = \begin{bmatrix} a + b\theta & c + d\theta \\ i(c + d\bar{\theta}) & a + b\bar{\theta} \end{bmatrix} \in \mathcal{O}$$

Then the matrix $Q_k = \frac{1}{\sqrt{5}}(B_r)^k A_l \Phi_{\mathcal{O}} = (B_r)^k R$ maps $\mathbf{u} = (a, b, c, d) \in \mathbb{Z}[i]^4$ to $\phi\left(\frac{1}{\sqrt{5}} AWB^k\right)$. $Q_k\mathbb{Z}[i]^4$ is a sublattice of $R\mathbb{Z}[i]^4$, of the form $RS_k\mathbb{Z}[i]^4$, and $S_k = R^{-1}Q_k = R^H Q_k = R^H(B_r)^k R$ is a generator matrix for the lattice associated to $\mathcal{G}_k$.

Lemma 6.21. *Let* $\mathbf{X} = (X_1, \ldots, X_L)$ *with* $X_1, \ldots, X_L \in \mathcal{G}_k$, $k \in \{0, \ldots, 4\}$. *Then*

$$\det(\mathbf{X}\mathbf{X}^H) \geq \frac{2^k N^2}{5},$$

where $N = w_H(\mathbf{X}) = \#\{i \in \{1, \ldots, L\} \mid X_i \neq 0\}$ *is the Hamming weight of the block* $\mathbf{X}$.

Proof. From Proposition 6.15 we get:

$$\begin{aligned}
&\det(\mathbf{X}\mathbf{X}^H) = |\det(X_1)|^2 + \ldots + |\det(X_L)|^2 + \sum_{j>i} \left\| \widetilde{X}_j X_i \right\|_F^2 = \\
&= \frac{1}{25}\Big(\left|\det(AW_1B^k)\right|^2 + \ldots + \left|\det(AW_LB^k)\right|^2 + \sum_{j>i} \left\| \widetilde{B}^k \widetilde{W}_j \widetilde{A} A W_i B^k \right\|_F^2 \Big) = \\
&= \frac{1}{25} |N(\alpha)|^2 |\det B|^{2k} \left(|\det(W_1)|^2 + \ldots + |\det(W_L)|^2 + 2 \right) = \\
&= \frac{2^k}{5} \left(|\det(W_1)|^2 + \ldots + |\det(W_L)|^2 + 2 \right) \geq \frac{2^k}{5} \left(N + 2 \cdot \frac{N(N-1)}{2} \right)
\end{aligned}$$

In the last estimate we have used Remark 6.17 and Remark 6.11. Observe that in particular the Lemma holds for $\mathcal{G}_0 = \mathcal{G}$, with $k = 0$. □

6.5 Coding with cosets: a first example

We will focus on the first of the left principal ideals in the chain (6.17), that is $\mathcal{G}_1 = \mathcal{G}B$. The quotient group $\mathcal{G}/\mathcal{G}_1$ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$; we will consider block codes that are lifts of linear binary codes of length 2 and 3 over the quotient. We will compute the minimum determinants among the pre-images of

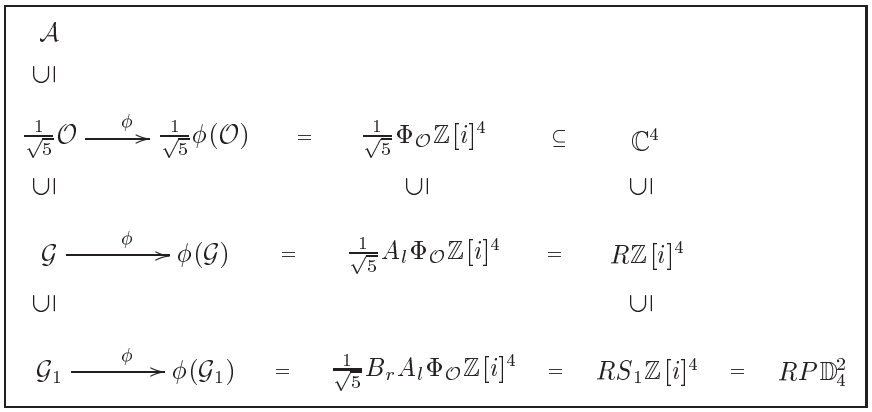

Figure 6.1: A summary of the relations described below.

each binary codeword; these are related to the minima of the Frobenius norms over the products of two cosets.

The lattice $\phi(\mathcal{G}_1)$ is spanned by RS_1, where R is the unitary matrix defined in Remark 6.13, and

$$S_1 = \begin{bmatrix} i & -i & 0 & i \\ -i & 0 & i & i \\ 1 & -1 & 0 & i \\ -1 & 0 & i & i \end{bmatrix}$$

This lattice is equivalent to the complex D_4^2 lattice (see [8], Chapter 7.8): in fact $P_1S_1U_1 = H_1$, where

$$H_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1+i & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1+i \end{bmatrix}$$

is the generator matrix of D_4^2, P_1 is a permutation, U_1 is integer unimodular:

$$P_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad U_1 = \begin{bmatrix} 0 & 0 & 0 & -i \\ 0 & -i & 0 & -i \\ i & i & -i & -i \\ -i & -i & 0 & 0 \end{bmatrix}$$

Then $\phi(\mathcal{G}_1) = RS_1\mathbb{Z}[i]^4 = RP_1^{-1}D_4^2 = RP_1D_4^2 \subset R\mathbb{Z}[i]^4$ is a rotated version of the complex D_4^2 lattice, that is a subgroup of index 4 of $R\mathbb{Z}[i]^4$. The quotient group $\mathcal{G}/\mathcal{G}_1$ is isomorphic to $R\mathbb{Z}[i]^4/D_4^2 \cong \mathbb{Z}_2 \times \mathbb{Z}_2$.

We want to study the behavior of $\left\|\widetilde{X}_1X_2\right\|_F^2$ when X_1, X_2 belong to different cosets of $\mathcal{G}_1$ in $\mathcal{G}$. The map ϕ is a group isomorphism and so the images of distinct cosets of $\mathcal{G}_1$ in $\mathcal{G}$ are distinct cosets (as $\mathbb{Z}[i]$-modules) of $RS_1\mathbb{Z}[i]^4$ in $R\mathbb{Z}[i]^4$, that is they are the images through R of the four cosets of $S_1\mathbb{Z}[i]^4$ in

$\mathbb{Z}[i]^4$.
Let $\{\mathbf{e}_1, \dots, \mathbf{e}_4\}$ denote the canonical basis of $\mathbb{Z}[i]^4$; then

$$C_{00} \div S_1\mathbb{Z}[i]^4 = (i\mathbf{e}_2 + \mathbf{e}_4, i\mathbf{e}_1 + \mathbf{e}_3, \mathbf{e}_2 + \mathbf{e}_4, \mathbf{e}_1 + \mathbf{e}_3) \tag{6.18}$$

To find the coset leaders, remark that $\mathbf{e}_1$ and $\mathbf{e}_2$ do not belong to C_{00} because its nonzero vectors have squared norm greater or equal to 2. Moreover, it is easy to check that $\mathbf{e}_1 - \mathbf{e}_2 \notin C_{00}, \mathbf{e}_1 + \mathbf{e}_2 \notin C_{00}$ and so $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_2$ belong to different cosets of C_{00}:

$$\begin{aligned} C_{01} &= C_{00} + \mathbf{e}_1, \\ C_{10} &= C_{00} + \mathbf{e}_2, \\ C_{11} &= C_{00} + \mathbf{e}_1 + \mathbf{e}_2 \end{aligned}$$

Therefore $\{RC_{00} = \phi(\mathcal{G}_1), RC_{01} = \phi(\mathcal{G}_1) + R\mathbf{e}_1, RC_{10} = \phi(\mathcal{G}_1) + R\mathbf{e}_2, RC_{11} = \phi(\mathcal{G}_1) + R(\mathbf{e}_1 + \mathbf{e}_2)\}$ is a decomposition of $R\mathbb{Z}[i]^4$ into cosets of $RS_1\mathbb{Z}[i]^4$.

Thus the cosets of $\mathcal{G}_1$ in $\mathcal{G}$ are:

$$\mathcal{C}_{00} = \mathcal{G}_1, \quad \mathcal{C}_{01} = \mathcal{G}_1 + \frac{A}{\sqrt{5}}, \quad \mathcal{C}_{10} = \mathcal{G}_1 - \frac{i\widetilde{A}}{\sqrt{5}}, \quad \mathcal{C}_{11} = \mathcal{G}_1 + \frac{1}{\sqrt{5}}(A - i\widetilde{A})$$

, and

$$\widetilde{C}_{00} \div R^{-1}\phi(\widetilde{\mathcal{G}}_1) = (\mathbf{e}_1 - \mathbf{e}_4, \mathbf{e}_2 + \mathbf{e}_3, i\mathbf{e}_1 + \mathbf{e}_4, i\mathbf{e}_2 - \mathbf{e}_3)$$

Unfortunately, the involution $X \mapsto \widetilde{X}$ does not preserve cosets.
The sum of two vectors in $\mathbb{Z}[i]^4$ with even squared norm has even squared norm: if $\mathbf{z} = (z_1, z_2, z_3, z_4)$ and $\mathbf{w} = (w_1, w_2, w_3, w_4)$ are such that $\sum_{i=1}^4 |z_i|^2 = 2n$, $\sum_{i=1}^4 |w_i|^2 = 2m$, then

$$\sum_{i=1}^4 |z_i + w_i|^2 = 2n + 2m + 2\Re(w_i\overline{z_i})$$

The vectors with even squared norm form a sublattice of index 2 of $\mathbb{Z}[i]^4$; it is easy to check that this is equal to $C_{00} \cup C_{11}$. We also observe that $\widetilde{C}_{00} \subset C_{00} \cup C_{11}$.
We want to find

$$\begin{aligned} &\min_{\substack{X\in\mathcal{C}_i,\, Y\in\mathcal{C}_j \\ (X,Y)\neq(0,0)}} \left\| \widetilde{X}Y \right\|_F^2 = \min_{\substack{\mathbf{x}\in C_i,\, \mathbf{y}\in C_j \\ (x,y)\neq(0,0)}} \left\| \phi^{-1}(R\widetilde{\mathbf{x}})\phi^{-1}(R\mathbf{y}) \right\|_F^2 = \\ &= \min_{\substack{\mathbf{x}\in C_i,\, \mathbf{y}\in C_j \\ (x,y)\neq(0,0)}} \left\| (\phi^{-1}R)(\widetilde{\mathbf{x}}\mathbf{y}) \right\|_F^2 = \min_{\substack{\mathbf{x}\in C_i,\, \mathbf{y}\in C_j \\ (x,y)\neq(0,0)}} \|\widetilde{\mathbf{x}} * \mathbf{y}\|^2 \end{aligned}$$

for each pair $(i, j) \in (\mathbb{Z}_2 \times \mathbb{Z}_2)^2$.

Lemma 6.22. *The minimum of the Frobenius norm over the products of the cosets of $\mathcal{G}_1$ is*

$$\min_{\substack{X\in\mathcal{C}_i, Y\in\mathcal{C}_j \\ (X,Y)\neq(0,0)}} \left\| \widetilde{X}Y \right\|_F^2 = \min_{\substack{\mathbf{x}\in C_i, \mathbf{y}\in C_j \\ (\mathbf{x},\mathbf{y})\neq(0,0)}} \|\widetilde{\mathbf{x}} * \mathbf{y}\|^2 = \frac{n(\mathcal{C}_i, \mathcal{C}_j)}{5},$$

where $n(\mathcal{C}_i,\mathcal{C}_j)$ takes the following values:

	$\mathcal{C}_{00}$	$\mathcal{C}_{01}$	$\mathcal{C}_{10}$	$\mathcal{C}_{11}$
$\mathcal{C}_{00}$	0	0	0	0
$\mathcal{C}_{01}$	0	2	3	3
$\mathcal{C}_{10}$	0	3	2	3
$\mathcal{C}_{11}$	0	3	3	2

(6.19)

Proof. The table in (6.19) is clearly symmetrical because $\left\|\widetilde{X}Y\right\|_F = \left\|\widetilde{\widetilde{X}Y}\right\|_F = \left\|\widetilde{Y}X\right\|_F$.
Let $x = R^{-1}\phi(X), y = R^{-1}\phi(Y)$.

- In the case $X \in \mathcal{C}_{00}$,$Y \in \mathcal{C}_i$, we can choose $X = 0$ and $Y \neq 0$ so that the product $\widetilde{X}Y$ is 0.
- From Remarks 6.11 and 6.17 it follows that $\left\|\widetilde{X}Y\right\|_F^2 \geq \frac{2}{5}$ for $X, Y \in \mathcal{G}$. When X, Y are both in $\mathcal{C}_i$, $i \neq 0$, we can exhibit directly an element of squared norm $\frac{2}{5}$:

  $$\widetilde{\mathbf{e}}_1 * \mathbf{e}_1 = i\mathbf{e}_2 * \mathbf{e}_1 = \frac{1}{\sqrt{5}}(\mathbf{e}_1 + i\mathbf{e}_2) \in \widetilde{C}_{01} * C_{01},$$
  $$\widetilde{\mathbf{e}}_2 * \mathbf{e}_2 = -i\mathbf{e}_1 * \mathbf{e}_2 = \frac{1}{\sqrt{5}}(-\mathbf{e}_1 - i\mathbf{e}_2) \in \widetilde{C}_{10} * C_{10},$$
  $$(\widetilde{\mathbf{e}}_1 + \widetilde{\mathbf{e}}_2) * (\mathbf{e}_1 + \mathbf{e}_2) = i(\mathbf{e}_2 - \mathbf{e}_1) * (\mathbf{e}_1 + \mathbf{e}_2) = \frac{1}{\sqrt{5}}(-\mathbf{e}_1 + i\mathbf{e}_2) \in \widetilde{C}_{11} * C_{11}$$

- When $X \in \widetilde{\mathcal{C}}_i$,$Y \in \mathcal{C}_j$, $i, j \neq 0$, $i \neq j$, $\mathbf{xy}$ is of the form

  $$(\widetilde{\mathbf{c}}_0 + \widetilde{\mathbf{e}}_i) * (\mathbf{c}'_0 + \mathbf{e}_j) = (\widetilde{\mathbf{c}}_0 + \widetilde{\mathbf{e}}_i) * \mathbf{c}'_0 + \widetilde{\mathbf{c}}_0 * \mathbf{e}_j + \widetilde{\mathbf{e}}_i * \mathbf{e}_j,$$

  where $\mathbf{c}_0, \mathbf{c}'_0 \in C_{00}$. Then $(\widetilde{\mathbf{c}}_0 + \widetilde{\mathbf{e}}_i) * \mathbf{c}'_0 \in \mathbb{Z}[i]^4 * C_{00} \subset \frac{1}{\sqrt{5}}C_{00}$ and $\widetilde{\mathbf{c}}_0 * \mathbf{e}_j \in \widetilde{C}_{00} * \mathbb{Z}[i]^4 \subset \frac{1}{\sqrt{5}}\widetilde{C}_{00} \subset \frac{1}{\sqrt{5}}(C_{00} \cup C_{11})$ both have even squared norm over 5, while $\widetilde{\mathbf{e}}_i * \mathbf{e}_j$ has odd norm over 5: therefore all the elements in $\widetilde{C}_i * C_j$ have odd norm over 5. Since we know that it is greater or equal to $\frac{2}{5}$, it is sufficient to exhibit an element of squared norm $\frac{3}{5}$:

  $$\widetilde{\mathbf{e}}_1 * \mathbf{e}_2 = i\mathbf{e}_2^2 = \frac{1}{\sqrt{5}}(i\mathbf{e}_1 + (1+i)\mathbf{e}_2) \in \widetilde{C}_{01} * C_{10},$$
  $$\widetilde{\mathbf{e}}_1 * (\mathbf{e}_1 - \mathbf{e}_2) = i\mathbf{e}_2 * (\mathbf{e}_1 - \mathbf{e}_2) = \frac{1}{\sqrt{5}}((1-i)\mathbf{e}_1 - \mathbf{e}_2) \in \widetilde{C}_{01} * C_{11},$$
  $$\widetilde{\mathbf{e}}_2 * (\mathbf{e}_1 + \mathbf{e}_2) = \frac{1}{\sqrt{5}}(-i\mathbf{e}_1 - (i+1)\mathbf{e}_2) \in \widetilde{C}_{10} * C_{11},$$

  which concludes the proof. □

6.5.1 Codes of length 2

For every $(i, j) \in (\mathbb{Z}_2 \times \mathbb{Z}_2)^2 \cong \mathbb{Z}_2^4$, define

$$d(\mathcal{C}_i, \mathcal{C}_j) = \min_{\substack{X_1 \in \mathcal{C}_i, X_2 \in \mathcal{C}_j, \\ (X_1, X_2) \neq (0,0)}} 5\left(|\det(X_1)|^2 + |\det(X_2)|^2 + \left\|\widetilde{X}_2 X_1\right\|_F^2\right) \qquad (6.20)$$

From Lemma 6.22, and recalling that

$$\min_{X\in\mathcal{C}_{00}} |\det(X)| = \frac{2}{5}, \qquad \min_{X\in\mathcal{C}_i} |\det(X)| = \frac{1}{5} \quad \text{for } i \neq 0,$$

it follows that $d(\mathcal{C}_i, \mathcal{C}_j)$ takes the following values:

	$\mathcal{C}_{00}$	$\mathcal{C}_{01}$	$\mathcal{C}_{10}$	$\mathcal{C}_{11}$
$\mathcal{C}_{00}$	2	1	1	1
$\mathcal{C}_{01}$	1	4	5	5
$\mathcal{C}_{10}$	1	5	4	5
$\mathcal{C}_{11}$	1	5	5	4

(6.21)

Let $\mathcal{S}$ be a subgroup of $\mathbb{Z}_2^4$: then we can define two *coset weight enumerator functions* as follows:

$$A_{\mathcal{S}}(D) = \sum_{(i,j)\in\mathcal{S}} D^{d(\mathcal{C}_i,\mathcal{C}_j)}, \quad \hat{A}_{\mathcal{S}}(D) = \sum_{(i,j)\in\mathcal{S}\setminus\{0\}} D^{d(\mathcal{C}_i,\mathcal{C}_j)}$$

Given a polynomial $p(D) = a_0 + a_1 D + \ldots + a_n D^n$, we define

$$\Delta(p) = \min\{i \geq 0 \mid a_i \neq 0\}$$

and we introduce the following order on the set of all polynomials:

$$p >_\Delta q \Leftrightarrow \exists k \geq 0 \ : \ a_i(p) = a_i(q) \text{ for } i = 0, \ldots, k-1, \ a_k(p) > a_k(q)$$

In particular, we are interested in maximizing $\Delta_{\min}(\mathcal{S}) = \Delta(A_{\mathcal{S}})$ and $\hat{\Delta}_{\min}(\mathcal{S}) = \Delta(\hat{A}_{\mathcal{S}})$ with respect to this order.

Linear codes induced by a permutation

Let σ be any permutation of $\mathbb{Z}_2 \times \mathbb{Z}_2$, and consider the code of length 2

$$\mathcal{G}_\sigma = \{(X_1, X_2) \in \mathcal{G} \times \mathcal{G} \mid X_1 \in \mathcal{C}_i, X_2 \in \mathcal{C}_{\sigma(i)} \text{ for some } i\} \tag{6.22}$$

In general, $\mathcal{G}_\sigma$ is not a linear code: given two words $X = (X_1, X_2) \in \mathcal{C}_i \times \mathcal{C}_{\sigma(i)}$, $Y = (Y_1, Y_2) \in \mathcal{C}_j \times \mathcal{C}_{\sigma(j)}$, we have $X - Y \in \mathcal{C}_{i+j} \times \mathcal{C}_{\sigma(i)+\sigma(j)}$. Therefore the code is linear if and only if

$$\sigma(i) + \sigma(j) = \sigma(i+j) \qquad \forall (i,j) \in (\mathbb{Z}_2 \times \mathbb{Z}_2)^2 \tag{6.23}$$

Thus a necessary condition for linearity is that $\sigma(i) + \sigma(0) = \sigma(i)$ for all i, that is $\sigma(0) = 0$. In this particular case it is also a sufficient condition: $\forall i, j \neq 0$ such that $i \neq j$, $i + j$ is equal to the only element k in $\{01, 10, 11\} \setminus \{i, j\}$. Then $\sigma(i) + \sigma(j) = \sigma(k) = \sigma(i + j)$.

Lemma 6.23. *Consider the code $\mathcal{G}_\sigma$ defined in (6.22), such that σ is a permutation keeping 0 fixed, and let $\mathcal{S} = \{(i, \sigma(i)) \mid i \in \mathbb{Z}_2 \times \mathbb{Z}_2\}$. Then we have three cases:*

1. *If 0 is the only fixed point of σ, $A_{\mathcal{S}}(D) = D^2 + 3D^5$, $\hat{A}_{\mathcal{S}}(D) = 3D^5$.*
2. *If σ has two fixed points, $A_{\mathcal{S}}(D) = D^2 + D^4 + 2D^5$, $\hat{A}_{\mathcal{S}}(D) = D^4 + 2D^5$.*
3. *If σ is the identity (so that $\mathcal{G}_\sigma$ is the repetition code), $A_{\mathcal{S}}(D) = D^2 + 3D^4$, $\hat{A}_{\mathcal{S}}(D) = 3D^4$.*

The proof is straightforward from the table (6.21). Clearly the first case is the best one with respect to the coding gain.

Non linear codes induced by a permutation

Actually the study of the non linear case can be reduced to the linear one. We know that in this case $\sigma(0) \neq 0$. Given $X = (X_1, X_2) \in \mathcal{C}_i \times \mathcal{C}_{\sigma(i)}$, $Y = (Y_1, Y_2) \in \mathcal{C}_j \times \mathcal{C}_{\sigma(j)}$, we have

$$\det(X-Y)(X-Y)^H = \det((X_1-Y_1)(X_1-Y_1)^H + (X_2-Y_2)(X_2-Y_2)^H) =$$
$$= |\det(X_1-Y_1)|^2 + |\det(X_2-Y_2)|^2 + \left\|(\widetilde{X}_2-\widetilde{Y}_2)(X_1-Y_1)\right\|_F^2,$$

where $X_1 - Y_1 \in \mathcal{C}_{i+j}$, $X_2 - Y_2 \in \mathcal{C}_{\sigma(i)+\sigma(j)}$.
We want to show that $(i+j, \sigma(i)+\sigma(j))$ is a permutation keeping 0 fixed (and thus (X_1-Y_1, X_2-Y_2) all belong to the same linear code as X, Y vary).
Let $\tau(i+j) = \sigma(i)+\sigma(j)$: τ is well-defined because the sum of two distinct elements $i, j \in \mathbb{Z}_2 \times \mathbb{Z}_2$ is equal to the sum of the elements in $\mathbb{Z}_2 \times \mathbb{Z}_2 \setminus \{i, j\}$. In fact if $i+j = k+m$, either $i+j = 0$ and $i = j, k = m$, so that $\tau(i+j) = 2\sigma(i) = 2\sigma(k) = 0$, or $\{i, j, k, m\} = \mathbb{Z}_2 \times \mathbb{Z}_2$, and $\sigma(i)+\sigma(j) = \sigma(k)+\sigma(m)$.
Moreover, τ is surjective because it is the original permutation σ shifted by $\sigma(0)$: $\tau(j) = \sigma(j)+\sigma(0)$. And $\tau(0) = \sigma(0)+\sigma(0) = 0$.
It follows that, from the point of view of the coding gain, our code is equivalent to the linear code induced by τ.

Other linear codes of length 2

In the case of linear codes not induced by permutations, $\mathcal{S}$ contains at least two points with the same first or second coordinate, of the form (i, j) and (i, k) or (j, i) and (k, i). But then $(0, t)$ or $(t, 0)$ belong to $\mathcal{S}$ for some t, and the coset weight enumerator polynomial has $\Delta_{\min} = 1$ (see table 6.21), so these codes have a worse performance than those we analyzed in the previous paragraphs.

6.5.2 Codes of length 3

Let $\mathcal{U} = \{(i, j, k) \mid i, j, k \in \mathbb{Z}_2 \times \mathbb{Z}_2\}$. For the sake of simplicity, we only consider linear codes $\mathcal{S} \subset \mathcal{U}$. Similarly to the case of length 2, for every $(i, j, k) \in \mathcal{U}$ we can define

$$d(\mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k) = \min_{\substack{X \in \mathcal{C}_i, Y \in \mathcal{C}_j, Z \in \mathcal{C}_k \\ (X,Y,Z) \neq (0,0,0)}} 5\left(|\det(X)|^2 + |\det(Y)|^2 + |\det(Z)|^2 + \right.$$
$$\left. + \left\|\widetilde{Y}X\right\|_F^2 + \left\|\widetilde{Z}X\right\|_F^2 + \left\|\widetilde{Z}Y\right\|_F^2\right),$$

and the coset weight enumerator polynomials

$$A_{\mathcal{S}}(D) = \sum_{(i,j,k) \in \mathcal{S}} D^{d(\mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k)}, \quad \hat{A}_{\mathcal{S}}(D) = \sum_{(i,j,k) \in \mathcal{S} \setminus \{0\}} D^{d(\mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k)}$$

For a given dimension n of the code as an $\mathbb{F}_2$-vector space, we search for the best possible coset weight enumerator polynomial with respect to $>_\Delta$.
The 64 codewords $(i, j, k) \in \mathcal{U}$ can be divided in several groups according to $d = d(\mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k)$. Let $a, b, c \in \mathbb{Z}_2 \times \mathbb{Z}_2$ be distinct and nonzero, and let π be a permutation; we can summarize the different cases as in Table 6.5.2.
If $\mathcal{S}$ is equal to the whole space $\mathcal{U}$, its coset weight generator function is

$$A_{\mathcal{S}}(D) = 9D + D^2 + 9D^4 + 18D^5 + 3D^9 + 18D^{11} + 6D^{12}$$

shape	# of codewords	$d(\mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k)$
$\pi(a,b,c)$	6	$d = 1+1+1+3+3+3 = 12$
$\pi(a,a,b)$	18	$d = 1+1+1+3+3+2 = 11$
$\pi(a,a,a)$	3	$d = 1+1+1+2+2+2 = 9$
$\pi(0,a,b)$	18	$d = 1+1+0+3+0+0 = 5$
$\pi(0,a,a)$	9	$d = 1+1+0+2+0+0 = 4$
$\pi(0,0,a)$	9	$d = 1$
$(0,0,0)$	1	$d = 2$

Table 6.2: A list of the codewords in $\mathcal{U}$ according to their shape.

Subgroups of order 4

In this case, it is clear that the code

$$\mathcal{S} = \{(00,00,00), (01,10,11), (10,11,01), (11,01,10)\}$$

gives rise to the best possible coset weight enumerator polynomial, that is

$$A_{\mathcal{S}(D)} = D^2 + 3D^{12}, \quad \hat{A}_{\mathcal{S}(D)} = 3D^{12}, \quad \hat{\Delta}(\mathcal{S}) = 12$$

Subgroups of order 8

In this case $\mathcal{S}$ has dimension 3 as an $\mathbb{F}_2$-vector space. We define the following subspaces of dimension 4 over $\mathbb{F}_2$: $\mathcal{I}_0 = \{(i,j,k) \mid i = 0\}$, $\mathcal{J}_0 = \{(i,j,k) \mid j = 0\}$, $\mathcal{K}_0 = \{(i,j,k) \mid k = 0\}$.
Then Grassmann's formula implies

$$\begin{aligned} 6 \geq \dim(\mathcal{I}_0 + \mathcal{S}) &= \dim(\mathcal{S}) + \dim(\mathcal{I}_0) - \dim(\mathcal{S} \cap \mathcal{I}_0) = 7 - \dim(\mathcal{S} \cap \mathcal{I}_0) \\ \Rightarrow \dim(\mathcal{S} \cap \mathcal{I}_0) &\geq 1 \end{aligned}$$

Similarly, $\dim(\mathcal{S} \cap \mathcal{J}_0) \geq 1$, $\dim(\mathcal{S} \cap \mathcal{K}_0) \geq 1$. So, even in the best case when $\dim(\mathcal{S} \cap \mathcal{I}_0) = \dim(\mathcal{S} \cap \mathcal{J}_0) = \dim(\mathcal{S} \cap \mathcal{K}_0) = 1$, we have at least three nonzero codewords with one digit equal to zero, giving at best a term $3D^5$ in the coset weight enumerator polynomial.
Now consider the code $\mathcal{S}$ generated by $(00,01,10), (01,10,00), (11,11,01)$:

$$\begin{aligned} \mathcal{S} = \{&(00,00,00), (00,01,10), (01,10,00), (11,11,01), (01,11,10), \\ &(11,10,11), (10,01,01), (10,00,11)\} \end{aligned}$$

It is easy to check that

$$A_{\mathcal{S}}(D) = D^2 + 3D^5 + 3D^{11} + D^{12}, \qquad \hat{\Delta}(\mathcal{S}) = 5$$

This is the best case for $\dim(\mathcal{S}) = 3$. Suppose by contradiction that there exists $\mathcal{S}'$, $\dim(\mathcal{S}') = 3$, such that $A_{\mathcal{S}'}(D) >_\Delta A_{\mathcal{S}}(D)$. The term D^2 is always present in a linear code, and we have already observed that the term $3D^5$ cannot be avoided. The only possibility would be to have (at least) one more word (a,b,c) such that $d(\mathcal{C}_a, \mathcal{C}_b, \mathcal{C}_c) = 12$.
Suppose we already have one such word (i,j,k), where i,j,k are all distinct and different from 00, and we want to add another, (a,b,c).

- If we choose $a = i$, then since $(a, b, c) \neq (i, j, k)$ we must choose $(a, b, c) = (i, k, j)$. But then $(i, j, k) + (i, k, j) = (i, i, i) \in \mathcal{S}'$, giving rise to the term D^9 in $A_{\mathcal{S}'}(D)$, and thus $A_{\mathcal{S}'} <_\Delta A_{\mathcal{S}}$.
- Then necessarily $a \neq i$. Without loss of generality (since the order of the digits is uninfluential) we can suppose $a = j$. Then (a, b, c) is either (j, i, k) or (j, k, i). In the first case, $(i, j, k) + (j, i, k) = (k, k, 0) \in \mathcal{S}'$, introducing the term D^4 in $A_{\mathcal{S}'}(D)$; but this is impossible because $A_{\mathcal{S}'} >_\Delta A_{\mathcal{S}}$ by hypothesis. Then the only possibility we have left is that $(i, j, k), (j, k, i), (i, j, k) + (j, k, i) = (k, i, j) \in \mathcal{S}'$.

Now we know that there is a nonzero element in $\mathcal{I}_0 \cap \mathcal{S}'$, of the form $(0, a, b)$; we have to discard the option $(a, b) \in \{(j, k), (k, i), (i, j)\}$ because in that case $\mathcal{S}'$ would contain a codeword with two digits equal to 00 (in fact $(i, j, k)+(0, j, k) = (i, 0, 0)$, $(j, k, i)+(0, k, i) = (j, 0, 0)$, and $(k, i, j)+(0, i, j) = (k, 0, 0)$ respectively). On the other hand, if $(a, b) \in \{(k, j), (i, k), (j, i)\}$, then the code would contain the codeword $(i, i, i) = (i, j, k)+(0, k, j)$, or $(j, j, j) = (j, k, i)+(0, i, k)$,$(k, k, k) = (k, i, j) + (0, j, i)$ respectively, and again we would have $A_{\mathcal{S}'} <_\Delta A_{\mathcal{S}}$.

Subgroups of order 16

We have $\dim(\mathcal{S}) = 4$. Grassmann's formula implies

$$6 \geq \dim(\mathcal{I}_0 + \mathcal{S}) = \dim(\mathcal{S}) + \dim(\mathcal{I}_0) - \dim(\mathcal{S} \cap \mathcal{I}_0) = 8 - \dim(\mathcal{S} \cap \mathcal{I}_0)$$
$$\Rightarrow \dim(\mathcal{S} \cap \mathcal{I}_0) \geq 2,$$

and in the same fashion we find that $\dim(\mathcal{S} \cap \mathcal{J}_0) \geq 2$, $\dim(\mathcal{S} \cap \mathcal{K}_0) \geq 2$. Then there are at least nine nonzero codewords with one digit equal to 00, and $A_{\mathcal{S}}(D)$ contains at the least the term $9D^5$.
Consider the code generated by $(00, 01, 10)$,$(00, 11, 01)$,$(11, 01, 00)$,$(10, 00, 01)$:

$$\mathcal{S} = \{(00, 00, 00), (00, 01, 10), (00, 11, 01), (11, 01, 00), (10, 00, 01), (00, 10, 11),$$
$$(11, 00, 10), (10, 01, 11), (11, 10, 01), (10, 11, 00), (01, 01, 01), (11, 11, 11),$$
$$(10, 10, 10), (01, 00, 11), (01, 10, 00), (01, 11, 10)\}$$

Its coset weight generator polynomial is

$$A_{\mathcal{S}}(D) = D^2 + 9D^5 + 3D^9 + 3D^{12}$$

and $\hat{\Delta}(\mathcal{S}) = 5$. Again, we want to show that this code is optimal under our conditions, that is, that the presence of the term $3D^9$ can't be avoided.
We need to choose three elements in $\mathcal{S} \cap \mathcal{I}_0$, without repeated digits; if the first is $(0, i, j)$, the second cannot be $(0, j, i)$ because then $(0, k, k) \in \mathcal{S}$ and $A_{\mathcal{S}}$ would contain the term D^4; it cannot be $(0, i, k)$ or $(0, k, j)$ because its sum with the first element would give one codeword with two digits equal to 00. So the three codewords have to be $\{(0, i, j), (0, j, k), (0, k, i)\}$.
Now let's consider the three nonzero codewords in $\mathcal{S} \cap \mathcal{J}_0$: as in the previous case, the choice of the first element determines the others and we can choose either the triple $\{(i, 0, j), (j, 0, k), (k, 0, i)\}$ or $\{(j, 0, i), (i, 0, k), (k, 0, j)\}$.
In the first case, $(0, i, j) + (i, 0, j) = (i, i, 0) \in \mathcal{S}$, and $A_{\mathcal{S}}$ would include a term D^4, so this option must be discarded.
In the second case, $(0, i, j) + (i, 0, k) = (i, i, i)$, $(0, j, k) + (j, 0, i) = (j, j, j)$, $(0, k, i)+(k, 0, j) = (k, k, k)$, which proves that the term $3D^9$ in the coset weight enumerator polynomial can't be avoided.

Subgroups of order 32

Let's consider the following subspaces of dimension 2 over $\mathbb{F}_2$:

$$\mathcal{I} = \{(i,j,k) \mid j = k = 0\},\ \mathcal{J} = \{(i,j,k) \mid i = k = 0\},\ \mathcal{K} = \{(i,j,k) \mid i = j = 0\}$$

Since $\dim(\mathcal{S}) = 5$, Grassmann's formula implies that $\mathcal{S} \cap \mathcal{I}$, $\mathcal{S} \cap \mathcal{J}$, $\mathcal{S} \cap \mathcal{K}$ have dimension at least one. Then $A_{\mathcal{S}}$ contains the term D, and these codes are not efficient from the point of view of the determinant.

6.6 Structure of the quotient rings of $\mathcal{G}$

In this section we describe the structure of the two-sided ideals of $\mathcal{G}$ and of the corresponding quotients. As we have seen, the multiplicative structure of $\mathcal{G}$ plays an important role when computing the mixed terms in the minimum determinant, and it is convenient for the quotient to have a ring structure.
In particular, we want to find all the two-sided ideals whose norm is a power of $1+i$; these can be especially useful to build a binary partition chain like the one described in §6.4.3. Unfortunately, we will see that the only two-sided ideals with this property are the trivial ones. We then analyze the structure of the quotient rings, and find that they are rings of matrices over non-integral rings.

First of all, we need some further notions from non-commutative algebra; we will see that the existence of two-sided ideals is related to the ramification of primes over the base field. We will also show that $\mathcal{O}$ is a maximal order of $\mathcal{A}$.

6.6.1 Ideals, valuations and maximal orders

6.24 (Prime ideals). Let Θ be an order, $\mathfrak{P}$ a two-sided ideal of Θ (that is, the left and right order of I coincide with Θ). $\mathfrak{P}$ is *prime* if it is nonzero and $\forall I, J$ integer two-sided ideals of Θ, $IJ \subset \mathfrak{P} \Rightarrow I \subset \mathfrak{P}$ or $J \subset \mathfrak{P}$.

The proofs of the following theorems can be found in Reiner's book [19]:

Theorem 6.25. *The two-sided ideals of an order Θ form a free group generated by the prime ideals.*

Theorem 6.26. *Let Θ be a maximal order in a quaternion algebra $\mathcal{H}$. Then the prime ideals of Θ coincide with the maximal two-sided ideals of Θ, and there is a one-to-one correspondence between the prime ideals $\mathfrak{P}$ in $\mathcal{H}$ and the prime ideals P of R, given by $P = R \cap \mathfrak{P}$.*
Moreover, $\Theta/\mathfrak{P}$ is a simple algebra over the finite field R/P.

6.27 (Valuations and local fields). A *valuation* v of K is a positive real function of K such that $\forall k, h \in K$,

1. $v(k) = 0 \Leftrightarrow k = 0$,
2. $v(kh) = v(k)v(h)$,
3. $v(k+h) \leq v(k) + v(h)$.

v is *non-archimedean* if $v(k+h) \leq \max(v(k), v(h))\ \forall k, h \in K$; it is *discrete* if $v(K^*)$ is an infinite cyclic group.
K can be endowed with a topology induced by v in the following way: a neighborhood basis of a point k is given by the sets

$$U_\varepsilon(k) = \{h \in K \mid v(h-k) < \varepsilon\}$$

K will be called *complete* if it is complete with respect to this topology.
If v is non archimedean, the set

$$R_v = \{k \in K \mid v(k) \leq 1\}$$

is a local ring, called the *valuation ring* of v. The quotient R_v/P_v, where P_v is the unique maximal ideal of R_v, is called the *field of residues* of K.
K is a *local field* if it is complete with respect to a discrete valuation v and if R_v/P_v is finite.

6.28 (Places). A *place* v of K is an immersion $i_v : K \to K_v$ into a local field K_v. If v is non-archimedean, we say that it is a *finite place*; otherwise, that it is an *infinite place*.

The finite places of K arise from discrete P-adic valuations of K, where P ranges over the maximal ideals in the ring of integers R of K. (Recall that the ring of integers in a number field is always a Dedekind domain, and so the maximal ideals coincide with the prime ideals).

Here we recall a few well-known facts about ramification in commutative extensions: let L be a separable extension of degree d of K, and P a prime ideal of K. In general, P is not a prime in L, and admits a unique decomposition with respect to the primes of L:

$$P = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r},$$

where $p_1, \ldots, p_r$ are distinct and are called the *primes over* P.

Proposition 6.29. *Let $\mathcal{O}_L$ be the ring of integers of L: then for each p_i over P, $\mathcal{O}_L/p_i$ is a finite extension of the finite field R/P. The degree f_i of this extension is called the* inertial degree *of p_i over P, and the following relation holds:*

$$d = \sum_{i=1}^{r} e_i f_i$$

6.30. We say that P

- is *ramified* if $\exists i$ such that $e_i > 1$
- is *totally ramified* if $r = 1$, $e_1 = d$, $f_1 = 1$, that is, $P = p^d$,
- is *unramified* if $e_i = 1\ \forall i = 1, \ldots, r$,
- *splits completely* if it is unramified and $r = d$, that is, $f_i = 1\ \forall i = 1, \ldots, r$,
- is *inert* if $f_i = 1\ \forall i = 1, \ldots, r$.

6.31 (Ramified places). Let $\mathcal{H}$ be a quaternion algebra over K, and P a place of K.
Consider the K-module $\mathcal{H}_P = \mathcal{H} \otimes_K K_P$; $\mathcal{H}_P$ is isomorphic to a matrix algebra $M_r(D)$ over a skew field D of center K_P and index m_P over K_P; m_P is called the *local index* of $\mathcal{H}$ at P. We say that P is *ramified* in $\mathcal{H}$ if $m_P > 1$.

Given a maximal order Θ, the set $\mathrm{Ram}(\mathcal{H})$ of ramified places of $\mathcal{H}$ is related to a particular two-sided ideal of Θ:

6.32 (Different and discriminant). Let Θ be an order. The set

$$\Theta^* = \{x \in \mathcal{H} \mid tr(x\Theta) \subset R\}$$

is a two-sided ideal, called the *dual* of Θ. Its inverse $\mathfrak{D} = (\Theta^*)^{-1}$ is a two-sided integral ideal, called the *different* of Θ. If $\{w_1, \ldots, w_4\}$ is a basis of Θ as a free R-module,

$$(n(\mathfrak{D}))^2 = R \det(\mathrm{tr}(w_i w_j))$$

The ideal $n(\mathfrak{D})$ of R is called the *reduced discriminant* of Θ and is denoted by $d(\Theta)$.

Proposition 6.33. *If Θ, Θ' are two orders and $\Theta' \subsetneq \Theta$, then $d(\Theta') \subsetneq d(\Theta)$.*

The notion of ramification for quaternion algebras is a generalization of the notion of ramification for field extensions:

Theorem 6.34. *Let Θ be a maximal order in $\mathcal{H}$. For each place P of K, let m_P be the local index of $\mathcal{H}$ at P, and let $\mathfrak{P}$ be the prime ideal of Θ corresponding to P (see Theorem 6.26). Then $m_P > 1$ only for a finite number of places P, and*

$$P\Theta = \mathfrak{P}^{m_P}, \qquad \mathfrak{D} = \prod_{P \in \mathrm{Ram}(\mathcal{H})} \mathfrak{P}^{m_P - 1}$$

Proposition 6.35. *Let $\mathcal{H}$ be a quaternion algebra unramified at infinity.*
A necessary and sufficient condition for an order Θ to be maximal is that

$$d(\Theta) = \prod_{P \in \mathrm{Ram}(\mathcal{H}) \setminus \infty} P$$

In the case of infinite places P, the P-adic completion can be $\mathbb{R}$ (*real primes*) or $\mathbb{C}$ (*complex primes*). Complex primes are never ramified. [19]

Theorem 6.36. *The two-sided ideals of a maximal order Θ form a commutative group generated by the ideals of R and the ideals of reduced norm P, where P varies over the prime ideals of R that are ramified in $\mathcal{H}$.*
Moreover, consider the normalizer $N(\Theta) = \{x \in \mathcal{H} \mid x^{-1}\Theta x = \Theta\}$ *of Θ. Then $\frac{N(\Theta)}{K^*\Theta^*} \cong \mathbb{Z}_2^m$, where m is the number of prime divisors of $\mathfrak{D}$.*

6.6.2 Structure of the quotient rings of $\mathcal{G}$

Recall the definition of the division algebra $\mathcal{A}$ in (6.3):

$$\mathcal{A} = \left\{ \begin{bmatrix} x_1 & x_2 \\ i\bar{x}_2 & \overline{x}_1 \end{bmatrix},\ x_1, x_2 \in \mathbb{Q}(i, \theta) \right\}$$

As we have seen in §6, $\mathcal{O} = \mathbb{Z}[i,\theta] \oplus \mathbb{Z}[i,\theta]j$ is a $\mathbb{Z}[i]$-order of $\mathcal{A}$, and the Golden Code is (up to a scaling factor $\frac{1}{\sqrt{5}}$) the ideal $\mathcal{G} = \alpha\mathcal{O}$, where $\alpha = 1 + i\bar{\theta}$; we have seen that it is a two-sided principal ideal.
$\mathcal{G}$ is also a prime ideal because of Theorem 6.26, since $\mathcal{G} \cap \mathbb{Z}[i] = (2+i)$ is a prime ideal of $\mathbb{Z}[i]$.
Then $\mathcal{O}/\sqrt{5}\mathcal{G}$ is a simple algebra over $\mathbb{Z}[i]/(2+i) \cong \mathbb{F}_5$.

Observe that the prime ideals $(2+i)$ and $(2-i)$ of $\mathbb{Z}[i]$ are both ramified in $\mathcal{A}$: in fact

$$(2+i) = (\alpha)^2, \text{ and } (2-i) = (\alpha')^2, \text{ where } \alpha' = 1 - i\bar{\theta}$$

(Remark that $\alpha = i\theta\bar{\alpha}$, $\alpha' = -i\bar{\theta}\bar{\alpha'}$).

Proposition 6.37. *$\mathcal{O}$ is a maximal order.*

Proof. In the case of $\mathcal{O}$, the infinite primes are complex because $\mathbb{Q}(i)$ is complex. Therefore the conditions of Proposition 6.35 are satisfied.
We can compute $\det(\mathrm{tr}(w_i w_j))$, for the basis $\{w_1 = 1, w_2 = \theta, w_3 = j, w_4 = \theta j\}$ of $\mathcal{O}$:

$$(w_i w_j)_{1\leq i,j\leq 4} = \begin{pmatrix} 1 & \theta & j & \theta j \\ \theta & \theta^2 & \theta j & \theta^2 j \\ j & \bar{\theta} j & i & i\bar{\theta} \\ \theta j & -j & \theta i & -i \end{pmatrix}$$

$$\det(\mathrm{tr}(w_i w_j)) = \det \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 0 & 0 & 2i & i \\ 0 & 0 & i & -2i \end{pmatrix} = 25$$

Therefore $d(\mathcal{O}) = 5\mathbb{Z}[i]$.
If $\mathcal{O}$ were strictly contained in another order $\mathcal{O}'$, $d(\mathcal{O}')$ would be strictly larger than $5\mathbb{Z}[i]$. But we know from Proposition 6.35 that $d(\mathcal{O}')$ would be the product of all ramified primes of $\mathcal{H}$; in particular it should be contained in $(2+i)$ and $(2-i)$. But then it would be contained in $5\mathbb{Z}[i]$, a contradiction. Therefore $\mathcal{O}$ is a maximal order, and $\mathcal{G}$ is a normal ideal. □

As a consequence of the fact that $\mathcal{O}$ is maximal, from Proposition 6.35 we also learn that $2+i$ and $2-i$ are the only ramified primes in $\mathcal{H}$.
Then Theorem 6.36 implies that the prime two-sided ideals of $\mathcal{O}$ are either of the form $p\mathcal{O}$, where p prime in $\mathbb{Z}[i]$, or belong to $\{\alpha\mathcal{O}, \alpha'\mathcal{O}\}$.
It follows that the only two-sided ideals of $\mathcal{G}$ whose norm is a power of $1+i$ are the trivial ideals of the form $(1+i)^k\mathcal{G}$.

The quotient ring $\mathcal{G}/(1+i)\mathcal{G}$

Consider the prime ideal $(1+i)\mathcal{O}$. It is easy to check that $\mathcal{G}$ and $(1+i)\mathcal{O}$ are *coprime* ideals, that is $\mathcal{G} + (1+i)\mathcal{O} = \mathcal{O}$ and as a consequence, $\mathcal{G} \cap (1+i)\mathcal{O} = \mathcal{G}(1+i)\mathcal{O} = (1+i)\mathcal{G}$. Recall the following basic result:

Theorem 6.38 (Third Isomorphism Theorem for rings). *Let I and J be ideals in a ring R. Then*

$$\frac{I}{I \cap J} \cong \frac{I+J}{J}$$

Putting $I = \mathcal{G}$ and $J = (1+i)\mathcal{O}$, we get

$$\frac{\mathcal{G}}{(1+i)\mathcal{G}} \cong \frac{\mathcal{O}}{(1+i)\mathcal{O}} \tag{6.24}$$

If $\pi_{\mathcal{G}} : \mathcal{G} \to \mathcal{G}/(1+i)\mathcal{G}$ and $\pi_{\mathcal{O}} : \mathcal{O} \to \mathcal{O}/(1+i)\mathcal{O}$ are the canonical epimorphisms, the ring isomorphism in (6.24) is simply given by $\phi_{\mathcal{G}}(g) \mapsto \phi_{\mathcal{O}}(g)$.
Theorem 6.26 implies that $\mathcal{O}/(1+i)\mathcal{O}$ is a simple algebra over $\mathbb{Z}[i]/(1+i) \cong \mathbb{F}_2$. We denote the image of $x \in \mathcal{O}$ through the canonical epimorphism $\mathcal{O} \to \mathcal{O}/(1+i)\mathcal{O}$ with $[x]$.

Lemma 6.39. *$\mathcal{O}/(1+i)\mathcal{O}$ is isomorphic to the ring $M_2(\mathbb{F}_2)$ of 2×2 matrices over $\mathbb{F}_2$.*

Proof. Recall the following well-known lemma [14]:

Lemma 6.40. *Let R be a ring with identity, I a proper ideal of R, M a free R-module with basis X and $\pi : M \to M/IM$ the canonical epimorphism. Then M/IM is a free R/I-module with basis $\pi(X)$ and $|\pi(X)| = |X|$.*

We know that $\mathcal{O}/(1+i)\mathcal{O}$ is a $\mathbb{Z}[i]$-module; the lemma implies that it is also a free $\mathbb{Z}[i]/(1+i)$-module, that is a vector space over $\mathbb{F}_2$, whose basis is $\{[1], [\theta], [j], [\theta j]\}$. We define an homomorphism of $\mathbb{F}_2$-vector spaces $\varphi : \mathcal{O}/(1+i)\mathcal{O} \to M_2(\mathbb{F}_2)$ by specifying the images of the basis:

$$\varphi([1]) = \mathbb{1}, \quad \varphi([\theta]) = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \varphi([j]) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \varphi([\theta j]) = \varphi([\theta])\varphi([j])$$

It is one-to-one since $\varphi([1]), \varphi([\theta]), \varphi([j]), \varphi([\theta j])$ are linearly independent.
In order to prove that φ is also a ring homomorphism, it is sufficient to verify that $\varphi(w_i w_j) = \varphi(w_i)\varphi(w_j)$ for all pairs of basis vectors w_i, w_j; the only non-trivial cases are

$$\varphi([\theta])^2 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}^2 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}^2 = \varphi([\theta+1]) = \varphi([\theta^2]),$$

$$\varphi([j])^2 = \mathbb{1} = \varphi([1]) = \varphi([i]),$$

$$\varphi([\bar{\theta} j]) = \varphi([(1+\theta)j]) = \varphi([j+\theta j]) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} = \varphi([j])\varphi([\theta])$$

□

Remark 6.41. Clearly $M_2(\mathbb{F}_2)$ has no proper two-sided ideals; its only left ideals are

$$I_0 = \left\{ \begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix} : \quad a, b \in \mathbb{F}_2 \right\},$$

$$I_1 = \left\{ \begin{pmatrix} a & a \\ b & b \end{pmatrix} : \quad a, b \in \mathbb{F}_2 \right\},$$

$$I_\infty = \left\{ \begin{pmatrix} a & 0 \\ b & 0 \end{pmatrix} : \quad a, b \in \mathbb{F}_2 \right\},$$

all of index 4.

Recall that $\mathcal{G}$ is isometric to $\sqrt{5}\mathbb{Z}[i]^4$, and a canonical basis is given by

$$\mathbf{e}_1 = \alpha, \qquad \mathbf{e}_2 = \alpha\theta, \qquad \mathbf{e}_3 = \alpha j, \qquad \mathbf{e}_4 = \theta\alpha j \tag{6.25}$$

The corresponding elements of $M_2(\mathbb{F}_2)$ are

$$\mathbf{e}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{e}_3 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{e}_4 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{6.26}$$

Remark that the Hamming weight of a word in $\mathcal{G}/(1+i)\mathcal{G}$ with respect to the basis (6.26) is equal to the minimum Euclidean norm over all the lattice points in the corresponding coset.
It is easy to check that the only invertible elements (the matrices with full rank) are

$$\mathbf{e}_1,\ \mathbf{e}_2,\ \mathbf{e}_3,\ \mathbf{e}_4,\ \mathbf{e}_1 + \mathbf{e}_2 = \mathbb{1},\ \mathbf{e}_3 + \mathbf{e}_4 = \varphi(j)$$

It is easy to see that their lifts to $\mathcal{G}$ of non-invertible elements have a higher determinant:

Remark 6.42. If $X \in M_2(\mathbb{F}_2) \setminus \{0\}$ is non-invertible,

$$\min_{Y \in \mathcal{G},\ [Y]=X} |\det(Y)|^2 \geq 2$$

Proof. $\pi_{\mathcal{G}}(Y)$ is non-invertible in $\mathcal{G}/(1+i)\mathcal{G}$ if and only if its determinant is non-invertible in $\mathbb{Z}[i]/(1+i)$, that is, $\det(Y) = \widetilde{Y}Y \in (1+i) \setminus \{0\}$ if $X \neq 0$, since $\mathcal{H}$ is a division ring.
Then $\left|\det(\widetilde{Y}Y)\right| = |\det(Y)|^2 \geq 2$. □

The quotient ring $\mathcal{G}/2\mathcal{G}$

Again, $\mathcal{G}$ and $2\mathcal{O}$ are coprime and so $\frac{\mathcal{G}}{2\mathcal{G}} \cong \frac{\mathcal{O}}{2\mathcal{O}}$.

Lemma 6.43. *$\mathcal{O}/2\mathcal{O}$ is isomorphic to the ring $M_2(\mathbb{F}_2[i])$ of 2×2 matrices over the ring $\mathbb{F}_2[i]$.*

Proof. First of all, Lemma 6.40 implies that $\mathcal{O}/2\mathcal{O}$ is a free $\mathbb{Z}[i]/2$-module, that is a free $\mathbb{F}_2[i]$-module, of dimension 4. As in the previous case, we can construct an explicit homomorphism of $\mathbb{F}_2[i]$-modules $\phi : \mathcal{O}/2\mathcal{O} \to M_2(\mathbb{F}_2[i])$:

$$\phi([1]) = \mathbb{1}, \quad \phi([\theta]) = \begin{pmatrix} 1+i & 1 \\ i & i \end{pmatrix}, \quad \phi([j]) = \begin{pmatrix} 0 & 1 \\ i & 0 \end{pmatrix}, \quad \phi([\theta j]) = \phi([\theta])\phi([j])$$

Again, it is easy to check that the images of the basis elements are linearly independent and therefore ϕ is one-to-one. It is also surjective since the cardinalities of the domain and codomain are the same.
Moreover, ϕ is a ring homomorphism:

$$\phi([j])^2 = i\mathbb{1} = \phi([j^2]),$$

$$\phi([\theta^2]) = \phi([\theta+1]) = \begin{pmatrix} i & 1 \\ i & i+1 \end{pmatrix} = \begin{pmatrix} 1+i & 1 \\ i & 1 \end{pmatrix}^2,$$

$$\phi([j])\phi([\theta]) = \begin{pmatrix} i & i \\ 1+i & i \end{pmatrix} = \phi([(1+\theta)j])$$

□

Remark 6.44. In order to find an explicit isomorphism between $\mathcal{G}/2\mathcal{G}$ and $M_2(\mathbb{F}_2)$, consider the following diagram, where $\phi : \mathcal{O}/2\mathcal{O} \to M_2(\mathbb{F}_2[i])$ is the mapping defined in Lemma 6.43:

$$\sqrt{5}\mathcal{G} \xrightarrow{\pi} \mathcal{G}/2\mathcal{G} \xrightarrow{\varphi} \mathcal{O}/2\mathcal{O} \xrightarrow{\phi} M_2(\mathbb{F}_2[i])$$

The basis $\{\alpha, \alpha\theta, \alpha j, \alpha\theta j\}$ of $\sqrt{5}\mathcal{G}$ as a $\mathbb{Z}[i]$-module is also a basis of $\mathcal{G}/2\mathcal{G}$ as an $\mathbb{F}_2[i]$-module. The isomorphism φ is simply the composition of the inclusion $\sqrt{5}\mathcal{G} \to \mathcal{O}$ and the quotient mod $(1+i)\mathcal{O}$. We can compute the images through the isomorphism ϕ of the basis vectors: recalling that

$$\alpha = 1 + i - i\theta, \quad \alpha\theta = \theta - i, \quad \alpha j = (1 + i - i\theta)j, \quad \alpha\theta j = (\theta - i)j,$$

we get

$$\begin{aligned}
\phi(\alpha) &= \phi(1+i) + i\phi(\theta) = \begin{pmatrix} 0 & i \\ 1 & i \end{pmatrix} \\
\phi(\alpha\theta) &= \phi(\theta) + \phi(i) = \begin{pmatrix} 1 & 1 \\ i & 0 \end{pmatrix} \\
\phi(\alpha j) &= \phi(\alpha)\phi(j) = \begin{pmatrix} 0 & i \\ 1 & i \end{pmatrix}\begin{pmatrix} 0 & 1 \\ i & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \\
\phi(\alpha\theta j) &= \phi(\alpha\theta)\phi(j) = \begin{pmatrix} 1 & 1 \\ i & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ i & 0 \end{pmatrix} = \begin{pmatrix} i & 1 \\ 0 & i \end{pmatrix}
\end{aligned}$$

Recall that there is a one-to-one correspondence between the ideals of $\mathcal{G}$ that contain $2\mathcal{G}$ and the ideals of $\mathcal{G}/2\mathcal{G}$. Therefore $M_2(\mathbb{F}_2[i])$ has only one proper two-sided ideal, $(1+i)M_2(\mathbb{F}_2[i])$.

Also in this case, the lifts Y of non-invertible elements in $\mathcal{G}$ will have non-invertible determinant, that is $|\det Y|^2 \geq 2$.

In general, we know that $\mathcal{O}/(1+i)^k\mathcal{O}$ is a free $Z[i]/(1+i)^k$- module of dimension 4 (see Lemma 6.40). The following proposition holds:

Proposition 6.45.

1. *When $k = 2n$ is even, $\frac{\mathbb{Z}[i]}{(1+i)^k} \cong \mathbb{Z}_{2^n}[i]$.*
2. *When $k = 2n+1$ is odd, $\frac{\mathbb{Z}[i]}{(1+i)^k} = \frac{\mathbb{Z}[y]}{(y^2-2y+2, 2^n y)} \cong \mathbb{Z}_{2^{n+1}} \oplus \mathbb{Z}_{2^n} y$, with $y^2 = 2y - 2$.*

Proof.

1. When $k = 2n$, from the Third Isomorphism Theorem we get:

$$\frac{\mathbb{Z}[i]}{(1+i)^k} = \frac{\mathbb{Z}[x]}{(x^2+1, 2^n)} \cong \frac{\mathbb{Z}[x]/2^n}{(x^2+1, 2^n)/2^n} = \frac{\mathbb{Z}_{2^n}[x]}{x^2+1} = \mathbb{Z}_{2^n}[i]$$

2. When $k = 2n+1$,

$$\frac{\mathbb{Z}[i]}{2^n(1+i)} \cong \frac{\mathbb{Z}[x]}{(x^2+1, 2^n(1+x))}$$

Putting $y = 1 + x$, this is isomorphic to

$$\frac{\mathbb{Z}[y]}{(y^2 - 2y + 2, 2^n y)} = \frac{\mathbb{Z}[y]}{(y^2 - 2y + 2, 2^n y, 2^{n+1})}$$

since $2^{n+1} = 2^n(y^2 - 2y + 2) - 2^n y(y - 2) \in (y^2 - 2y + 2, 2^n y)$. We have a surjective ring homomorphism from $\mathbb{Z}[y]/(y^2 - 2y + 2, 2^n y, 2^{n+1})$ to $\mathbb{Z}_{2^{n+1}} \oplus \mathbb{Z}_{2^n} y$: given a polynomial $p(y)$ we can take the remainder (mod $(y^2 - 2y + 2)$)- a polynomial of degree 1 - and reduce the coefficients of 1 and y by 2^{n+1} and 2^n respectively.
This map is also surjective, since the two rings have the same cardinality:

Remark 6.46.

$$\# \left(\frac{\mathbb{Z}[i]}{(1+i)^{k+1}} \right) = 2^k$$

It can be proved by induction on k:

$$\begin{aligned} &\frac{\mathbb{Z}[i]}{(1+i)^k} \cong \frac{\mathbb{Z}[i]/(1+i)^{k+1}}{(1+i)^k/(1+i)^{k+1}} \\ &\Rightarrow \#\left(\mathbb{Z}[i]/(1+i)^{k+1}\right) = \#\left(\mathbb{Z}[i]/(1+i)^{k+1}\right) \#\left((1+i)^k/(1+i)^{k+1}\right) = \\ &= \#\left(\mathbb{Z}[i]/(1+i)^{k+1}\right) \#\left(\mathbb{Z}[i]/(1+i)\right) \end{aligned}$$

since $\mathbb{Z}[i]/(1+i) \cong (1+i)^k/(1+i)^{k+1}$ (where the ring isomorphism is given by $y \mapsto y(1+i)^k$). □

We can find an explicit matrix representation of $\mathcal{O}/(1+i)^k\mathcal{O}$ over $\mathbb{Z}[i]/(1+i)^k\mathbb{Z}[i]$ also for $k = 3$ and $k = 4$:

Lemma 6.47. $\frac{\mathcal{O}}{4\mathcal{O}} \cong M_2(\mathbb{Z}_4[i])$.

Proof. As in the previous cases, it is enough to find matrix representations of θ and j such that $1, \theta, j, \theta j$ are linearly independent over $\mathbb{Z}_4[i]$:

$$\theta = \begin{pmatrix} i-1 & 1 \\ -i & 2-i \end{pmatrix}, \qquad j = \begin{pmatrix} 0 & 1 \\ i & 0 \end{pmatrix}$$

In fact

$$\begin{aligned} \theta^2 &= \begin{pmatrix} i-1 & 1 \\ -i & 2-i \end{pmatrix} = \theta + \mathbb{1}, \\ \theta j &= \begin{pmatrix} i & i-1 \\ 2i+1 & -i \end{pmatrix} = j(1-\theta) \end{aligned}$$

Since $M_2(\mathbb{Z}_4[i])$ and $\frac{\mathcal{O}}{4\mathcal{O}}$ have the same cardinality because of Lemma 6.40, this representation is a ring isomorphism. □

The same matrix representations of θ and j can be used for $k = 3$, recalling that $i = y - 1$:

Proposition 6.48. $\frac{\mathcal{O}}{2(1+i)\mathcal{O}} \cong M_2(\mathbb{Z}_4 \oplus \mathbb{Z}_2[y])$, *where* $y^2 = 2y - 2$.

Proof. The proof is similar to the previous one. We define

$$\theta = \begin{pmatrix} y-2 & 1 \\ 1+y & -1+y \end{pmatrix}, \qquad j = \begin{pmatrix} 0 & 1 \\ y-1 & 0 \end{pmatrix}$$

Computing products mod $(4, 2y)$, we get

$$\theta^2 = \begin{pmatrix} -1+y & 1 \\ 1+y & y \end{pmatrix} = \theta + 1,$$
$$\theta j = \begin{pmatrix} -1+y & y+2 \\ -1 & 1+y \end{pmatrix} = j\bar{\theta}$$

□

6.7 The repetition code

In this paragraph we want to illustrate with a simple example how the properties of the minimum determinant described in §6.4.1 influence the actual code performance. We consider a block code of length 2, the lift of the repetition code over the cosets of $(1+i)\mathcal{G}$: if $\pi : \mathcal{G} \to \mathcal{G}/(1+i)\mathcal{G}$ is the projection on the quotient ring, we define

$$\mathcal{C} = \{\mathbf{X} = (X_1, X_2) \in \mathcal{G}^2 \mid \pi(X_1) = \pi(X_2)\}$$

The fact that the codewords of Hamming weight 1 belong to the 0 coset ensures that the $\Delta_{\min}$ for $\mathcal{C}$ is equal to the minimum square determinant in $(1+i)\mathcal{G}$, which is 4: in fact, if $\pi(X_1) = \pi(X_2) \neq 0$, $\det(\mathbf{X}\mathbf{X}^H) = \det(X_1)^2 + \det(X_2)^2 + \left\| \widetilde{X}_2 X_1 \right\|_F^2 \geq (|\det(X_1)| + |\det(X_2)|)^2 \geq 4$ because of Lemma 6.15 and Remark 6.17.
A simple variation of the repetition scheme consists in choosing any bijection h of $\mathcal{G}/(1+i)\mathcal{G}$ and defining

$$\mathcal{C}_h = \{\mathbf{X} = (X_1, X_2) \in \mathcal{G}^2 \mid \pi(X_2) = h(\pi(X_1))\}$$

In the case of the repetition code, suppose that $\pi(X_1) = \pi(X_2) = C_i$. If C_i is an invertible element in $M_2(\mathbb{F}_2)$, then

$$\widetilde{C}_i C_i = \det(C_i)\mathbb{1} = \mathbb{1} = \mathbf{e}_1 + \mathbf{e}_2$$

in the basis (6.26), and so the minimum determinant of a "lifted" codeword $X \in \pi^{-1}(\widetilde{C}_i C_i)$ is also 1, and the minimum of $\|X\|_F^2$ is 2.
If on the other side C_i corresponds to a non-invertible, nonzero element in $M_2(\mathbb{F}_2)$, then $\min_{X \in \pi^{-1}(\widetilde{C}_i C_i)} |\det(X)| \geq 2$ (see Remark 6.42). Thus in the first case $\det(XX^H) \geq 4$, in the second $\det(XX^H) \geq 8$.
This remark suggests that it might be more convenient to consider a group homomorphism $h : M_2(\mathbb{F}_2) \to M_2(\mathbb{F}_2)$ which maps invertible elements into non-invertible elements, raising the minimum determinant for $C_i \neq 0$ to 9. Such a function is not difficult to define: for example, recalling the definition of the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}$ in equation (6.26), we can take

$$h(\mathbf{e}_1) = \mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_4,\ h(\mathbf{e}_2) = \mathbf{e}_2 + \mathbf{e}_3 + \mathbf{e}_4,\ h(\mathbf{e}_3) = \mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3,\ h(\mathbf{e}_4) = \mathbf{e}_1 + \mathbf{e}_3 + \mathbf{e}_4$$

In the case of $4-QAM$ modulation, an exhaustive search on the 65536 points in the finite lattice shows that $\mathcal{C}_h$ is indeed better than $\mathcal{C}$.[2] The asymptotic coding gain estimate is the same for both codes: using 4-QAM constellations, the choice of a coset requires 4 information bits, while the choice of a point in a given coset requires 4 more bits. Each codeword then carries 12 information bits, yielding a spectral efficiency of 3 bpcu (bits per channel use).
Then we can compare these schemes with the uncoded Golden Code at 3 bpcu, using 4-QAM constellations for the symbols a, c and BPSK (Binary Phase Shift Keying) constellations, consisting of the two points $\{-\frac{1}{2}, \frac{1}{2}\}$, for the symbols b, d in each Golden codeword. The average energy per symbol is $E_{\mathcal{S}} = 0.5(0.5 + 0.25) = 0.375$.

$$\gamma_{\text{as}} = \frac{\sqrt{\Delta_{\min,1}}/E_{\mathcal{S},1}}{\sqrt{\Delta_{\min,2}}/E_{\mathcal{S},2}} = \frac{2/0.5}{1/0.375} = 1.5,$$

This computation gives a theoretical gain of at least $10\log_{10}(1.5)\,\text{dB} = 1.7\,\text{dB}$.

Simulation results

Figure 6.2 shows the performance of the codes $\mathcal{C}_{Id}$ and $\mathcal{C}_h$, which gain 2.4 dB and 2.9 dB respectively over the uncoded scheme at 3 bpcu at the frame error rate of 10^{-3}, supposing that the channel is constant for 2 time blocks.

6.8 Golden Reed-Solomon Codes

We now go back to the original problem stated in §6.4, that is, how to improve the performance of the Golden Code in the *slow fading* setting, using block codes over $\mathcal{G}$. We would like to compensate the diversity loss due to the slow changing of the channel with an increase of the Hamming distance of the code over the alphabet $\mathcal{G}$. We will combine the choice of a modulation scheme and of a maximum-distance separable error-correcting code.

Remark 6.49. As we have seen in the previous sections, in addition to the minimum Hamming distance, also the multiplicative structure and the minimum number of non-invertible components have a significant influence on the coding gain of a block code design. Thus, an optimal solution in order to keep track of these parameters and take advantage of the ring structure would be to consider error-correcting codes based on $M_2(\mathbb{F}_2[i])$. However, such a project, albeit interesting, would be difficult to implement, since at present very little is known about codes over non-commutative rings, and no efficient decoding algorithms are available.

[2] In fact we can compute the function

$$\Theta_h(q) = \sum_{\mathbf{X}\in\mathcal{C}_h} = q^{Det(\mathbf{X}\mathbf{X}^H)}$$

In the case of the repetition code, the first terms in the series are

$$\Theta_{Id}(q) = 1 + 66q^4 + 120q^8 + 48q^{10} + 202q^{16} + \dots$$

while for the function h just defined,

$$\Theta_h(q) = 1 + 24q^4 + 61q^8 + 24q^9 + 8q^{10} + 74q^{12} + 58q^{13} + 74q^{14} + 108q^{16} + \dots$$

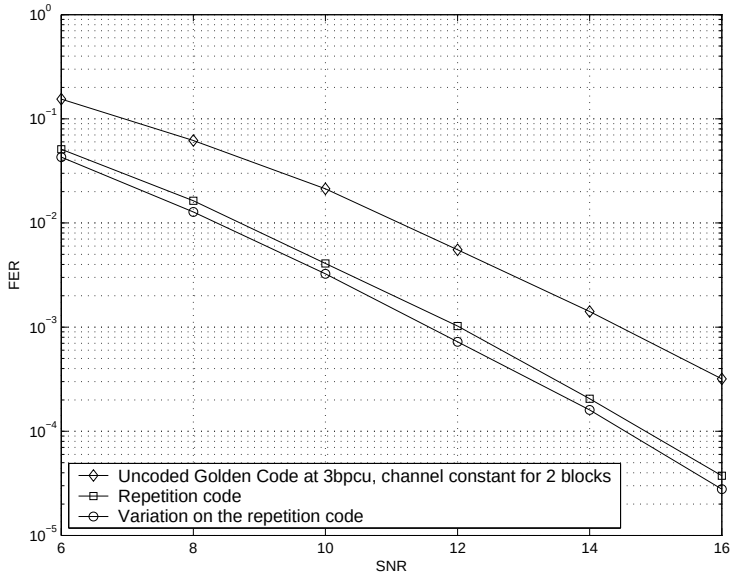

Figure 6.2: Performance of the repetition code $\mathcal{C}_{Id}$ and of the variation $\mathcal{C}_h$ at 3 bpcu compared with the uncoded Golden Code scheme with the same spectral efficiency. The channel is supposed to be constant for 2 time blocks.

We also remark that in the case of codes over rings, a distinction must be made between the *rank* (minimum number of generators as a module) and the *free rank* (maximum rank of the free submodules of the code), so that in general the number of codewords might be smaller than the power of the cardinality of the ring to the rank of the code. In the case of non-commutative rings, the rank might not even be well-defined.
We choose shortened Reed-Solomon codes instead because they are maximum distance separable and their implementation is very simple; we will restrict our attention to the additive structure, defining a group isomorphism between $\mathcal{G}/2\mathcal{G}$ and the finite field $\mathbb{F}_{256}$.

6.8.1 The 4-QAM case

Using 4-QAM constellations (see Figure 4.2) to modulate each of the 4 symbols a, b, c, d in a Golden codeword

$$X = \frac{1}{\sqrt{5}} \begin{bmatrix} \alpha(a+b\theta) & \alpha(c+d\theta) \\ \bar{\alpha}i(c+d\bar{\theta}) & \bar{\alpha}(a+b\bar{\theta}) \end{bmatrix},$$

give a total of 256 codewords, one in each coset of $\mathcal{G}/2\mathcal{G} \cong M_2(\mathbb{F}_2[i])$. In this case, simply by combining an (n,k,d) error correcting code with the quotient $\mathcal{G}/2\mathcal{G}$, we can be sure to achieve minimum Hamming distance d. On the contrary, if we have more than one point per coset and consider the lifts of linear codes on the quotient, we would get blocks of Hamming weight 1 that are lifts of the codeword $\mathbf{0}$ in the error-correcting code.

We consider an (n, k, d) Reed-Solomon code over $\mathbb{F}_{256}$. We recall that these codes are maximum distance separable, that is $k = n - 1 + d$.
Each quadruple (a, b, c, d) of 4-QAM signals corresponds to 8 information bits; each block of n Golden codewords will carry $8k$ information bits.
We describe in detail each step of the encoding and decoding procedure:

1. Reed-Solomon encoding:
Each information byte can be seen as a binary polynomial of degree ≤ 8, that is, an element of the Galois Field $\mathbb{F}_{256}$. A random information message of k bytes, seen as a vector $\mathbf{U} = (U_1, \ldots, U_k) \in \mathbb{F}_{256}^k$ is encoded into a codeword $\mathbf{V} = (V_1, \ldots, V_n) \in \mathbb{F}_{256}^n$ using the $\mathrm{RS}(n, k, d)$ shortened code $\mathcal{C}$.
In order to obtain the generator matrix for the shortened code $\mathcal{C}$, we start with the "long" code $\mathrm{RS}(255, 255 - d + 1, d)$ [16]. As a generator polynomial we can take

$$g(x) = \prod_{i=1}^{d-1} (x - \alpha^i) = c_0 + c_1 x + \ldots + c_{d-2} x^{d-1} + x^{d-1}, \quad c_i \in \mathbb{F}_{256} \forall i$$

where α is a *primitive element*, that is a generator of the multiplicative group $\mathbb{F}_{256}^*$. α is a root of an irreducible (*primitive*) polynomial p of degree 8.[3] The corresponding generator matrix is

$$G = \begin{bmatrix} c_0 & c_1 & c_2 & \cdots & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & c_0 & c_1 & c_2 & \cdots & 1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & c_0 & c_1 & c_2 & \cdots & 1 & 0 & 0 & 0 \\ \vdots & & & \ddots & \ddots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & \ddots & & & \ddots & \vdots \\ 0 & \cdots & \cdots & & \cdots & c_0 & c_1 & c_2 & \cdots & 1 \end{bmatrix}$$

Choosing only the rows whose first $255 - n$ components are equal to 0 and deleting the null columns, we obtain the generator matrix G for the (n, k, d) shortened Reed-Solomon code.[4]

2. From the Galois field $\mathbb{F}_{256}$ to the matrix ring $M_2(\mathbb{F}_2[i])$:
We now have a vector in $\mathbb{F}_{256}^n$; we want to translate each component into an element of $M_2(\mathbb{F}_2[i])$.

[3] There are 16 primitive polynomials of degree 8. For this simulation I have arbitrarily chosen $p(x) = x^8 + x^6 + x^5 + x + 1$.
The coefficients of the generator polynomial g in the Galois Field $\mathbb{F}_{256}$ can be easily computed using any symbolic manipulation software, such as Maple.

[4] In the case of a $(4, 2, 3)$ shortened Reed-Solomon code, the generator matrix is

$$G = \begin{bmatrix} \alpha^3 & \alpha^2 + \alpha & 1 & 0 \\ 0 & \alpha^3 & \alpha^2 + \alpha & 1 \end{bmatrix}$$

For our purposes, it is much better to obtain a *systematic* version of the code, that is one that preserves the first k bits of the input. This equivalent version can be obtained simply by performing the Gauss reduction algorithm over $\mathbb{F}_{256}$, yielding the matrix

$$G = \begin{bmatrix} 1 & 0 & 1 + \alpha + \alpha^2 + \alpha^3 + \alpha^4 + \alpha^5 + \alpha^6 + \alpha^7 & 1 + \alpha + \alpha^3 \\ 0 & 1 & \alpha^3 + \alpha^4 + \alpha^6 & 1 + \alpha^2 + \alpha^6 + \alpha^7 \end{bmatrix}$$

We remark that in order to speed up the computation of products over $\mathbb{F}_{256}$, a table storing the conversions between the representation as a polynomial of degree less than 7 in α and the representation as a power of α can be computed once and for all.

We can represent the elements of $M_2(\mathbb{F}_2[i])$ as bytes, simply by vectorising each matrix and separating real and imaginary parts: for example

$$\begin{pmatrix} a+bi & e+fi \\ c+di & g+hi \end{pmatrix} \mapsto (a,b,c,d,e,f,g,h) \in \{0,1\}^8$$

Since we are only working with the additive structure, we can identify $\mathbb{F}_{256}$ and $M_2(\mathbb{F}_2[i])$, which are both $\mathbb{F}_2$-vector spaces of dimension 8. According to our simulation results, it seems that the choice of the linear identification has very little influence on the code performance.

3. From the matrix ring $M_2(\mathbb{F}_2[i])$ to the quotient ring $\mathcal{G}/2\mathcal{G}$:
For this step we make use of the isomorphism of $\mathbb{F}_2[i]$-modules $\varphi \circ \phi : \mathcal{G}/2\mathcal{G} \to M_2(\mathbb{F}_2[i])$ described in Remark 6.44. In vectorized form, with respect to the bases $\{\alpha, \alpha\theta, \alpha j, \alpha\theta j\}$ and

$$\left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\},$$

this is given by the matrix

$$A = \begin{pmatrix} 0 & 1 & 1 & i \\ 1 & i & 1 & 0 \\ i & 1 & 0 & 1 \\ i & 0 & 1 & i \end{pmatrix}$$

In practice, it is sufficient to send each matrix $M \in M_2(\mathbb{F}_2[i])$ to

$$A^{-1}(M) = (\Re(a), \Im(a), \Re(b), \Im(b), \Re(c), \Im(c), \Re(d), \Im(d)),$$

with $a,b,c,d \in \mathbb{Z}_2[i] = \mathbb{Z}[i]/2\mathbb{Z}[i]$. Taking the corresponding coset leaders in $\mathbb{Z}[i]$, we obtain the Golden codeword

$$\frac{1}{\sqrt{5}} \begin{pmatrix} \alpha(a+b\theta) & \alpha(c+d\theta) \\ i\bar{\alpha}(c+d\bar{\theta}) & \bar{\alpha}(a+b\bar{\theta}) \end{pmatrix}$$

4. Golden Code encoding:
For each of the n vector components, the symbols a,b,c,$d \in \mathbb{Z}_2[i]^4$ are modulated into four 4-QAM signals, and then encoded into a Golden codeword using the (vectorized, real) generator matrix R of Remark 6.13. Thus we have obtained a Golden block $\mathbf{X} = (X_1, X_2, \ldots, X_n) = \xi(\mathbf{V})$, where $\xi : \mathbb{F}_{256}^n \to \mathcal{G}^n$ is injective.

5. Channel simulation:
We suppose the channel matrix H to be constant during the transmission of the n Golden codewords. This assumption can be considered realistic for a slow fading channel if $n < 100$ (see §4.2.1). The received signal is

$$\mathbf{Y} = H\mathbf{X} + \mathbf{W},$$

where $\mathbf{W}$ is the Gaussian noise.

6. Soft decoding:
In a first phase, for each component $i = 1, \ldots, n$ of the received vector $\mathbf{Y}$ and for each modulated point $Z^{(j)}$, the Euclidean distance

$$d(i,j) = \left\| HZ^{(j)} - Y_i \right\|^2$$

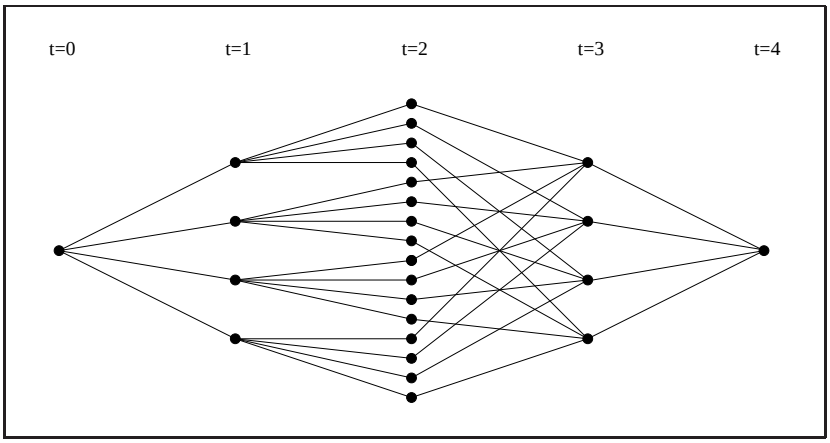

Figure 6.3: A simplified diagram showing the trellis associated to the $(4, 2, 3)$-Reed-Solomon code.

is computed and stored in memory.
In a second phase, ML decoding or *soft decoding* is performed (see equation (4.4)): we search for the minimum of the Euclidean distance

$$\sum_{i=1}^{n} \|HZ_i - Y_i\|^2$$

over all the images $\mathbf{Z} = \xi(\mathbf{V}')$ of Reed-Solomon codewords. The *Viterbi algorithm* over the trellis diagram for the RS code may be used for this search (for a reference see for example [2]): if the generator matrix of the RS code is systematic, the partial distance $\sum_{i=1}^{k} \|HZ_i - Y_i\|^2$ can be obtained directly without computing the whole Reed-Solomon codeword, and the points for which this distance is too big can be discarded.
In the case of a RS(4,2,3) code, the dimensions k_t of the state spaces Z_t at time t in the trellis as $\mathbb{F}_{256}$-vector spaces are respectively $k_1 = 1$, $k_2 = 2$, $k_3 = 1$, $k_4 = 0$. By exchanging the fourth and second coordinate in the trellis, the decoding process is reduced to finding

$$\hat{\mathbf{X}} = \underset{\overline{X}_2}{\operatorname{argmin}} \left(\left(\min_{\mathbf{X}=(X_1,\overline{X}_2,X_3,X_4)} \sum_{i\neq 2} \|Y_i - HX_i\|^2 \right) + \left\|Y_2 + H\overline{X}_2\right\|^2 \right)$$

over all the blocks $\mathbf{X} = \xi(\mathbf{V})$ arising from Reed-Solomon codewords (see figure 6.3).

Performance

In the 4-QAM case, the spectral efficiency of the Golden Reed-Solomon codes is given by

$$\frac{8k \text{ bits}}{2n \text{ channel uses}} = \frac{4k}{n} \text{ bpcu}$$

From Proposition 6.20, we get a lower bound for $\Delta_{\min}$: for a Golden Reed-Solomon code of minimum Hamming distance d in $\mathcal{G}$, $\Delta_{\min} \geq \frac{d^2}{5}$. Thus we

obtain an estimate of the *asymptotic coding gain* for these codes (see §4.2.1), by comparing them with the uncoded Golden Code with the same spectral efficiency. In the case of 2 bpcu, we can consider a BPSK constellation on the real axis.

- **2 bpcu**
  If $k = \frac{n}{2}$, the spectral efficiency is 2bpcu. Comparing the 4-QAM, (n, k, d) Golden-RS design ($E_{\mathcal{S}} = 0.5$) with the uncoded Golden Code using BPSK ($E_{\mathcal{S}} = 0.25$), we get an asymptotic coding gain of:

$$\gamma_{\text{as}} = \frac{\sqrt{\Delta_{\min,1}}/E_{s,1}}{\sqrt{\Delta_{\min,2}}/E_{s,2}} = \frac{d/0.5}{1/0.25} = \frac{d}{2} \tag{6.27}$$

- **3 bpcu**
  If $k = \frac{3}{4}n$, the spectral efficiency is 3 bpcu. Comparing the 4-QAM, (n, k, d) Golden-RS design ($E_{\mathcal{S}} = 0.5$) with the uncoded Golden Code using BPSK for the symbols a and c and 4-QAM for the symbols b and d ($E_{\mathcal{S}} = 0.5(0.5 + 0.25) = 0.375$), we get an asymptotic coding gain of:

$$\gamma_{\text{as}} = \frac{\sqrt{\Delta_{\min,1}}/E_{\mathcal{S},1}}{\sqrt{\Delta_{\min,2}}/E_{\mathcal{S},2}} = \frac{d/0.5}{1/0.375} = \frac{3d}{4} \tag{6.28}$$

Simulation results

Figure 6.4 shows the performance comparison of the Golden-RS code $(4, 2, 3)$ with the uncoded scheme at the spectral efficiency of 2 bpcu.
Assuming the channel to be constant for 4 blocks, the Golden-RS code outperforms the uncoded scheme by 6.1 dB.
This gain is unexpectedly high compared with the theoretical coding gain (6.27) for $d = 3$, which is $10\log_{10}\left(\frac{d}{2}\right)\text{dB} = 1.7\,\text{dB}$. The rough estimate (6.27) is based on the worst possible occurrence, that of a codeword of Hamming weight 3 in which all three non-zero components correspond to invertible elements in the quotient.
However, we can verify empirically that in the 4-QAM case and with our choice of the code, this event does not take place and in fact the actual value for $\Delta_{\min}$ found by computer search is $\sqrt{34}$, giving a rough estimate for the gain of 4.6 dB, which is much closer to the observed value.
This favorable behavior might be due to the fact that the chosen constellation contains only one point in each coset, so that the codewords of Hamming distance 3 are few.

The soft decoding method has the drawback of being slow, which makes it unsuitable to use with longer Reed-Solomon codes. A faster (if suboptimal) soft decoding algorithm, such as *stack decoding*, could make up for this loss of speed while still retaining most of the coding gain.

Hard decoding case

In an early version of the algorithm described in §6.8.1 we replaced Step 6 with the following steps:

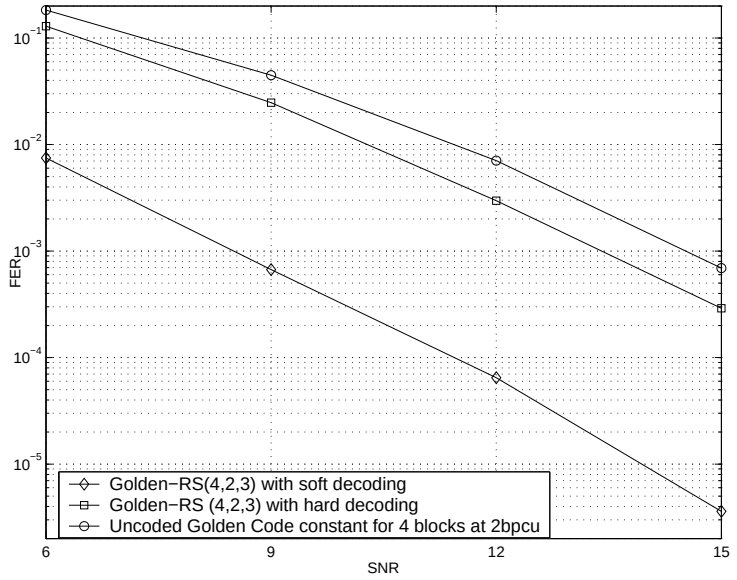

Figure 6.4: Comparison between hard and soft decoding for the RS$(4, 2, 3)$ code at 2 bpcu. The first method achieves a gain of only 1.1 dB over the uncoded case, compared to the 6.1 dB of the second.

6. n SEPARATE SPHERE DECODERS:
We apply an 8-bit Sphere Decoder separately on each of the n received words $(Y_1, \ldots, Y_n)$, obtaining the estimate $(\hat{X}_1, \ldots, \hat{X}_n)$. The signal is then demodulated, and we apply to each byte the inverse mappings of Steps 3 and 2 successively, obtaining a vector $(\hat{V}_1, \ldots, \hat{V}_n)$ in $\mathbb{F}_{256}^n$.

7. REED-SOLOMON DECODING:
The received sequence $(\hat{V}_1, \ldots, \hat{V}_n)$ doesn't necessarily belong to the RS code, so we still need to perform RS decoding, yielding the estimate $(\hat{U}_1, \ldots, \hat{U}_k)$. Finally, we compute the error probability, comparing $(\hat{U}_1, \ldots, \hat{U}_k)$ with the initial message $(U_1, \ldots, U_n)$ of Step 1, and record one word error whenever they are different.

This "hard" decoding has the advantage of speed and allows to use longer Reed-Solomon codes with high minimum distance. However it is highly suboptimal, since it substitutes the decision on the Euclidean distance with poor partial decisions on each coordinate. Performance simulations show that with this method the coding gain is almost entirely cancelled out (see figure 6.4).

Simulation results:

- **2 bpcu:** Figure 6.5 shows the performance comparison of the Golden-RS codes with hard decoding with the uncoded scheme at the spectral efficiency of 2 bpcu.
  Assuming the channel to be constant for 4, 8 and 12 blocks respectively, the $(4, 2, 3)$, $(8, 4, 5)$ and $(12, 6, 7)$ Golden-RS codes outperform the uncoded scheme at the same spectral efficiency by 1.1 dB, 1.7 dB and 2.8 dB at

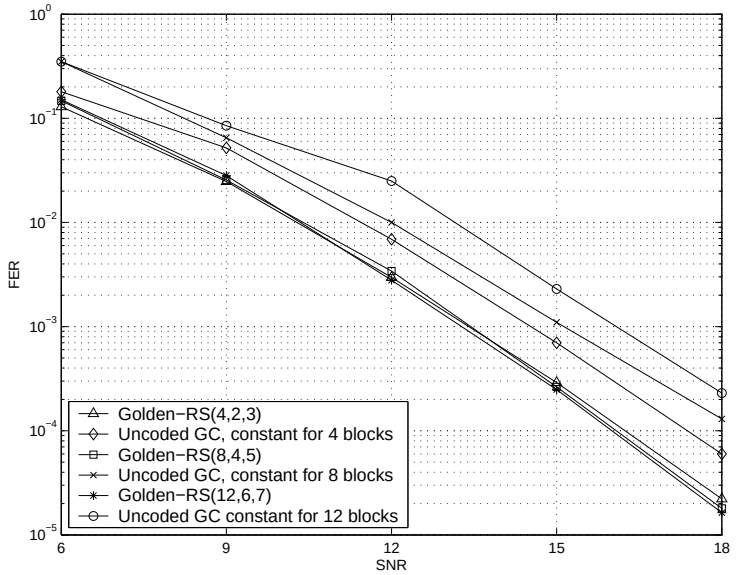

Figure 6.5: Performance of $(4, 2, 3)$, $(8, 4, 5)$, and $(12, 6, 7)$ Golden Reed-Solomon codes with "hard decoding" at 2 bpcu compared with the uncoded Golden Code scheme with the same spectral efficiency.

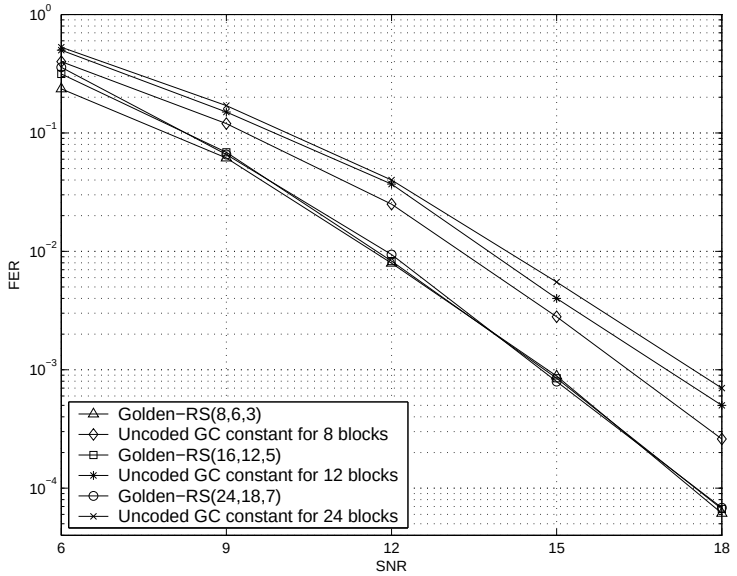

Figure 6.6: Performance of $(8, 6, 3)$, $(16, 12, 5)$, and $(24, 18, 7)$ Golden Reed-Solomon codes with "hard decoding" at 3 bpcu compared with the uncoded Golden Code scheme with the same spectral efficiency.

the FER of 10^{-3}.
The Golden-RS schemes seem to be more robust on slow fading channels; in fact the performances of the Golden-RS(n,k,d) codes on a channel which is constant for n blocks remain almost unchanged (the variation is less than 0.2 dB) when n varies between 4 and 12, while the uncoded Golden Code has a loss of almost 1.5 dB.

- **3 bpcu:** Assuming the channel to be constant for 8, 16 and 24 blocks respectively, the $(8,6,3)$, $(16,12,5)$ and $(24,18,7)$ Golden-RS codes gain 1.5 dB, 2.2 dB and 2.8 dB over the uncoded scheme at the FER of 10^{-3} (see Figure 6.6).
Similarly to the previous case, the Golden-RS(n,k,d) codes lose less than 0.3 dB when n varies between 8 and 24, while the Golden Code has a loss of 1.1 dB.

6.8.2 The 16-QAM case

If we use a 16-QAM constellation for each symbol a,b,c,d in a Golden codeword, we have $16^4 = 2^{16} = 65536$ available Golden codewords. Recalling that $\#(\mathcal{G}/2\mathcal{G}) = 256$, we have 256 words for each of the 256 cosets of $2\mathcal{G}$ in $\mathcal{G}$.
In this case, the coding gain depends on the minimum Hamming distance inside each coset in addition to the minimum Hamming distance in the quotient.
As in the 4-QAM case, we consider block codes which are lifts of Reed-Solomon codes on the quotient $\mathcal{G}/2\mathcal{G}$. Intuitively, the minimum distance of the Reed-Solomon code "protects" the cosets from being decoded wrongly; if this choice is correct, the estimate for the right point in the coset is "protected" by the minimum determinant in $2\mathcal{G}$.
We consider the lift of an (n,k,d) Reed-Solomon code $\mathcal{C}$ on the quotient. The total information bits transmitted are $8k+8n$; they will be encoded into $8n+8n = 16n$ bits.

- The code $\mathcal{C}$ outputs $8n$ bits, which are used to encode the first two bits of $4n$ 16-QAM constellations, that is the bits which identify one of the four cosets of $2\mathbb{Z}[i]$ in $\mathbb{Z}[i]$; each byte corresponds to a different coset configuration of (a,b,c,d) (see Figure 6.8).

- the other $8n$ bits, left uncoded, are used to choose the last two bits of each 16-QAM signal.

In total, we have $4n$ 16-QAM symbols, that is n Golden codewords $\mathbf{X} = (X_1,\dots,X_n)$. The resulting spectral efficiency is

$$\frac{8(k+n)\text{ bits}}{2n\text{ channel uses}} = \frac{4(k+n)}{n}\text{ bpcu}$$

$\mathcal{C}$ acts as a code over $\mathcal{G}/2\mathcal{G} \cong M_2(\mathbb{F}_2[i])$: if $\{W_0, W_1, \dots, W_{255}\}$ are the coset leaders of $2\mathcal{G}$ in $\mathcal{G}$, then $\forall j = i,\dots,n$,

$$X_i = W_{j_i} + Z_i, \quad Z_i \in 2\mathcal{G}, \quad (W_{j_1},\dots,W_{j_n}) \in \mathcal{C} \tag{6.29}$$

Clearly, if $(W_{j_1},\dots,W_{j_n}) = \mathbf{0}$, then $(X_1,\dots,X_n) = (Z_1,\dots,Z_n) \in (2\mathcal{G})^n$ and for $\mathbf{X} \neq \mathbf{0}$, $\det(\mathbf{X}\mathbf{X}^H) \geq 16$.

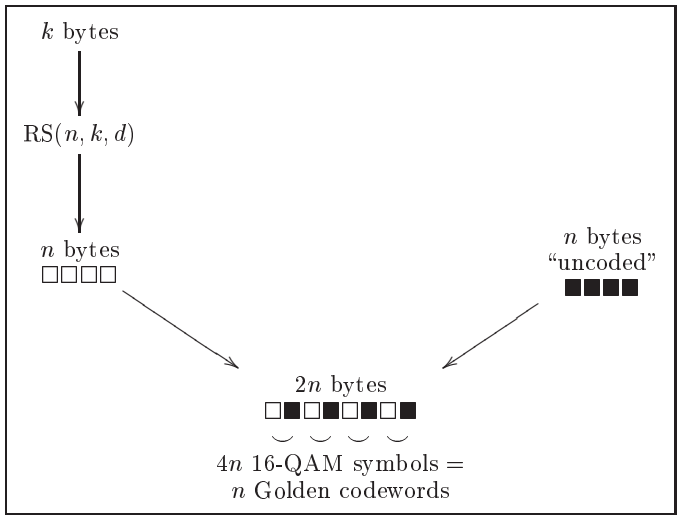

Figure 6.7: The output of the Reed-Solomon code and the uncoded bits are "mingled" before modulation.

If on the contrary $(W_{j_1}, \ldots, W_{j_n}) \neq \mathbf{0}$, then there are at least d components of $\mathbf{X}$ which do not belong to $2\mathcal{G}$, and consequently are nonzero, so that $\det(\mathbf{X}\mathbf{X}^H) \geq d^2$. In conclusion, we have

$$\Delta_{\min} \geq \min(16, d^2) \tag{6.30}$$

With an error-correcting code of rate $k = \frac{n}{2}$, we obtain a spectral efficiency of 6 bpcu.

- If $d \geq 4$, we have $\gamma_{\text{as}} = \frac{4/2.5}{1/1.5} = 2.4$, leading to an approximate gain of 3.8 dB. Thus it does not seem worthwhile to use long codes with a high minimum distance with this scheme.
- If $d = 3$, $\gamma_{\text{as}} = \frac{3/2.5}{1/1.5} = 1.8$, making for a gain of 2.5 dB.

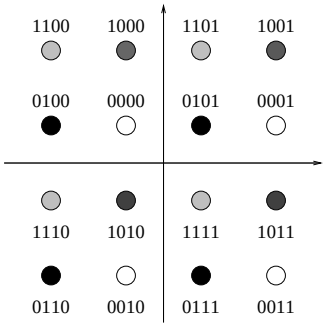

Figure 6.8: The labelling of the 16-QAM constellation used for performance simulations. The first and second bit identify one of the four cosets of $2\mathbb{Z}[i]$ in $\mathbb{Z}[i]$ (drawn in different shades of gray); the third and fourth bit identify one of the four points in the coset. We remark that this type of labelling cannot be a Gray mapping.

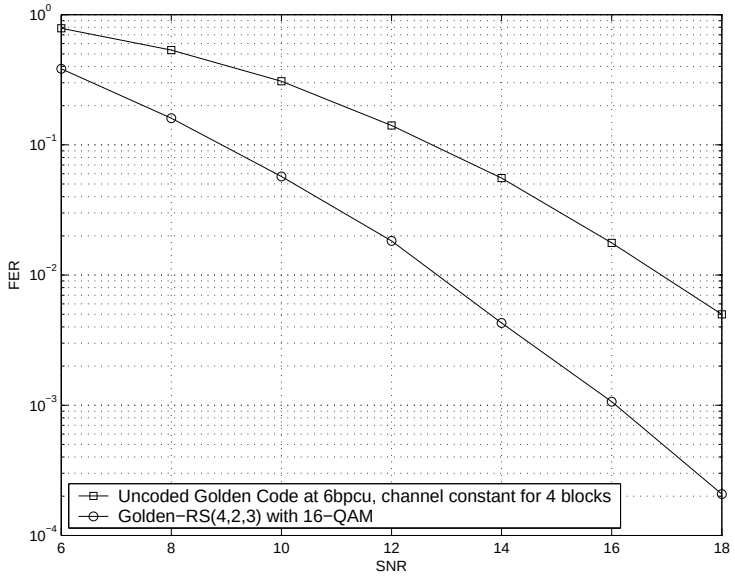

Figure 6.9: Performance of the $(4, 2, 3)$ Golden Reed-Solomon code with soft decoding at 6 bpcu compared with the uncoded Golden Code scheme with the same spectral efficiency.

Decoding

The soft decoding procedure for the 16-QAM case requires only a slight modification with respect to Step 6 illustrated in §6.8.1. In the first phase, for each component $i = 1, \ldots, n$ and for each coset leader W_j, $j = 0, \ldots, 255$, we determine the closest point in that coset to the received component Y_i, that is

$$\hat{Z}_{i,j} = \operatorname*{argmin}_{Z \in 2\mathcal{G}} \|Y_i - H(Z + W_j)\|^2$$

Computing HZ and HW_j separately allows to perform only 512 products instead of 256^2. The second phase can be performed as in the 4-QAM case, and the search is limited to the "closest points" $\hat{Z}_{i,j} + W_j$ determined in the previous phase:

$$\hat{\mathbf{X}} = \operatorname*{argmin}_{(\hat{Z}_{1,j_1}+W_{j_1},\ldots,\hat{Z}_{n,j_n}+W_{j_n})} \sum_{i=1}^{n} \left\| H(\hat{Z}_{i,j_i} + W_{j_i}) - Y_i \right\|^2$$

over all the images $(W_{j_1}, \ldots, W_{j_n})$ of Reed-Solomon codewords.

Simulation results

In the 16-QAM case, the $(4, 2, 3)$ Golden Reed-Solomon code achieves a gain of 3.8 dB over the uncoded scheme at 6 bpcu at the frame error rate of 10^{-2}, supposing that the channel is constant for 4 time blocks (see figure 6.9).

Bibliography for Part I

[1] J.F. Alves, K. Oliveira, A. Tahzibi, "On the continuity of the SRB entropy for endomorphisms", *J. Stat. Phys.* vol. 123 n. 4 (2006)

[2] J.F. Alves, M. Viana, "Statistical stability for robust classes of maps with non-uniform expansion", *Ergodic Theory Dynam. Systems* 22 (2002) 1–32

[3] J. Bourdon, B. Daireaux, B. Vallée, "Dynamical analysis of α-Euclidean algorithms", *J. Algorithms* 44 (2002), 246-285

[4] W. Bosma, "Optimal continued fractions", *Indag. Math*, A90, 1987, 353–379

[5] A. Broise, "Transformations dilatantes de l'intervalle et théorèmes limites", *Astérisque* 238 (1996) 5–109.

[6] A. Cassa, "Dinamiche caotiche e misure invarianti", Tesi di Laurea, Facoltà di Scienze Matematiche, Fisiche e Naturali, University of Florence, Italy, 1995

[7] I. P. Cornfeld, S. V. Fomin, Ya. G. Sinai, *Ergodic Theory*, Springer-Verlag 1982

[8] M. Denker, G. Keller, M. Urbanski, "On the uniqueness of equilibrium states for piecewise monotone mappings", Studia Math. 97 (1990), 27–36

[9] E. Giusti, *Minimal surfaces and functions of bounded variations*, Birkäuser Verlag, Basel-Boston, 1984

[10] S. Katok, "Coding of closed geodesics after Gauss and Morse", *Geom. Dedicata* 63 (1996), no. 2, 123–145

[11] C. Kraaikamp, "A new class of continued fraction expansions", *Acta Arith.* 57 (1991), no. 1, 1–39

[12] C. Kraaikamp, "Maximal S-expansions are Bernoulli shifts", *Bull. Soc. Math. France*, 121 no. 1 (1993), 117–131

[13] L. Luzzi, S. Marmi, "On the entropy of Japanese continued fractions", preprint, 2006 (to appear in *Discr. Cont. Dyn. Syst.*)

[14] L. Lhote, "Modélisation et approximation des sources", Rapport de stage de DEA de l'Université de Caen, 2002

[15] P. Moussa, A. Cassa, S. Marmi, "Continued fractions and Brjuno functions", *J. Comput. Appl. Math.* 105 (1999), 403–415

[16] H. Nakada, "Metrical theory for a class of continued fraction transformations and their natural extensions", *Tokyo J. Math.* 4, 1981

[17] R. Natsui, "On the isomorphism problem of α-Farey maps", *Nonlinearity* 17 (2004), 2249–2266

[18] H. Nakada, S. Ito, S. Tanaka, "On the invariant measure for the transformations associated with some real continued-fractions", *Keio Eng. Rep.* 30 (1977), 159–175

[19] V. Rohlin, "Exact endomorphisms of a Lebesgue space", *Amer. Math. Soc. Transl.* 39(2) (1964), 1–36

[20] F. Schweiger, *Ergodic theory of fibred systems and metric number theory*, Oxford Sci. Publ. Clarendon Press, Oxford, 1995

[21] C. Series, "Symbolic dynamics for geodesic flows", *Acta Math.* 146 (1981), 103–128

[22] M. Thaler, "Transformations on $[0, 1]$ with infinite invariant measures", *Israel J. of Math.* 46, 1983

[23] B. Vallée, "Dynamical analysis of a class of Euclidean algorithms", *Theoret. Comput. Sci.* 297 (2003), 447–486

[24] M. Viana, *Stochastic dynamics of deterministic systems*, Instituto de Matematica Pura e Applicada (IMPA), Rio de Janeiro, 1997

[25] R. Zweimüller, "Ergodic structure and invariant densities of non-Markovian interval maps with indifferent fixed points", *Nonlinearity* 11 (1998), 1263–1276

[26] R. Zweimüller, "Invariant measures for generalized induced transformations", *Proc. AMS*, 133 n.8, 2283–2295

Bibliography for Part II

[1] S. Alamouti, "A Simple Transmit Diversity Technique for Wireless Communications", IEEE Journal on select areas in communications, vol 16, n. 8, 1998

[2] S. Benedetto, E. Biglieri, *Principles of Digital Transmission with Wireless Applications*, Kluwer 1999

[3] E. Biglieri, *Coding for wireless channels*, Springer, 2005

[4] J-C. Belfiore, Y. Hong, E. Viterbo, "Golden Space-Time trellis coded modulation", *IEEE Trans. Inform. Theory*, vol 53 n. 5, 2007

[5] J-C. Belfiore, G. Rekaya, E. Viterbo, "The Golden Code: a 2×2 full-rate Space-Time Code with non-vanishing determinants", *IEEE Trans. Inform. Theory*, vol 51 n.4, 2005

[6] J. Boutros, E. Viterbo, C. Rastello, J.-C. Belfiore, "Good lattice constellations for both Rayleigh fading and Gaussian channels.", *IEEE Trans. Inform. Theory* vol 42 n.2, 1996

[7] D. Champion, J.-C. Belfiore, G. Rekaya and E. Viterbo, "Partitioning the Golden Code: A framework to the design of Space-Time coded modulation", *Canadian Workshop on Information Theory*, 2005

[8] J. H. Conway, N.J.A. Sloane, *Sphere packings, lattices and groups*, Springer-Verlag 1999

[9] M. Damen, N. Beaulieu, "On two high-rate algebraic space-time codes", *IEEE Trans. Inform. Theory*, vol. 49, pp. 1059 – 1063, 2003

[10] M. Damen, A. Tewfik, J-C. Belfiore, "A construction of a space-time code based on number theory", *IEEE Trans. Inform. Theory*, vol. 48 n. 3, 2002

[11] M. Damen, H. El Gamal, "Universal Space-Time Coding", *IEEE Trans. Inform. Theory*, vol 49 n. 5, 2003

[12] P. Elia, P. V. Kumar, S. A. Pawar, R. R. Kumar, B. S. Rajan, H. F. Lu, "Diversity-multiplexing tradeoff analysis of a few algebraic space-time constructions", *Proc. Allerton Conf. Comm. Control and Computing*, 2004

[13] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas", *Bell Labs Technical Journal*, vol 1, no. 2, pp. 41–59, 1996

[14] T. W. Hungerford, *Algebra*, Springer-Verlag 1974

[15] D. A. Marcus, *Number Fields*, Springer-Verlag

[16] F. J. MacWilliams, N. J. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1978

[17] F. Oggier, G. Rekaya, J.-C. Belfiore, E. Viterbo, "Perfect Space-Time Blocks Codes", *IEEE Trans. Inform. Theory*, vol. 52 n.9, 2006

[18] F. Oggier, E. Viterbo, *Algebraic number theory and code design for Rayleigh fading channels*, Now Publishers, 2004

[19] I. Reiner, *Maximal Orders*, Clarendon Press, Oxford 2003

[20] G. Rekaya, "Nouvelles constructions algébriques de codes spatio-temporels atteignant le compromis "multiplexage-diversité"", Thèse de l'Ecole Nationale Supérieure des Télécommunications, Paris, 2004

[21] A. Shokrollahi, "A note on double antenna diagonal space-time codes", `http://mars.bell-labs.com/cm/ms/what/mars/papers/2ant/`

[22] B.A. Sethuraman, B. S. Rajan, V. Shashidhar, "Full-Diversity, high-rate Space-Time Block Codes from division algebras", *IEEE Trans. Inform. Theory*, vol. 49 n.10, 2003

[23] I. N. Stuart, D. O. Tall, *Algebraic Number Theory*, Chapman and Hall, 1979

[24] V. Tarokh, N. Seshadri, A. Calderbank, "Space-Time codes for high data rate wireless communications: performance criterion and code construction", *IEEE Trans. Inform. Theory*, vol 44 n. 2, 1998

[25] M-F. Vignéras, *Arithmétique des Algèbres de Quaternions*, Lecture Notes in Mathematics, Springer Verlag 1980

[26] L. Zheng, D. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple antenna channels", *IEEE Trans. Inform. Theory*, vol 49 (2003), pp 1073–96

Index for Part I

Index for Part II