European Women in Mathematics – Marseille 2003

Proceedings of the 11th conference of EWM

K. Dajani, J. von Reis (eds.)

2000 Mathematics Subject Classification:

11K16	Normal numbers, radix expansions, etc.
11F06	Structure of modular groups and generalizations; arithmetic groups
11A55	Continued fractions
11K50	Metric theory of continued fractions
28D20	Entropy and other invariants
37B10	Symbolic dynamic
42A38	Fourier and Fourier-Stieltjes transforms and other transforms of Fourier type
42C40	Wavelets
65M12	Stability and convergence of numerical methods
93B05	Controllability

ISBN 90 6196 527 6

NUGI-code: 811

Printed in the Netherlands

Contents

From the editors

These proceedings of the 11th general meeting of the EWM contain a selection of the talks given at the conference as well as information about the activities and workings of the organization itself. The meeting, held in November 2003 at Luminy contained three sessions of invited talks, as well as a poster session and activities related to the governance and development of the EWM.

All of the talks included here have been through a refereeing process. We particularly thank the referees for their advice and insight and the contributors who made this publication possible.

The Center International de Rencontres Mathematiques in Luminy, France provided a scenic and practical setting for the meeting, and we would like to thank the staff of CIRM for ensuring a pleasant and effective conference.

Finally, we extend our thanks to our publisher CWI, the Center for Mathematics and Computer Science in Amsterdam for their understanding and assistance. We also would like to thank Cor Kraaikamp for his valuable Latex expertise and help.

The editors, Karma Dajani and Jennifer von Reis

Preface

This volume contains reports on some of the talks given at the 11th General Meeting of the association "European Women in Mathematics" (EWM). The meeting took place near Marseille, France, November 3–7 2003 and gathered fourty-four women from France, Italy, the UK, Germany, Denmark, Sweden, Norway, Finland, The Netherlands, Serbia and Montenegro, Russia, Ukraine, the Czech Republic, Algeria, Morocco and the USA.

Since it started in 1986, EWM has developped a tradition of gathering women mathematicians from all specialities, pure and applied, to meet and share mathematics and reflections on issues related to being a woman in mathematics. Because we wish to discuss mathematics together across specialties, we have been led to plan talks that are meant to be accessible to non-specialists, and nonetheless present the latest research.

Proceedings for the 1991 and 1995 meetings in Luminy and in Madrid, as well as for the interdisciplinary workshop on Renormalization in 1996 were copied and distributed by the association. For the Trieste and Hanover meetings (1997 and 1999), the Hindawi Publishing Corporation produced both a paper version and an electronic version, freely available at http://www.hindawi.com, whilst the Malta meeting in 2001 was published as a book by World Scientific. The present journal edition marks a new step towards a wider distribution among mathematicians.

The mathematical program consisted of three series of talks. The session in Pure mathematics, on Functional Analysis and Spectral Theory (with a orientation towards Ergodic Theory), chaired by Karma Dajani, with
Karma Dajani: *Measures of maximal entropy for random expansions in non-integer bases,*
Paola Loreti:*Ingham type theorems and applications,*
Svetlana Katok:*Livshitz Theorem for the unitary frame flow and its applications,*
Kathy Merrill: *Constructing Wavelets from Generalized Conjugate Mirror Filters,*
Martine Queffelec: *Fourier analysis and continued fractions,*
Anne Siegel: *Spectral theory for dynamical systems arisen by substitutions.*

The applied session devoted to Biomathematics was chaired by Alessandra Carbone, with talks by Natasha Jonoska: *DNA nanotechnology,*
Marie-France Sagot: *Bioinformatic,*
Rebecca Wade: *From protein structure to drug via the computer?*
Susan Holmes: *Using distances in Multidimensional Statistics.*

The interdisciplinary session devoted to Numerical Methods, chaired by Rosa Maria Spitaleri, with
Michelle Schatzman: *Preconditioning and partial differential equations:*
cross-fertilization of numerical analysis and PDE theory in unusual ways,
Rosa Maria Spitaleri: *Grid generation and partial differential equations: numerical methods and applications,*
Tatyana Kozubskaya: *Mathematical Models and Numerical Techniques in Euler Based Computational Aeroacoustics,*
Tatiana Vasileyva: *Historical aspects of numerical methods for ODEs,*
Zorica Uzelac: *A spline collocation method and a layer adapted mash for a singularity perturbed convection-diffusion problem,*

Dahbia Boukari: *Free surface flow over an obstacle: a numerical study.*
Abstracts of all the talks can be found at
http://www.math.helsinki.fi/ewm/meetings/luminy03.html

The new EWM web-based mentoring scheme formed the basis for a discussion session on the theme of mentoring. In addition, a poster session was held, where all the participants were invited to present themselves and their work. Abstracts of the posters are presented in this volume.

The organizing committee consisted of Laura Fainsilber (Sweden), chair, Valeri Berth (France), Elisabeth Remy (France), Aviva Szpirglas (France), Tatiana Ivanova (Russia), Sheung Tsun Tsou (United Kingdom), Irene Sciriha (Malta).

The venue for the conference was the CIRM (Centre International de Rencontres Mathematiques) in Luminy, between the city of Marseille and the "Calanques" coastal cliffs, a very pleasant and welcoming center for mathematical conferences. We are very grateful for sponsoring from the CIRM, the Institut Mathematique de Luminy, the city of Marseille, the region Provence-Alpes-Cote d'Azur, and donations from participants, Michelle Schatzman and others.

During the meeting, we showed the EWM video *Women and Mathematics across Cultures.* The exhibit *Women in mathematics, Why not you* produced by the French association *femmes et mathematiques* with beautiful portraits of 16 contemporary women mathematicians was hanged in the library of the CIRM and stayed there for one month to be seen also by participants in other conferences. This exhibit (in French or in English) is available to be shown in other places. (Contact: *femmes et mathematiques* http://www.femmes-et-maths.fr.fm/

As these proceedings go to press, we are planning the 12th General Meeting of EWM in Volgograd, Russia in the fall of 2005.

Warm thanks to all the participants, and especially to Karma Dajani and Jennifer Von Reis, editors of the proceedings.

Laura Fainsilber, Gteborg, September 2004.

Participants

Agranovich, Polina, Institute for Low Temperature Physics

email: agranovich@ilt.kharkov.ua

Berthe, Valerie, LIRMM

email: berthe@lirmm.fr

Boukari, Dahbia

email:

Carbone, Alessandra, University of Paris VI

email: carbone@inhes.fr

Castella, Sofie, Roskilde University

email: sic@ruc.dk

Charretton, Christine, University of Lyon 1

email: Christine.Charretton@univ-lyon1.fr

Dajani, Karma, Universiteit Utrecht

email: dajani@math.uu.nl

Demlova, Marie, Czech Technical University

email: demlova@math.feld.cvut.cz

Fainsilber, Laura, Chalmers University and Göteborg University

email: laura@math.chalmers.se

Flodén, Liselott, Mid Sweden University

email: lotta.floden@mh.se

Guillopé, Colette, University of Paris XII

email: guilope@univ-paris12.fr

Götmark, Elin, Göteborg University

email: elin@math.chalmers.se

Holmes, Susan, Stanford University

email: susan@stat.stanford.edu

Ivanova, Tatiana, BLTP JINR, Dubna, Russia

email: jonoska@math.usf.edu

Katok, Svetlana, Pennsylvania State University

email: Katok s@math.psu.edu

Knutson, Unger Johanne, Agder University

email: inger.j.knuson@hia.no

Kozubskaya, Tatyana Institute for Mathematical Modeling, Moscow

email: tata@imamod.ru

Laakso, Mari-Anna, University of Helsinki

email: mari.x.laakso@helsink.fi

Lipponen-Salhi, Marjo, University of Turku

email: marlip@utu.fi

Loreti, Paola, Università degli Studi di Roma

email: loreti@dmmm.uniroma1.it

Martinsson, Hanna

email: Hannam@math.chalmers.se

Merrill, Kathy, Colorado College

email: kmerill@coloradocollege.edu

Olsson, Marianne, Mid Sweden University

email: Marianne.Olsson@mh.se

Queffelec, Martine, University of Lille I email:martine@agat.univ-lille.fr

Rajae, Bemtaher, University My Ismail
email:bentaher@fsmek.ac.ma

Sagot, Marie-France, Inria Rhone-Alpes
email: Marie-France.Sagot@inria.fr

Schatzman,Michelle, U.CB.L.
email: schatz@maply.univ-lyon1.fr

Siegel, Anne, IRISA (CNRS) Rennes
email: Anne.Seigel@irisa.fr

Silfver, Jeanette, Mid Sweden University
email: jeanette.silfver@mh.se

Slavova, Angela, Bulgarian Academy of Science
email: slavova@math.bas.bg

Spitaleri, Rosa Maria,
email: spitaleri@iac.rm.cnr.it

Szpirglas, Aviva, IUFM Poitou-Charentes
email: aviva@math.univ-poitiers.fr

Tedeschini-Lalli, Laura, University of Rome III
email: tedeschi@mat.uniroma3.it

Tegnander. Laura, Norwegian University for Science
email: cathrine@math.ntnu.no

Timmerman, Stine, Roskilde University
email: stinet@ruc.dk

Tsou, Sheung Tsun, University of Oxford
email: tsou@maths.ox.ac.uk

Twarock, Reidun, Center for Mathematical Science
email: r.twarock@city.ac.uk

Vallilieva, C Essia, Universiaet zu Koeln
email: vaska 99@mail.ru

Vasileyva, Tatiana, Volgograd State University
email: tatiana vas@mail.ru

Vinerean, Mireal, Karlstad University
email: mirela.vinerean@kau.se

von Reis, Jennifer, University of Turku
email: jenvon@utu.fi

Wade, Rebecca, EML Research
email: rebecca.wade@eml-r.villa-bosche.de

Wulcan, Elizabeth, Chalmers University
email: wulcan@math.chalmers.se

Constructing Wavelets from Generalized Filters

Kathy Merrill

Department of Mathematics
Colorado College
Colorado Springs, CO 80903-3294
USA
kmerrill@coloradocollege.edu

Abstract. Over the past twenty years, wavelets have gained popularity as bases for transforms used in image and signal processing. We begin by showing how wavelets arise naturally in this context. Classical construction techniques using Fourier analysis are then presented. The paper concludes with recent extensions of these techniques employing the tools of abstract harmonic analysis and spectral multiplicity theory.

1. Introduction

Wavelets arise naturally in efforts to store images efficiently. To capture a black and white image on a 1600 by 1200 pixel computer screen we might first try storing a gray scale number between 0 and 255 for each of the 1,920,000 pixels. However, pixel by pixel storage is not very efficient, because it does not take advantage of regions in which the darkness does not change. For example, there are clearly more efficient ways to store an image of a black rectangle covering half of the screen, than to keep 960,000 copies of the number 0 and 960,000 copies of the number 255. Even a photograph of a face usually has large regions of constant darkness.

To overcome the inefficiency of pixel by pixel storage, we would like to use different levels of resolution in different regions of the image. In areas where darkness is highly variable, we need a higher level of resolution than in areas where it stays constant. As a first step toward this goal, we capture the whole image at different levels of resolution as follows: First we record the average gray scale on the whole image, which for convenience we think of as occupying the unit square. (For more general images, we can think of averaging over each of the 1×1 squares whose vertices are lattice points.) We call this the 0^{th} level of resolution. Then we record the average on each $\frac{1}{2} \times \frac{1}{2}$ subsquare, which we call the 1^{st} level of resolution. We can proceed to the resolution of single pixels by successively averaging our image and recording that average on each of the $\frac{1}{2^j} \times \frac{1}{2^j}$ subsquares (called the j^{th} level of resolution), for larger and larger j. This process yields a sequence of approximations to our image. We will have captured our image completely accurately at the j^{th} level if it was of constant darkness on all of the subsquares of a $\frac{1}{2^j} \times \frac{1}{2^j}$ grid.

Mathematically, we can describe this process in terms of a sequence of closed subspaces of $L^2(\mathbb{R}^2)$ given by $V_j =$ functions constant on $\frac{1}{2^j} \times \frac{1}{2^j}$ squares. Our approximation at the j^{th} level of resolution is simply the closest L^2 approximation to our image in the subspace V_j. If we allow ourselves to both zoom in and zoom out arbitrarily far, i.e. to consider $-\infty < j < \infty$, we will have a structure of the following type, first defined by S. Mallat [**13**]:

Definition 1. *A* Multiresolution Analysis (MRA) *in* $L^2(\mathbb{R}^n)$ *is a collection of closed subspaces* V_j *that have the following properties:*

(1) $V_j \subset V_{j+1}$

(2) $V_{j+1} = \{\delta(f) \equiv 2f(2x)\}_{f \in V_j}$
(3) $\cup V_j$ *is dense in* $L^2(\mathbb{R}^n)$ *and* $\cap V_j = \{0\}$
(4) V_0 *has a* scaling function ϕ *whose translates form an orthonormal basis for* V_0

Property 2 explicitly defines a dilation operator on $L^2(\mathbb{R}^n)$ that takes us between different levels of resolution. The normalization factor of 2 makes this dilation a unitary operator. The first three properties together describe how the different levels of resolution are related in a way that reflects the successive capturing of our image. The final property describes how we can use a second unitary operator of translation to move around at the 0^{th} level (and thus at any fixed level if we conjugate by dilation). In our image example, ϕ is the characteristic function of the unit square.

By using an MRA, we have achieved our preliminary goal of capturing our image at different levels of resolution. However, we have not yet gained efficiency over pixel by pixel storage unless our image is, like the rectangle, an element of one of the V_j spaces. Indeed, if we continue our process down to the level of pixel by pixel resolution, we will have all the inefficiency we started with, together with information from all the previous levels of resolution as well. The problem is that we are starting over at each level, so that there is redundancy in the information stored at successive levels. To see an explicit example of this, notice that in going from the 0^{th} level to the 1^{st}, we already know the overall average gray scale value, and thus would only need to record the averages on three of the four subsquares to have total information about all four subsquare averages.

To overcome the redundancy, instead of storing all of the V_1 information in addition to V_0's, we write $V_1 = V_0 \oplus W_0$ and seek an orthonormal basis for W_0. In our example, we let q_1, q_2, q_3, and q_4 be the upper left, upper right, lower left, and lower right quadrants of the unit square respectively, and let

$$\psi_1 = \chi_{q_1 \cup q_2} - \chi_{q_3 \cup q_4},$$
$$\psi_2 = \chi_{q_1 \cup q_3} - \chi_{q_2 \cup q_4}$$

and

$$\psi_3 = \chi_{q_1 \cup q_4} - \chi_{q_2 \cup q_3},$$

where χ_A denotes the characteristic function of the set A. Then the translates of ψ_1, ψ_2, ψ_3 and ϕ form an orthornormal basis for V_1. In fact, positive and negative dilates of translates of just ψ_1, ψ_2 and ψ_3 form an orthonormal basis for $L^2(\mathbb{R}^2)$. Storing the coefficients of our image in terms of its coefficients for the orthonormal basis given by the dilates and translates of the ψ's does finally achieve the image compression we were seeking. At each level, the new information given by the coefficients of the further dilated ψ's can be thought of as correction terms to update the information from the previous level of resolution. In regions of the image where darkness does not change, these coefficients will all eventually be 0. Thus we achieve *lossless* compression from the savings of storing sequences containing lots of zeroes. We can accomplish further compression with the least loss of accuracy in the image by throwing away the coefficients that are smallest in absolute value. Since our dilation operator normalizes at each step, the coefficients will give an accurate measure of the relative importance of the correction terms.

The ψ's are called a wavelet. In general we have:

Definition 2. $\{\psi_k\}_{k=1\ldots r} \subset L^2(\mathbb{R}^n)$ *is an* orthonormal wavelet *for dilation by an integral expansive matrix* D *if* $\{\psi_{j,k,l} \equiv \sqrt{|\det D|}^j \psi_k(D^j x - l)\}_{j,l \in \mathbb{Z}; k=1\ldots r}$ *form an orthonormal basis for* $L^2(\mathbb{R}^n)$.

We began our description of wavelets in $L^2(\mathbb{R}^2)$ in order to show their relationship to the problem of image compression, but the simplest place to study wavelets is in 1-dimension. Three well-known and simple examples of wavelets for dilation by 2 in $L^2(\mathbb{R})$ are the Haar wavelet [**11**],

$$\psi = \chi_{[0,\frac{1}{2})} - \chi_{[\frac{1}{2},1)},$$

the Shannon wavelet, for which

$$\widehat{\psi} = \chi_{[-1,-\frac{1}{2})\cup[\frac{1}{2},1)}$$

and the Journé wavelet [**13**], with

$$\widehat{\psi} = \chi_{[-\frac{16}{7},-2)\cup[-\frac{1}{2},-\frac{2}{7})\cup[\frac{2}{7},\frac{1}{2}]\cup[2,\frac{16}{7})}.$$

We can think of the 2-dimensional example we developed above as being built out of tensor products of the 1-dimensional Haar wavelet. The other 1-dimensional examples given here are described in terms of their Fourier transform $\widehat{\psi}$. They are interesting as wavelets because of the simplicity of these transforms. In particular, the Shannon wavelet is so-named because of its relationship to the Shannon sampling formula [**15**]. Note that on the Fourier transform side, translation becomes multiplication by exponentials. We will see in the next section that this fact makes the Fourier transform very useful in the study of wavelets.

Two basic questions arise in looking at the examples given above:

(1) How do we find new wavelets with desirable properties? The examples of wavelets we have described so far all have their drawbacks. The Haar wavelet in either dimension 1 or 2 is discontinuous, and thus is not the best basis to use to capture smooth images. The Shannon and Journé wavelets have Fourier transforms that are discontinuous, and thus are not well localized. Can we find a smooth wavelet for dilation by 2 in $L^2(\mathbb{R})$ that still has compact support? Another natural question about finding new wavelets concerns the number of ψ_k's required. The fact that our examples so far consist of single wavelets ($r = 1$ in Definition 2) in 1 dimension, but a 3-wavelet in 2 dimensions also raises the question of whether we can find a 1-wavelet or 2-wavelet for dilation by 2 in $L^2(\mathbb{R}^2)$.

(2) How strong is the connection between wavelets and MRA's? Although we motivated the definition of wavelet using the idea of an MRA, it turns out that some wavelets (for example, the Journé wavelet above) have no associated MRA's.

The first question is easiest to answer if we assume we have an MRA for dilation by 2 in $L^2(\mathbb{R})$. This is the setting in which Meyer [**21**] and Daubechies [**10**] carried out their famous construction of wavelets using filters. We describe that work in Section 2 below. Their answer can then be generalized to situations where no MRA is possible. This leads to the work of Baggett, Courter, Jorgensen, Medina, Packer and Merrill, which is described in Section 3.

2. Building MRA wavelets from filters in $L^2(\mathbb{R})$

Suppose we have a single wavelet ψ for dilation by 2 in $L^2(\mathbb{R})$, with an associated MRA, and so a scaling function ϕ whose translates form an orthonormal basis for V_0.

Because $\widehat{V_0} \subset \widehat{V_1}$, and $\widehat{W_0} \subset \widehat{V_1}$, we can write $\widehat{\phi} \in \widehat{V_0}$ and $\widehat{\psi} \in \widehat{W_0}$ in terms of exponentials times the dilate of $\widehat{\phi}$. That is, there must exist periodic functions (with period 1) h and g such that

$$\widehat{\phi}(x) = \frac{1}{\sqrt{2}} h(\frac{x}{2})\widehat{\phi}(\frac{x}{2}) \tag{1}$$

and

$$\widehat{\psi}(x) = \frac{1}{\sqrt{2}} g(\frac{x}{2})\widehat{\phi}(\frac{x}{2}). \tag{2}$$

EXAMPLES: For the Shannon wavelet, where $\widehat{\psi} = \chi_{[-1,-\frac{1}{2})\cup[\frac{1}{2},1)}$ and $\widehat{\phi} = \chi_{[-\frac{1}{2},\frac{1}{2})}$, we have

$$h = \sqrt{2}\chi_{[-\frac{1}{4},\frac{1}{4})} \text{ and } g = \sqrt{2}\chi_{[-\frac{1}{2},-\frac{1}{4})\cup[\frac{1}{4},\frac{1}{2})}.$$

For the Haar wavelet, where $\phi = \chi_{[0,1)}$ and $\psi = \chi_{[0,\frac{1}{2})} - \chi_{[\frac{1}{2},1)}$, we have

$$h = \frac{1}{\sqrt{2}}(1 + e^{2\pi i x}) \text{ and } g = \frac{1}{\sqrt{2}}(e^{2\pi i x} - 1).$$

The functions h and g are called **low and high pass filters**. Notice that in the Shannon example in particular, h and g do indeed act by filtering out all but low (for h) or high (for g) frequencies. Because of the orthonormality conditions satisfied by translates of ϕ and ψ, all filters defined by (1) and (2) must satisfy orthonormality-like conditions:

$$|h(x)|^2 + |h(x+\frac{1}{2})|^2 = 2 \tag{3}$$

$$|g(x)|^2 + |g(x+\frac{1}{2})|^2 = 2 \tag{4}$$

and

$$h(x)\overline{g(x)} + h(x+\frac{1}{2})\overline{g(x+\frac{1}{2})} = 0. \tag{5}$$

The reason filters are useful is that we can reverse this process of finding filters from wavelets. First note that we can easily build a high pass filter to go with any low pass filter. Indeed, if h is any periodic function that satisfies (3), we can take

$$g(x) = e^{2\pi i x}\overline{h(x+\frac{1}{2})},$$

and the other two orthonormality conditions (4) and (5) will be satisfied as well. (Other choices for g are possible.) Given h and g, under appropriate conditions, we can then build $\widehat{\phi}$ by iterating equation (1). The appropriate conditions are exactly those that are needed to make the resulting infinite product converge.

THEOREM 1. *Let h and g be C^1 functions that satisfy the orthonormality conditions (3), (4), and (5). Suppose, in addition, that h is nonvanishing on $[-\frac{1}{4},\frac{1}{4})$, and $|h(0)| = \sqrt{2}$. Then:*

$$\widehat{\phi}(x) = \prod_{j=1}^{\infty} \frac{1}{\sqrt{2}} h(2^{-j}x)$$

is a scaling function for an MRA, and

$$\widehat{\psi}(x) = \frac{1}{\sqrt{2}} g(\frac{x}{2})\widehat{\phi}(\frac{x}{2}).$$

is an orthonormal wavelet.

This technique was developed by Mallat [**13**] and Meyer [**21**], and used by Daubechies [**10**] to build C^r wavelets with compact support. (It can be shown that there are no C^∞ wavelets with compact support.) Daubechies' construction uses powers of the trigonometric identity $\sin^2(x) + \cos^2(x) = 1$, which fits naturally into equation 3 to find the lowpass filter h. A good introductory description of these constructions appears in [**16**].

The theorem's requirements that h satisfy $|h(0)| = \sqrt{2}$ and h be in C^1 are natural restrictions in order to make the infinite product converge. The condition that h is nonvanishing on $[-\frac{1}{4}, \frac{1}{4}]$ appears less natural; it is used in the proof to ensure L^2 convergence of the infinite product and thus the orthonormality of the translates of ϕ. A famous example due to A. Cohen [**6**] showed that removing the condition h nonvanishing on $[-\frac{1}{4}, \frac{1}{4})$ can in fact lead to functions ϕ and ψ whose translates are not orthonormal. Cohen took $h = \frac{1+e^{-6\pi i x}}{\sqrt{2}}$, which resulted in a stretched out version of the Haar scaling function and wavelet, $\phi = \frac{1}{3}\chi_{[0,3)}$ and $\psi = \frac{1}{3}(\chi_{[-\frac{1}{2},1)} - \chi_{[1,\frac{5}{2})})$.

However, the classical Theorem 1 can be extended to accommodate this and similar examples if we generalize our definition of wavelet.

Definition 3. *$\{\psi_{j,k,l}\}$ is a* normalized tight frame *for $L^2(\mathbb{R}^n)$ if for each $f \in L^2$ we have $\|f\|^2 = \sum_{j,k,l} |\langle f|\psi_{j,k,l}\rangle|^2$.*

$\{\psi_k\} \subset L^2(\mathbb{R}^n)$ is a frame wavelet *for dilation by an integral expansive matrix D if $\{\psi_{j,k,l} \equiv \sqrt{|\det D|}^j \psi_k(D^j x - l)\}$ form a normalized tight frame for $L^2(\mathbb{R}^n)$.*

Note that a normalized tight frame can exhibit redundancy, and therefore need not be a basis. Indeed, it can include 0 as one of its elements. However, a normalized tight frame $\{f_j\}$ does have the property that every $f \in L^2$ can be recaptured from its coefficients, $f = \sum\langle f, f_j\rangle f_j$. It turns out that a normalized tight frame of unit vectors must be an orthonormal basis.

By broadening our definition of wavelet to include frame wavelets, we get the following generalization of Theorem 1, which appears in [**5**] and was proven independently in [**12**] for $d = 2$:

Theorem 2. *Suppose $h, g_1, \cdots g_{d-1}$ are periodic Lipschitz continuous function in $L^2(\mathbb{R})$, which satisfy $|h(0)| = \sqrt{d}$ and the filter equations*

(1) $\sum_{l=0}^{d-1} |h(x + \frac{l}{d})|^2 = d$
(2) $\sum_{l=0}^{d-1} h(x + \frac{l}{d})\overline{g_i(x + \frac{l}{d})} = 0$
(3) $\sum_{l=0}^{d-1} g_i(x + \frac{l}{d})\overline{g_j(x + \frac{l}{d})} = d\delta_{i,j}$

then the construction $\widehat{\phi}(x) = \prod_{j=1}^{\infty} \frac{1}{\sqrt{d}} h(d^{-j}x)$ produces an L^2 function ϕ (whose translates are not necessarily orthogonal), and the $d - 1$ functions

$$\widehat{\psi_k}(x) = \frac{1}{\sqrt{d}} g_k(\frac{x}{d})\widehat{\phi}(\frac{x}{d})$$

form a frame wavelet for dilation by d in $L^2(\mathbb{R})$.

This extension starts with a filter from an MRA orthonormal wavelet and produces a frame wavelet that need not be associated with an MRA. Thus it suggests that we look again at the connection between wavelets and MRA's, and consider the possibility of more general filters.

3. Generalized Multi-resolution Analyses and Generalized Filters

As mentioned in the introduction, even orthonormal wavelets need not be associated with an MRA. The reason for this lies in the MRA requirement of the existence of a scaling function. Given an orthonormal wavelet $\{\psi_k\} \subset L^2(\mathbb{R}^n)$, if we let $V_j =$ the closed linear span of $\{\psi_{i,k,l}\}_{i<j}$, then the subspaces $\{V_j\}$ do determine a *generalized* MRA, according to the definition below:

Definition 4. *A* Generalized Multiresolution Analysis (GMRA) *is a collection of closed subspaces* $\{V_j\}_{j\in\mathbb{Z}}$ *of* $L^2(\mathbb{R}^n)$ *such that:*

(1) $V_j \subset V_{j+1}$
(2) $V_{j+1} = \{\delta_D(f) \equiv \sqrt{|\det D|} f(Dx)\}_{f\in V_j}$
(3) $\cup V_j$ *dense in* $L^2(\mathbb{R}^n)$ *and* $\cap V_j = \{0\}$
(4) V_0 *is invariant under translation.*

The definitions of MRA and GMRA differ only in condition (4): An MRA requires that V_0 has a scaling function ϕ such that translates of ϕ form an orthonormal basis for V_0, while a GMRA requires only that V_0 be invariant under translation by the integer lattice. In spite of this difference, it is shown in [**3**] that a GMRA has almost as much structure as an MRA. Translation is a unitary representation of $\mathbb{Z}^n$ on V_0, and thus is completely determined by a multiplicity function $m : [-\frac{1}{2}, \frac{1}{2})^n \mapsto \{0, 1, 2, \cdots, \infty\}$ describing how many times each character occurs as a subrepresentation. A GMRA is an MRA iff $m \equiv 1$. Journé's famous non-MRA wavelet example for dilation by 2 in $L^2(\mathbb{R})$ has

$$m(x) = \begin{cases} 2 & x \in [-\frac{1}{7}, \frac{1}{7}) \\ 1 & x \in \pm[\frac{1}{7}, \frac{2}{7}) \cup \pm[\frac{3}{7}, \frac{1}{2}) \\ 0 & \text{otherwise} \end{cases}$$

In any GMRA, we write $V_1 = V_0 \oplus W_0$, just as we did in the MRA case. Representation theory can then be used (see [**3**]) to show that the GMRA has an associated orthonormal wavelet if and only if the multiplicity function satisfies a *consistency equation*:

$$m(x) + (\text{number of wavelets}) = \sum m(\text{preimages of } x \text{ under } D^t).$$

Taking $m \equiv 1$, this consistency equation determines that the number of wavelets must be 3 for an MRA wavelet for dilation by 2 in $L^2(\mathbb{R}^2)$, so 1-wavelets cannot be found there using the MRA filter technique. However, examples of non-MRA, GMRA orthonormal wavelets whose Fourier transforms are characteristic functions can be built directly from the consistency equation. We have used this technique (in [**3**]and [**1**]) to make 1-wavelets and a 2-wavelet in $L^2(\mathbb{R}^2)$, and a 4-wavelet in $L^2(\mathbb{R}^3)$. (By the consistency equation, MRA wavelets require 7-wavelets in $L^2(\mathbb{R}^3)$.) Other examples of non-MRA 1-wavelets in $L^2(\mathbb{R}^2)$ appear in, e.g. [**9**] and [**4**].

In every GMRA, there is a unitary equivalence between translation on V_0 and multiplication by exponentials on $\oplus L^2(S_j)$, where $S_j = \{x : m(x) \geq j\}$. This unitary equivalence plays a role here similar to that of the Fourier transform in the classical MRA case. It ensures that in a GMRA we can find *generalized scaling functions*, $\phi_1, \phi_2, \cdots$ such that $\{\phi_i(x - l)\}$ form a normalized tight frame for V_0. It also enables us to develop a generalized notion of low and high pass filters.

Just as in the classical case, we begin by building filters from wavelets, then see if we can reverse the process. Suppose we have a non-MRA orthonormal k-wavelet $\psi_1, \cdots, \psi_k$ for dilation by D in $L^2(\mathbb{R}^n)$. We then have generalized scaling functions $\phi_1, \cdots, \phi_c$. Since $V_0 \subset V_1$ and $W_0 \subset V_1$, we can show ([**1**]) there exist periodic functions $h_{i,j}$ and $g_{l,j}$, supported on the periodization of S_j, such that

$$\widehat{\phi_i}(x) = \frac{1}{\sqrt{|\det D|}} \sum_{j=1}^{c} h_{i,j}((D^t)^{-1}x)\widehat{\phi_j}((D^t)^{-1}x) \tag{6}$$

and

$$\widehat{\psi_l}(x) = \frac{1}{\sqrt{|\det D|}} \sum_{j=1}^{c} g_{l,j}((D^t)^{-1}x)\widehat{\phi_j}((D^t)^{-1}x). \tag{7}$$

These *generalized filters* $g_{i,j}$ and $h_{i,j}$ satisfy orthonormality-like conditions that are generalizations of the classical conditions (3),(4) and (5):

$$\sum_{j=1}^{c} \sum_{l=0}^{|\det D|-1} h_{i,j}(x_l)\overline{h_{k,j}(x_l)} = (\det D)\delta_{i,k}\chi_{S_i}(x), \tag{8}$$

$$\sum_{j=1}^{c} \sum_{l=0}^{|\det D|-1} g_{i,j}(x_l)\overline{g_{k,j}(x_l)} = (\det D)\delta_{i,k}, \tag{9}$$

and

$$\sum_{j=1}^{c} \sum_{l=0}^{|\det D|-1} h_{i,j}(x_l)\overline{g_{k,j}(x_l)} = 0, \tag{10}$$

where the x_l are the preimages of x under D^t mod 1.

In the case of the Journé wavelet, equations (6) and 7 simplify to:

$$\widehat{\phi_1}(x) = \frac{1}{\sqrt{2}}\left(h_{1,1}((\frac{x}{2})\widehat{\phi_1}(\frac{x}{2}) + h_{1,2}((\frac{x}{2})\widehat{\phi_2}(\frac{x}{2})\right)$$

$$\widehat{\phi_2}(x) = \frac{1}{\sqrt{2}}\left(h_{2,1}((\frac{x}{2})\widehat{\phi_1}(\frac{x}{2}) + h_{2,2}((\frac{x}{2})\widehat{\phi_2}(\frac{x}{2})\right),$$

and

$$\widehat{\psi}(x) = \frac{1}{\sqrt{2}}\left(g_1((\frac{x}{2})\widehat{\phi_1}(\frac{x}{2}) + g_2((\frac{x}{2})\widehat{\phi_2}(\frac{x}{2})\right),$$

which look very similar to the equations (1) and (2) that define classical low and high pass filters. We can use these to find the Journé generalized filters from the wavelet

$$\widehat{\psi} = \chi_{[-\frac{16}{7},-2)\cup[-\frac{1}{2},-\frac{2}{7})\cup[\frac{2}{7},\frac{1}{2}]\cup[2,\frac{16}{7})},$$

and generalized scaling functions

$$\widehat{\phi_1}(x) = \chi_{[-\frac{4}{7},-\frac{1}{2})\cup[-\frac{2}{7},\frac{2}{7})\cup[\frac{1}{2},\frac{4}{7})}, \widehat{\phi_2}(x) = \chi_{[-\frac{8}{7},-1)\cup[1,\frac{8}{7})}.$$

We obtain (see [7])

$$h_{1,1} = \sqrt{2}\chi_{[-\frac{2}{7},-\frac{1}{4})\cup(-\frac{1}{7},\frac{1}{7})\cup[\frac{1}{4},\frac{2}{7})}\ , h_{1,2} = 0,$$

$$h_{2,1} = \sqrt{2}\chi_{[-\frac{4}{7},-\frac{1}{2})\cup[\frac{1}{2},\frac{4}{7})}\ , h_{2,2} = 0$$

$$g_1 = \sqrt{2}\chi_{[-\frac{1}{4},-\frac{1}{7})\cup[\frac{1}{7},\frac{1}{4})}\ , \text{and } g_2 = \sqrt{2}\chi_{[-\frac{1}{7},\frac{1}{7})}.$$

To use generalized filters to build wavelets, we now wish to reverse this procedure, just as we did in the classical case. In order to first build filters, we use functions on the disjoint union of the S_j's whose values are $\sqrt{\det D}$ times unitary matrices, with different dimensions for different values of x. We need the values of the filters to be $\sqrt{\det D}$ times unitary matrices in order to satisfy the generalized orthonormality conditions (8), (9), and (10). The matrices of filter values have different dimensions depending on how many of the sets S_j the point x and its preimages are in. Once we have the filters, we build the generalized scaling function using an infinite product of matrices that comes from the iteration of equation (6). The wavelet is then produced by equation(7). Conditions that make this possible are described in the following generalization (see [**2**]) of the Bratteli-Jorgensen theorem:

THEOREM 3. *Suppose $\{h_{i,j}\}$ and $\{g_{k,j}\}$ are periodic functions that are supported on the periodization of S_j, Lipschitz continuous in a neighborhood of the origin, and that satisfy the three generalized orthonormality conditions (8),(9), and (10). Suppose in addition the $h_{i,j}$ satisfy the generalized lowpass conditions $h_{i,j} = 0$ for $j > i$ and $|h_{i,j}(0)| = \sqrt{(|\det D|)}\delta_{(i,1)}\delta_{(j,1)}$. Write H for the matrix $(h_{i,j})$. Then the components of $\prod_{k=1}^{\infty} \frac{1}{\sqrt{|\det D|}} H((D^t)^{-k}x)$ converge pointwise to $P_{i,j} \in L^2(\mathbb{R}^n)$. If we let $\widehat{\phi_i} = P_{i,1}$, then*

$$\widehat{\psi_k}(x) \equiv \frac{1}{\sqrt{|\det D|}} \sum_j g_{k,j}((D^t)^{-1}x)\widehat{\phi_j}((D^t)^{-1}x)$$

are the Fourier transforms of a frame wavelet on $L^2(\mathbb{R}^n)$.

The proof, like that of [**5**], proceeds by using matrices of values of the filters to define partial isometries that satisfy relations similar to those defining a Cuntz algebra [**8**].

Using this procedure, we have built wavelets with interesting properties, for example, a non-MRA orthonormal wavelet on $L^2(\mathbb{R})$ whose Fourier transform is C^∞ on an arbitrarily large interval (see [**1**]), and a non-MRA frame wavelet on $L^2(\mathbb{R})$ whose Fourier transform is C^∞ on all of $\mathbb{R}$ (see [**2**]). These examples are somewhat surprising since it is known that compactly supported wavelets on $\mathbb{R}$ must be MRA wavelets.

References

[1] L. W. Baggett, J. E. Courter, and K. D. Merrill, *The construction of wavelets from generalized conjugate mirror filters in $L^2(\mathbb{R}^n)$*, Appl. Comput. Harmon. Anal. **13** (2002), 201-223.

[2] L. Baggett, P. Jorgensen, K. Merrill, and J. Packer *Construction of Parseval wavelets from redundant systems of filters*, preprint.

[3] L. W. Baggett, H. A. Medina, and K. D. Merrill, *Generalized multi-resolution analyses and a construction procedure for all wavelet sets in $\mathbb{R}^n$*, J. Fourier Anal. Appl. **5** (1999), 563-573.

[4] J. Benedetto and M. Leon, *The construction of multiple dyadic minimally supported frequency wavelets on $\mathbb{R}^d$*, in *The Functional and Harmonic Analysis of Wavelets and Frames (San Antonio, TX)*, L. W. Baggett and D. R. Larson, eds., Contemp. Math. **247**, Amer. Math. Soc., Providence, RI, 1999, 43-74.

[5] O. Bratteli and P. Jorgensen, *Wavelets Through a Looking Glass: the World of the Spectrum*, Birkäuser, Boston-Basel-Berlin, 2002.

[6] A. Cohen, *Wavelets and Multiscale Signal Processing*, translated by R. Ryan, Chapman and Hall, U.K., 1995.

[7] J. Courter, *Construction of dilation d wavelets*, in *The Functional and Harmonic Analysis of Wavelets and Frames (San Antonio, TX)*, L. W. Baggett and D. R. Larson, eds., Contemp. Math. **247**, Amer. Math. Soc., Providence, RI, 1999, 183-205.

[8] J. Cuntz, *Simple C^*-algebras generated by isometries*, Comm. Math. Phys. **57** (1977), 173-185.

[9] X. Dai, D. R. Larson, and D. Speegle, *Wavelet Sets in $\mathbb{R}^n$*, J. Fourier Anal. Appl., **3** (1997), 451-456.

[10] I. Daubechies, *Ten Lectures on Wavelets*, American Mathematical Society, Providence, RI, 1992.

[11] A. Haar, *Zur theorie der orthogonalen funktionen systems*, Math. Ann. **69** (1910), 331-371.

[12] W. M. Lawton, *Tight frames of compactly supported affine wavelets*, J. Math. Phys. **31** (1990), 1898-1901.

[13] S. Mallat, *Multiresolution approximations and wavelet orthonormal bases of* $L^2(\mathbb{R})$, Trans. Amer. Math. Soc. **315** (1989), 69-87.

[14] Y. Meyer, *Wavelets and Operators*, Cambridge Studies in Advanced Mathematics v. 37, Cambridge University Press, Cambridge, England, 1992.

[15] C. E. Shannon, *Communications in the presence of noise*, Proc. Inst. Radio Eng. **37** (1949), 10-21.

[16] R. Strichartz, *How to make wavelets*, Amer. Math. Monthly **100** (1993), 539-556.

Spectral theory for dynamical systems arising from substitutions

Anne Siegel

IRISA CNRS-UMR 6074
Campus de Beaulieu
35042 Rennes Cedex
France
`Anne.Siegel@irisa.fr`

Abstract. Symbolic dynamical systems were first introduced to better understand the dynamics of geometric maps; particularly to study dynamical systems for which past and future are disjoint as for instance toral automorphisms or Pseudo-Anosov diffeomorphisms of surfaces. Self-similar systems are defined to be topologically conjugate to their own first return map on a given subset. A basic idea is that, as soon as self-similarity appears, a substitution is hidden behind the original dynamical system. In this lecture, we first illustrate this idea with concrete examples, and then, try to understand when symbolic codings provide a good representation. A natural question finally arises: which substitutive dynamical systems are isomorphic to a rotation on a compact group? Partial answers have been given by many authors since the early 60's. Then, we will see how a spectral analysis problem finally reduces to a combinatorial problem, whose partial answers imply Euclidean geometry and even some arithmetics.

1. What is a substitution ?

Let $\mathcal{A}$ be a finite alphabet and $\mathcal{A}^*$ the set of finite words on $\mathcal{A}$. The empty word is denoted ε.

1.1. Substitution on finite words. A *substitution* or *iterated morphism* is a combinatorial object that simply replaces letters in $\mathcal{A}$ by nonempty finite words. An example on the three-letters alphabet $\mathcal{A} = \{1, 2, 3\}$ is given by σ defined by $1 \mapsto 12$, $2 \mapsto 3$, $3 \mapsto 1$.

As dynamicians, our aim is to iterate this substitution. Hence we formally define a substitution as an endomorphism of the free monoid $\mathcal{A}^*$ endowed with the concatenation (defined by $\sigma(uv) = \sigma(u)\sigma(v)$), such that the image of each letter of $\mathcal{A}$ is nonempty, and such that for at least one letter, say a, the length of the successive iterations $\sigma^n(a)$ tends to the infinity (these two conditions ensure that the substitution can be iterated infinitely).

Then, the successive iterations of the example σ previously defined applied on the letter 1 give

1

12

123

1231

123112

123112123

1231121231231

12311212312311231 12

123112123123112311212311 2123

12311212312311231121231123231231121231231

One should notice that for all n, each word $\sigma^n(1)$ starts with the preceeding one $\sigma^n(1)$. Roughly, it is natural to call the infinite iteration of these words a *fixed point* of σ. However, such a fixed point appears to be an infinite word defined as a limit, so that we need to introduce now a couple of formal definitions about infinite sequences and topology.

1.2. Extension of a substitution to infinite words. A (finite or infinite) *word* on $\mathcal{A}$ is denoted $w = w_0w_1\dots$. The metrizable topology of the set of infinite words $\mathcal{A}^{\mathbb{N}}$ is the product topology of the discrete topology on each copy of $\mathcal{A}$. A *cylinder* of $\mathcal{A}^{\mathbb{N}}$ is a closed-open set of the form: $[W] = \{(w_i)_i \in \mathcal{A}^{\mathbb{N}} | w_0 \dots w_{|W|-1} = W\}$ for $W \in \mathcal{A}^*$.

A substitution naturally extends by concatenation to the set of infinite words $\mathcal{A}^{\mathbb{N}}$:

$$\sigma(w_0w_1\dots) = \sigma(w_0)\sigma(w_1)\dots$$

A *periodic point* of a substitution σ is an infinite word $u = (u_i)_{i\in\mathbb{N}} \in \mathcal{A}^{\mathbb{N}}$ that satisfies $\sigma^\nu(u) = u$ for some $\nu > 0$. If $\sigma(u) = u$, then u is a *fixed point* of σ. A simple combinatorial proof states that a substitution may not always have a fixed point, but it always admits at least one periodic point [**25**].

1.3. What is symbolic dynamics? A specialist in dynamical systems always looks for maps acting on objects. Dealing with infinite sequences, a natural map immediatly appears, that is, the deletion of the first letter of the word. Formally, we denote by S the *shift map* on $\mathcal{A}^{\mathbb{N}}$ defined by $S((w_i)_{i\in\mathbb{N}}) = (w_{i+1})_{i\in\mathbb{N}}$.

Symbolic dynamics consists in studying the shift map on a closed set of infinite sequences of $\mathcal{A}^{\mathbb{N}}$, which is supposed to be invariant through the action of the shift map. We are particularly interested in symbolic sets that are minimal, that is, that do not contain a strictly smaller closed invariant subset.

1.4. Symbolic dynamical system associated with a substitution. Dealing with a substitution, a natural process to associate with it a symbolic dynamical consists in first building a fixed point (or a periodic point if a fixed point does not exist) by iteration, then shifting this infinite word infinitely many often (then one gets the *orbit* of the sequence through the shift point), and finally considering the closure of this orbit. However, this process should be interesting provided that when further periodic points do exist, they generate the same symbolic system.

Formally, the *symbolic dynamical system* generated by a word u is the pair (X_u, S), where X_u denotes the closure in $\mathcal{A}^{\mathbb{N}}$ of the orbit $\{S^n u,\ n \in \mathbb{N}\}$ of u under the shift map. The shift map S is an homeomorphism on this compact subset of $\mathcal{A}^{\mathbb{N}}$.

We call a substitution σ *primitive* if there exists an integer ν (independent of the letters) such that, for each pair $(a, b) \in \mathcal{A}^2$, the word $\sigma^\nu(a)$ contains at least one occurrence of the letter b.

THEOREM 4 (see [**25, 13**]). *Let σ be a primitive substitution. If u is a periodic point for σ, then X_u does not depend on u and we denote by (X_σ, S) the* symbolic dynamical system generated by σ. *The system (X_σ, S) is* minimal *and* uniquely ergodic*: X_σ contains no non-empty closed shift-invariant subset and there exists a unique shift-invariant probability measure μ_{X_σ} on X_σ.*

Notice that the property of minimality has a combinatorial interpretation in this case: (X_σ, S) is minimal if and only if any every word occurring in a periodic point u appears in an infinite number of positions with bounded gaps.

2. From geometric dynamics to symbolic dynamics

Historically, symbolic dynamics has been introduced to better understand the dynamics of geometric maps. Indeed, by coding the orbits of a dynamical system with respect to a cleverly chosen finite partition indexed by the alphabet $\mathcal{A}$, one can replace the initial dynamical system, which may be difficult to understand, by a simpler dynamical system, that is, the shift map on a subset of $\mathcal{A}^{\mathbb{N}}$.

This old idea was used intensively, up to these days, particularly to study dynamical systems for which past and future are disjoint, such as toral automorphisms or pseudo-Anosov diffeomorphisms of surfaces. These systems with no memory, whose entropy is strictly positive, are coded by subshifts of finite type, defined by a finite number of forbidden words, and belong to the *Markov* framework. Some very important literature has been devoted to their many properties (see [**21**]). The partitions which provide a good description for a topological dynamical system, leading to a subshift of finite type, are called *Markov partitions.*

2.1. An example of the use of symbolic dynamics: The Morse sequence. In 1920, M. Morse was studying *geodesics*, that is, the curves realizing the minimum distance between two points, on connected surfaces with constant negative curvature. He was looking at infinite geodesics which remain within a small part of the space. More precisely, a geodesics is said to be *recurrent* if every point of the geodesics lies at a given distance (whatever small it can be) of a point in every long enought segment of the geodesics. Hence, closed geodesics are recurrent or periodic. An intricate question is the existence of non-closed recurrent geodesics.

FIGURE 1. Two examples of connected surfaces with constant negative curvature.

To answer this question, in [**24**], using a method initiated by Hadamard, Morse did a *coding* of geodesics, by infinite sequences of 0's and 1's, according to which boundary of the surface they meet: thus, we arrive in the space $\{0,1\}^{\mathbb{N}}$ of infinite symbolic sequences. To advance along a geodesic translates into looking at the next element of the sequence. The coding sends under suitable conditions the topology of the surface onto the product topology in $\{0,1\}^{\mathbb{N}}$.

Properties of geodesics are then easy to check: a closed geodesic corresponds to a periodic sequence. In the same way, by replacing points by elementary segments, the reader shall be able to check that a recurrent geodesic corresponds to what is now called a *minimal* sequence: every word occurring in u appears an infinite number of positions with bounded gaps.

Thanks to this coding, Morse proved the existence of a closed and recurrent geodesics:

THEOREM 5 ([**24**]). *A minimal and nonperiodic sequence is given by the (Prouhet-Thue)-Morse sequence,*

$$01101001100101101001011001101001100101100110100101101001100101101\ldots$$

defined as the fixed point (starting with 0*) of the Morse substitution* $\sigma : 0 \mapsto 01 \quad 1 \mapsto 10$.

A full study of the Morse sequence is made in [**13**].

2.2. Self-similar dynamics and substitutions. Dealing with a dynamical system, a usual problem is to try to understand the local structure of its orbits. A classical method to study this problem is to consider the *first return map* (Poincaré map) over an appropriate neighborhood $\mathcal{N}$ of a given point. For some systems such as toral quadratic rotations or some interval exchanges with parameters living in a quadratic extension, the system defined by the first return map on some subset is topologically conjugated to the original system. One can say that the original dynamical system has a *self-similar structure.* A basic idea is that, in general, as soon as self-similarity occurs, a substitution is hidden behind the original dynamical system: the trajectories of points in the neighborhood $\mathcal{N}$ before they come back into $\mathcal{N}$, define a substitution. Then, the trajectories of the points of the full system belong to the symbolic dynamical system associated with the substitution. Let us immediatly illustrate this idea with a simple example.

2.3. Example: addition of the golden ratio. Let φ denote the *addition of the golden ratio* $\alpha = 1 - \alpha^2 = 1,61...$ on the one-dimensional torus $\mathbb{T}$:

$$\varphi : x \in \mathbb{T} = \mathbb{R}\backslash\mathbb{Z} \quad \mapsto \quad x + \alpha \bmod 1 \in \mathbb{T}.$$

This map has two intervals of continuity:

$$\mathbb{T} = I_2 \cup I_1, \text{ with } I_2 = [0, 1-\alpha[, \quad I_1 = [1-\alpha, 1[.$$

Let ψ denote the first return map of φ on the largest interval of continuity I_1, that is,

$$\forall x \in [1-\alpha, 1[, \quad \psi(x) = \varphi^{\min\{k \in \mathbb{N}^*,\, \varphi(x) \in [1-\alpha,1[\}}(x).$$

We are going to prove thanks to a short computation that ψ is equal to φ itself, up to a reversal of the orientation and a renormalization.

Indeed, Let us consider the following partition of I_1:

$$I_1 = J_1 \cup J_2, \text{ with } J_1 = [1-\alpha, 2-2\alpha[, \quad J_2 = [2-2\alpha, 1[.$$

Then a simple computation yields that ψ restricted to J_1 is equal to φ^2:

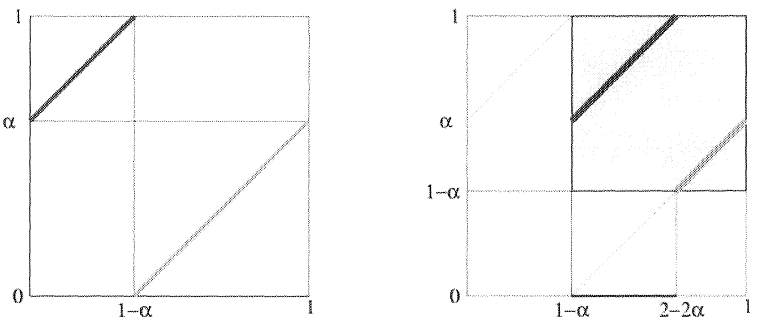
- $J_1 = [1-\alpha, 2-2\alpha[\subset I_1$,
- $\varphi(J_1) = [0, 1-\alpha[\not\subset I_1$,
- $\varphi^2(J_1) = [\alpha, 1[\subset I_1$.

Similarly, since $J_2 = [2-2\alpha, 1[\subset I_1$ and $\varphi(J_1) = [1-\alpha, \alpha[\subset I_1$, ψ restricted to J_2 is equal to φ.

A graphical representation of φ and ψ is given in Figure 2: the two graphics appear to be equal up to a reversal of the orientation and a renormalization. Formally, there is no difficulty to prove that φ and ψ are homeomorphic through the conjugacy map $\tau : x \in [0,1[\mapsto (1-\alpha)x + 1 \in [1-\alpha, 1[$.

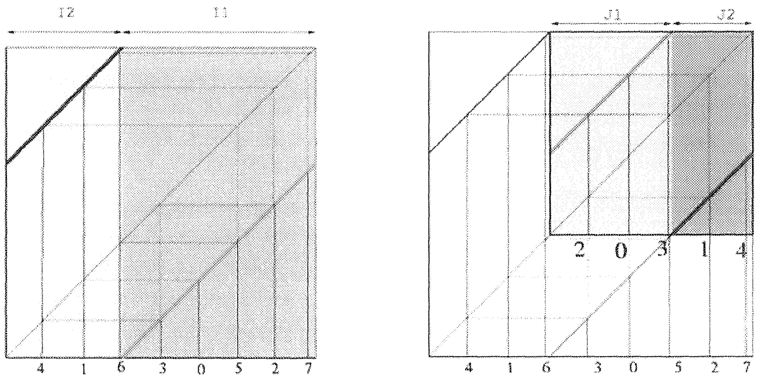
The interest of such a coding is that we are now able to code the trajectories of a point in $[1-\alpha, 1[$ for both the addition φ of the golden ratio and its first return map ψ. Let us study the example shown in Fig. 3. Indeed, the point α mod 1 (denoted by 0 on each figure) belongs to the largest interval I_1. Then one sees that $\varphi(\alpha)$ (denoted by 1) belongs to I_2, $\varphi^2(\alpha) \in I_1$, etc. Then the trajectory of α is coded by $I_1 I_2 I_1 I_1 I_2 I_1 I_2 I_1$.

Similarly, computing the trajectory of α for the first return map ψ gives $J_1 J_2 J_1 J_1 J_2$.



Representation of φ First return map ψ on $[1-\alpha, 1[$

FIGURE 2. The addition of the golden ratio is equal to its first return map up to a reversal of the orientation and a renormalization.



Trajectory through φ: $I_1I_2I_1I_1I_2I_1I_2I_1$.

Trajectory through ψ: $J_1J_2J_1J_1J_2$

FIGURE 3. Trajectories of the point α relatively to intervals of continuity of φ and its first return map ψ

The main point is that, since ψ is defined as the first return map of φ, there is a relationship between the two codings introduced here. Indeed, as soon as a point x lies in J_1, then we know that

- x belongs to I_1,
- $\varphi(x) \in I_2$
- $\psi(x) = \varphi^2(x)$.

Hence, coding a point x by J_1 according to ψ implies that the trajectory of the same point will be coded by I_1I_2 according to φ. Similarly, coding a point x by J_2 according to ψ implies that the trajectory of the same point will be coded by I_1 according to φ. We thus deduce that the trajectory of a point x through φ can be obtained by mapping the trajectory of a point x through ψ thanks to the map:

$$J_1 \to I_1I_2; \quad J_2 \to I_1.$$

One should remember now that we stated that φ and ψ were conjugate through the map τ. However, α is a fixed point for τ, and the partition $J_1 \cup J_2$ is the image of $I_1 \cup I_2$ through τ. Hence, the trajectories of α have the same coding through φ

and ψ. Consequently, this coding must be nothing else than the fixed point of the following substitution, called the Fibonacci substitution

$$1 \mapsto 12; \quad 2 \mapsto 1.$$

One finally proves that the addition of the golden ratio is very well represented as a symbolic shift map:

THEOREM 6 (see a general proof in [4]). *The coding of the trajectory of α mod 1 through the addition φ by the golden ratio α on $\mathbb{T}$ according to the intervals of continuity I_1 and I_2 is the fixed point of the Fibonacci substitution $1 \to 12$, $2 \to 1$:*

$$u = 121121211211212112121121121211211212112121121121211212112121121121211212112112121\ldots$$

The set of codings of all the points of $\mathbb{T}$ is equal the symbolic dynamical system associated with the Fibonacci substitution. The coding map is a semi-topological conjugacy between the shift map on the symbolic system and the addition by the golden ratio.

REMARK 7. For the example of the toral addition by the golden ratio, we can define an inverse map, from the symbolic system onto the torus. It is proved that this map is continuous, 2-to-1, and 1-to-1 except on a countable set; this is the best possible result, given the fact that one of the sets is connected and the other one a Cantor set.

3. From symbolic dynamics to geometry ?

As shown in Section 2.3, Poicaré's method defines a coding map from the geometric system onto the substitutive symbolic dynamical system. A natural question is: how far is this map from being a bijection? We have just seen that a precise answer has been given to this question for the Fibonacci substitution (Remark 7). For other examples, the question can be much more difficult. It is natural then to focus on the reverse question: given a substitution, which self-similar geometric actions are coded by this substitution?

For the Morse substitution, it is proved that the symbolic dynamical system associated with this substitution is a two-point extension of the dyadic odometer, that is, the group $\mathbb{Z}_2$ of 2-adic integers ([**9**] and also [**13**], chapter 2).

The three-letter equivalent of the Fibonacci substitution is the Tribonacci substitution $1 \mapsto 12$, $2 \mapsto 13$, $3 \mapsto 1$. G. Rauzy, with methods from number theory, proved in 1981 that the symbolic dynamical system associated with this substitution is measure-theoretically isomorphic, by a continuous map, to a domain exchange on a self-similar compact subset of $\mathbb{R}^2$ called the *Rauzy fractal* [**27**]. Tiling properties of the Rauzy fractal yield an isomorphism between the substitutive system and a translation on the two-dimensional torus. This example will be studied in more details in Section 3.2.

These examples emphasize the connection between searching for a geometric interpretation of a symbolic dynamical systems and understanding whether this dynamical system is already known up to an isomorphism. Since substitutive dynamical systems are deterministic, i.e., of zero entropy, they are very different from subshifts of finite type. Hence, the following question is natural: which substitutive dynamical systems are isomorphic to a translation on a compact group? More generally, what is their maximal equicontinuous factor, that is, the largest translation on a compact group that topologically embeds into this symbolic system?

Let us introduce now the point of view of spectral theory. Indeed, to a dynamical system (X, S) is associated the unitary operator $U : \; f \in L^2(X_\sigma, S) \mapsto f \circ S \in$

$L^2(X_\sigma, S)$ [**34**]. One usual calls *eigenvalues* of the dynamical system the eigenvalues λ's of U; their modulus is equal to one, so that the word *eigenvalue* sometimes also holds for every $x \in [0, 1[$ such that $\lambda = e^{2i\pi x}$. The *eigenfunctions* of the dynamical system are the eigenfunctions of U; they appear to be functions $f \in L^2(X_\sigma, S)$.

From this point of view, the maximum equicontinuous factor of a dynamical system is proved to be the unique abelian compact group translation with the same eigenvalues than the dynamical system. Hence, it uniquely determined by the eigenvalues [**34**].

Starting from a geometrical and combinatorial question, we naturally come to a question of spectral theory, that is, computing the eigenvalues of a dynamical system.

3.1. Substitution of constant length. During the seventies, a precise answer to this question has been obtained for substitutions of constant length (the images of each letters in the alphabet share the same length) [**18, 23, 10**]. This caracterization implies some p-adic groups $\mathbb{Z}_p$, also called p-adic odometer, obtained as the completion of $\mathbb{Z}$ for the p-adic topology [**14**].

THEOREM 8 (Dekking [**10**]). *Let σ be a substitution of constant length n. Let $u = (u_n)_{n \in \mathbb{N}}$ be a periodic point for σ. We call* height *of the substitution the greatest integer m which is coprime with n and divides all the strictly positive ranks of occurrence of the letter u_0 in u. The height is less that the cardinality of the alphabet.*

The maximal equicontinuous factor of the substitutive dynamical system associated with σ is the addition of $(1, 1)$ on the abelian group $\mathbb{Z}_n \times \mathbb{Z}/m\mathbb{Z}$, where $\mathbb{Z}_n$ denotes the product of the p-adic groups $\mathbb{Z}_p$ for every prime p that divides n.

As an example, the letter 1 appears at rank 3 and 5 in the fixed point

$$u = 1221211221121221211 21...$$

of the Morse substitution so that this sustitution has height 1. Hence, the maximal equicontinuous factor of the associated substitutive system is the 2-adic group $\mathbb{Z}_2$.

An example of a substitution with an height different from 1 is given by $1 \mapsto 121 \quad 2 \mapsto 312 \quad 3 \mapsto 213$: the letter 1 appears at every even rank in the fixed point

$$u = 121312121213121312121312$$

so that the height is 2 and the maximal equicontinuous factor is $\mathbb{Z}_3 \times \mathbb{Z}/2\mathbb{Z}$.

Dekking also provides a necessary and sufficient condition for a measure-theoretic isomorphism between such a substitutive system of constant length and its maximal equicontinuous factor. This condition is purely combinatorial: a substitution σ is said to satisfy the *coincidence condition* if there exists n such that the image of each letter under a power σ^k has the same n-th letter. We have:

THEOREM 9 (Dekking [**10**]). *Let σ be a substitution of constant length and of height 1. The substitutive dynamical system associated with σ has a purely discrete spectrum if and only if the substitution σ satisfies the condition of coincidence.*

As an example, the substitution $1 \mapsto 12 \quad 2 \mapsto 23 \quad 3 \mapsto 13$ has a pure discrete spectrum dynamical system since its three fixed points contain a 1 at rank 6:

$$\begin{array}{l}1223231231312132313121 31...\\ 2313121312231213122323 13...\\ 1223231323131213122323 13...\end{array}$$

Conversely, the two fixed points of the Morse substitution have no coincidence so that the associated dynamical system is not isomorphic to the dyadic odometer.

In the case when the height of the substitution is different from 1, it is possible to recode the substitution into a substitution with height 1 and to check the coincidence condition on this last substitution. As an example of application, this allows one to prove that the substitution $1 \mapsto 121 \quad 2 \mapsto 312 \quad 3 \mapsto 213$ introduced previously has a pure discrete spectrum dynamical system (see [**10**] and [**13**], Chap. 7 for details).

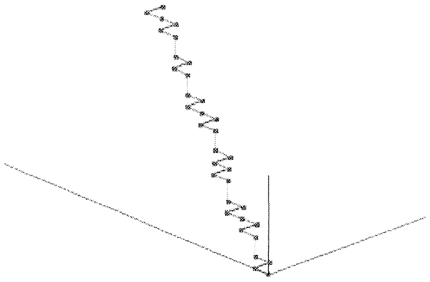
3.2. A first step towards the study of substitution of nonconstant length: The Tribonnacci substitution. G. Rauzy generalized in [**27**] the dynamical properties of the Fibonacci substitution to a three-letter alphabet substitution, called the Tribonacci substitution or Rauzy substitution, and defined by

$$\sigma(1) = 12 \qquad \sigma(2) = 13 \qquad \sigma(3) = 1.$$

Broken line associated with the substitution - Let $u =$ denote the unique infinite fixed point of σ:

$$u = 1213121121312121312112131213121121312121312\text{1...}$$

Let us embed this infinite word u as a broken line in $\mathbb{R}^3$ by replacing succesively each letter of u by the corresponding vector in the canonical basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ in $\mathbb{R}^3$.



An interesting property of this broken line is that it remains at a bounded distance of a line, turning around it. One states that this axis if nothing else that the expanding direction of the incidence matrix of the substitution, that is, the matrix that contains in each column j the number of occurences of each letter i in $\sigma(j)$.

$$\mathbf{M}_\sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Notice that the reason why σ is called the *Tribonacci substitution* is that the characteristic polynomial of $\mathbf{M}_\sigma$ is $X^3 - X^2 - X - 1$ so that its roots satisfy $\alpha^3 = \alpha^2 + \alpha + 1$, and are called Tribonacci numbers in reference to the Fibonacci number. One root is strictly greater than 1 and is associated with an expanding eigenline; the two other roots are complex conjugates of modulus less than 1. They generate a contracting plane.

Definition of the Rauzy fractal - When one projects the vertices of the broken line onto the contracting plane of $\mathbf{M}_\sigma$, along the expanding direction, then one obtains a bounded set in a two-dimensional vector space. The closure of this set of points is a compact set denoted by $\mathcal{R}$ and called the *Rauzy fractal* (see Fig. 4).

To be more precise, denote by π the linear projection in $\mathbb{R}^3$, parallel to the expanding direction of $\mathbf{M}_\sigma$, on the contracting plane of $\mathbf{M}_\sigma$, identified with the complex plane $\mathbb{C}$. If $u = (u_i)_{i\in\mathbb{Z}}$ is the periodic point of the substitution, then the Rauzy fractal is

$$\mathcal{R} = \overline{\left\{\pi\left(\sum_{i=0}^{n} \mathbf{e}_{u_i}\right) \; ; \; n \in \mathbb{Z}\right\}}.$$

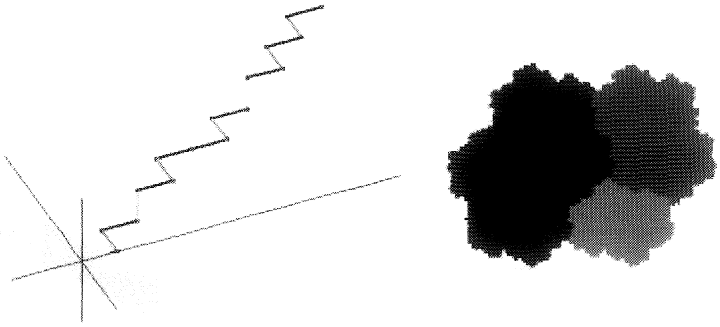


FIGURE 4. The projection method to get the Rauzy fractal for the Tribonnacci substitution.

Partition of the Rauzy fractal – As shown in Fig.4, three subsets of the Rauzy fractal can be distinguished. Indeed, for each letter $j = 1, 2, 3$, the *cylinder* $\mathcal{R}_j$ is defined to be the closure of the set of ends of any segment on the broken line which is parallel to the canonical vector $\mathbf{e}_j$:

$$\mathcal{R}_j = \overline{\left\{\pi\left(\sum_{i=0}^{n} \mathbf{e}_{u_i}\right) \; ; \; n \in \mathbb{Z}, \; u_{n+1} = j\right\}}.$$

The union of these three cylinders covers the compact $\mathcal{R}$, and G. Rauzy proved in [**27**] that their intersection has zero measure.

Dynamics on the Rauzy fractal – One should notice that it is possible to move on the broken line, from a vertex to the following one, thanks to a translation by one of the three canonical vectors $\mathbf{e}_1$, $\mathbf{e}_2$ or $\mathbf{e}_3$. In the contracting plane, this means that each cylinder $\mathcal{R}_i$ can be translated by a given vector, i.e., $\pi(\mathbf{e}_i)$, without going out of the Rauzy fractal: $\mathcal{R}_i + \pi(\mathbf{e}_i) \subset \mathcal{R}$.

Thus, the following map φ, called a *domain exchange* (see Fig. 5) is well defined for any point of the Rauzy fractal which belongs to only one set $\mathcal{R}_j$. Since the cylinders intersect on a set of measure zero, this map is defined almost everywhere on the Rauzy fractal:

$$\forall x \in \mathcal{R}, \quad \varphi(x) = x + \pi(\mathbf{e}_i), \quad \text{if } x \in \mathcal{R}_i.$$

It is natural to code, up to the partition defined by the 3 cylinders, the action of the domain exchange φ over the Rauzy fractal $\mathcal{R}$. G. Rauzy proved in [**27**] that the coding map, from $\mathcal{R}$ into the three-letter alphabet full shift $\{1, 2, 3\}^{\mathbb{Z}}$ is almost everywhere one-to-one. Moreover, this coding map is onto the substitutive system associated with the Tribonacci substitution. Thus we have the following result:

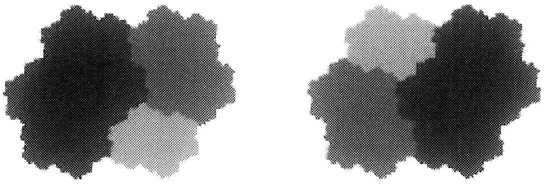


FIGURE 5. Domain exchange over the Rauzy fractal.

THEOREM 10 (Rauzy, [**27**]). *The domain exchange φ defined on the Rauzy fractal $\mathcal{R}$ is semi-topologically conjugate to the shift map on the symbolic dynamical system associated with the Tribonacci substitution.*

Factorization onto a torus - The domain exchange φ is defined only almost everywhere, which prevents us to define a continuous dynamics on the Rauzy fractal. A solution to this problem consists in factorizing the Rauzy fractal by the lattice $\mathcal{L} = \mathbb{Z}\,\pi(\mathbf{e}_1 - \mathbf{e}_3) + \mathbb{Z}\,\pi(\mathbf{e}_2 - \mathbf{e}_3)$. Indeed, this quotient map sends the contracting plane onto a two-dimensional torus; the three vectors $\pi(\mathbf{e}_1)$, $\pi(\mathbf{e}_2)$ and $\pi(\mathbf{e}_3)$ map onto the same vector on the torus. Thus, the factorization of the domain exchange φ on the quotient is a toral translation.

G. Rauzy proved in [**27**] that the restriction of the quotient map to the Rauzy fractal is onto and almost everywhere one-to-one. Consequently, we get that the domain exchange on the Rauzy fractal, which is known to be semi-topologically conjugate to the Tribonacci substitutive dynamical system, is also measure-theoretically isomorphic to a minimal translation on the two-dimensional torus $\mathbb{T}^2$. Finally, by mixing dynamics, self-similarity and number theory, we get the two following equivalent results:

THEOREM 11 (Rauzy, [**27**]). *The symbolic dynamical system generated by the Tribonacci substitution is measure-theoretically isomorphic to a toral translation, that is, it has a purely discrete spectrum.*

3.3. A significant advance towards the understanding of the spectrum of substitutive systems. B. Host made a significant contribution to the understanding of ergodic properties of substitutive systems; in [**16**], any class of eigenfunctions is proved to contain a continuous eigenfunction (see [**34**] for usual definition about ergodic theory). Thus, the two main dynamical classifications (up to measure-theoretic isomorphism and topological conjugacy) are equivalent for primitive substitutive systems.

Coboundaries - In the continuation of this, the notion of *coboundaries* introduced by B. Host allows one to better understand the structure of the spectrum of a substitutive system. A *coboundary* of a substitution σ is defined as a map $h : \mathcal{A} \to \mathbb{U}$ (where $\mathbb{U}$ denotes the unit circle) such that there exists a map $f : \mathcal{A} \to \mathbb{U}$ with $f(b) = f(a)h(a)$ for every word ab of length 2 that appears in a periodic point for σ. The coboundary defined by $h(a) = 1$ for every letter a (that is, $f(a) = f(b)$ for every ab in the language) is called the *trivial coboundary*. For substitutions of constant length, nontrivial coboundaries are related to the finite group contained in the maximal equicontinuous factor described in Theorem 8. Details can be found in [**13**], Chap. 7.

In the most simplest cases the only coboundary is the trivial one, that is, the constant function equal to 1. However, there exist some substitutions with nontrivial coboundaries such as $1 \mapsto 1231$, $2 \mapsto 232$, $3 \mapsto 3123$. Indeed, words of length 2 that appear in the fixed point of this substitution begining with 1 are 12, 23, 31 and 32. Hence for every $\lambda \in [0,1[$, the function $h(1) = 1$, $h(2) = e^{2\pi\lambda} = 1/h(3)$ defines a non-trivial couboundary associated with the function $f(1) = 1 = f(2)$, $f(3) = e^{2\pi\lambda}$.

Structure of the spectrum – Coboundaries allow Host to describe precisely the structure of the spectrum defined in Section 3.

THEOREM 12 (Host [**16**]). *Let σ be a primitive substitution over the alphabet $\mathcal{A}$. A complex number $\lambda \subset \mathbb{U}$ is an eigenvalue of (X_σ, S) if and only if there exists $p > 0$ such that for every $a \in \mathcal{A}$, the limit $h(a) = \lim_{n\to\infty} \lambda^{|\sigma^{pn}(a)|}$ is well defined, and h is a coboundary of σ.*

Hence, the spectrum of a substitutive system can be divided into two parts.

Arithmetic spectrum: incidence matrix – Since the constant function equal to one 1 is always a coboundary, a sufficient condition is the following: if there exists $p \in \mathbb{N}$ such that $\lambda \in \mathbb{C}$ satisfies $\lim \lambda^{|\sigma^{pn}(a)|} = 1$ for every letter a of the alphabet, then λ is an eigenvalue of the substitutive dynamical system associated with σ.

Such eigenvalues are said to be *arithmetic* since they are computable (the condition $\lim \lambda^{|\sigma^{pn}(a)|} = 1$ can be interpreted in terms of scalar product) and depend only on the incidence matrix of the substitution. Especially, two substitutions that differ only by the order of occurencies of the letters in images of the letters have the same arithmetical spectrum (see [**13**], Chapter 7).

Combinatorial spectrum: return words – Conversely, the eigenvalues for non-trivial coboundaries are "non-commutative": they depend heavily on the combinatorics of the substitution. Durand [**11**], Ferenczi [**12**] and Livshits [**22**] established that they depend on *return words*, playing the role of the height that was defined for substitutions of constant length. Roughly, a *return word* is a word $W = a_1 \ldots a_k$ such that Wa_1 is in a factor of the periodic point of the substitution, and $a_i \neq a_1$ for all i. A more precise definition should be found in [**13**].

A condition for no combinatorial spectrum: coincidences – A combinatorial condition is related to the existence of only a trivial coboundary. This condition is called strong coincidence condition and generalizes the condition of Dekking. It was defined by Host, Hollander and formalized by Arnoux and Ito [**5**]. Formally, σ is said to satisfy the *strong coincidences conditions* if for every pair of letters b_1, b_2, there exists a letter a and $P_1, S_1, P_2, S_2 \in \mathcal{A}^*$ such that

$$\sigma^n(b_1) = P_1\, a\, S_1 \quad \sigma^n(b_2) = P_2\, a\, S_2.$$

Coincidences are related to coboundaries by the following result (see a proof in [**13**], Chapter 7).

LEMMA 13 (Host). *Let σ be a substitution with a nontrivial coboundary $g : \mathcal{A} \to \mathbb{U}$. Let f be the function of modulus 1 which satisfies $f(b) = g(a)f(a)$ as soon as the word ab belongs to the language of a periodic point of the substitution. If there exist two letters a and b and a rank k such that*

- *$f(a) \neq f(b)$,*
- *$\sigma^k(a)$ begins with a and $\sigma^k(b)$ begins with b,*

then σ does not satisfy the coincidence condition on prefixes.

Roughtly, this lemma means that a substitutive system with coincidences do not have a combinatorial spectrum. However, we are unable to prove this last result in general, but only for substitutions of Pisot type (see Section 4.2).

4. Applications

4.1. Properties of the spectrum of substitutive systems. From the end of the 80's, many papers have provided conditions for a substitutive dynamical system to have a purely discrete spectrum [**22, 33, 31, 15**]. Some are necessary conditions, others are sufficient conditions. Let us focus on some typical examples of applications.

- Weakly mixing examples of substitutive systems are derived From Host's results, as $1 \mapsto 12121$, $2 \mapsto 112$, since 1 is the only eigenvalue of the associated substitutive system.
- Refinements of Host's theory allowed Livshits to define conditions for pure discrete spectrum or partially continuous spectrum, as a mix of the coincidence condition and return words. Hence, the system associated with $1 \mapsto 23$, $2 \mapsto 12$, $3 \mapsto 23$, has as a continuous spectral component but is not weakly mixing [**22, 33**].
- An important result is stated by Solomyak in the case when the incidence polynomial of a substitution is irreducible: the existence of discrete spectrum depends on the expanding eigenvalues of the incidence matrix of the substitution. Indeed, if there exist $P \in \mathbb{Z}[X]$ and $C \in \mathbb{R}$ such that $P(\alpha) = C$ for every expanding eigenvalue α of the matrix, then $exp(2\pi iC)$ is an eigenvalue of (X_σ, S) [**31**]. A partial converse was established by Ferenczi, Mauduit, Nogueira [**12**]. This allows one to compute explicitly the spectrum of some substitutive systems, such as $1 \mapsto 1244$, $2 \mapsto 23$, $3 \mapsto 4$, $4 \mapsto 1$, whose spectrum is $exp(2\pi i\mathbb{Z}\sqrt{2})$.

4.2. A specific class of substitutions: substitutions of Pisot type. A substitution σ is *of Pisot type* if every non-dominant eigenvalue λ of its incidence matrix $\mathbf{M}$ satisfies $0 < |\lambda| < 1$. We deduce that the characteristic polynomial of the incidence matrix of such a substitution is irreducible over $\mathbb{Q}$. Consequently, the dominant eigenvalue α is a Pisot number and the other eigenvalues λ are its algebraic conjugates and substitutions of Pisot type are primitive (see the proofs in [**13**]). A substitution σ is *unimodular* if $\det \mathbf{M} = \pm 1$.

The spectrum of substitutive systems of Pisot type has some important properties:

- such systems are never weakly mixing since they have only one expanding eigenvalue so that they satisfy the conditions of Solomyak given in Section 4.1.
- Their arithmetical spectrum can be computed thanks to Host's method. In the unimodular case, the arithmetic spectrum is generated by the frequencies of the letters in the fixed point. In the non-unimodular case, additional rational eigenvalues have to be computed.
- Substitutions of Pisot type never has a nontrivial coboundary [**7**]. Hence, their spectrum is equal to their arithmetic spectrum, which is explicit as explained in the preceeding item.

From these properties, one naturally wonders whether substitutions of Pisot type have a pure discrete spectrum. Unfortunately, a positive answer is not so easy to give.

The case of substitutions on a two-letters alphabet is completely studied. We first know from the work of Host and Solomyak-Hollander that substitutions that are of of Pisot type with coincidences on two letters all have a pure discrete spectrum dynamical system [**15**]. Then, Barge and Diamond proved that substitutions of

Pisot type on two letters always have coincidences [**6**]. This yields the following theorem:

THEOREM 14. *All substitutive systems of Pisot type on two letters have a pure discrete spectrum.*

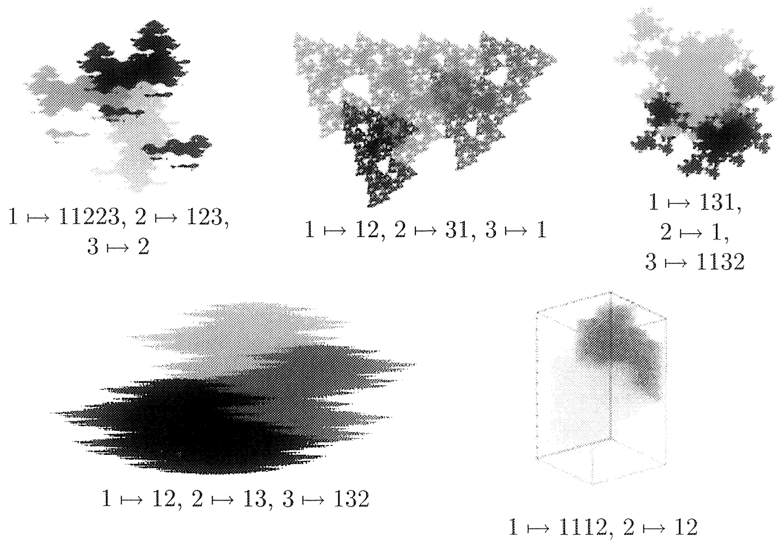
However, on more than three letters, the methods used before are not successful anymore. More intricate results have to be proved in the flavour of Rauzy's work for the Tribonacci substitution.

4.3. Rauzy fractals. Starting for a substitution of Pisot type, nothing prevents one from computing a Rauzy fractal as done for the Tribonacci substitution:

(1) one can build a broken line from a periodic point of the substitution. Since the substitution is of Pisot type, the broken line turns around a one-dimensional direction and projects onto a compact set called the *Rauzy fractal* of the substitution. If the substitution is not unimodular, then the projection space should take into account an arithmetic part. More precisely, the space of projection is a product of the Euclidean space with finite extensions of p-adic spaces that has a non-zero Haar measure [**29**].

(2) A piece on the Rauzy fractal is associated with each letter of the alphabet. The strong coincidence condition means that the pieces are disjoint in measure [**5**]. Finally, the *Rauzy fractal* of a Pisot type substitution with strong coincidences appears to be *self-similar* and compact.

(3) Shifting the fixed point, that is moving on the broken line, factorize onto an exchange of domains on the Rauzy fractal. Arnoux and Ito proved that the shift map and the domain exchange are equivalent from a spectral point of view, as stated in Theorem 15.



$1 \mapsto 11223,\ 2 \mapsto 123,\ 3 \mapsto 2$

$1 \mapsto 12,\ 2 \mapsto 31,\ 3 \mapsto 1$

$1 \mapsto 131,\ 2 \mapsto 1,\ 3 \mapsto 1132$

$1 \mapsto 12,\ 2 \mapsto 13,\ 3 \mapsto 132$

$1 \mapsto 1112,\ 2 \mapsto 12$

FIGURE 6. Example of Rauzy fractals for substitutions of Pisot type.

THEOREM 15. *Let σ be substitution of Pisot type over a d-letter alphabet which satisfies the condition of coincidence. Then the substitutive dynamical system associated with σ is measure-theoretically isomorphic to the exchange of d domains* *defined almost everywhere on the Rauzy fractal of σ, that is, a self-similar compact set on a product of the Euclidean space with finite extensions of p-adic spaces that has a non-zero Haar measure.*

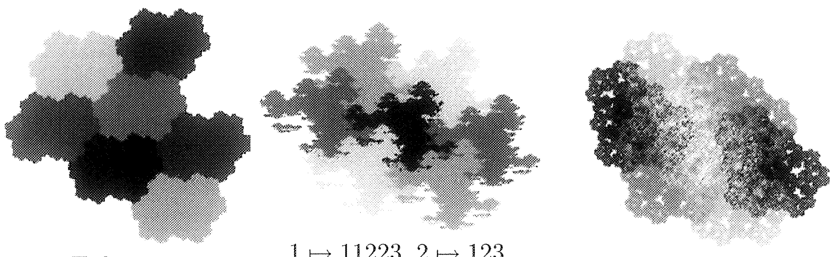
Notice that we do not know any example of a substitution of Pisot type with no strong coincidence.

As for the Tribonacci substitution, there is no problem to factorize the Rauzy fractal through a lattice on an compact abelian group, so that the exchange of domains reduces to a group translation. The question is the same as before: is this representation one-to-one? Unfortunately, the methods used for the Tribonacci substitution are quite specific and cannot be generalized. Anyway, some researches on that direction allow to deduce from the factorization of Rauzy fractal on compact abelian groups some combinatorial conditions for pure discrete spectrum. These conditions are based either on graphs [**30, 32**] or on the notion of balanced pairs [**7, 26**]. The problem is that the conditions are not general and need to be checked by hand on each substitution.

Since each example of a substitution of Pisot type that have been tested has a pure discrete spectrum, the point now is to exhibit some families of substitutions that provide a pure discrete spectrum dynamical system.

5. Conclusion

As a conclusion, we would like to emphasize the fact that the results exposed here mainly deal with spectral theory but can be also be expressed in more geometrical terms. Indeed, pure discrete spectrum has a nice geometrical equivalent in the unimodular case: thanks to the geometrical representation with Rauzy fractal, it is proved that a substitution of Pisot type with coincidence has a pure discrete spectrum if and only if its Rauzy fractals generates a periodic tiling of the plane [**30, 26, 7**]. Hence, conditions for pure discrete spectrum discussed above allows one to prove that the Rauzy fractals generated by the Tribonacci substitution, the substitution $1 \mapsto 11223$, $2 \mapsto 123$, $3 \mapsto 2$, or the substitution $1 \mapsto 12$, $2 \mapsto 3$, $3 \mapsto 1$ generate a periodic tiling. More generally, all the Rauzy fractal showed before do generate a periodic tiling.



Tribonacci substitution

$1 \mapsto 11223$, $2 \mapsto 123$, $3 \mapsto 2$

$1 \mapsto 12$, $2 \mapsto 3$, $3 \mapsto 1$

FIGURE 7. Periodic tilings generated by Rauzy fractals.

Hence, substitutions have relations with a quite large number of mathematical domains (further illustrations are given in [**13**]). Combination of combinatorics, spectral theory, geometry and number theory will allow now to consider and apply this simple combinatorial object (a substitution) in different directions:

- proving general results on discrete spectrum and tilings;

- application to β-numeration and diophantine analysis [**8, 1**];
- Generation of discrete planes [**2, 3**];
- Models for quasi-crystals [**28, 20**];
- Construction of explicit Markov partitions for toral automorphisms [**17, 19**].

References

[1] Akiyama, S. — *Self affine tiling and Pisot numeration system*, In: Number theory and its applications (Kyoto, 1997), 7–17.Kluwer Acad. Publ., Dordrecht, 1999.

[2] Arnoux, P. and Berthé, V. and Ito, S. — *Discrete planes, $\mathbb{Z}^2$-actions, Jacobi-Perron algorithm and substitutions*, Ann. Inst. Fourier (Grenoble), **52**, (2002), 2, 305–349.

[3] Arnoux, P. and Berthé, V. and Siegel, A. — *Two-dimensional iterated morphisms and discrete planes*, Theoret. Comput. Sci., **319**, (2004), 145–176.

[4] Arnoux, P. and Ferenczi, S. and Hubert, P. — *Trajectories of rotations*, Acta Arith., **87**, (1999), 3, 209–217.

[5] Arnoux, P. and Ito, S. — *Pisot substitutions and Rauzy fractals*, Journées Montoises (Marne-la-Vallée, 2000), Bull. Belg. Math. Soc. Simon Stevin, **8**, (2001), 2, 181–207.

[6] Barge, M. and Diamond, B. — *Coincidence for substitutions of Pisot type*, Bull. Soc. Math. France, **130**, (2002), 619-626.

[7] Barge, M. and Kwapisz, J. — *Geometric theory of unimodular Pisot substitutions*, Preprint, 2004.

[8] Bassino, F. — *Beta-expansions for cubic Pisot numbers*, In: LATIN 2002: Theoretical informatics (Cancun), Springer Lecture Notes in Comput. Sci., **2286**, 2002, 141–152.

[9] del Junco, Andrés — *A transformation with simple spectrum which is not rank one*, Canad. J. Math., **29**, (1977), 3, 655–663.

[10] Dekking, F. M. — *The spectrum of dynamical systems arising from substitutions of constant length*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, **41**, (1977/78), 3, 221–239.

[11] Durand, F. — *A characterization of substitutive sequences using return words*, Discrete Math., **179**, (1998), 1-3, 89–101.

[12] Ferenczi, S. and Mauduit, C. and Nogueira, A. — *Substitution dynamical systems: algebraic characterization of eigenvalues*, Ann. Sci. École Norm. Sup., **29**, (1996), 4, 519–533.

[13] Pytheas-Fogg, N. — *Substitutions in Dynamics, Arithmetics and Combinatorics*, Lectures Notes in Mathematics 1794, Springer-Verlag, 2002.

[14] Gouvêa, F. Q. — *p-adic numbers*, Universitext, Springer-Verlag, Berlin, 1997.

[15] Hollander, M. and Solomyak, B. — *Two-symbol Pisot substitutions have pure discrete spectrum*, Ergodic Theory Dynam. Systems, **23**, (2003), 533-540.

[16] Host, B. — *Valeurs propres des systèmes dynamiques définis par des substitutions de longueur variable*, Ergodic Theory Dynam. Systems, **6**, (1986), 4, 529–540.

[17] Ito, S. and Ohtsuki, M. — *Modified Jacobi-Perron algorithm and generating Markov partitions for special hyperbolic toral automorphisms*, Tokyo J. Math., **16**, 2, (1993), 441–472.

[18] Kamae, T. — *Spectrum of a substitution minimal set*, J. Math. Soc. Japan, **22**, (1970), 567–578.

[19] Kenyon, R. and Vershik, A. — *Arithmetic construction of sofic partitions of hyperbolic toral automorphisms*, Ergodic Theory Dynam. Systems, **18**, (1998), 2, 357–372.

[20] Lagarias, J.C. — *Geometric models for quasicrystals I. Delone sets of finite type*, Discrete Comput. Geom. **21**, (1999), 2, 161–191.

[21] Lind, D. and Marcus, B. — *An introduction to symbolic dynamics and coding*, Cambridge University Press, Cambridge, 1995.

[22] Livshits, A. N. — *On the spectra of adic transformations of Markov compact sets*, Uspekhi Mat. Nauk, **42**, (1987), 3(255), 189–190. (English translation: *Russian Math. Surveys* 42(3): 222–223, 1987).

[23] Martin, J. C. — *Substitution minimal flows*, Amer. J. Math., **93**, (1971), 503–526.

[24] Morse, H. M. — *Recurrent geodesics on a surface of negative curvature*, Trans. Amer. Math. Soc., **22**, (1921), 1, 84–100.

[25] Queffélec, M. — *Substitution dynamical systems—spectral analysis*, Lecture Notes in Mathematics, 1294. Springer-Verlag, Berlin, 1987.

[26] Rao, H. and Ito, S. — *On super-coincidence condition*, Preprint.

[27] Rauzy, G. — *Nombres algébriques et substitutions*, Bull. Soc. Math. France, **110**, 2, (1982), 147–178.

[28] Senechal, Marjorie — *Quasicrystals and geometry*, Cambridge University Press, Cambridge, 1995.

[29] Siegel, A. — *Représentation des systèmes dynamiques substitutifs non unimodulaires*, Ergodic Theory Dynam. Systems, (2003), 23, 1247–1273.

[30] Siegel, A. — *Pure discrete spectrum dynamical system and periodic tiling associated with a substitution*, Ann. Inst. Four., **54**, (2004), 2.

[31] Solomyak, B. — *On the spectral theory of adic transformations*, In: Representation theory and dynamical systems, 217–230, Amer. Math. Soc., Providence, RI, 1992.

[32] Thuswaldner, J. — *Unimodular Pisot substitutions and their associated tiles*, Preprint.

[33] Vershik, A. M. and Livshits, A. N. — *Adic models of ergodic transformations, spectral theory, substitutions, and related topics*, In: Representation theory and dynamical systems, 185–204, Amer. Math. Soc., Providence, RI, 1992.

[34] Walters, P. — *An introduction to ergodic theory*, Springer-Verlag, New York, 1982.

Properties of expansions to non-integer bases: a survey

Karma Dajani

Universiteit Utrecht
Mathematisch Instituut
P.O. Box 80.000
3508 TA Utrecht
the Netherlands
dajani@math.uu.nl

Abstract. In this survey article we give some of the known results describing the arithmetic and ergodic properties of algorithms generating expansions to non-integer bases.

AMS classification: 28D05, 11K55

1. Integer Versus Non-integer

Given any integer $r > 1$, any $x \in [0,1)$ can be developed in a series expansion of the form

$$x = \sum_{k=1}^{\infty} \frac{c_k}{r^k} = .c_1c_2 \ldots c_n \ldots , \tag{11}$$

where $c_k \in \{0,1,\ldots,r-1\}$. Furthermore, every $x \in [0,1)$ has a unique series expansion; only rationals p/q of the form m/r^k for some $k \geq 1$ and $m = 0,1,\ldots,r^k-1$ have two different expansions of the form (11), one of them being finite while the other expansion ends in an infinite string of $r-1$'s. Dynamically expansions in base r are generated by iterating the map $T_r : [0,1) \to [0,1)$ given by

$$T_r(x) = rx \pmod 1,$$

and the digits $c_k = c_k(x)$, $k \geq 1$, are given by

$$c_k = \lfloor rT_r^{k-1}(x) \rfloor, \quad k \geq 1,$$

where $\lfloor \xi \rfloor$ denotes the largest integer not exceeding ξ, and T_r^K is the k-fold composition of T_r.

It well-known that the Lebesgue measure λ is T_r-invariant, i.e., $\lambda(T_r^{-1}(A)) = \lambda(A)$, for every Borel set A in $[0,1)$, and T_r is related to the Bernoulli-shift on r symbols, with uniform product measure.

On the other hand, if $\beta > 1$ is a non-integer, then almost every $x \in [0, \lfloor\beta\rfloor/(\beta-1)]$ has a continuum number of expansions of the form

$$x = \sum_{k=1}^{\infty} \frac{a_k}{\beta^k}, \quad a_k \in \{0,1,\ldots,\lfloor\beta\rfloor\}, \tag{12}$$

see [EJK], [Si], [DV].

We assume from now on that $\beta > 1$ is a non-integer, and we identify the expansion (12) with the infinite sequence $(a_1, a_2, \ldots)$. For each $x \in [0, \lfloor\beta\rfloor/(\beta-1)]$, we order the set of all possible expansions of x of the form (12) using the lexicographical ordering. Then, the largest expansion lexicographically is called the *greedy expansion* of x, and the smallest is called the *lazy expansion*. One main advantage of these two extreme cases is that they can be generated dynamically by iterating an appropriate map, this makes it possible to use the tools of ergodic theory in order to understand the dynamical and statistical properties of these maps; see [P], [G], [DK1], [DK2], [R], [KL], [EJK].

2. Greedy Versus Lazy

The greedy expansion was introduced in 1957 by A. Rényi [R1]. Originally the greedy expansion was studied for points on $[0,1)$, and it is obtained by iterating the transformation T_β defined on $[0,1]$ by

$$T_\beta x = \beta x \pmod 1.$$

Rényi studied the statistical properties of these expansions. He showed that T_β is ergodic with respect to λ, i.e., any Borel set A satisfying $T_\beta^{-1}(A) = A$ has measure 0 or 1. He also showed that λ is equivalent to a T_β-invariant probability measure μ_β with density h_β satisfying

$$1 - \frac{1}{\beta} \le h_\beta(x) \le \frac{1}{1 - \dfrac{1}{\beta}}.$$

Independently, A.O. Gel'fond [G] (in 1959) and W. Parry [P] (in 1960) showed that

$$h_\beta(x) = \frac{1}{F(\beta)} \sum_{n=0}^{\infty} \frac{1}{\beta^n} 1_{[0,T^n(1))}(x)\,, \tag{13}$$

where $F(\beta) = \int_0^1 (\sum_{x<T^n(1)} \frac{1}{\beta^n})dx$ is a normalizing constant. After Parry the ergodic properties of T_β were studied by several authors. E.g., M. Smorodinsky [Sm] "closed the gap" between the ergodic properties of T_β for $\beta \in \mathbb{Z}$ and $\beta \notin \mathbb{Z}$, by showing that for each non-integer $\beta > 1$ the system $(\,[0,1), \mu_\beta, T_\beta\,)$ is weak-Bernoulli, which roughly means that digits b_k in the 'far future' are independent of digits b_ℓ in the 'far past'.

Since we are interested in expanding all points in $[0, \lfloor\beta\rfloor/(\beta-1)]$, we will extend the definition of the greedy map T_β to all points in $[0, \lfloor\beta\rfloor/(\beta-1)]$ by

$$T_\beta(x) = \begin{cases} \beta x \pmod 1, & 0 \le x < 1, \\ \\ \beta x - \lfloor\beta\rfloor, & 1 \le x \le \lfloor\beta\rfloor/(\beta-1). \end{cases}$$

It is easy to see that the interval $[0,1)$ is an attractor for the map T_β. This allows us to extend the measure μ_β defined above by simply setting $h_\beta(x) = 0$ on $[1, \lfloor\beta\rfloor/(\beta-1)]$. The new measure obtained is invariant with respected to the extended greedy transformation, which from now on we will refer to as simply the greedy transformation.

In the last decade an interest in expansions to non-integer bases $\beta > 1$ other than the greedy expansion has developed. In particular in papers by P. Erdös, M. and I. Joo, V. Komornik, P. Loreti, F. Schnitzer and others, the so-called *lazy expansion* to base $\beta \in (1,2)$ has been studied, see e.g.[EJK], [KL1],[KL2], [JS]. In particular in these (and other) papers the lazy expansion of 1, and its relation to the greedy expansion of 1 has been thoroughly investigated. The dynamical properties of the lazy expansion, as well as the interconnection with the greedy expansion has been studied in [DK1].

Dynamically the lazy expansion is obtained by iterating the map L_β defined on $[0, \lfloor\beta\rfloor/(\beta-1)]$ by

$$L_\beta(x) = \beta x - d \quad \text{for } x \in \Delta(d),$$

where

$$\Delta(0) = \left[0, \frac{\lfloor\beta\rfloor}{\beta(\beta-1)}\right],$$

and

$$\begin{aligned}\Delta(d) &= \left(\frac{\lfloor\beta\rfloor}{\beta-1} - \frac{\lfloor\beta\rfloor-d+1}{\beta}, \frac{\lfloor\beta\rfloor}{\beta-1} - \frac{\lfloor\beta\rfloor-d}{\beta}\right] \\ &= \left(\frac{\lfloor\beta\rfloor}{\beta(\beta-1)} + \frac{d-1}{\beta}, \frac{\lfloor\beta\rfloor}{\beta(\beta-1)} + \frac{d}{\beta}\right], \quad d \in \{1, 2, \ldots, \lfloor\beta\rfloor\}.\end{aligned}$$

The greedy map T_β and the lazy map L_β are strongly related. If one defines the map $\psi : [0, \lfloor\beta\rfloor/(\beta-1)] \to [0, \lfloor\beta\rfloor/(\beta-1)]$ by

$$\psi(x) = \frac{\lfloor\beta\rfloor}{\beta-1} - x,$$

then ψ is a continuous (hence measurable) bijection and $\psi T_\beta = L_\beta \psi$. As a consequence of this one sees that if the greedy expansion of x is given by (12), then the lazy expansion of $\psi(x)$ is given by

$$\psi(x) = \sum_{k=1}^{\infty} \frac{\lfloor\beta\rfloor - a_k}{\beta^k}.$$

As mentioned above the greedy transformations has an attractor the set $[0, 1)$. Likewise, the interval $(\psi(1), \lfloor\beta\rfloor/(\beta-1)]$ is an attractor for L_β. Hence, any invariant measure must be supported on the corresponding attractor.

Using the map ψ, any T_β-invariant measure on $[0, \lfloor\beta\rfloor/(\beta-1))$ gives rise to an L_β-invariant on the same space. Hence, the measure $\rho_\beta = \mu_\beta \circ \psi$ is L_β-invariant. It was shown by F. Hofbauer [Ho] that μ_β is the unique measure of maximal entropy of entropy $\log\beta$. Using the map ψ one concludes that the measure $\rho_\beta = \mu_\beta \circ \psi$ is the unique of measure of maximal entropy for the map L_β. Using the tools of ergodic theory, one can give a complete description of the distribution of the digits generated and their statistical properties in general (see [DK1]).

3. Intermediate β-Expansions

In the previous section, two transformations were given, the greedy and the lazy, whose iterations generated expansions in base β for points in $[0, \lfloor\beta\rfloor/(\beta-1)]$. In [DK1], a family of transformations, defined on $[0, \lfloor\beta\rfloor/(\beta-1)]$, were given whose iterations generate *intermediate* expansions in base β that are neither greedy nor lazy. To do so, we first super-impose the greedy map and the corresponding lazy map on $[0, \lfloor\beta\rfloor/(\beta-1)]$. One then gets a natural partition of $[0, \lfloor\beta\rfloor/(\beta-1)]$ into two types of sets *switch regions* $\{S_1, S_2, \cdots, S_{\lfloor\beta\rfloor}\}$, and *equality regions* $\{E_0, E_1, \ldots, E_{\lfloor\beta\rfloor}\}\}$, where

$$S_k = \left[\frac{k}{\beta}, \frac{\lfloor\beta\rfloor}{\beta(\beta-1)} + \frac{k-1}{\beta}\right], \quad k = 1, \ldots, \lfloor\beta\rfloor,$$

and

$$E_k = \left(\frac{\lfloor\beta\rfloor}{\beta(\beta-1)} + \frac{k-1}{\beta}, \frac{k+1}{\beta}\right), \quad k = 1, \ldots, \lfloor\beta\rfloor - 1,$$

$$E_0 = \left[0, \frac{1}{\beta}\right) \quad \text{and} \quad E_{\lfloor\beta\rfloor} = \left(\frac{\lfloor\beta\rfloor}{\beta(\beta-1)} + \frac{\lfloor\beta\rfloor-1}{\beta}, \frac{\lfloor\beta\rfloor}{\beta-1}\right].$$

On S_k, the greedy map assigns the digit k, while the lazy map assigns the digit $k-1$. Outside these switch regions both maps are identical, and hence they assign the same digits. This means that new algorithms can be defined based on what one decides to do in the switch regions. In [DK1] the case when each overlapping interval is divided in the same proportion was considered. To be more precise, for each

$$\alpha \in \left[0, \frac{\lfloor\beta\rfloor}{\beta-1} - 1\right]$$

define a map $N_{\beta,\alpha} : [0, \lfloor\beta\rfloor/(\beta-1)] \to [0, \lfloor\beta\rfloor/(\beta-1)]$ by

$$N_{\beta,\alpha}(x) := \begin{cases} \beta x, & x \in [0, m_1), \\ \beta x - i, & x \in [m_i, m_{i+1}),\ 1 \le i < \lfloor\beta\rfloor, \\ \beta x - \lfloor\beta\rfloor, & x \in [m_{\lfloor\beta\rfloor}, \frac{\lfloor\beta\rfloor}{\beta-1}), \end{cases}$$

where

$$m_i := \frac{\alpha + i}{\beta}, \quad i = 1, \ldots, \lfloor\beta\rfloor.$$

It is not hard to see that the interval $[\alpha, \alpha+1)$ is an attractor for the transformation $N_{\beta,\alpha}$. Just as the greedy map T_β and the lazy map L_β, iterations of the map $N_{\beta,\alpha}$ generate series expansion in base β of the form (12).

In order to understand the dynamical properties of $N_{\beta,\alpha}$ consider the map $\psi^* : [\alpha, \alpha+1) \to [0, 1]$, given by $\psi^*(x) := \alpha + 1 - x$. Setting

$$T^*(x) = \psi^*(N_{\beta,\alpha}(\psi^{*^{-1}}(x))).$$

THEOREM 1 (DK1). Let $\beta > 1$, $\beta \notin \mathbb{Z}$, and let $\alpha \in [0, \frac{\lfloor\beta\rfloor}{\beta-1} - 1)$. Then

$$T^*(x) = \beta x + \alpha^* \ (\mathrm{mod}\ 1),$$

where $\alpha^* = \lfloor\beta\rfloor - (\alpha+1)(\beta-1)$.

Remark 1. Maps of the form $T_{\beta,\alpha}(x) = \beta x + \alpha$ (mod 1) were first introduced and studied by Parry in [P1]. Parry showed that $T_{\beta,\alpha}$ is ergodic with respect to the Lebesgue measure λ, and that there exists a unique $T_{\beta,\alpha}$-invariant probability measure $\tau\ (= \tau_{\beta,\alpha}) \ll \lambda$, with density

$$h_\tau(x) = M\left(\sum_{x<T^n_{\beta,\alpha}(1)} \frac{1}{\beta^n} - \sum_{x<T^n_{\beta,\alpha}(0)} \frac{1}{\beta^n}\right) 1_{[0,1)}(x),$$

where $M = M_{\beta,\alpha}$ is a normalizing constant. In [Wi], Wilkinson showned that $T_{\beta,\alpha}$ is weak-Bernoulli with respect to Parry's measure τ when $\beta > 2$, a result which was first extended in [P2] to $\beta > \sqrt{2}$, and then by R. Palmer [Pa]. In [Pa] all pairs (β, α) are characterized for which $T_{\beta,\alpha}$ is weakly-Bernoulli.

4. Random β-Expansions

It is natural to seek a *nice* map whose iterations generate **all** possible expansions in base β. In order to do so, the following random procedure was introduced in [DK2] and studied further in [DK2], [DV]. The expansions generated are random mixtures of greedy and lazy expansions, and are obtained by randomizing the choice of the map used in the switch regions. So, whenever x belongs to a switch region flip a coin to decide which map will be applied to x (greedy or lazy), and hence which digit will be assigned. To be more precise consider $\Omega = \{0,1\}^{\mathbb{N}}$ with product σ-algebra $\mathcal{A}$. Let

$\sigma : \Omega \to \Omega$ be the left shift, and define $K_\beta : \Omega \times [0, \lfloor\beta\rfloor/(\beta-1)] \to \Omega \times [0, \lfloor\beta\rfloor/(\beta-1)]$ by

$$(14) \quad K_\beta(\omega, x) = \begin{cases} (\omega, \beta x - k) & x \in E_k,\ k = 0, 1, \dots, \lfloor\beta\rfloor, \\ (\sigma(\omega), \beta x - k) & x \in S_k \text{ and } \omega_1 = 1,\ k = 1, \dots, \lfloor\beta\rfloor, \\ (\sigma(\omega), \beta x - k + 1) & x \in S_k \text{ and } \omega_1 = 0,\ k = 1, \dots, \lfloor\beta\rfloor. \end{cases}$$

We call the map K_β the *random β-transformation*, and the elements of Ω represent the coin tosses ('heads'=1 and 'tails'=0) used every time the orbit hits a switch region. In order to see that iterations of K_β generate expansions in base β, we will rewrite K_β as follows. Let $E = \bigcup_{k=0}^{\lfloor\beta\rfloor} E_k$, and $S = \bigcup_{k=1}^{\lfloor\beta\rfloor} S_k$. Define

$$d_1 = d_1(\omega, x) = \begin{cases} k & \text{if } x \in E_k,\ k = 0, 1, \dots, \lfloor\beta\rfloor, \\ & \text{or } (\omega, x) \in \{\omega_1 = 1\} \times S_k,\ k = 1, 2, \dots, \lfloor\beta\rfloor, \\ k - 1 & \text{if } (\omega, x) \in \{\omega_1 = 0\} \times S_k,\ k = 1, 2, \dots, \lfloor\beta\rfloor, \end{cases}$$

then

$$K_\beta(\omega, x) = \begin{cases} (\omega, \beta x - d_1) & \text{if } x \in E, \\ (\sigma(\omega), \beta x - d_1) & \text{if } x \in S. \end{cases}$$

Set $d_n = d_n(\omega, x) = d_1\left(K_\beta^{n-1}(\omega, x)\right)$. The sequence $\{d_i\}$ is referred to as the *random digits* of x in base β. Let $\pi_2 : \Omega \times [0, \lfloor\beta\rfloor/(\beta-1)] \to [0, \lfloor\beta\rfloor/(\beta-1)]$ be the canonical projection onto the second coordinate. Then

$$\pi_2\left(K_\beta^n(\omega, x)\right) = \beta^n x - \beta^{n-1} d_1 - \cdots - \beta d_{n-1} - d_n,$$

and rewriting yields

$$x = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_n}{\beta^n} + \frac{\pi_2\left(K_\beta^n(\omega, x)\right)}{\beta^n}.$$

Since $\pi_2\left(K_\beta^n(\omega, x)\right) \in [0, \lfloor\beta\rfloor/(\beta-1)]$, it follows that

$$x - \sum_{i=1}^{n} \frac{d_i}{\beta^i} = \frac{\pi_2\left(K_\beta^n(\omega, x)\right)}{\beta^n} \to 0 \qquad \text{as } n \to \infty.$$

This shows that for all $\omega \in \Omega$ and for all $x \in [0, \lfloor\beta\rfloor/(\beta-1)]$ one has that

$$x = \sum_{i=1}^{\infty} \frac{d_i}{\beta^i} = \sum_{i=1}^{\infty} \frac{d_i(\omega, x)}{\beta^i}.$$

The random procedure just described shows that with each $\omega \in \Omega$ corresponds an algorithm that produces expansions in base β. Further, if we identify the point (ω, x) with $(\omega, (d_1(\omega, x), d_2(\omega, x), \dots))$, then the action of K_β on the second coordinate corresponds to the left shift. In [DV] it was shown that the map K_β has three essential properties.

(i) It preserves the lexicographical ordering on the set of all possible random expansions.

(ii) It captures all possible expansions in base β.

(iii) It gives a characterization of unique expansions.

To be more precise, et $<_{lex}$ and $\leq_{lex}$ denote the lexicographical ordering on both Ω, and $\{0, 1, \ldots, \lfloor\beta\rfloor\}^{\mathbb{N}}$. For each $x \in [0, \frac{\lfloor\beta\rfloor}{\beta-1}]$, consider the set

$$D_x = \{(d_1(\omega, x), d_2(\omega, x), \ldots) : \omega \in \Omega\}\,.$$

The following theorem shows that the elements of D_x are ordered by the lexicographical ordering on Ω.

THEOREM 2. [DV] Suppose $\omega, \omega' \in \Omega$ are such that $\omega <_{lex} \omega'$, then

$$(d_1(\omega, x), d_2(\omega, x), \ldots) \leq_{\text{lex}} (d_1(\omega', x), d_2(\omega', x), \ldots).$$

Theorem 2 gives another proof of the fact that among all possible β-expansions of a point $x \in [0, \lfloor\beta\rfloor/(\beta-1)]$, the greedy expansion is the largest in lexicographical order (it corresponds to the largest element $(1, 1, \ldots)$ of Ω), and the lazy is the smallest (it corresponds to the smallest element $(0, 0, \ldots)$ of Ω).

THEOREM 3. [DV] Let $x \in [0, \lfloor\beta\rfloor/(\beta-1)]$, and let $x = \sum_{i=1}^{\infty} a_i/\beta^i$ with $a_i \in \{0, 1, \ldots \lfloor\beta\rfloor\}$ be a representation of x in base β. Then there exists an $\omega \in \Omega$ such that $a_i = d_i(w, x)$.

The above theorem shows that the map K_β captures all algorithms producing β-expansions. In fact the proof shows that there are three possibilities based on how often does the sequence $\{\sum_{l=1}^{\infty} \frac{a_{j+l-1}}{\beta^l} : j \geq 1\}$ hit the S-region.

- If it hits S infinitely often, then there is a unique $\omega \in \Omega$ such that $d_i(\omega, x) = a_i$ for all $i \geq 1$.
- If it hits S finitely many times only, then there is a cylinder C in Ω such that $d_i(\omega, x) = a_i$ for all $i \geq 1$, and all $\omega \in C$.
- If it never hits S, then $d_i(\omega, x) = a_i$ for all $i \geq 1$, and all $\omega \in \Omega$. In this case x has a unique representation in base β, and the greedy expansion of x is the only representation of x in base β. Furthermore, for all $n \geq 1$, $x_n = T_\beta^{n-1}x = S_\beta^{n-1}x$.

The following theorem can be easily proved from the structure of the map K_β see [DV]. We remark that this theorem was obtained independently for the case $x = 1$, and via other methods in [**KL**], Theorem 3.1.

THEOREM 4. Suppose x has an infinite greedy expansion of the form $x = a_1/\beta + a_2/\beta^2 + \ldots$. Then, x has a unique expansion in base β if and only if for all $n \geq 0$ with $a_{n+1} \geq 1$, we have $T_\beta^{n+1}x > \frac{\lfloor\beta\rfloor}{\beta-1} - 1$.

Remark 2. The question of unique expansion was initiated in the 1990's by a group of Hungarian mathematicians led by Paul Erdös [**EJ**], [**EJK**], [**EK**], [**JS**], [**KL**]. Their initial investigation was for $1 < \beta < 2$, and their interest was in how "large" (in the measure theoretic and topological sense) is the set of points with a unique expansion. They showed that the set of points with unique expansion has zero Lebesgue measure, and if $\beta < \dfrac{1+\sqrt{5}}{2}$, then every x has a continuum of expansions. The smallest $1 < \beta < 2$ for which 1 has a unique expansion was obtained by Komornik and Loretti [**KL**], which is defined as the unique solution of the equation $\sum_{k=1}^{\infty} t_n x^{-n+1} = 1$, where $t = (t_n)$ is the well-known Thue-Morse sequence

$$t = 0110\,1001\,1001\,0110\,1001\,0110\,0110\,1001\cdots.$$

In [**AC**], Allouche and Cosnard proved that this number is transcendental. Questions on the Hausdorff dimensions of this set were studied in [**GS**].

5. Ergodic Properties of K_β

In [**DV**] it was shown that the map K_β on $\Omega \times [0, \frac{\lfloor\beta\rfloor}{\beta-1}]$ can be essentially identified with the left shift on $\{0, \ldots, \lfloor\beta\rfloor\}^{\mathbb{N}}$. This allows K_β to inherit all the nice dynamical properties of the shift map. As a consequence one can identify the maximal entropy of the map K_β, and show that there is a unique K_β invariant measure that has this maximal entropy. In this section we summarize the results obtained in [**DV**].

Let $D = \{0, \ldots, \lfloor\beta\rfloor\}^{\mathbb{N}}$ be equipped with the product σ-algebra $\mathcal{D}$, and the uniform product measure $\mathbb{P}$. Let σ' be the left shift on D.. On the set $\Omega \times [0, \lfloor\beta\rfloor/(\beta-1)]$ we consider the product σ-algebra $\mathcal{A} \times \mathcal{B}$, where $\mathcal{B}$ is the Borel σ-algebra on $[0, \lfloor\beta\rfloor/(\beta-1)]$, and $\mathcal{A}$ the product σ-algebra on Ω. Define the function $\varphi : \Omega \times [0, \frac{\lfloor\beta\rfloor}{\beta-1}] \to D$ by

$$\varphi(\omega, x) = (d_1(\omega, x), d_2(\omega, x), \ldots).$$

It is easily seen that φ is measurable, and $\varphi \circ K_\beta = \sigma' \circ \varphi$. Furthermore, Theorem 3 implies that φ is surjective. Unfortunately the map φ is not injective (see the paragraph following Theorem 3). However, the restriction φ to an appropriate K_β-invariant subset is in fact invertible, and the image of this set under φ has full $\mathbb{P}$ measure. To be more precise, let

$$Z = \{(\omega, x) \in \Omega \times [0, \frac{\lfloor\beta\rfloor}{\beta-1}] : K_\beta^n(\omega, x) \in \Omega \times S \text{ infinitely often}\},$$

and

$$D' = \{(a_1, a_2, \ldots) \in D : \sum_{i=1}^{\infty} \frac{a_{j+i-1}}{\beta^i} \in S \text{ for infinitely many } j\}.$$

Then, $\varphi(Z) = D'$, $K_\beta^{-1}(Z) = Z$ and $(\sigma')^{-1}(D') = D'$. Let φ' be the restriction of the map φ to Z.

Theorem 5. [DV] The map $\varphi' : Z \to D'$ is a bimeasurable bijection, and $\mathbb{P}(D') = 1$.

Now, consider the K_β-invariant measure ν_β defined on $\mathcal{A} \times \mathcal{B}$ by $\nu_\beta(A) = \mathbb{P}\left(\varphi(Z \cap A)\right)$.

Theorem 6. [DV] Let $\beta > 1$ be a non-integer. Then the map $\varphi : (\Omega \times [0, \frac{\lfloor\beta\rfloor}{\beta-1}], \mathcal{A} \times \mathcal{B}, \nu_\beta, K_\beta) \to (D, \mathcal{D}, \mathbb{P}, \sigma')$ is a measurable isomorphism.

The above theorem implies that $h_{\nu_\beta}(K_\beta) = \log(1+\lfloor\beta\rfloor)$. Furthermore, one can show that any K_β-invariant measure μ with $\mu(Z^c) > 0$, one has $h_\mu(K_\beta) < \log(1 + \lfloor\beta\rfloor)$. Using this and the fact that $\mathbb{P}$ is the unique shift invariant measure of maximal entropy on D, one arrives at the following theorem.

Theorem 7. [**DV**] The measure ν_β is the unique K_β-invariant measure of maximal entropy.

An interesting consequence of the above theorems is that if $\beta, \beta' > 1$ are non-integers, then

$$\lfloor\beta\rfloor = \lfloor\beta'\rfloor \text{ if and only if } (K_\beta, \nu_\beta) \text{ is isomorphic to } (K_{\beta'}, \nu_{\beta'}).$$

There is an intimate connection between the measure ν_β and the so called Erdös measure on $[0, \frac{1}{\beta-1}]$, which is an infinite convolution of Bernoulli measure. In [**DV**] it was shown that the second marginal of ν_β is exactly the Erdös measure. To be more precise, let $\pi_2 : \Omega \times [0, \frac{\lfloor\beta\rfloor}{\beta-1}] \to [0, \frac{\lfloor\beta\rfloor}{\beta-1}]$ be the natural projection $\pi_2(\omega, x) = x$, and let $\nu_\beta \circ \pi_2^{-1}$ be the measure defined on $[0, \frac{\lfloor\beta\rfloor}{\beta-1}]$ by $\nu_\beta \circ \pi_2^{-1}(A) = \nu_\beta(\pi_2^{-1}A)$. Consider the purely discontinuous measures $\{\delta_i\}_{i\geq 1}$ defined on $\mathbb{R}$ as follows:

$$\delta_i(\{0\}) = \frac{1}{\lfloor\beta\rfloor + 1}, \ldots, \delta_i(\{\lfloor\beta\rfloor\beta^{-i}\}) = \frac{1}{\lfloor\beta\rfloor + 1}.$$

So δ_i is concentrated on the set

$$\{0, \beta^{-i}, \ldots, \lfloor\beta\rfloor\beta^{-i}\}.$$

Let δ_β be the corresponding infinite Bernoulli convolution,

$$\delta_\beta = \lim_{n\to\infty} \delta_1 * \ldots * \delta_n.$$

Theorem 8. [**DV**] $\nu_\beta \circ \pi_2^{-1} = \delta_\beta$.

Remark 3. If $\beta \in (1,2)$ then δ is an Erdös measure on $[0, \frac{1}{\beta-1}]$, and lots of things are already known. For example, if β is a Pisot number, then δ is singular with respect to Lebesque measure; [**E1**], [**E2**], [**S**]. Further, for almost all $\beta \in (1,2)$ the measure δ is equivalent to Lebesgue measure; [**So**], [**MS**]. There are many generalizations of these results to the case of an arbitrary digit set (see [**PSS**] for more references and results).

6. The Markov property of K_β

In the previous section it was shown that the dynamical system $(\Omega\times[0, \frac{\lfloor\beta\rfloor}{\beta-1}], \mathcal{A}\times\mathcal{B}, \nu_\beta, K_\beta)$ is isomorphic to a Bernoulli shift, dynamically a very desirable property. Unfortunately, the natural partition

$$\mathcal{E} = \{E_0, S_1, E_1, S_2 \ldots, S_{\lfloor\beta\rfloor}, E_{\lfloor\beta\rfloor}\}$$

is not the generating Bernoulli partition. As a result it is quite hard to calculate the ν_β measure of measurable sets, even the basic ones such as $C \times E_i$ or $C \times S_i$ with C a cylinder in Ω. In [**DV**] (see also [**DK2**]), it was shown that for a certain Pisot values of β, one can find a refinement of $\mathcal{E}$ which is a generating Markov partition for the dynamical system $(\Omega \times [0, \frac{\lfloor\beta\rfloor}{\beta-1}], \mathcal{A}\times\mathcal{B}, \nu_\beta, K_\beta)$.

Let $\beta > 1$ be such that the greedy expansion of 1 in base β has the form

$$1 = b_1/\beta + b_2/\beta^2 + \ldots + b_n/\beta^n$$

with $b_i \geq 1$, and $n \geq 2$ (notice that $\lfloor\beta\rfloor = b_1$). We describe briefly how one finds a Markov refinement of $\mathcal{E}$ such that the dynamics of K_β can be identified with a subshift of finite type with an irreducible adjacency matrix. This permits one to define several K_β invariant Markov measures, one of which is the measure ν_β, the measure of maximal entropy. We refer the reader to [DV] for more details.

The crucial ingredient in obtaining the Markov partition is the following simple proposition that gives a complete description of the K_β-orbits of the points $(\omega, 1)$ and $(\omega, \psi(1))$ for all $\omega \in \Omega$. Recall that $\psi(1) = \dfrac{\lfloor\beta\rfloor}{\beta-1} = \dfrac{b_1}{\beta-1}$.

Proposition 1. [DV] Suppose 1 has a finite greedy expansion of the form

$$1 = b_1/\beta + b_2/\beta^2 + \ldots + b_n/\beta^n.$$

If $b_j \geq 1$ for $1 \leq j \leq n$, then

(i) $T_\beta^i 1 = L_\beta^i 1 \in E_{b_{i+1}}$, $i = 0, 1, \ldots, n-2$,

(ii) $T_\beta^{n-1}1 = L_\beta^{n-1}1 = \frac{b_n}{\beta} \in S_{b_n}$, $T_\beta^n 1 = 0$, and $L_\beta^n 1 = 1$.

(iii) $T_\beta^i\psi(1) = L_\beta^i\psi(1) \in E_{b_1-b_{i+1}}$, $0 \le i \le n-2$,

(iv) $T_\beta^{n-1}\psi(1) = L_\beta^{n-1}\psi(1) = \frac{b_1}{\beta(\beta-1)} + \frac{b_1-b_n}{\beta} \in S_{b_1-b_n+1}$, $T_\beta^n\psi(1) = \frac{b_1}{\beta-1} - 1$, and $L_\beta^n\psi(1)) = \frac{b_1}{\beta-1}$.

We use the points $\{T_\beta^i 1, T_\beta^i\psi(1) : i = 0, \dots, n-2\}$ to refine the partition

$$\mathcal{E} = \{E_0, S_1, E_1, S_2 \dots, S_{b_1}, E_{b_1}\}.$$

We obtain a refinement

$$\mathcal{C} = \{C_0, C_1, \dots, C_L\}.$$

We choose $\mathcal{C}$ to satisfy the following. For $0 \le i \le n-2$,

- $T_\beta^i 1 \in C_j$ if and only if $T_\beta^i 1$ is a left end-point of C_j,

- $T_\beta^i\psi(1) \in C_j$ if and only if $T_\beta^i\psi(1)$ is a right end-point of C_j.

From the dynamics of K_β on this refinement, one reads the following properties of $\mathcal{C}$.

p1- $C_0 = [0, \psi(1)]$ and $C_L = \left[1, \frac{b_1}{\beta-1}\right]$.

p2- For $i = 0, 1, \dots, b_1$, E_i can be written as a finite disjoint union of the form $E_i = \cup_{j \in M_i} C_j$ with $M_0, M_1, \dots, M_{b_1}$ disjoint subsets of $\{0, 1, \dots L\}$. Further, the number of elements in M_i equals the number of elements in M_{b_1-i}.

p3- For each S_i there corresponds exactly one $j \in \{0, 1, \dots, L\} \setminus \cup_{k=0}^{b_1} M_k$ such that $S_i = C_j$. This is possible since the T_β-orbits of 1 and $\psi(1)$ never hit the interior of $\cup_{i=1}^{b_1} S_i$.

p4- If $C_j \subset E_i$, then $T_\beta(C_j) = S_\beta(C_j)$ is a finite disjoint union of elements of $\mathcal{C}$, say $T_\beta(C_j) = C_{i_1} \cup \dots \cup Ci_l$. Since $\psi(C_j) = C_{L-j} \subset E_{b_1-i}$, it follows that $T_\beta(C_{L-j}) = C_{L-i_1} \cup \dots \cup C_{L-i_l}$.

p5- If $C_j = S_i$, then $T_\beta(C_j) = C_0$ and $L_\beta(C_j) = C_L$.

From **p4** and **p5** we conclude that $\mathcal{C}$ is a Markov partition underlying the map K_β.

To define the underlying subshift of finite type associated with the map K_β, we consider the $(L+1) \times (L+1)$ matrix $A = (a_{i,j})$ with entries in $\{0, 1\}$ defined by

$$a_{i,j} = \begin{cases} 1 & \text{if } i \in \cup_{k=0}^{b_1} M_k \text{ and } C_j \subseteq T_\beta(C_i), \\ 0 & \text{if } i \in \cup_{k=0}^{b_1} M_k \text{ and } \lambda(C_i \cap T_\beta^{-1} C_j) = 0, \\ 1 & \text{if } i \in \{0, \dots, L\} \setminus \cup_{k=0}^{b_1} M_k \text{ and } j = 0, L, \\ 0 & \text{if } i \in \{0, \dots, L\} \setminus \cup_{k=0}^{b_1} M_k \text{ and } j \ne 0, L. \end{cases} \tag{15}$$

The matrix A is irreducible, and can be seen as defining a graph with vertices $\{0, 1, \dots, L\}$. There is an edge from vertex i to vertex j if and only if $C_j \subseteq T_\beta(C_i)$.

The set Y of all infinite paths on this graph is the topological Markov chain determined by A. That is, $Y = \{y = (y_i) \in \{0, 1, \dots L\}^{\mathbb{N}} : a_{y_i y_{i+1}} = 1\}$. We let σ_Y be the left shift on Y. In [DV], the topological entropy of (Y, σ_Y) was shown to be $h(Y) = \log(b_1 + 1)$. The corresponding unique Markov measure of maximal entropy Q can be calculated following Parry's recipe. Namely, Q is generated by the transition matrix $P = (p_{i,j})$, where $p_{i,j} = a_{i,j} \frac{v_j}{(b_1+1)v_i}$, and stationary distribution $p = v$ which is the right positive eigenvector of A satisfying $\sum_{i=0}^{L} v_i = 1$.

To see that the dynamics of K_β and σ_Y are essentially the same, we first need to find a map α from Y to $\Omega \times [0, \frac{\lfloor\beta\rfloor}{\beta-1}]$ that commutes the actions of K_β and σ_Y. For ease of notation, we denote by $s_1, s_2, \dots, s_{b_1}$ the states $j \in \{0, 1, \dots, L\} \setminus \cup_{k=0}^{b_1} M_k$ corresponding to the switch regions $S_1, S_2, \dots, S_{b_1}$ respectively.

For each $y \in Y$, one can easily associate an $x \in \left[0, \frac{b_1}{\beta-1}\right]$. One first associates a sequence $(e_i) \in \{0, 1, \dots, b_1\}^{\mathbb{N}}$ as follows,

$$e_j = \begin{cases} i & \text{if } y_j \in M_i, \\ i & \text{if } y_j = s_i \text{ and } y_{j+1} = 0, \\ i-1 & \text{if } y_j = s_i \text{ and } y_{j+1} = L. \end{cases} \tag{16}$$

Now set

$$x = \sum_{j=1}^{\infty} \frac{e_j}{\beta^j}. \tag{17}$$

To define a point $\omega \in \Omega$ corresponding to y, one needs that $y_i \in \{s_1, \dots s_l\}$ infinitely often. For this reason it is not possible to define α on all of Y, but only on an invariant subset. To be more precise, let

$$Y' = \{y = (y_1, y_2, \cdots) \in Y : y_i \in \{s_1, \dots, s_{b_1}\} \text{ for infinitely many } i\text{'s}\}.$$

Given $y \in Y'$, we first locate the indices $n_i = n_i(y)$ where the realization y of the Markov chain is in state s_ℓ for some $\ell \in \{1, \dots, b_1\}$. That is, let $n_1 < n_2 < \cdots$ be the indices such that $y_{n_i} = s_\ell$ for some $\ell = 1, \dots, b_1$. Define

$$\omega_j = \begin{cases} 1 & \text{if } y_{n_j+1} = 0, \\ 0 & \text{if } y_{n_j+1} = L. \end{cases}$$

Now set $\alpha(y) = (\omega, x)$.

Theorem 9. [DV] The map

$$\alpha : (Y, \mathcal{G}, Q, \sigma_Y) \to \left(\Omega \times \left[0, \frac{b_1}{\beta - 1}\right], \mathcal{D} \times \mathcal{B}, Q \circ \alpha^{-1}, K_\beta\right)$$

is a measurable isomorphism.

The above theorem implies that $h_{Q\circ\alpha^{-1}}(K_\beta) = \log(b_1 + 1) = \log(\lfloor\beta\rfloor + 1)$. By the uniqueness of the measure ν_β described in the previous section, we see that $Q \circ \alpha^{-1} = \nu_\beta$. Recall that the projection of ν_β in the second coordinate is the Erdös measure $\delta_\beta = \nu_\beta \circ \pi_2^{-1}$. One easily calculates that $\delta_\beta(E_i) = \nu_\beta(\Omega \times E_i) = \sum_{i \in M_i} v_i$, and $\delta_\beta(S_i) = \nu_\beta(\Omega \times S_i) = v_{s_i}$.

In [DV] it was shown that the projection of $Q \circ \alpha^{-1} = \nu_\beta$ on the first coordinates is the uniform Bernoulli measure on Ω. If in the switch regions we decide to flip a biased coin, with $0 < \mathbb{P}(\text{Heads}) = p < 1$, in order to decide whether to use the greedy or the lazy map, then the measure of maximal entropy does not reflect this fact. A natural invariant measure that preserves this property is obtained by considering the Markov measure on Y with transition probabilities $p_{i,j}$ given by,

$$p_{ij} = \begin{cases} \lambda(C_i \cap T_\beta^{-1} C_j)/\lambda(C_i) & \text{if } i \in \cup_{k=0}^{b_1} M_k \\ p & \text{if } i \in \{0,1,\ldots,L\} \setminus \cup_{k=0}^{b_1} M_k \text{ and } j=0, \\ 1-p & \text{if } i \in \{0,1,\ldots,L\} \setminus \cup_{k=0}^{b_1} M_k \text{ and } j=L, \end{cases}$$

and initial distribution the corresponding stationary distribution. Another interesting feature of this measure is that if $p = 1$, then one gets the Parry measure μ_β, and if $p = 0$, then one gets the lazy measure ρ_β.

References

[AC] Allouche, J.-P.; Cosnard, M. Non-integer bases, iteration of continuous real maps, and an arithmetic self-similar set. Acta Math. Hungar. 91 (2001), no. 4, 325–332.

[DK1] Dajani, K., Kraaikamp, C. – *From Greedy to Lazy Expansions, and their driving dynamics*, Expo. Math. **20** (2002), 315-327.

[DK2] Dajani, K., Kraaikamp, C. – *On random β-expansions*, Ergodic Theory and Dynam. Sys. **23** (2003), 461-479.

[DV] Dajani, K., de Vries, C. – *Measures of maximal entropy for random β-expansions*, To appear in the Journal of the Eorpean Math. Soc.

[E1] Erdös, P., – *On a family of symmetric Bernoulli convolutions*, Amer. J. Math. **61** (1939), 974-976.

[E2] Erdös, P., – *On the smoothness properties of Bernoulli convolutions*, Amer. J. Math. **62** (1940), 180-186.

[EJ] Erdös, P.; Jo, I. – *On the number of expansions* $1 = \sum_{i=1}^{\infty} q^{-n_i}$ Ann. Univ. Sci. Budapest. Etvs Sect. Math. 36 (1993), 229–233.

[EJK] Erdös, P., Joó, I. and Komornik, V. – *Characterization of the unique expansions* $1 = \sum_{i=1}^{\infty} q^{-n_i}$ *and related problems*, Bull. Soc. Math. France **118** (1990), no. 3, 377–390. MR 91j:11006

[EK] Erdös, P., Komornik,V. – *Developments in non-integer bases*, Acta. Math. Hungar. **79** (1998), 57-83.

[JS] Joó, I., Schnitzer, F.J. – *Expansions with respect to non-integer bases*, Grazer Mathematische Berichte, **329**. Karl-Franzens-Universität Graz, Graz, 1996. MR 98e:11090

[GS] Glendinning, P.; Sidorov, N. – *Unique representations of real numbers in non-integer bases*, Math. Res. Lett. 8 (2001), no. 4, 535–543.

[G] Gel'fond, A.O. – *A common property of number systems*, Izv. Akad. Nauk SSSR. Ser. Mat. **23** (1959), 809–814. MR 22 #702

[KL] Komornik V., Loreti P. – *Subexpansions, superexpansions and uniqueness properties in non-integer bases*, Period. Math. Hungar. 44 (2002), no. 2, 197–218.

[MS] Mauldin R.D., Simon K. – *The equivalence of some Bernoulli convolutions to Lebesgue measure*, Proc. Amer. Math. Soc. Vol. 126, no. 9, (1998), 2733-2736.

[Pa] Palmer, M.R. – *On the classification of measure preserving transformations of Lebesgue spaces, Ph.D. Thesis*, University of Warwick.

[P] Parry, W. – *On the β-expansions of real numbers*, Acta Math. Acad. Sci. Hungar. **11** (1960), 401–416. MR 26 #288

[P1] Parry, W. – *Representations for real numbers*, Acta Math. Acad. Sci. Hungar. **15** (1964), 95–105

[P2] Parry, W. – *The Lorenz attractor and a related population model*, Ergodic theory (Proc. Conf., Math. Forschungsinst., Oberwolfach, 1978), 169–187, Lecture Notes in Math., 729, Springer, Berlin.

[PSS] Peres Y., Schlag W., Solomyak B. – *Sixty years of Bernoulli convolutions*, Fractal Geometry and Stochastic II, Prog. Prob. **46** (1998), 39-65.

[R1] Rényi, A. – *Representations for real numbers and their ergodic properties*, Acta Math. Acad. Sci. Hung. **8** (1957), 472-493. MR 20 #3843

[R2] Rényi, A. – *On algorithms for the generation of real numbers*, Magyar Tud. Akad. Mat. Fiz. Oszt. Kzl. **7** (1957), 265–293. MR 20 #4113

[S] Salem R. – *Algebraic numbers and Fourier Analysis*, Heath (1963).

[Si] Sidorov N. – *Almost every number has a continuum of β-expansion*, Amer. Math. Monthly, **110**, 2003.

[Sm] Smorodinsky, M. – *β-automorphisms are Bernoulli shifts*, Acta Math. Acad. Sci. Hungar. **24** (1973), 273–278.

[So] Solomyak B. – *On the random series $\sum \pm\lambda^i$ (an Erdös problem)*, Annals of Math. **142** (1995), 611-625.

[W] Walters, P. – *An Introduction to Ergodic Theory*, Springer-Verlag, New York (1982).

[Wi] Wilkinson, K.M. –*Ergodic properties of a class of piecewise linear transformations*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **31** (1974/75), 303–328.

On some gap theorems

Paola Loreti[1]

Dipartimento di Metodi e Modelli
Matematici per le Scienze Applicate
Università degli Studi di Roma "La Sapienza"
Via A. Scarpa, 16
00161 Roma,
Italy
loreti@dmmm.uniroma1.it

Abstract. We describe a method developed in collaboration with C. Baiocchi and V. Komornik [**1**], [**2**]. This new method allows us to solve different control problems (see [**8**] for a general description). In particular we investigate this approach by discussing some former results due to A. E. Ingham [**5**] and to Kahane [**6**].

1. Introduction

Motivated by many engineering applications, control problems have been studied for a long time. To describe a general class of control problems let us consider a system, whose evolution is governed by the equation

$$x' = Ax + Bu, \qquad x(0) = x_0,$$

where A is a densely defined, closed linear operator in a Hilbert space H, and B is a densely defined, closed linear operator from another Hilbert space S into H. Here B is called a *control operator* and u is a *control.* In the applications A is usually an elliptic linear partial differential operator.

The exact controllability problem for the system can be stated in the following way: given a positive real number T (time), can we steer the system to the rest at the time T by a suitable choice of the control $u \in L^2(0,T;S)$?

The problem of controllability is equivalent to a problem of observability concerning the *dual system*

$$\phi' = -A^*\phi, \qquad \phi(0) = \phi_0, \qquad \psi = B^*\phi,$$

where A^* and B^* denote respectively the adjoints of A and B. Here B^* is called an *observability operator* and ψ is an *observation.*

The problem of observability in time T for this system can be stated whether two different initial data always lead to two different observations.

We refer to Russell [**12**] for a basic review of the subject of controllability, observability, and stabilizability for linear partial differential equations. In particular [**12**] contains the analysis of the equivalence of the controllability of a linear system and the observability of the associated linear observed system.

From a mathematical point of view the observability problem leads to the research of suitable estimates. The way to prove these estimates depends on the problem we are considering.

If we consider hyperbolic or more general time reversible systems, then the Hilbert Uniqueness Method (HUM), introduced by J.-L. Lions [**11**], applies to a

[1] The author thanks the organizers of the conference "11eme colloque international de l'EWM" for the invitation.

rather large class of partial differential equations and it can be used in general domains. In this approach the estimates are established by the so-called multiplier method.

Alternatively, the so-called harmonic (or non harmonic) analysis method can be applied to problems on special domains. It is based on an accurate study of the spectrum of the spatial operator of the dual problem, and it gives fine estimates of the controllability time. The fields of applicability of the two methods are somewhat complementary. For a general exposition of this method we refer to [**8**].

Let us recall that many linear problems can be solved easily by using Fourier's method of the separation of variables. In the simplest one-dimensional case the desired estimates can then be obtained by applying the theory of Fourier series. This method was extended by Wiener [**13**] to more general series with nonharmonic exponents. Subsequently, his results were generalized by Ingham [**5**], Beurling [**3**], Kahane [**6**] and many others in different directions. These generalizations made it possible to solve a great number of linear problems in one space dimension and on symmetric domains in several dimensions. The object of this work is to describe a method developed in [**1**], [**2**] by discussing the result of Ingham [**5**]. We also give a generalization of a result contained in [**1**] and [**2**].

2. A Theorem of Ingham

Proposition 1. Let $(\omega_n)_{n\in M}$ be a family of real numbers, satisfying the gap condition

$$\inf_{n\neq m} |\omega_n - \omega_m| \geq \gamma$$

for some $\gamma > 0$. Then all sums

$$x(t) := \sum_{n\in M} x_n e^{i\omega_n t}$$

with square summable complex coefficients x_n satisfy the estimate

$$\int_I |x(t)|^2\, dt \leq c \sum_{n\in M} |x_n|^2$$

for all intervals I, with a constant c depending only on γ and the length of I.

Proof. Let us fix an integrable, nonnegative function k and consider the integral

$$\int_{-\infty}^{\infty} k(t)|x(t)|^2\, dt.$$

Introducing the Fourier transform K of k by the usual formula

$$K(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} k(t)e^{-itx}\, dt,$$

this integral can be rewritten in the form

$$\int_{-\infty}^{\infty} k(t)|x(t)|^2\, dt = \sqrt{2\pi} \sum_{n,m} x_n \overline{x_m} K(\omega_n - \omega_m).$$

Here and in the sequel, in all sums the indices n and m run over M.

If the function K vanishes outside the interval $(-\gamma, \gamma)$, then thanks to the gap condition this formula reduces to

$$\int_{-\infty}^{\infty} k(t)|x(t)|^2\, dt = \sqrt{2\pi} K(0) \sum_{n} |x_n|^2.$$

Now, if k has a positive lower bound β on some interval J, then we deduce from this equality the following estimate:

$$\int_J |x(t)|^2 \, dt \leq \frac{\sqrt{2\pi}K(0)}{\beta} \sum_n |x_n|^2.$$

This inequality remains valid for every translate $J + t_0$ of J. Indeed, putting

$$y(t) := x(t + t_0) = \sum_n \big(x_n e^{i\omega_n t_0}\big) e^{i\omega_n t}$$

we have

$$\int_{J+t_0} |x(t)|^2 \, dt = \int_J |y(t)|^2 \, dt \leq \frac{\sqrt{2\pi}K(0)}{\beta} \sum_n \big|x_n e^{i\omega_n t_0}\big|^2 = \frac{\sqrt{2\pi}K(0)}{\beta} \sum_n |x_n|^2.$$

Now every interval I can be covered by a finite number of translates $J + t_1, \dots,$ $J + t_p$ of J. Hence the desired estimate follows with $c = \sqrt{2\pi} pK(0)$:

$$\int_I |x(t)|^2 \, dt \leq \sum_{j=1}^{p} \int_{J+t_j} |x(t)|^2 \, dt \leq \frac{\sqrt{2\pi}K(0)p}{\beta} \sum_n |x_n|^2.$$

It remains to find a function k having the above properties, that is:

- k is integrable, nonnegative and it has a positive lower bound β on some interval J;
- its Fourier transform K vanishes outside the interval $(-\gamma, \gamma)$.

Let us choose an arbitrary nonzero, even, real-valued, square summable function H, which vanishes outside the interval $\left(-\frac{\gamma}{2}, \frac{\gamma}{2}\right)$, and set $K = H * H$, where the convolution is defined by the formula

$$(H * H)(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} H(x - y)H(y) \, dy.$$

It is clear that K is also even and real-valued. Furthermore, we have the following properties:

- The inverse Fourier transforms h and k of H and K are also even and real-valued. This follows from the definition of h and k and from the analogous properties of H and K by the formulae

  $$h(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} H(x) e^{itx} \, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} H(x) \cos tx \, dx$$

  and

  $$k(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} K(x) e^{itx} \, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} K(x) \cos tx \, dx.$$

- Since h is real-valued and since $k = h^2$ by the usual rule of the Fourier transform of a convolution, k is nonnegative.
- Since H is square summable, h is also square summable by Plancherel's theorem. Then $k = h^2$ is integrable by Hölder's theorem.
- Since k is integrable, its Fourier transform K is continuous.
- K vanishes outside the interval $(-\gamma, \gamma)$, because H vanishes outside the interval $\left(-\frac{\gamma}{2}, \frac{\gamma}{2}\right)$ and $K = H * H$.
- We have seen that K is continuous and it has a compact support. Hence K is integrable, and therefore its inverse Fourier transform k is continuous.

- Since $H \not\equiv 0$ by assumption and
$$H(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(x) e^{-itx} \, dx,$$
$h \not\equiv 0$, and thus $k = h^2 \not\equiv 0$.
- Since k is nonnegative, continuous and not identically zero, it has a positive lower bound β on some interval J.

A simple choice for H is the characteristic function of the interval $\left(-\frac{\gamma}{2}, \frac{\gamma}{2}\right)$. □

Proposition 2. Let $(\omega_n)_{n\in M}$ be again a family of real numbers, satisfying the gap condition
$$\inf_{n\neq m} |\omega_n - \omega_m| \geq \gamma$$
for some $\gamma > 0$. Then all sums
$$x(t) := \sum_{n\in M} x_n e^{i\omega_n t}$$
with square summable complex coefficients x_n satisfy the estimate
$$\sum_{n\in M} |x_n|^2 \leq c \int_I |x(t)|^2 \, dt$$
for all intervals I of length $> 2\pi/\gamma$, with a constant c only depending on γ and the length of I.

Proof. By translation invariance it suffices to establish the estimate for the intervals $I = (-R, R)$ with $R > \pi/\gamma$.

Let us choose two functions k, K as at the beginning of the proof of the previous proposition, so as to have the identity
$$\int_{-\infty}^{\infty} k(t)|x(t)|^2 \, dt = \sqrt{2\pi} K(0) \sum_n |x_n|^2.$$
If k is negative outside an interval I, then k is bounded from above by some constant α (because k is continuous), and we deduce from this identity the following inequality:
$$\sqrt{2\pi} K(0) \sum_n |x_n|^2 \leq \alpha \int_I |x(t)|^2 \, dt.$$
If $K(0)$ is positive, then we can conclude by a translation argument.

It remains to find a function k having the required properties, that is:

- k is integrable, bounded from above and nonpositive outside I,
- $K(0)$ is stricly positive and K vanishes outside the open interval $(-\gamma, \gamma)$.

We try to find such a function k with $I = (-R, R)$ as short as possible.

Write $I_\gamma := \left(-\frac{\gamma}{2}, \frac{\gamma}{2}\right)$ for brevity. Let us choose an arbitrary nonzero, even, real-valued function H, which belongs to the Sobolev space $H_0^1(I_\gamma)$, and define the functions K and k by the formulae
$$K = R^2 H * H + H' * H',$$
where R is a positive constant to be chosen later, and by
$$k(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} K(x) e^{itx} \, dx.$$
Then H and H' are square summable, and hence k is integrable by the same arguments as before. Moreover, it follows from the equality
$$k(t) = (R^2 - t^2) h^2(t)$$

that $k(t)$ is nonpositive if t is outside of the interval $(-R,R)$. Finally, since H is even and thus H' is odd, we have

$$\sqrt{2\pi}(H*H)(0)=\int_{-\infty}^{\infty}H(x)H(-x)\ dx=\int_{-\infty}^{\infty}H^2(x)\ dx$$

and

$$\sqrt{2\pi}(H'*H')(0)=\int_{-\infty}^{\infty}H'(x)H'(-x)\ dx=-\int_{-\infty}^{\infty}(H')^2(x)\ dx.$$

Since both H and H' vanish outside I_γ, it follows that

$$\sqrt{2\pi}K(0)=R^2\int_{I_\gamma}H^2(x)\ dx-\int_{I_\gamma}(H')^2(x)\ dx.$$

Since the first integral is positive, we conclude that $K(0)$ is positive if R is large enough.

We give an example of a function for which the above properties are satisfied. Let G be the characteristic function of the interval $\left(-\frac{\gamma}{4},\frac{\gamma}{4}\right)$; then $H=G*G$ satisfies the above conditions. In order to see it, let us compute the value of

$$K=R^2H*H+H'*H'$$

in 0. Since

$$H(x)=\frac{\gamma}{2}-|x|$$

if $x\in I_\gamma$ and 0 otherwise, we have:

$$\int_{I_\gamma}H^2(x)\ dx=\frac{\gamma^3}{12}\quad\text{and}\quad\int_{I_\gamma}(H')^2(x)\ dx=\gamma.$$

Hence the corresponding function K has all the required properties if

$$R>\frac{2\sqrt{3}}{\gamma}.$$

However this simple choice does not lead to the shortest possible intervals $(-R,R)$.

The above form of $K(0)$ shows that the optimal choice for K is the first eigenfunction of the operator $-\Delta$ in $H_0^1(I_\gamma)$. Indeed, this is the function which minimizes the fraction

$$\frac{\int_{I_\gamma}(H')^2(x)\ dx}{\int_{I_\gamma}H^2(x)\ dx}$$

in $H_0^1(I_\gamma)$.

We shall thus use the function $H:\mathbb{R}\to\mathbb{R}$ defined by

$$H(x):=\begin{cases}\cos\frac{\pi x}{\gamma} & \text{if } x\in I_\gamma,\\ 0 & \text{otherwise.}\end{cases}$$

We have

$$\int_{I_\gamma}H^2(x)\ dx=\int_{I_\gamma}\cos^2\frac{\pi x}{\gamma}=\frac{\gamma}{2}$$

and

$$\int_{I_\gamma}(H')^2(x)\ dx=\int_{I_\gamma}\frac{\pi^2}{\gamma^2}\sin^2\frac{\pi x}{\gamma}=\frac{\pi^2}{\gamma^2}\cdot\frac{\gamma}{2},$$

so that

$$\sqrt{2\pi}K(0)=R^2\frac{\gamma}{2}-\frac{\pi^2}{\gamma^2}\cdot\frac{\gamma}{2}=\left(R^2-\frac{\pi^2}{\gamma^2}\right)\cdot\frac{\gamma}{2}$$

is positive if and only if

$$R > \frac{\pi}{\gamma}.$$

This proves the proposition. □

3. On a Theorem of Kahane

In [**6**] Kahane extended Ingham's theorem to several variables. The following propositions generalize to the L^p case some results obtained in [**2**] and also described in [**8**].

Let B_R denote the Euclidean ball of radius R in $\mathbb{R}^N$, centered at the origin:

$$B_R := \{x \in \mathbb{R}^N \ : \ \|x\|_2 < R\}.$$

Fix a number $1 \leq p \leq \infty$.

Proposition 3. Let $(\omega_n)_{n\in M}$ be a family of vectors in $\mathbb{R}^N$, satisfying the gap condition

$$\inf_{n\neq m} \|\omega_n - \omega_m\|_p \geq \gamma$$

for some $\gamma > 0$. Then all sums

$$x(t) := \sum_{n\in M} x_n e^{i\omega_n t}$$

with square summable complex coefficients x_n satisfy the estimate

$$\int_{B_R} |x(t)|^2 \, dt \leq c \sum_{n\in M} |x_n|^2$$

for all $R > 0$.

In the preceding result the value of p has not a particular rule. In order to formulate a theorem concerning the converse inequality, let us denote by μ_p the first eigenvalue of $-\Delta$ in $H_0^1(B^p_{\gamma/2})$ where $B^p_{\gamma/2}$ denotes the ball of radius $\gamma/2$ in $\mathbb{R}^N$ with respect to the p-norm:

$$B^p_{\gamma/2} := \{x \in \mathbb{R}^N \ : \ \|x\|_p < \gamma/2\}.$$

Proposition 4. Let $(\omega_n)_{n\in M}$ be again a family of vectors in $\mathbb{R}^N$, satisfying the gap condition

$$\inf_{n\neq m} \|\omega_n - \omega_m\|_p \geq \gamma$$

for some $\gamma > 0$ and for some $p \geq 1$. Then all sums

$$x(t) := \sum_{n\in M} x_n e^{i\omega_n t}$$

with square summable complex coefficients x_n satisfy the estimate

$$\sum_{n\in M} |x_n|^2 \leq c \int_{B_R} |x(t)|^2 \, dt$$

for all $R > \sqrt{\mu_p}$.

Idea of the proof. We can repeat the proofs given before, by choosing K to be the first eigenfunction of $-\Delta$ in $H_0^1(B^p_{\gamma/2})$, and by choosing $K = (R^2 + \Delta)H * H$ with $R > \sqrt{\mu_p}$. □

References

[1] C. Baiocchi, V. Komornik and P. Loreti, *Théorèmes du type Ingham et application à la théorie du contrôle*, C. R. Acad. Sci. Paris Sér. I Math. 326 (1998), 453–458.

[2] C. Baiocchi, V. Komornik and P. Loreti, *Ingham type theorems and applications to control theory*, Boll. Un. Mat. Ital. B (8), II - B, n. 1, Febbraio 1999, 33–63.

[3] J. N. J. W. L. Carleson and P. Malliavin, editors, *The Collected Works of Arne Beurling*, Volume 2, Birkhäuser, 1989.

[4] R. Courant and D. Hilbert, *Methods of Mathematical Physics I*, John Wiley & Sons, New York, 1989.

[5] A. E. Ingham, *Some trigonometrical inequalities with applications in the theory of series*, Math. Z. 41 (1936), 367–379.

[6] J.-P. Kahane, *Pseudo-périodicité et séries de Fourier lacunaires*, Ann. Sci. de l'E.N.S. 79 (1962), 93–150.

[7] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, Masson, Paris and John Wiley & Sons, Chicester, 1994.

[8] V. Komornik and P. Loreti, *Fourier Series in Control Theory*, Springer, New York, to appear.

[9] J.-L. Lions, *Contrôle des systèmes distribués singuliers*, Gauthier-Villars, Paris, 1983.

[10] J.-L. Lions, *Exact controllability, stabilizability, and perturbations for distributed systems*, Siam Rev. 30 (1988), 1–68.

[11] J.-L. Lions, *Contrôlabilité exacte et stabilisation de systèmes distribués*, Vol. 1, Masson, Paris, 1988.

[12] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions*, SIAM Rev. 20 (1978), 639–739.

[13] N. Wiener, *A class of gap theorems*, Ann. Scuola Norm. Sup. Pisa (2), 3 (1934), 367–372.

[14] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

Fourier Analysis and Continued Fractions

Martine Queffelec

UFR de Mathe'matiques Pures et Appliquées
Université de Lille I
59655 Villeneuve d'Ascq
France
martine@agat.univ-lille.fr

1. Introduction

If we want to study the arithmetical or dynamical properties of real numbers we usually associate to them various expansions such as:

–q-adic-expansions;

–Continued fraction expansions;

–β-expansion; and many others (Engel's Series expansion,...), which lead to different classifications and results; no simple link exists from one to another and it is quite natural to ask about properties involving two different expansions. For example what can we say about normal numbers to base q but non-normal to base p? or does there exist a normal number (to any base) with bounded partial quotients?

The first question refers to p-adic and q-adic expansions, while the second one refers to any adic expansion and to regular continued fraction (RCF) expansion.

After recalling several facts around expansions, we concentrate on the second problem, mentioned by Montgomery in his book "Ten lectures on the interface between Analytic Number Theory and Harmonic Analysis" (1994) [**22**] and try to prove that the set of normal numbers with bounded partial quotients is rather big (in a sense to be precised).

2. Uniform Distribution and Normal Numbers

2.1. Uniform distribution.

DEFINITION 5. *The real sequence $(u_n)_n$ is said to be uniformly distributed mod 1 if*

$$\forall 0 \le a < b \le 1,\ \frac{1}{N}|\{n \le N\ ;\ \{u_n\} \in [a, b]\}| \to b - a,$$

where $\{x\}$ is the fractional part of x.

For example, for any irrational α, the sequence $(n\alpha)_n$ is uniformly distributed mod 1 because of the following [**27**]:

Weyl's criterion: *$(u_n)_n$ is uniformly distributed mod 1 if and only if*

$$\forall k \ne 0,\ \frac{1}{N}\sum_{n \le N} e(ku_n) \to 0,$$

where $e(x) := e^{2i\pi x}$.

The definition can be equivalently expressed in the following way, which allows interesting generalizations: we consider the set of all complex Borel measures $M(\mathbb{T})$ as the dual space of the set of continuous functions $C(\mathbb{T})$ and we have the

Definition 6. *The real sequence $(u_n)_n$ is uniformly distributed mod 1 if the sequence of probability measures on $\mathbb{T}$ (identified to $[0,1)$):*

$$\frac{1}{N}\sum_{n\le N}\delta_{\{u_n\}} \to \lambda \ weak^*$$

where δ_x is the unit mass at x and λ the Lebesgue measure on $[0,1]$.

In case of convergence to some probability measure $\nu \neq \lambda$ on $\mathbb{T}$, we speak of *Weyl-distributed* sequence [**19**]. We can also consider other summability methods.

A first general result on uniform distribution is the following:

Theorem 1. (Weyl) If $(n_k)_k$ is an increasing sequence of positive integers, then the real sequence $(n_k x)_k$ is uniformly distributed mod 1 for λ-almost all x.

Proof: Consider $f_N(x) = \frac{1}{N^2}\sum_{k=0}^{N^2-1} e(jn_kx)$, $j \neq 0$. Now

$$\begin{aligned}\int_{\mathbf{T}} |f_N|^2(x)\,dx &= \tfrac{1}{N^4}\textstyle\sum_{0\le k,l<N^2}\int_{\mathbf{T}} e(j(n_k-n_l)x)\,dx\\ &= O(\tfrac{1}{N^2})\end{aligned}$$

because $n_k \neq n_l$ if $k \neq l$.

It follows that $\sum\int_{\mathbf{T}}|f_N|^2 = \int_{\mathbf{T}}\sum|f_N|^2 < +\infty$ and $\sum|f_N|^2$ converges almost everywhere. This implies that $\lim_N f_N = 0\ \lambda - ae.$

To finish the proof, for $M > 0$ we peak N such that $N^2 \le M < (N+1)^2$, and write

$$\begin{aligned}\tfrac{1}{M}\textstyle\sum_0^{M-1} &= \tfrac{1}{M}(\textstyle\sum_0^{N^2-1}+\sum_{N^2}^{M-1})\\ &= \tfrac{N^2}{M}(\tfrac{1}{N^2}\textstyle\sum_0^{N^2-1}) + \tfrac{1}{M}\textstyle\sum_{N^2}^{M-1}\\ &= \tfrac{N^2}{M}f_N + O(\tfrac{1}{N})\end{aligned}$$

for $0 \le M - N^2 \le 2N$, and this tends to 0. □

Of course the interesting set is the one, attached to (n_k), of those x for which uniform distribution does not happen.

If now, instead of fixing the sequence (n_k), we fix the set E of real numbers and look for sequences of integers ensuring uniform distribution for the elements of E, which property of E is relevant?

In this direction we have the following:

Proposition 1. Let μ be a probability measure supported on $E \subset \mathbb{T}$ satisfying

$$\hat{\mu}(n) = O(|n|^{-\delta}) \tag{18}$$

where $\delta > 0$; then, for every increasing sequence of integers (s_n), $\{s_n x\}$ is uniformly distributed for μ-almost all $x \in E$.

Proof: It turns out to be a simple consequence of the very well-known and very useful result of Davenport-Erdös-LeVeque (see [**6**], [**23**], and [**25**] for various applications).

Theorem 2. (Davenport-Erdös-LeVeque): Let μ be a probability measure on X, (X_n) a sequence of bounded random variables and $S_N = \frac{1}{N}\sum_{n=1}^N X_n$; if

$$\sum_{N\ge1}\frac{1}{N}\int|S_N|^2 d\mu < \infty,$$

then $S_N \to 0$ a.e.-μ.

It remains to verify the hypothesis of the DEL theorem with $X_n = e(-ks_n)$ and μ. If $k \neq 0$,

$$\begin{aligned}\sum_{m,n=1}^{N} \hat{\mu}(k(s_n - s_m)) &= N + \sum_{m,n\leq N,\ m\neq n} \hat{\mu}(k(s_n - s_m)) \\ &\leq N + C \sum_{m,n\leq N,\ m\neq n} |k(s_n - s_m)|^{-\delta} \\ &\leq N + 2C \sum_{m=2}^{N} \sum_{n=1}^{m-1} |(s_n - s_m)|^{-\delta}.\end{aligned}$$

When $m > n$, $s_m - s_n = s_m - s_{m-1} + s_{m-1} - \ldots + s_{n+1} - s_n \geq m - n$, and $\sum_{m,n=1}^{N} \hat{\mu}(k(s_n - s_m)) \leq N + 2C \sum_{m=2}^{N} \sum_{n=1}^{m-1} (m-n)^{-\delta}$; now,

$$\begin{aligned}\sum_{m=2}^{N} \sum_{n=1}^{m-1} (m-n)^{-\delta} &= \sum_{m=2}^{N} \sum_{n=1}^{m-1} (n)^{-\delta} \\ &\leq N \sum_{n=1}^{N-1} (n)^{-\delta} \\ &= O(N^{2-\delta})\end{aligned}$$

Finally

$$\sum_{N\geq 1} \frac{1}{N^3} \sum_{m,n=1}^{N} \hat{\mu}(k(s_n - s_m)) < \infty$$

since $\delta > 0$. □

Property (1) required in the proposition is rather restrictive and gives very few information about the size of the set E; it is not shared by arbitrary sets of positive Lebesgue measure, or even of full Lebesgue measure and actually it is correlated with the shape of the set E.

We begin our investigation with weaker properties.

2.2. Classification of measures on $\mathbb{T}$. If $\mu \in M(\mathbb{T})$, μ is uniquely determined by its Fourier coefficients

$$\hat{\mu}(n) = \int_{\mathbb{T}} e(-nx)\, d\mu(x),\ n \in \mathbb{Z}.$$

$L^1(\mathbb{T})$ can be identified with the set of absolutely continuous measures (with respect to Lebesgue measure) and $M_0(\mathbb{T})$ consists of the measures μ whose Fourier transform tends to 0: $\lim_{|n|\to\infty} \hat{\mu}(n) = 0$.

By the Riemann-Lebesgue lemma, $L^1(\mathbb{T}) \subset M_0(\mathbb{T})$, and Mensov (1916) constructed the first singular measure in $M_0(\mathbb{T})$ (a variant of the Cantor measure) [**21**].

If we denote by $M_c(\mathbb{T})$ the set of continuous measures ($\mu(\{x\}) = 0\ \forall x \in \mathbb{T}$), as a consequence of Wiener's lemma:

$$\lim_N \frac{1}{N+1} \sum_{|n|\leq N} |\hat{\mu}(n)|^2 = \sum_{x\in\mathbb{T}} |\mu\{x\}|^2,$$

we have $M_0(\mathbb{T}) \subset M_c(\mathbb{T})$; Riesz products give examples of continuous measures which are not in $M_0(\mathbb{T})$ [**17**]. This can be summarized in

$$L^1(\mathbb{T}) \subsetneq M_0(\mathbb{T}) \subsetneq M_c(\mathbb{T}).$$

For a singular measure, it is difficult to have precise knowledge of both its support and its Fourier transform (Heisenberg's uncertainty principle).

Continuous measures are, by definition, characterized by a class of annihilated sets: indeed countable sets; and, thanks to Wiener's lemma, they can be characterized by the behaviour of their Fourier transform. Whence the question: Is there some class of sets $\mathcal{C}$ such that the measures in $M_0(\mathbb{T})$, well-described by their Fourier transform, put no mass on the sets in $\mathcal{C}$? Such a class should be intermediate between the countable sets and the sets of Lebesgue measure zero. Its existence has been discovered by R. Lyons [**19**].

A set E is called a *Weyl set*, if *there exists a non-decreasing sequence* (n_k) *such that, for any* x,

$$\frac{1}{K}\sum_{k\leq K}\delta_{\{n_k x\}} \to \nu_x \textit{ weak}^*$$

where $\nu_x \in M(\mathbb{T})$ *different from* λ $(\{n_k x\} \sim \nu_x \neq \lambda)$. We have the following result.

THEOREM 3. Let $\mu \in M(\mathbb{T})$.

$$\mu \in M_0(\mathbb{T}) \Longleftrightarrow \mu(E) = 0 \ \forall \text{ Weyl set } E.$$

2.3. Normal numbers. We begin with a somewhat algorithmic description of normality with respect to the base q, involving the q-expansion of numbers ([**27**]).

Let q be an integer > 1. A real number $x \in [0,1]$ is normal to the base q (q-normal) if, when $x = 0.x_1x_2\ldots$ is written in base q, every digit $0 \leq d < q$ appears equally often:

$$\lim_N \frac{1}{N}|\{n \leq N,\ x_n = d\}| = \frac{1}{q},$$

every pair of digits d_1d_2 appears equally often:

$$\lim_N \frac{1}{N}|\{n \leq N,\ x_n = d_1, x_{n+1} = d_2\}| = \frac{1}{q^2},$$

and so on, whence the definition:

DEFINITION 7. *The real number* $x \in [0,1]$ *is* q*-normal if each word* $w = d_1d_2\ldots d_r$ *on the alphabet* $\{0,1,\ldots,q-1\}$ *appears in the* q*-adic expansion of* x *with expected frequency, namely:* $\forall r \geq 1$, $\forall w$ *word of length* r,

$$\frac{1}{N}|\{n \leq N,\ x_n x_{n-1}\ldots x_{n+r-1} = w\}| \to \frac{1}{q^r}.$$

By using this definition, explicit q-normal numbers and non-normal numbers can be exhibited ; for example Champernowne's number whose expansion consists in concatenation of all consecutive words to base q is q-normal:

$$0.01234567891011121314151617\ldots$$

when $q = 10$.

Nevertheless, almost nothing is known concerning the question of whether classical arithmetical constants (π, e, $\sqrt{2}$, $\zeta(n), \ldots$) are normal numbers to a fixed base, say $q = 2$; it is unknown whether any irrational algebraic number is normal to any integer base; even weaker assertions are unresolved. For example it is not known whether $\sqrt{2}$ has arbitrarily long blocks of zeros appearing in its binary expansion i.e. ,$\liminf_{n\to\infty}\{2^n\sqrt{2}\} = 0$?

We develop another point of view: the first definition is equivalent to the following condition: if $I \subset \mathbb{T}$ is any q-adic interval $[\frac{a}{q^k}, \frac{a+1}{q^k}[$,

$$\lim_N \sum_{n=1}^{N} \mathbf{1}_I(q^{n-1}x) = \lambda(I),$$

where $\mathbf{1}_I$ is the characteristic function of I. Since any interval can be approximated by subsets and supersets which are finite unions of q-adic intervals, we have equivalently:

DEFINITION 8. *The real number x is normal to base q (q-normal) if the sequence $(q^n x)_n$ is uniformly distributed mod* 1.

By the preceding Weyl's theorem, N_q, the set of q-normal numbers in $[0,1]$, has full measure. It is also a consequence of Birkhoff's ergodic theorem: if we denote by T_q the transformation of $[0,1)$ $x \to qx$ mod 1 (the q-transformation), the Lebesgue measure is invariant under T_q and ergodic (that means $T_q^{-1}A = A \Longrightarrow \lambda(A) =$ 0 ou 1) so that, if $k \neq 0$, $\frac{1}{N}\sum_{n<N} e_k \circ T_q^n \to \int_{\mathbb{T}} e_k \, d\lambda = 0$ λ-almost everywhere ([**4**]).

It follows that the set $N = \cap_q N_q$ of normal numbers to any base has full measure.

2.4. Negligible sets. So what more can be said about N_q^c, N^c? To answer this problem we first need the following definition (see also [**20**]).

DEFINITION 9. *If $\Lambda = (n_k)_k$, $W^*(\Lambda)$ is the set of x such that $(\{n_k x\})_k$ is not uniformly distributed (sometimes called non-normal set)*

(Note that the associated-Weyl set $W(\Lambda) \subset W^*(\Lambda)$.)

So what can we say about $W^*(\Lambda)$?

Classically one uses Hausdorff dimension to compare set of Lebesgue measure zero. We give a brief overview of the underlying theory.

The α-dimensional Hausdorff measure ($\alpha > 0$) is in fact the outer measure in scale x^α. Let E be a subset of $\mathbb{R}$; for $\varepsilon > 0$,

$$H^\alpha_\varepsilon(E) = \inf\{\sum |I_n|^\alpha, \ E \subset \cup I_n\}$$

the infimum being taken over all possible covers (I_n) of E consisting of intervals of diameter $\leq \varepsilon$.

H^α_ε is a non-decreasing function of ε and we put

$$H^\alpha(E) = \lim_{\varepsilon \downarrow 0} H^\alpha_\varepsilon(E).$$

It is easily seen that $H^\alpha(E) = +\infty$ when $\alpha \sim 0$, $H^\alpha(E) = 0$ when $\alpha \sim \infty$ and $\sup\{\alpha \ / \ H^\alpha(E) = +\infty\} = \inf\{\alpha \ / \ H^\alpha(E) = 0\}$; this value is the Hausdorff dimension of E ($\dim_H(E)$).

Note that countable sets have zero Hausdorff dimension, and that

$$0 \leq \dim_H(E) \leq 1, \ \forall E \subset \mathbb{R}.$$

If the set E supports a probability measure μ such that

$$\mu(I) \leq C|I|^s, \ \forall I \text{ interval } \subset \mathbb{R},$$

for some constants $C > 0$, $s \in]0,1]$, then

$$1 = \mu(E) \leq \sum \mu(I_j) \leq C \sum |I_j|^s,$$

and $H_\varepsilon^s(E) \geq \frac{1}{C}$, $H^s(E) \geq \frac{1}{C}$, $\dim_H(E) \geq s$. In this way we get a lower bound for the Hausdorff dimension.

Erdös and Taylor proved the following [**8**]:

Theorem 4. If $\Lambda = (n_k)$ is a lacunary sequence of integers: $\dfrac{n_{k+1}}{n_k} \geq \rho > 1$, then $\dim_H(W^*(\Lambda)) = 1$.

Thus $W^*(q) := W^*((q^k))$ or N_q^c is a "big set" in the sense where $\dim_H(W^*(q)) = 1$.

The size of a negligible set E can also be related to the nature of the (singular) probability measures μ supported on it and to the behavior of their Fourier transform [**28**]:

Proposition 2. If there exist $C > 0, \eta > 0$ such that $|\hat{\mu}(t)| \leq C|t|^{-\eta}$, then $\dim_H(E) \geq 2\eta$.

If $\mu \neq 0$ is a singular probability measure, invariant and ergodic with respect to the q-transform (there are plenty of such measures), then for some $k \neq 0$,

$$\frac{1}{N} \sum_{n<N} e(kq^n x) \to \hat{\mu}(k) \neq 0 \ \mu - ae,$$

and $W^*(q)$ supports a continuous singular measure (by considering a Riesz product). Due to the invariance, this measure does not belong to $M_0(\mathbb{T})$ and it was conjectured that for any probability measure $\mu \in M_0(\mathbb{T})$: "μ-almost every number in $[0, 1)$ is normal to base q" (as for Lebesgue measure).

But in 1986, R. Lyons ([**20**]) proved that this is not the case.

Theorem 5. Suppose that ϕ is a non-increasing function on the non-negative integers such that

1) $\displaystyle\sum_2^\infty \frac{\phi(n)}{n \log n} = \infty$ (with an additional technical condition). Then there exists a probability measure μ concentrated on N_2^c such that $|\hat{\mu}(n)| \leq \phi(n)$.
2) $\displaystyle\sum_2^\infty \frac{\phi(n)}{n \log n} < \infty$. Then $\mu(N_2^c) = 0$ for any positive measure with $|\hat{\mu}(n)| \leq \phi(n)$.

$W^*(2)$ thus supports a probability measure $\mu \in M_0(\mathbb{T})$ with prescribed optimal decay rate.

Actually these two points of view (measure and dimension) are dependent in a non-obvious way: the set of Liouville numbers has zero Hausdorff dimension and supports a probability measure in $M_0(\mathbb{T})$ [**3**], while this is not the case for the triadic Cantor set, which has yet the positive Hausdorff dimension $\log 2/\log 3$.

As an illustration of this, let us mention a striking result, due to Erdös and Salem, attached to Cantor sets E_ξ, $0 < \xi < 1/2$ [**7**].

Theorem 6. E_ξ supports a probability measure $\mu \in M_0(\mathbb{T}) \iff \dfrac{1}{\xi} \notin S$, where S is the set of Pisot numbers.

Turning back to normal numbers, W. Schmidt [**29**] proved the existence of q-normal numbers which are not p-normal if and only if p and q are multiplicatively independent integers. Pollington [**24**] showed that the set of such numbers has full Hausdorff dimension. Later, Brown, Moran & Pierce [**5**] constructed a continuous singular probability measure supported on $N_q \cap N_p^c$. Generalizations have been obtained by Feldman & Smorodinski [9], B. Host [H],.... But none of these measures do belong to $M_0(\mathbf{T})$.

3. The Regular Continued Fraction

3.1. The RCF algorithm. We recall now the classical notations and results for the regular continued fraction algorithm [11].

∙ Given a real number $x \in [0,1)$, x is the limit of the sequence of rational numbers

$$\frac{P_k}{Q_k} = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots \cfrac{1}{a_{k-1} + \cfrac{1}{a_k}}}},$$

shortly $[0; a_1, \ldots, a_k]$. The rationals $\frac{P_k}{Q_k}$ are called the **convergents** of x and $a_1, a_2, \ldots$ the sequence of its **partial quotients** ($a_k \geq 1$ if $k \geq 1$).

So we have

$$P_0 = 0,\ Q_0 = 1,\ P_1 = 1,\ Q_1 = a_1,$$

and for $k \geq 2$

$$\begin{cases} P_k &= a_k P_{k-1} + P_{k-2}, \\ Q_k &= a_k Q_{k-1} + Q_{k-2}, \end{cases}$$

This can be easily established with help of the following matrices

$$A_i = \begin{pmatrix} 0 & 1 \\ 1 & a_i \end{pmatrix},$$

and the identity:

$$M_k = A_k \ldots A_1 = \begin{pmatrix} P_{k-1} & Q_{k-1} \\ P_k & Q_k \end{pmatrix}. \tag{19}$$

From these recurrence relations, it can be seen that P_k and Q_k in fact are polynomials in $a_1, \cdots, a_k$, connected by the relation:

$$P_{k-1}Q_k - Q_{k-1}P_k = (-1)^k.$$

By taking the transpose in (2),

$$Q_k(a_1, \cdots, a_k) = Q_k(a_k, \ldots, a_1),$$

and

$$P_k(a_1, \cdots, a_k) = Q_{k-1}(a_2, \ldots, a_k),$$

whence

$$\frac{Q_{k-1}}{Q_k} = [0; a_k, \ldots, a_1].$$

This means that a pair of consecutive denominators contains the total prior history of the continued fraction expansion.

The convergents are good rational approximations of x and we recall that

$$x - \frac{P_k}{Q_k} = \frac{(-1)^k}{(x_{k+1}Q_k + Q_{k-1})Q_k}$$

where

$$x_{k+1} = [a_{k+1}; a_{k+2}, \ldots],$$

so that

$$|x - \frac{P_k}{Q_k}| \leq \frac{1}{Q_{k+1}Q_k}.$$

⊛ The q-transformation on $\mathbb{T}$ is topologically half-conjugated to the q-shift on $\mathcal{A}^{\mathbb{N}}$ with $\mathcal{A} = \{0, 1, \ldots, q-1\}$. The shift on the RCF expansion is conjugated to an expanding transformation, the Gauss map, defined on $X = [0,1]\backslash\mathbb{Q}$ by $Tx = \dfrac{1}{x}$ mod 1; thus

$$a_1 = [\frac{1}{x}],\ a_{k+1} = a_k(Tx)$$

and

$$T^k x = [0; a_{k+1}, a_{k+2}, \ldots].$$

⊛ The absolutely continuous probability measure μ on X (called the Gauss measure) defined by $\mu(A) = \frac{1}{\log 2}\int_A \frac{dx}{1+x}$ for every Borel-set A, is preserved by T and T-ergodic [**2**].

Applying Birkhoff's ergodic theorem, we get the following result.

Proposition 3. The set BAD of numbers in $[0, 1)$ with bounded partial quotients has zero Lebesgue measure.

The notation BAD comes from another characterization of these numbers: these are the badly approximable numbers [**25**].

Proof: Observe first that the function $a_1 \notin L^1(X, \mu)$:

$$\begin{aligned} \textstyle\int_0^1 [\frac{1}{x}]\ d\mu(x) &= \textstyle\sum_k \int_{\frac{1}{k+1}}^{\frac{1}{k}} \frac{k}{\log 2(1+x)}\ dx \\ &= \textstyle\frac{1}{\log 2}\sum_k k \log \frac{(k+1)^2}{k(k+2)} = \infty \end{aligned}$$

Let be $A > 0$; applying now the ergodic theorem with $f = a_1 \mathbf{1}_{a_1 \leq A} \in L^1(X, \mu)$ we obtain

$$\frac{1}{n}\sum_{j \leq n} a_j \mathbf{1}_{a_j \leq A} \to \int_{a_1 \leq A} a_1\ d\mu = \alpha(A) \quad \mu - a.e$$

so that, for every $A > 0$,

$$\liminf_n \frac{1}{n}\sum_{j\leq n} a_j \geq \liminf_n \frac{1}{n}\sum_{j\leq n} a_j \mathbf{1}_{a_j \leq A} = \alpha(A) \quad \mu - a.e$$

and the results follows, letting $A \to \infty$. □

However, in terms of dimension, BAD is maximal: a result of Jarnik [**13**] states that $\dim_H BAD = 1$ as we shall see below.

3.2. Sets $F(\mathcal{A})$. Mahler and many others considered the sets $F(N)$ of real numbers in $[0, 1)$ with partial quotients bounded by $N \geq 2$:

$$F(N) = \{x \in [0, 1);\ x = [0; a_1, a_2, \ldots] \text{ with } a_i \leq N\ \forall i \geq 1\}$$

More generally, if $\mathcal{A}$ is a finite alphabet of integers ≥ 1, $|\mathcal{A}| \geq 2$,

$$F(\mathcal{A}) = \{x \in [0, 1);\ x = [0; a_1, a_2, ...] \text{ with } a_i \in \mathcal{A}\ \forall i \geq 1\}$$

All these sets are zero Lebesgue measure Cantor-type sets with positive Hausdorff dimension. The first estimates of $d(N) = \dim_H(F(N))$ are due to Jarnik: for $N > 8$

$$1 - \frac{4}{N\log 2} \leq d(N) \leq 1 - \frac{8}{N \log N}.$$

In 1941 Good [**10**] proved that α_m, the solution of the equation

$$\sum_{a_1,\ldots,a_m \in \mathcal{A}} Q_m(a_1, \ldots, a_m)^{-2\alpha} = 1,$$

tends to $\dim_H(F(\mathcal{A}))$ as $m \to \infty$, which leads to better estimations.

Recently, a joint work of Jenkinson and Pollicott [**14**] yields an algorithm to compute these dimensions: it is interesting to note that

$$\dim_H(F(2)) = 0.531280\ldots \quad \dim_H(F(3)) = 0.68\ldots$$

while

$$\dim_H(F(\mathcal{A})) < 1/2 \;\text{ if } \mathcal{A} = \{1,4\}.$$

3.3. A question of Montgomery. Montgomery [**22**] in his book proposes the following classical problem: Find a normal number whose partial quotients are bounded.

In 1980, R. Kaufman [**18**], using the structure of the sets $F(N)$, obtained the following deep result:

THEOREM 7. Suppose that $F(N)$ has Hausdorff dimension $> 2/3$.Then $F(N)$ carries a probability measure μ such that $|\hat{\mu}(t)| \leq c|t|^{-\eta}$ for a certain $\eta > 0$.

He shows in his construction that one may take $\eta = 0.0007$.

When Kaufman's paper appeared, R.C. Baker [**1**] observed that combining this with Jarnik's computations and Davenport-Erdós-Leveque's theorem gives the existence of a normal number in BAD. Such a number can be taken from $F(3)$ since $F(N)$ fulfills Kaufman's hypothesis for $N \geq 3$; but $F(2)$ does not.

Actually, Kaufman's estimations can be improved to give [**26**]:

THEOREM 8. Let $\mathcal{A}$ be a finite alphabet of integers ≥ 1, $|\mathcal{A}| \geq 2$. Suppose that $\dim_H(F(\mathcal{A})) > 1/2$. Then for all $\varepsilon > 0$ and $1/2 < \delta < \dim_H(F(\mathcal{A}))$, we are able to construct a probability measure $\mu = \mu_{\varepsilon,\delta}$, supported by $F(\mathcal{A})$ such that

(1) $\mu(I) \leq C_1|I|^{\delta}$, $\forall I$ interval $\subset [0,1)$;

(2) $|\hat{\mu}(t)| \leq C_2(1+|t|)^{8\varepsilon + \frac{\delta(1-2\delta)}{(1+2\delta)(4-\delta)}}$, $\forall t > 0$.

As a consequence there exist infinitely many normal numbers with partial quotients $\in \{1,2\}$; but the question remains open for alphabets yielding low Hausdorff dimension such that $\{1,4\}$ for example.

Sketch of Proof: Suppose $\mathcal{A} \subset [1,\ldots,N]$.

1. We first observe that the set $F(\mathcal{A})$ may be endowed with two topological structures: the metric structure of $[0,1]$ and the one of $\mathcal{A}^{\mathbb{N}^*}$, which can easily be compared on $F(\mathcal{A})$:

a) If $x, y \in F(\mathcal{A})$ are such that $a_j(x) = a_j(y)$, $1 \leq j \leq k$ and $a_{k+1}(x) \neq a_{k+1}(y)$, then

$$|x-y| \leq \frac{N^2}{Q_{k+1}^2(x)}.$$

b) If t and $t+h \in [0,1)$ and $0 < h < \frac{1}{N+2}$ then the interval $[t, t+h] \subset [a_1,\ldots,a_\ell]$, the cylinder $\{x \in [0,1);\ a_j(x) = a_j,\ 1 \leq j \leq \ell\}$, where ℓ is such that $Q_\ell(a_1,\ldots,a_\ell) \geq \frac{1}{(N+2)h^{1/2}}$.

2. After Kaufman, we consider $\mathcal{A}^{\mathbb{N}^*}$ as the product of infinitely many blocks $\mathcal{A}^{J_0}$, J_0 to be specified, ν some discrete measure on $\mathcal{A}^{J_0}$ and $\mu = \nu \times \nu \times \ldots$ in such a way the quantity $\log Q_J$ to be equal $a.e. - \mu$ to its mean value $E_\mu(\log Q_J) =: \log Q$, for J multiple of J_0.

The condition (1) in the theorem can be seen to be satisfied.

3. It remains to evaluate

$$\hat{\mu}(u) = \int_0^1 e(ut)\, d\mu(t), \quad u > 0.$$

Put $J = kJ_0$ and decompose any x into

$$x = [0; a_1(x), \dots, a_J(x) + t] = \frac{P_J(x) + tP_{J-1}(x)}{Q_J(x) + tQ_{J-1}(x)};$$

μ in turn can be decomposed into $\rho_k \times \mu$ with $\rho_k = \underbrace{\nu \times \dots \times \nu}_{k}$ so that

$$\hat{\mu}(u) = \int_0^1 (\int e(u\frac{P_J(x) + tP_{J-1}(x)}{Q_J(x) + tQ_{J-1}(x)})d\rho_k(x))d\mu(t),$$

that we write

$$\hat{\mu}(u) = \int_0^1 F(t)d\mu(t)$$

with $F(t) = \int e(u\frac{P_J + tP_{J-1}}{Q_J + tQ_{J-1}})\, d\rho_k$.

A lemma, due to Kaufman, allows us to compare the μ-integral of F with the λ-one:

Lemma 1. Let F be a function C^1 on $[0,1]$ such that $|F(t)| \leq 1$ and $|F'(t)| \leq M$. Denote by $m_2 = \int_0^1 |F(t)|^2 dt$. Let now λ be a probability measure on $[0,1]$ and denote by $\Lambda(u)$ the maximum of $\lambda[t, t+u]$ for $t \in [0, 1-u]$. Then we have

$$\int_0^1 |F(t)|\, d\lambda \leq 2r + \Lambda(r/M)(1 + m_2 M r^{-3}),$$

for any $r > 0$.

So we are led now to estimate the following

$$\int_0^1 |F(t)|^2\, dt = \int \int (\int e(u\frac{P_J + tP_{J-1}}{Q_J + tQ_{J-1}} - u\frac{P_J + tP_{J-1}}{Q_J + tQ_{J-1}})\, dt)\, d\rho_k d\rho_k$$

which contains an oscillating integral of the form $\int_0^1 e(f(t))\, dt$.

We plan to apply to it classical lemmas involving the behavior of f', by distinguishing two cases:

Lemma 2. If f is C^2 on $[0,1]$, satisfying $|f'(t)| \geq a$ and $|f''(t)| \leq b$, then we have

$$\left|\int_0^1 e(f(t))dt\right| \leq \frac{1}{a} + \frac{b}{a^2}.$$

Lemma 3. If f is C^2 on $[0,1]$ and $f'(t) = (\alpha t + \beta)g(t)$ where g satisfies $|g(t)| \geq a$ and $|g'(t)| \leq b$ with $b \geq a$, then we have

$$\left|\int_0^1 e(f(t))dt\right| \leq 6\frac{b}{a^{3/2}|\alpha|^{1/2}}.$$

4. Given $\varepsilon > 0$ we choose J_0, ν and μ as in the first step. Now we fix $u > 0$; by using the three lemmas we get an estimation of $|\hat{\mu}(u)|$ in terms of r, Q and u. Finally we choose k (or J) great enough to optimize this quantity in parameters Q and r. This gives (2). ◇◇◇

References

[1] BAKER R.C.: Non publié, cf référence **[22]** ci-dessous, problème numéro 45.

[2] BILLINGSLEY P.: *Ergodic Theory and Information*, New York: Wiley (1964).

[3] BLUHM C.: *Liouville numbers, Rajchman measures and small Cantor sets*, Proc. A.M.S., (2000) **128**, 2637–2640.

[4] BOREL E.: *Les probabilités dénombrables et leurs applications arithmétiques*, Rend. Circ. Mat. Palermo, (1909) **27**, 247–271.

[5] BROWN G., MORAN W., PEARCE C.: *Riesz products and normal numbers*, J. London Math. Soc., (1985) **32**, 12–18.

[6] BUGEAUD Y.: *Nombres de Liouville et nombres normaux*, C. R. Acad. Sci. Paris (2002) **335** , 117–120.

[7] ERDÖS P.: *On a family of symmetric Bernoulli convolutions*, Amer. J. Math., (1939) **61**.

[8] ERDÖS P., TAYLOR S.J.: *On the set of points of convergence of a lacunary trigonometrics series and the equidistribution properties of related sequences*, Proc. Lond. Math. Soc., (1957) **7**, 598–615.

[9] FELDMAN J., SMORODINSKY M.: *Normal numbers from independent processes*, Ergod. Th. Dynam. Sys., (1992) **12**, 707–712.

[10] GOOD I. J.: *The fractional dimensional theory of continued fractions*, Proc. Cambridge Phil. Soc. (1941) **37**, 199–228.

[11] HARDY G.H., WRIGHT E.M.: *An introduction to the theory of numbers* Clarendon Press, Oxford Univ. Press, 1979.

[12] HOST B.: *Nombres normaux, entropie, translations*, Israel J. Math., (1995) **91**, 419–428.

[13] JARNIK I.: *Zur metrischen Theorie der diophantischen Approximationen*, Proc. Math. Fyz. (1928) **36**, 91–106.

[14] JENKINSON O., POLLICOTT M.: *Computing the dimension of dynamically defined sets: E_2 and bounded continued fractions*, Ergod. Th. Dynam. Sys., (2001) **21**, 1429–1445.

[15] KAHANE J.P., SALEM. R.: *Distribution modulo* 1 *and sets of uniqueness*, Bull. Am. Math. Soc., (1964) **70**, 259–261.

[16] KAHANE J.P., SALEM. R.: *Ensembles parfaits et séries trigonométriques*, Hermann, nouvelle édition 1994.

[17] KATZNELSON Y.: *An introduction to harmonic analysis*, Dover, 1976.

[18] KAUFMAN R.: *Continued fractions and Fourier transforms*, Mathematika (1980) **27**, 262–267.

[19] LYONS R.: *Fourier-Stieltjes coefficients and asymptotic distribution modulo 1*, Ann. math., (1985) **122**, 155–170.

[20] LYONS R.: *The measure of non-normal sets*, Invent. math., (1986) **83**, 605–616.

[21] MENSOV D.E.: *Sur l'unicité du développement trigonométrique*, C. R. Acad. Sci. Paris (1916) **163**, 433–436.

[22] MONTGOMERY H. L.: *Ten lectures on the interface between Analytic Number Theory and Harmonic Analysis*, CBMS regional conf. series in Math. 84, A.M.S., Providence RI, 1994.

[23] MORAN W., POLLINGTON P.: *The discrimination theorem for normality to non-integer bases*, Israel J. Math. (1997) **100**, 339–347.

[24] POLLINGTON P.: *On the density of sequences* $\{n_k\xi\}$, Illinois J. Math. (1979) **23**, 511–515.

[25] POLLINGTON P., VELANI S.: *On a problem in simultaneous Diophantine approximation: Littlewood's conjecture*, Acta Math. (2000) **185**, 287–306.

[26] QUEFFELEC M., RAMARE O.: *Analyse de Fourier des fractions continues à quotients restreints*, l'Enseignement Math. (2003) (to appear).

[27] RAUZY G.: *Propriétés statistiques de suites arithmétiques*, Presses Universitaires de France (1976).

[28] SALEM R.: *Sets of uniqueness and sets of multiplicity*, Trans. Amer. Math. Soc. (1943) **54**, (1944) **56**, (1948) **63**.

[29] SCHMIDT W.: *On normal numbers*, Pacific J. Math. (1960) **10**, 661–672.

[30] WEYL H.: *Über die Gleichverteilung von Zahlen mod. Eins*, Math. Ann. (1916) **77**, 313–352.

Arithmetic coding of geodesics on the modular surface via continued fractions

Svetlana Katok and Ilie Ugarcovici

Department of Mathematics
The Pennsylvania State University
University Park, PA 16802
U.S.A.
katok_s@math.psu.edu, idu@math.psu.edu

Abstract. In this article we present three arithmetic methods for coding oriented geodesics on the modular surface using various continued fraction expansions and show that the space of admissible coding sequences for each coding is a one-step topological Markov chain with countable alphabet. We also present conditions under which these arithmetic codes coincide with the geometric code obtained by recording oriented excursions into the cusp of the modular surface.

Introduction

Let $\mathcal{H} = \{z = x + iy : y > 0\}$ be the upper half-plane endowed with the hyperbolic metric, $F = \{z \in \mathcal{H} : |z| \geq 1,\ |\mathrm{Re}\, z| \leq \frac{1}{2}\}$ be the standard fundamental region for the modular group $PSL(2,\mathbb{Z}) = SL(2,\mathbb{Z})/\{\pm I\}$, and $M = PSL(2,\mathbb{Z})\backslash\mathcal{H}$ be the modular surface which topologically is a sphere with one puncture (the cusp) and two singularities (fixed points of elliptic elements). Let $S\mathcal{H}$ denote the unit tangent bundle of $\mathcal{H}$. Then the quotient space $PSL(2,\mathbb{Z})\backslash S\mathcal{H}$ can be identified with the unit tangent bundle of M, SM, although the structure of the fibered bundle has singularities at the elliptic fixed points (see [**K1**, §3.6] for details). Let $\pi : S\mathcal{H} \to SM$ be the projection of the unit tangent bundles. In all our considerations, we assume implicitly that an oriented geodesic on M is endowed with a unit tangent (direction) vector at each point and therefore is an orbit of the geodesic flow $\{\varphi^t\}$ on M, which is defined as an $\mathbb{R}$-action on the unit tangent bundle SM (see e.g. [**KH**, §5.3, 5.4]). For an oriented geodesic γ on M, its lift to $\mathcal{H}$ is any oriented geodesic γ' on $\mathcal{H}$ such that $\pi(\gamma') = \gamma$.

In this article we will consider only oriented geodesics which do not go to the cusp of M in either direction. The corresponding geodesics in F contain no vertical segments, and both end points of all their lifts to $\mathcal{H}$ are irrational. In what follows, when we say "every oriented geodesic", we refer to every geodesic from this set. The set of excluded geodesics is insignificant from the measure-theoretic point of view, more precisely, the set of vectors tangent to the excluded geodesics $E \subset SM$ is invariant under the geodesic flow $\{\varphi^t\}$ and $\mu(E) = 0$ for any Borel probability measure μ invariant under $\{\varphi^t\}$. This can be seen from the decomposition of this set $E = E^+ \cup E^-$, so that $\varphi^t(E^+)$ (respectively, $\varphi^{-t}(E^-)$) escape to the cusp as $t \to +\infty$. For any compact $K \subset E^\pm$ there exists $T > 0$ such that $K \cap \varphi^{\pm t}(K) = \emptyset$ for any $t > T$. $\mu(E^\pm) > 0$ would then contradict the Poincaré Recurrence Theorem (see [**KH**, §4.1]).

Oriented geodesics on the modular surface $M = PSL(2,\mathbb{Z})\backslash\mathcal{H}$ can be symbolically coded in two different ways. The geometric code with respect to the fundamental region F is obtained by recording the successive sides of F cut by the geodesic, and can be presented by a bi-infinite sequence of non-zero integers by assigning an integer, positive or negative, depending on the orientation, to each excursion to the cusp. Another method (we call it arithmetic) is to use the boundary expansions of the end points of the geodesic at infinity and a certain "reduction theory". This method was first introduced by Artin [**Ar**] for the modular group who used simple continued fractions for the boundary expansions to prove the existence of a dense geodesic on M. Artin's method was used by Hedlund [**Hed**] to prove ergodicity of the geodesic flow on M. If applied literally, this method gives a $GL(2,\mathbb{Z})$-invariant code, but it does not classify geodesics on the modular surface. Artin's method has been modified by Series in [**S1**] to eliminate this problem, and further developed in [**BS, S2, S3**] for other Fuchsian groups. Related work on coding geodesics can be also found in [**AF1, AF2, AF3, Arn, GL, S4**].

In this article we give a unified approach for construction of arithmetic codes for geodesics on the modular surface using *generalized minus continued fractions.* Any irrational number x can be expressed uniquely in the form

$$x = n_0 - \cfrac{1}{n_1 - \cfrac{1}{n_2 - \cfrac{1}{\ddots}}}$$

which we will denote by $x = (n_0, n_1, \cdots)$ for short. The "digits" n_i are non–zero integers determined recursively by $n_{i+1} = (x_{i+1})$, $x_{i+1} = -\frac{1}{x_i - n_i}$, starting with $n_0 = (x)$ and $x_1 = -\frac{1}{x-n_0}$, where $(\cdot)$ is a certain integer-valued function. The function $x \mapsto \lceil x \rceil = \lfloor x \rfloor + 1$ (where $\lfloor x \rfloor$ is the integer part of x, or the floor function, i.e. the largest integer $\leq x$) gives the minus continued fraction expansion first used for the arithmetic code in [**K2, GK**], although the notations in the present paper are different from [**K2, GK**] where only one arithmetic code was studied. (Notice that $\lceil x \rceil$ is the smallest integer greater than x, and differs at integers from the commonly used ceiling function.) This coding procedure for closed geodesics is exactly the Gauss reduction theory for indefinite integral quadratic forms translated into matrix language [**K2**], therefore we will refer to the above code as the *Gauss arithmetic code (G-code).* We review it in Section 1. Using appropriate functions $(\cdot)$ we reinterpret the classical Artin code *(A-code)* in these terms in Section 2 and describe an arithmetic code based on the nearest integer continued fraction expansions of the end points in Section 3. The latter expansions were developed and used by Hurwitz [**H**] in order to establish a reduction theory for indefinite real quadratic forms, therefore we call the third code *Hurwitz arithmetic code (H-code).*

All three coding procedures are actually reduction algorithms which may be considered as generalized reduction theories for real indefinite quadratic forms translated into matrix language. Although they follow the same general scheme, the notion of reduced geodesic is different in each case, and so are the estimates in Theorems 1.1, 2.1, and 3.1.

The most elegant of the three codings is the Gauss arithmetic code obtained in [**K2, GK**] using minus continued fraction expansions of the end points, and interpreted in [**GK**] via a particular "cross-section" of SM. The set of such arithmetic coding sequences was identified in [**GK**]: it is a symbolic Bernoulli system on the infinite alphabet $\mathcal{N} = \{n \in \mathbb{Z}, n \geq 2\}$, i.e. it consists of all bi-infinite sequences constructed with symbols of the alphabet $\mathcal{N}$. We give similar interpretations for the Artin and the Hurwitz codes, and show that the space of admissible sequences for each code is given by a set of simple rules which can be described with the help of a transition matrix of zeros and ones, and constitutes a one-step topological Markov chain with countable alphabet. An explicit canonical Markov partition of the corresponding cross-section is presented for each arithmetic code. Symbolic representation of the geodesic flow on M as a special flow for each code is given in Section 4.

In contrast, the set of admissible geometric coding sequences is quite complicated, and, as has been proved in [**KU**], is not a finite-step topological Markov chain (see [**KH**, §1.9] for exact definitions). Therefore, there are geodesics whose geometric code differs from any arithmetic code. It is worth noting that the H-code comes closest to the geometric code: we show that for the class of geometrically Markov geodesics—identified in [**KU**] as the maximal one-step topological Markov chain in the set of all admissible geometric codes—the H-code coincides with the geometric code.

Acknowledgments. We thank David Fried for bringing to our attention the paper by Hurwitz [**H**], and the referee for helpful suggestions which improved the presentation of this article. The second author acknowledges summer support from NSF grant DMS-9704776.

1. Minus continued fraction coding (Gauss coding)

In this section we review the arithmetic coding procedure for geodesics on the modular surface, using minus (or, backward) continued fraction expansions which we call here *G-expansions.* Every real number α has a unique G-expansion $\alpha = \lceil n_0, n_1, n_2, \dots \rceil$ with $n_0 \in \mathbb{Z}$ and $n_1, n_2, \dots \geq 2$, by setting $n_0 = \lceil \alpha \rceil$ (the smallest integer greater than α), $\alpha_1 = -\frac{1}{\alpha - n_0}$, and, inductively,

$$n_i = \lceil \alpha_i \rceil \quad , \quad \alpha_{i+1} = -\frac{1}{\alpha_i - n_i} \, .$$

Conversely, any infinite sequence of integers $n_0, n_1, n_2, \dots$ with $n_i \geq 2$ for $i \geq 1$ defines a real number whose G-expansion is $\lceil n_0, n_1, n_2, \dots \rceil$. The following properties are satisfied (see [**Z, K2**], and [**K3**] for the proofs):

(G1) α is rational if and only if the tail of its G-expansion consists only of 2's, i.e., there exists a positive integer l such that $n_k = 2$ for all $k \geq l$;

(G2) α is a quadratic irrationality, i.e. a root of a quadratic polynomial with integer coefficients, if and only if its G-expansion is eventually periodic, $\alpha = \lceil n_0, n_1, \dots, n_k, \overline{n_{k+1}, \dots, n_{k+m}} \rceil$ (with the periodic part being anything but a tail of 2's);

(G3) A quadratic irrationality α has a purely periodic G-expansion if and only if $\alpha > 1$ and $\alpha' \in (0,1)$, where α' is conjugate to α, i.e. α' and α are roots of the same quadratic polynomial with integer coefficients;

(G4) If $\alpha = \lceil \overline{n_1, \dots, n_k} \rceil$, then $1/\alpha' = \lceil \overline{n_k, \dots, n_1} \rceil$;

(G5) Two irrationals α, β are $PSL(2,\mathbb{Z})$-equivalent if and only if their G-expansions have the same tail, that is $\alpha = \lceil n_0, n_1, \dots \rceil$ and $\beta = \lceil m_0, m_1, \dots \rceil$ with $n_{i+k} = m_{i+l}$ for some integers k, l and all $i \geq 0$.

From the theory of G-expansions (see [**K3**]), we have that if $\alpha = \lceil n_0, n_1, \dots \rceil$, then the convergents $r_k = \lceil n_0, n_1, \dots, n_k \rceil$ can be written as p_k/q_k where p_k and q_k are obtained inductively as:

$$\begin{aligned} p_{-2} &= 0 \, , \; p_{-1} = 1 \, ; \; p_k = n_k p_{k-1} - p_{k-2} \quad \text{for } k \geq 0 \\ q_{-2} &= -1 \, , \; q_{-1} = 0 \, ; \; q_k = n_k q_{k-1} - q_{k-2} \quad \text{for } k \geq 0 \, . \end{aligned}$$

Proposition 4. The following properties are satisfied:

(i) $1 = q_0 < q_1 < q_2 < \dots$;

(ii) $p_{k-1} q_k - p_k q_{k-1} = 1$, for all $k \geq 0$;

(iii) Let $T(z) = z+1$, $S(z) = -1/z$ be the generating transformations for $PSL(2,\mathbb{Z})$, then for any $z \in \mathcal{H} \cup \mathbb{R} \cup \{\infty\}$

$$T^{n_0}ST^{n_1}S \dots T^{n_k}S(z) = \lceil n_0, n_1, \dots, n_k, z \rceil = \frac{p_k z - p_{k-1}}{q_k z - q_{k-1}}\,;$$

(iv) The sequence $\{r_k\}$ is monotone decreasing, converges to α and

$$p_k/q_k - \alpha \leq 1/q_k\,;$$

(v) If α is irrational, then there is a sequence of denominators $\{q_{k_j}\}$ such that $\dfrac{q_{k_j}}{q_{k_{j-1}}} > 2$.

Proof. Proofs of the properties (i)–(iv) can be found in [**K3**]. We give a proof of (v) here. Since α is irrational, its G-expansion contains infinitely many entries strictly greater than 2, hence we can find a sequence k_j, such that $n_{k_j} \geq 3$. But this implies that

$$q_{k_j} = n_{k_j} q_{k_j - 1} - q_{k_j - 2} > 3q_{k_j - 1} - q_{k_j - 1} = 2q_{k_j - 1}.$$

Using (i) we obtain

$$\frac{q_{k_j}}{q_{k_{j-1}}} \geq \frac{q_{k_j}}{q_{k_j - 1}} > 2.$$

□

Definition 1. An oriented geodesic on $\mathcal{H}$ is called *G-reduced* if its repelling and attracting end points, denoted by u and w, respectively, satisfy $0 < u < 1$ and $w > 1$.

To a G-reduced geodesic γ, one associates a bi-infinite sequence of positive integers $\lceil\gamma\rceil = \lceil \dots, n_{-2}, n_{-1}, n_0, n_1, n_2, \dots \rceil$, called its *G-code*, by juxtaposing the G-expansions of $1/u = \lceil n_{-1}, n_{-2}, \dots \rceil$ and $w = \lceil n_0, n_1, n_2, \dots \rceil$.

Reduction algorithm. We present the procedure of reducing any geodesic to a G-reduced one. This will help us extend the symbolic coding to all geodesics on $\mathcal{H}$.

Theorem 1.1. Every oriented geodesic on $\mathcal{H}$ is $PSL(2,\mathbb{Z})$-equivalent to a G-reduced geodesic.

Proof. Let γ be an arbitrary geodesic on $\mathcal{H}$ with irrational end points u and w, and $\lceil n_0, n_1, n_2, \dots \rceil$ be the G-expansion of w. We construct the following sequence of real pairs $\{(u_k, w_k)\}$ ($k \geq 0$) defined by $u_0 = u$, $w_0 = w$ and:

$$w_{k+1} = ST^{-n_k} \dots ST^{-n_1}ST^{-n_0}w\,, \quad u_{k+1} = ST^{-n_k} \dots ST^{-n_1}ST^{-n_0}u\,.$$

Since w is irrational, $w_{k+1} = \lceil n_{k+1}, n_{k+2}, \dots \rceil > 1$. By Proposition 4 (iii),

$$w = T^{n_0}ST^{n_1}S \dots T^{n_k}S(w_{k+1}) = \frac{p_k w_{k+1} - p_{k-1}}{q_k w_{k+1} - q_{k-1}}$$

$$u = T^{n_0}ST^{n_1}S \dots T^{n_k}S(u_{k+1}) = \frac{p_k u_{k+1} - p_{k-1}}{q_k u_{k+1} - q_{k-1}}\,,$$

hence

$$u_{k+1} = \frac{q_{k-1}u - p_{k-1}}{q_k u - p_k} = \frac{q_{k-1}}{q_k} + \frac{1}{q_k^2(p_k/q_k - u)} = \frac{q_{k-1}}{q_k} + \varepsilon_k \tag{1.1}$$

where $\varepsilon_k \to 0$. Moreover, using property (iv), we have $p_k/q_k \searrow w$, hence, for large enough k, $|p_k/q_k - u| > \frac{1}{2}|w-u|$ and

$$|\varepsilon_k| = \frac{1}{q_k^2|p_k/q_k - u|} < \frac{2}{q_k^2|w-u|}\,.$$

Property (i) and the previous relation imply that, for large enough k, $|\varepsilon_k| < 1/q_k$ and

$$0 < \frac{q_{k-1}}{q_k} - \frac{1}{q_k} < u_{k+1} = \frac{q_{k-1}}{q_k} + \varepsilon_k < \frac{q_{k-1}}{q_k} + \frac{1}{q_k} \leq 1\,.$$

Therefore, we can find a positive integer l such that $0 < u_{l+1} < 1$. The geodesic with end points u_{l+1} and w_{l+1} is G-reduced and $PSL(2,\mathbb{Z})$-equivalent to γ. □

Remark 1. (i) The proof of Theorem 1.1 gives also the algorithm for G-reducing a geodesic γ: one has to construct the sequence $\{(u_k, w_k)\}$ inductively until $0 < u_k < 1$; (ii) any further application of the reduction algorithm to a reduced geodesic yields reduced geodesics whose G-codes are left shifts of the G-code of the first reduced one.

Now we associate to any oriented geodesic γ on $\mathcal{H}$ the G-code of a reduced geodesic $PSL(2,\mathbb{Z})$-equivalent to γ, e.g. obtained by the reduction algorithm described in the proof of Theorem 1.1. Since our goal is to define a symbolic coding for the geodesic flow on M, we need to show that this code is $PSL(2,\mathbb{Z})$-invariant, i.e. that two oriented geodesics on $\mathcal{H}$ are $PSL(2,\mathbb{Z})$-equivalent if and only if their G-codes coincide up to a shift. We are going to present a geometric proof of this fact in Corollary 1, by constructing a cross-section of the geodesic flow on M, directly related to the notion of G-reduced geodesics.

Construction of the cross-section. A cross-section for the geodesic flow is a subset of the unit tangent bundle SM which each geodesic (maybe with some exceptions) visits infinitely often both in the future and in the past. We construct a cross-section C_G for the geodesic flow on M, such that successive returns of a geodesic γ to C_G correspond to left-shifts in the G-code of γ. We define $C_G = P \cup Q$ to be a subset of SM, where P consists of all tangent vectors with base points in the circular side of F and pointing inward such that the corresponding geodesic on $\mathcal{H}$ is G-reduced, i.e. $0 < u < 1$ and $w > 1$ and Q consists of all tangent vectors with base points on the right vertical side of F pointing inwards, such that if γ is the corresponding geodesic, then $TS(\gamma)$ is G-reduced (Figure 8). This is a clarification of the definition given in [**GK**]. Notice that $C_G = \pi(C_g)$ where C_g is the set all unit tangent vectors with base points on the unit semi-circle $|z| = 1$ and pointing outward such that the associated geodesic on $\mathcal{H}$ is G-reduced.

Theorem 1.2. *C_G is a cross-section for the geodesic flow on M.*

Proof. Let γ be an oriented geodesic on M. It is presented as a bi-infinite sequence of $PSL(2,\mathbb{Z})$-equivalent geodesic segments on F. Any segment, extended to a geodesic on $\mathcal{H}$, can be reduced according to Theorem 1.1. Thus, there exists a G-reduced geodesic γ' on $\mathcal{H}$ such that $\pi(\gamma') = \gamma$. Notice that γ' intersects the right half of the unit semicircle $|z| = 1$ in such a way that the unit tangent vector of γ' at the intersection point belongs to the set C_g. Therefore either $\gamma' \cap F$ or $ST^{-1}(\gamma') \cap F$ is one of the segments of γ on F. In either case γ intersects C_G at least once. Denote this intersection point by $\mathbf{x}_0 \in C_G \subset SM$, and let us follow the geodesic on M from this starting point. If $\mathbf{x}_0 \in P$, then the corresponding geodesic γ' on $\mathcal{H}$ from u to w is G-reduced. In order to prove that γ intersects C_G again, it is enough to notice that γ' intersects the left-half of the semi-circle $|z - n_0| = 1$,

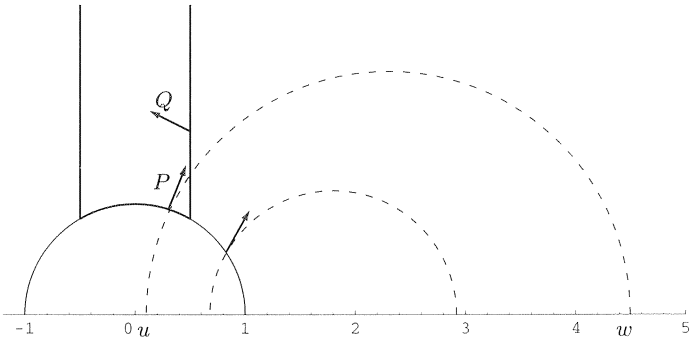


FIGURE 8. The cross-section $C_G = P \cup Q$

where $n_0 = \lceil w \rceil$, such that for the unit tangent vector at the intersection point $\mathbf{x}_1' \in S\mathcal{H}$, we have $ST^{-n_0}(\mathbf{x}_1') \in C_g$. Hence $\mathbf{x}_1 = \pi(\mathbf{x}_1') \in C_G$, and γ intersects C_G at $\mathbf{x}_1$. Moreover, $\mathbf{x}_1 \in C_G$ is the next intersection point of γ with C_G after $\mathbf{x}_0$. One obtains a similar property in the case where $\mathbf{x}_0 \in Q$, by studying the G-reduced geodesic γ' on $\mathcal{H}$ corresponding to $TS(\mathbf{x}_0)$. □

Every oriented geodesic γ on M can be represented as a bi-infinite sequence of segments σ_i between successive returns to C_G. To each segment σ_i we associate the corresponding G-reduced geodesic γ_i on $\mathcal{H}$. Thus we obtain a sequence of reduced geodesics $\{\gamma_i\}_{i=-\infty}^{\infty}$ representing the geodesic γ. If one associates to γ_i its G-code, $\lceil \gamma_i \rceil = \lceil \dots, n_{-2}, n_{-1}, n_0, n_1, n_2, \dots \rceil$ then $\gamma_{i+1} = ST^{-n_0}(\gamma_i)$ and the coding sequence is shifted one symbol to the left. Thus all G-reduced geodesics γ_i in the sequence produce the same, up to a shift, bi-infinite coding sequence, which we call the *G-code* of γ and denote by $\lceil \gamma \rceil$. The following Corollary shows that the G-code is well-defined.

COROLLARY 1. The G-code is $PSL(2,\mathbb{Z})$-invariant, i.e. two geodesics γ, γ' on $\mathcal{H}$ are $PSL(2,\mathbb{Z})$-equivalent if and only if for some integer l and all integers i one has $n_i' = n_{i+l}$, where $\lceil \gamma \rceil = \lceil n_i \rceil_{i=-\infty}^{\infty}$ and $\lceil \gamma' \rceil = \lceil n_i' \rceil_{i=-\infty}^{\infty}$.

PROOF. Let γ, γ' be $PSL(2,\mathbb{Z})$-equivalent. Then $\pi(\gamma) = \pi(\gamma')$ is the same oriented geodesic on M. By choosing the same starting point $\mathbf{x}_0$, one obtains the same bi-infinite sequence of segments σ_i between successive returns to C_G and hence the same G-code up to a left shift. Conversely, a left shift of a G-code corresponds to an application of ST^{-n_0} to the end points of the geodesic, i.e., it produces a $PSL(2,\mathbb{Z})$-equivalent reduced geodesic. □

EXAMPLE 16. Let γ be a geodesic on $\mathcal{H}$ from $u = \sqrt{5}$ to $w = -\sqrt{3}$. The G-expansions are

$$w = \lceil -1, 2, \overline{2, 3} \rceil, \quad 1/u = \lceil 1, \overline{2, 6, 2, 2} \rceil .$$

First, we need to find an equivalent G-reduced geodesic. For this we use the algorithm described in the proof of Theorem 1.1 to construct the sequence (u_1, w_1), $(u_2, w_2), \dots$, until we obtain a G-reduced pair equivalent to (u, w). We have

$$w_1 = ST(w) = (1+\sqrt{3})/2, \quad u_1 = ST(u) = (1-\sqrt{5})/4,$$
$$w_2 = ST^{-2}(w_1) = 1 + 1/\sqrt{3}, \; u_2 = ST^{-2}(u_1) = (7-\sqrt{5})/11$$

and the pair (u_2, w_2) is already G-reduced. The minus continued fraction expansions of $1/u_2$ and w_2 are

$$w_2 = \lceil \overline{2,3} \rceil\,, \quad 1/u_2 = \lceil 3, \overline{2,2,6,2} \rceil\,,$$

hence $\lceil \gamma \rceil = \lceil \overline{2,6,2,2}, 3, \overline{2,3} \rceil = \lceil \dots, 2,2,6,2,2,2,6,2,2,3,2,3,\dots \rceil$. This corrects a misprint in the Example of [**GK**].

Let $\mathcal{N}_G^{\mathbb{Z}}$ be the Bernoulli space on the infinite alphabet $\mathcal{N}_G = \{n \in \mathbb{Z}, n \geq 2\}$. We proved that each oriented geodesic which does not go to the cusp of M in either direction corresponds to its G-code, $\lceil \gamma \rceil \in \mathcal{N}_G^{\mathbb{Z}}$. Conversely, each bi-infinite sequence $x \in \mathcal{N}_G^{\mathbb{Z}}$ which does not have an infinite tail of 2's in either direction produces a geodesic on $\mathcal{H}$ from $u(x)$ to $w(x)$ (irrational end points), where

$$w(x) = \lceil n_0, n_1, \dots \rceil \quad , \quad \frac{1}{u(x)} = \lceil n_{-1}, n_{-2}, \dots \rceil\,.$$

This correspondence will extend to all oriented geodesics on M if we extend the notion of G-reduced geodesic to those with $0 < u < 1$ and $w \geq 1$, as can be easily seen from the proof of Theorem 1.1. For example, a geodesic which goes from the cusp down to the point $i \in \partial F$ and back to the cusp will be coded by the sequence $\lceil \bar{2}, 3, \bar{2} \rceil$. Thus the set of all oriented geodesics on M can be described symbolically as the Bernoulli space (minus one point) $X_G = \mathcal{N}_G^{\mathbb{Z}} \setminus \lceil \bar{2} \rceil$.

The partition of the cross-section. The infinite partition of the cross-section C_G corresponding to the G-code can be constructed as follows. We parameterize the cross-section C_g by (ϕ, θ), where $\phi \in [0, \pi/2]$ parameterizes the circle arc (counterclockwise) and $\theta \in [-\pi/2, \pi/2]$ is the angle the unit vector makes with the positive horizontal axis (counterclockwise). The angle θ depends on the position ϕ and is determined by the condition that the corresponding geodesic is G-reduced.

The partition of C_g (and that of C_G obtained by projection) corresponding to the arithmetic G-code ("the horizontal triangles") and its iteration under the first return map R to the cross-section C_g ("the vertical triangles") is shown on Figure 9. Its elements ("the horizontal triangles") are labeled by the symbols of the alphabet $\mathbb{N}_G$, $C_g = \sqcup_{n \in \mathbb{N}_G} C_n$ and are defined by the following condition: C_n consists of all tangent vectors $\mathbf{x}$ in C_g such that the corresponding geodesic in $\mathcal{H}$ goes from $0 < u < 1$ to $n - 1 < w < n$, i.e., if x is its coding sequence, then $n_0(x) = n$.

We also observe that the elements C_m and $R(C_n)$ intersect transversally for all $n, m \geq 2$, thus, according to Theorem 7.9 of [**Ad**], the infinite partition is Bernoulli. This gives an alternative geometric way to see that all arithmetic coding sequences are realized.

When does the G-code coincide with the geometric code? The relation between the geometric code and the arithmetic G-code of an oriented geodesic on M was established in [**K2, GK**]: *the geometric code and the arithmetic G-code of a geodesic γ on M coincide if and only if $\frac{1}{n_i} + \frac{1}{n_{i+1}} \leq \frac{1}{2}$, where $\lceil \gamma \rceil = \lceil n_i \rceil_{i=-\infty}^{\infty}$.*

2. Alternating continued fraction coding (Artin coding revisited)

In this section we describe the arithmetic coding of geodesics on the modular surface, using alternating continued fraction expansions which we call *A-expansions.*

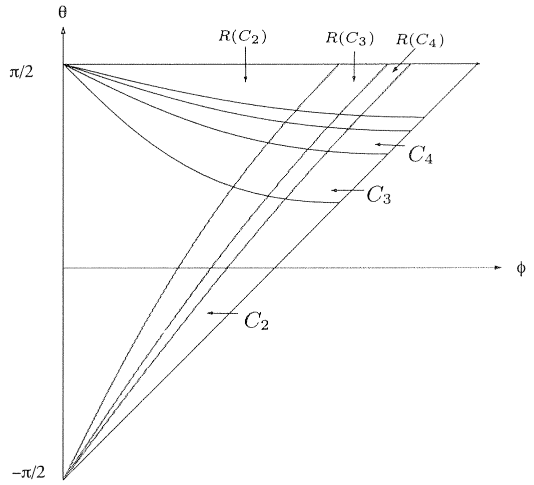


FIGURE 9. Infinite partition for the G-code and its image under the return map R

The result will be a modified Artin code, as described by Series [**S1**]. Every irrational number α has a unique A-expansion

$$\alpha := \lceil n_0, n_1, n_2, \dots \rfloor = n_0 - \cfrac{1}{n_1 - \cfrac{1}{n_2 - \cfrac{1}{\ddots}}}$$

with $n_0 \in \mathbb{Z}$ and $|n_i| \geq 1$, by setting $n_0 = \lceil \alpha \rfloor$, $\alpha_1 = -\frac{1}{\alpha - n_0}$, and, inductively,

$$n_i = \lceil \alpha_i \rfloor \,, \; \alpha_{i+1} = -\frac{1}{\alpha_i - n_i} \,, \qquad \text{where } \lceil \alpha \rfloor = \begin{cases} \lfloor \alpha \rfloor & \text{if } \alpha > 0 \\ \lceil \alpha \rceil & \text{if } \alpha < 0 \,. \end{cases}$$

Notice that $n_i n_{i+1} < 0$, hence the use of terminology of alternating continued fractions. Conversely, any infinite sequence of nonzero integers with alternating signs $n_0, n_1, n_2, \dots$ defines a real number whose A-expansion is $\lceil n_0, n_1, n_2, \dots \rfloor$.

REMARK 2. The properties of A-expansions can be easily established if one notices the relation with simple continued fraction expansions: if $\alpha > 0$ then

$$\alpha = \lceil n_0, n_1, n_2, \dots \rfloor = n_0 - \cfrac{1}{n_1 - \cfrac{1}{n_2 - \cfrac{1}{\ddots}}} = n_0 + \cfrac{1}{-n_1 + \cfrac{1}{n_2 + \cfrac{1}{\ddots}}} = \lfloor m_0, m_1, \dots \rfloor$$

where $m_i = (-1)^i n_i$, and $\lfloor m_0, m_1, \dots \rfloor$ is the simple continued fraction expansion of α.

The following properties are proved using the corresponding properties of the simple continued fractions, see e.g. [**O**]:

(A1) α is a quadratic irrationality if and only if its A-expansion is eventually periodic, $\alpha = \lceil n_0, n_1, \dots, n_k, \overline{n_{k+1}, \dots, n_{k+m}} \rfloor$;

(A2) A quadratic irrationality α has a purely periodic A-expansion if and only if $|\alpha| > 1$ and $-1 < \operatorname{sgn}(\alpha)\alpha' < 0$, where α' is conjugate to α;

(A3) If $\alpha = \lceil \overline{n_1, \dots, n_k} \rfloor$, then $1/\alpha' = \lceil \overline{n_k, \dots, n_1} \rfloor$;

(A4) Two irrationals α, β are $PSL(2, \mathbb{Z})$-equivalent if and only if their A-expansions have the same tail.

Similarly to the theory of minus continued fraction expansions, if $\alpha = \lceil n_0, n_1, \dots \rfloor$, then the partial fractions $r_k = \lceil n_0, n_1, \dots, n_k \rfloor$ can be written as p_k/q_k, where p_k and q_k are obtained inductively as:

$$\begin{aligned} p_{-2} &= 0\,,\ p_{-1} = 1\,;\ p_k = n_k p_{k-1} - p_{k-2} \quad \text{for } k \geq 0 \\ q_{-2} &= -1\,,\ q_{-1} = 0\,;\ q_k = n_k q_{k-1} - q_{k-2} \quad \text{for } k \geq 0\,. \end{aligned}$$

PROPOSITION 5. The following properties are satisfied:

(i) $1 = q_0 \leq |q_1| < |q_2| < \dots$;

(ii) $p_{k-1}q_k - p_k q_{k-1} = 1$, for all $k \geq 0$;

(iii) Let $T(z) = z + 1$, $S(z) = -1/z$ be the generating transformations for $PSL(2, \mathbb{Z})$, then for any $z \in \mathcal{H} \cup \mathbb{R} \cup \{\infty\}$

$$T^{n_0} S T^{n_1} S \dots T^{n_k} S(z) = \lceil n_0, n_1, \dots, n_k, z \rceil = \frac{p_k z - p_{k-1}}{q_k z - q_{k-1}}\,;$$

(iv) The sequence $\{r_k\}$ converges to α and $|p_k/q_k - \alpha| \leq 1/q_k^2$;

(v) If $\alpha > 0$, then either $\frac{q_{2k}}{q_{2k-1}} \geq \sqrt{2}$ or $\frac{q_{2k+1}}{q_{2k}} \leq -\sqrt{2}$; if $\alpha < 0$, then either $\frac{q_{2k}}{q_{2k-1}} \leq -\sqrt{2}$ or $\frac{q_{2k+1}}{q_{2k}} \geq \sqrt{2}$.

PROOF. The properties (i)–(iv) are proved similarly to those in Proposition 4, so we prove (v). We assume that $\alpha > 0$ (the case $\alpha < 0$ can be treated in a similar way). Then

$$1 = q_0 \leq -q_1 < -q_2 < q_3 < q_4 < -q_5 < -q_6 < q_7 < q_8 < \dots\,,$$

thus $0 < q_{2k-1}/q_{2k} < 1$, and $-1 < q_{2k}/q_{2k+1} < 0$. Taking into consideration the order of the signs and the fact that $q_k = n_k q_{k-1} - q_{k-2}$, one obtains

$$|q_k| = |n_k| \cdot |q_{k-1}| + |q_{k-2}|\,. \tag{2.1}$$

Indeed, if $k = 4m$, then $q_k > 0$, $q_{k-1} > 0$, $q_{k-2} < 0$, $n_k > 0$, and (2.1) follows. (The cases $k = 4m+1, 4m+2, 4m+3$ can be treated similarly.) From (2.1) and property (i), we get

$$|q_k| \geq |n_k| \cdot |q_{k-2}| + |q_{k-2}| = |q_{k-2}|(|n_k| + 1) \geq 2|q_{k-2}| \ \Rightarrow\ \frac{|q_{k-2}|}{|q_k|} \leq \frac{1}{2}$$

and since $\dfrac{q_{k-2}}{q_k} < 0$,

$$-\frac{1}{2} \leq \frac{q_{k-2}}{q_k} = \frac{q_{k-2}}{q_{k-1}} \cdot \frac{q_{k-1}}{q_k} < 0.$$

Therefore we either have

$$0 < \frac{q_{2k-1}}{q_{2k}} \leq \frac{1}{\sqrt{2}} \quad \text{or} \quad -\frac{1}{\sqrt{2}} \leq \frac{q_{2k}}{q_{2k+1}} < 0.$$

□

DEFINITION 2. An oriented geodesic on $\mathcal{H}$ is called *A-reduced* if its repelling and attracting end points, denoted by u and w, respectively, satisfy $|w| > 1$ and $-1 < \operatorname{sgn}(w)u < 0$.

To an A-reduced geodesic γ, one associates a bi-infinite sequence of nonzero integers (with alternating signs) $\lceil\gamma\rfloor = \lceil\dots, n_{-2}, n_{-1}, n_0, n_1, n_2, \dots\rfloor$, called its *A-code*, by juxtaposing the A-expansions of $1/u = \lceil n_{-1}, n_{-2}, \dots\rfloor$ and $w = \lceil n_0, n_1, n_2, \dots\rfloor$.

Reduction algorithm. The following theorem extends this symbolic coding to all geodesics on $\mathcal{H}$.

THEOREM 2.1. Every oriented geodesic on $\mathcal{H}$ is $PSL(2,\mathbb{Z})$-equivalent to an A-reduced geodesic.

PROOF. Let γ be an arbitrary geodesic on $\mathcal{H}$, with irrational end points u and w. Let $\lceil n_0, n_1, n_2 \dots\rfloor$ be the alternating continued fraction expansion of w. We construct the following sequence of real pairs $\{(u_k, w_k)\}$ $(k \geq 0)$ defined by $u_0 = u$, $w_0 = w$ and:

$$w_{k+1} = ST^{-n_k}\dots ST^{-n_1}ST^{-n_0}w\,, \quad u_{k+1} = ST^{-n_k}\dots ST^{-n_1}ST^{-n_0}u\,.$$

Notice that $w_{k+1} = \lceil n_{k+1}, n_{k+2}, \dots\rfloor$, $|w_{k+1}| > 1$, and by Proposition 5 (iii),

$$w = T^{n_0}ST^{n_1}S\dots T^{n_k}S(w_{k+1}) = \frac{p_k w_{k+1} - p_{k-1}}{q_k w_{k+1} - q_{k-1}}$$

$$u = T^{n_0}ST^{n_1}S\dots T^{n_k}S(u_{k+1}) = \frac{p_k u_{k+1} - p_{k-1}}{q_k u_{k+1} - q_{k-1}}$$

hence

$$u_{k+1} = \frac{q_{k-1}u - p_{k-1}}{q_k u - p_k} = \frac{q_{k-1}}{q_k}\cdot\frac{u - \frac{p_{k-1}}{q_{k-1}}}{u - \frac{p_k}{q_k}} = \frac{q_{k-1}}{q_k}\cdot\delta_k \tag{2.2}$$

where $\delta_k \to 1$. If $w > 0$, we have $w_{2k} > 1$ and $w_{2k+1} < -1$. By Proposition 5 (v), one can find a positive integer l such that either $u_{2l} \in (-1, 0)$ or $u_{2l+1} \in (0, 1)$. Then either a geodesic from u_{2l} to w_{2l} or a geodesic from u_{2l+1} to w_{2l+1} is A-reduced and $PSL(2,\mathbb{Z})$-equivalent to γ. The case $w < 0$ is treated similarly. □

REMARK 3. (i) The proof of Theorem 2.1 gives also the algorithm for A-reducing a geodesic γ: one has to construct inductively the sequence $\{(u_k, w_k)\}$ until $|w_k| > 1$ and $\operatorname{sgn}(w_k)u_k \in (-1, 0)$; (ii) any further application of the reduction algorithm to an A-reduced geodesic yields reduced geodesics whose A-codes are left shifts of the A-code of the first reduced one.

Similarly to the situation in Section 1, we define the A-code of an oriented geodesic γ on $\mathcal{H}$ to be the A-code of a reduced geodesic $PSL(2,\mathbb{Z})$-equivalent to γ, and prove its $PSL(2,\mathbb{Z})$-invariance by constructing a cross-section of the geodesic flow on M, directly related to the notion of A-reduced geodesics.

Construction of the cross-section. We describe the cross section C_A for the geodesic flow on M, such that successive returns to the cross section correspond to left-shifts in the arithmetic A-code. Let $C_A = P \cup Q_1 \cup Q_2$ be a subset of the unit tangent bundle SM, where P consists of all tangent vectors with base points in the circular side of F and pointing inward such that the corresponding geodesic is A-reduced; Q_1 consists of all tangent vectors with base points on the right vertical side of F pointing inwards, such that if γ is the corresponding geodesic, then $TS(\gamma)$ is A-reduced; Q_2 consists of all tangent vectors with base points on the left vertical side of F pointing inwards, such that if γ is the corresponding geodesic, then $T^{-1}S(\gamma)$ is A-reduced. Notice that $C_A = \pi(C_a)$ where C_a is the set all unit tangent vectors with base points on the unit semi-circle $|z| = 1$ and pointing outward such that the associated geodesic on $\mathcal{H}$ is A-reduced (Figure 10).

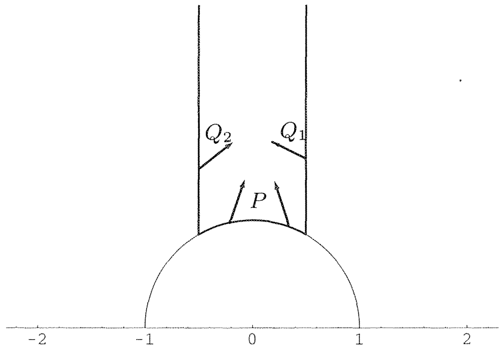


FIGURE 10. The cross-section $C_A = P \cup Q_1 \cup Q_2$

One can show similarly to the proof of Theorem 1.2 that $C_A = P \cup Q_1 \cup Q_2$ is indeed a cross-section for the geodesic flow on M, hence every geodesic γ can be represented as a bi-infinite sequence of segments σ_i between successive returns to C_A. To each segment σ_i is associated the corresponding A-reduced geodesic γ_i, so that $\lceil \gamma_{i+1} \rfloor$ differs from $\lceil \gamma_i \rfloor$ by a left shift. Thus we associate to γ a bi-infinite coding sequence, defined up to a shift, which we call the*A-code* of γ and denote by $\lceil \gamma \rfloor$. The argument of Corollary 1 shows that the A-code is $PSL(2,\mathbb{Z})$-invariant.

The set of all oriented geodesics on M can be described symbolically as a countable 1-step Markov chain $X_A \subset \mathcal{N}_A^{\mathbb{Z}}$ with the infinite alphabet $\mathcal{N}_A = \{n \in \mathbb{Z}, n \neq 0\}$ and transition matrix A,

$$A(n,m) = \begin{cases} 1 & \text{if } nm < 0, \\ 0 & \text{otherwise}. \end{cases} \tag{2.3}$$

Each oriented geodesic γ corresponds to its A-code, $\lceil \gamma \rfloor \in X_A$ and each bi-infinite sequence of nonzero integers with alternating signs $x \in X_A$ produces a geodesic on $\mathcal{H}$ from $u(x)$ to $w(x)$, where

$$w(x) = \lceil n_0, n_1, \dots \rfloor \quad , \quad \frac{1}{u(x)} = \lceil n_{-1}, n_{-2}, \dots \rfloor \, .$$

The partition of the cross-section. The infinite partition of the cross-section C_A corresponding to the A-code can be constructed as follows. We parameterize the cross-section C_a by (ϕ, θ), where $\phi \in [0, \pi]$ parameterizes the unit semicircle (counterclockwise) and $\theta \in [-\pi/2, (3\pi)/2]$ is the angle the unit vector makes with the positive horizontal axis (counterclockwise). The angle θ depends on ϕ and is determined by the condition that the corresponding geodesic is A-reduced.

The partition of C_a (and therefore of C_A by projection) corresponding to the arithmetic A-code ("the horizontal triangles") and its iteration under the first return map R to the cross-section C_a ("the vertical triangles") is shown on Figure 11. Its elements ("the horizontal triangles") are labeled by the symbols of the alphabet $\mathbb{N}_A$, $C_a = \sqcup_{n \in \mathbb{N}_A} C_n$ and are defined by the following condition: $C_n = \{\mathrm{x} \in C_a, n_0(\mathrm{x}) = n\}$, i.e. it consists of all tangent vectors $\mathbf{x}$ in C_a such that the coding sequence $x \in X_A$ of the corresponding geodesic with this initial vector has its first symbol in the A-code $n_0(x) = n$. Thus, for $n \geq 1$, C_n consists of all tangent vectors $\mathbf{x} \in C_a$ such that the corresponding geodesic goes from $-1 < u < 0$ to $n-1 < w < n$. We call this part of the cross section the positive part, and denote it by C_a^+ (and

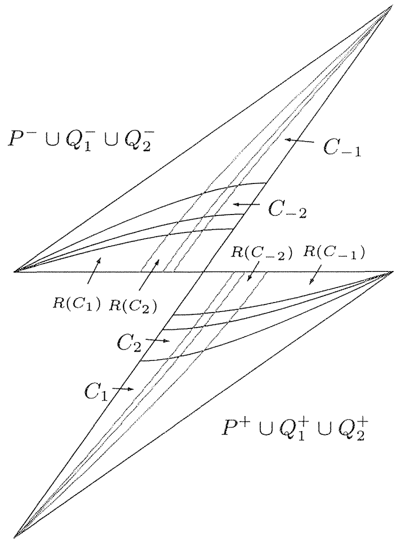


FIGURE 11. Infinite partition for the A-code and its image under the return map R

let $P^+ \cup Q_1^+ \cup Q_2^+$ denote its corresponding projection on C_A). For $n \leq -1$, C_n consists of all tangent vectors $\mathbf{x} \in C_a$ such that the corresponding geodesic goes from $0 < u < 1$ to $n-1 < w < n$. We call this part of the cross section the negative part, and denote it by C_a^- (with $P^- \cup Q_1^- \cup Q_2^-$ denoting $\pi(C_a^-)$).

Some results of this section can be illustrated geometrically since the Markov property of the partition is equivalent to the Markov property of the shift space. If $n_0(\mathbf{x}) = n$ and $n_1(\mathbf{x}) = m$ for some $\mathbf{x} \in C_A$, then $R(C_n) \cap C_m \neq \emptyset$. Therefore, as follows from Figure 11, the signs in the A-code must alternate, because $R(C_n) \cap C_m \neq \emptyset \Leftrightarrow nm < 0$. Moreover, all intersections are transversal, hence, according to Theorem 7.9 of [**Ad**], the partition is Markov.

When does the A-code coincide with the geometric code? The next theorem gives a sufficient condition for the geometric code and the arithmetic A-code of a geodesic γ on M to coincide.

THEOREM 2.2. *The geometric code and the arithmetic A-code of a geodesic γ coincide if $|n_i| \geq 2$ and $n_i n_{i+1} < 0$ where $\lceil \gamma \rfloor = \lceil n_i \rfloor_{i=-\infty}^{\infty}$.*

PROOF. Let $x = \{\dots, n_{-2}, n_{-1}, n_0, n_1, \dots\}$ be a sequence of integers with $|n_i| \geq 2$ and $n_i n_{i+1} < 0$. Consider the geodesic $\gamma(x)$ on $\mathcal{H}$, from

$$u(x) = \frac{1}{(n_{-1}, n_{-2}, \dots)} = \frac{1}{n_{-1} - \cfrac{1}{n_{-2} - \cfrac{1}{\ddots}}} \text{ to } w(x) = (n_0, n_1, \dots) = n_0 - \cfrac{1}{n_1 - \cfrac{1}{\ddots}}.$$

Since $x \in X_A$, the A-code of $\gamma(x)$ is $\lceil \gamma(x) \rfloor = \lceil n_{-2}, n_{-1}, n_0, n_1, \dots \rfloor$. We showed in [**KU**, Theorem 1.4], that if a sequence x satisfies $|n_i| \geq 2$ and $|\frac{1}{n_i} + \frac{1}{n_{i+1}}| \leq \frac{1}{2}$, then such geodesic $\gamma(x)$ from $u(x)$ to $w(x)$ has the geometric code $[\gamma(x)] = [\dots, n_{-1}, n_0, n_1, n_2, \dots]$. Therefore the geometric code and the A-code of $\pi(\gamma(x))$ coincide (up to a shift). □

3. Nearest integer continued fraction coding (Hurwitz coding)

In this section we describe the arithmetic coding procedure for geodesics on the modular surface, using the nearest integer continued fraction expansions, and the corresponding reduction theory for real quadratic forms with positive discriminant (indefinite real quadratic forms) developed by Hurwitz [**H**] (see also [**F**]). Every irrational number α has a unique H-expansion $\alpha = \langle n_0, n_1, n_2, \dots \rangle$ with $n_0 \in \mathbb{Z}$ and $|n_i| \geq 2$ for $i \geq 1$, by setting $n_0 = \langle \alpha \rangle$ (the nearest integer to α), $\alpha_1 = -\frac{1}{\alpha - n_0}$, and, inductively,

$$n_i = \langle \alpha_i \rangle \quad , \quad \alpha_{i+1} = -\frac{1}{\alpha_i - n_i} \, .$$

Notice that if $n_i = \pm 2$, then $n_i n_{i+1} < 0$. Conversely, any infinite sequence of integers $n_0, n_1, n_2, \dots$ with $|n_i| \geq 2$ for $i \geq 1$ which does not contain the pairs $\{2, p\}$ and $\{-2, -p\}$ for $p \geq 2$ defines an irrational number whose H-expansion is $\langle n_0, n_1, n_2, \dots \rangle$. The following properties are satisfied (see [**H**] for more details):

(H1) α is a quadratic irrationality, i.e. a root of a quadratic polynomial with integer coefficients, if and only if its H-expansion is eventually periodic, $\alpha = \langle n_0, n_1, \dots, n_k, \overline{n_{k+1}, \dots, n_{k+m}} \rangle$;

(H2) A quadratic irrationality α has a purely periodic H-expansion if and only if $|\alpha| > 2$ and $\operatorname{sgn}(\alpha)\alpha' \in [r-1, r]$, where α' is conjugate to α and $r = (3 - \sqrt{5})/2$;

(H3) Two irrationals α, β are $PSL(2, \mathbb{Z})$-equivalent if and only if their H-expansions have the same tail or one has $1/r = \langle \overline{3} \rangle$ as a tail, and the other has $-1/r = \langle \overline{-3} \rangle$ as a tail.

Definition 3. An oriented geodesic on $\mathcal{H}$ is called *H-reduced* if its repelling and attracting end points, denoted by u and w, respectively, satisfy $|w| > 2$ and $\operatorname{sgn}(w)u \in [r-1, r]$, where $r = (3 - \sqrt{5})/2$.

We remark that the H-expansion satisfies an asymmetric restriction (if $n_i = \pm 2$, then $n_i n_{i+1} < 0$), and the statement "if $\alpha = \langle \overline{n_1, \dots, n_k} \rangle$, then $1/\alpha' = \langle \overline{n_k, \dots, n_1} \rangle$" is not always true. For example, if one considers the conjugate quadratic irrationalities $\alpha = (15 + 12\sqrt{2})/7$ and $\alpha' = (15 - 12\sqrt{2})/2$, then $\langle \alpha \rangle = \langle \overline{5, 2, -3} \rangle$, but $\langle \alpha' \rangle = \langle \overline{-4, -2, 4} \rangle$. For that reason, we cannot construct a meaningful symbolic sequence for an H-reduced geodesic just by juxtaposing the H-expansions of w and $1/u$. In order to associate to an H-reduced geodesic a bi-infinite sequence of integers, we use a different expansion for $1/u$ introduced by Hurwitz, and called the *H-dual expansion*. Every irrational α has a unique H-dual expansion $\alpha = \langle\langle n_0, n_1, n_2, \dots \rangle\rangle$ with $n_0 \in \mathbb{Z}$ and $|n_i| \geq 2$ for $i \geq 1$, given by $n_0 = \langle\langle \alpha \rangle\rangle$, $\alpha_1 = -\frac{1}{\alpha - n_0}$ and, inductively,

$$n_i = \langle\langle \alpha_i \rangle\rangle \quad , \quad \alpha_{i+1} = -\frac{1}{\alpha_i - n_i} \, ,$$

where

$$\langle\langle \alpha \rangle\rangle = \begin{cases} \langle \alpha \rangle - \operatorname{sgn}(\alpha) & \text{if} \quad \operatorname{sgn}(\alpha)(\langle \alpha \rangle - \alpha) > r \\ \langle \alpha \rangle & \text{otherwise} \end{cases}$$

Notice that if $n_{i+1} = \pm 2$, then $n_i n_{i+1} < 0$, and moreover if $\alpha = \langle \overline{n_1, \dots, n_k} \rangle$, then $1/\alpha' = \langle\langle \overline{n_k, \dots, n_1} \rangle\rangle$.

If $\alpha = \langle n_0, n_1, \dots \rangle$, then the convergents $r_k = \langle n_0, n_1, \dots, n_k \rangle$ can be written as p_k/q_k where p_k and q_k are obtained inductively as:

$$p_{-2} = 0 \, , \; p_{-1} = 1 \, ; \; p_k = n_k p_{k-1} - p_{k-2} \;\; \text{for } k \geq 0$$
$$q_{-2} = -1 \, , \; q_{-1} = 0 \, ; \; q_k = n_k q_{k-1} - q_{k-2} \;\; \text{for } k \geq 0 \, .$$

The proof of the following Proposition is contained in [**H**].

Proposition 6. The following properties are satisfied:

(i) $1 = q_0 < |q_1| < |q_2| < \dots$;

(ii) $p_{k-1}q_k - p_kq_{k-1} = 1$, for all $k \geq 0$;

(iii) Let $T(z) = z + 1$, $S(z) = -1/z$ be the generating transformations for $PSL(2,\mathbb{Z})$, then for any $z \in \mathcal{H} \cup \mathbb{R} \cup \{\infty\}$

$$T^{n_0}ST^{n_1}S \dots T^{n_k}S(z) = \lceil n_0, n_1, \dots, n_k, z \rceil = \frac{p_kz - p_{k-1}}{q_kz - q_{k-1}} ;$$

(iv) The sequence $\{r_k\}$ converges to α and $|p_k/q_k - \alpha| \leq 1/q_k^2$;

(v) $\frac{q_k}{q_{k-1}} \in [n_k - r, n_k + 1 - r]$ if $n_k > 0$, and $\frac{q_k}{q_{k-1}} \in [n_k - 1 + r, n_k + r]$ if $n_k < 0$; in particular, $|\frac{q_k}{q_{k-1}}| \geq 2 - r = \frac{1+\sqrt{5}}{2}$.

To an H-reduced geodesic γ, one associates a bi-infinite sequence of integers $\langle\gamma\rangle = \langle \dots n_{-1}, n_0, n_1, \dots \rangle$, by juxtaposing the H-dual expansion of $1/u$ and the H-expansion of w. Observe that $|n_i| \geq 2$ and the only additional restriction on n_i's is that if $n_i = \pm 2$, then $n_in_{i+1} < 0$.

Reduction algorithm. We describe the reduction procedure of any geodesic to an H-reduced one.

Theorem 3.1. Every oriented geodesic on $\mathcal{H}$ is $PSL(2,\mathbb{Z})$-equivalent to an H-reduced geodesic.

Proof. For the sake of completeness, we present the proof following Hurwitz [**H**]. Let γ be an arbitrary geodesic on $\mathcal{H}$, with irrational end points u and w, and assume that $u < w$. Let $\langle n_0, n_1, n_2, \dots \rangle$ be the H-expansion of w, and suppose that its tail is different from $\langle \bar{3} \rangle$ (the situation when the tail of w coincides with $\langle \bar{3} \rangle$ can be treated similarly). We construct the following sequence of real pairs $\{(u_k, w_k)\}$ $(k \geq 0)$ defined by $u_0 = u$, $w_0 = w$ and:

$$w_{k+1} = ST^{-n_k} \dots ST^{-n_1}ST^{-n_0}w\,, \quad u_{k+1} = ST^{-n_k} \dots ST^{-n_1}ST^{-n_0}u\,.$$

Notice that $w_{k+1} = \langle n_{k+1}, n_{k+2}, \dots \rangle$ and by Proposition 6 (iii),

$$w = \frac{p_kw_{k+1} - p_{k-1}}{q_kw_{k+1} - q_{k-1}}\,, \; u = \frac{p_ku_{k+1} - p_{k-1}}{q_ku_{k+1} - q_{k-1}}\,.$$

Hence

$$u_{k+1} = \frac{q_{k-1}u - p_{k-1}}{q_ku - p_k} = \frac{q_{k-1}}{q_k} + \frac{1}{q_k^2(p_k/q_k - u)} = \frac{q_{k-1}}{q_k} + \varepsilon_k\,,$$

where $\varepsilon_k > 0$ (for large enough k) and $\varepsilon_k \to 0$. For infinitely many k's, $n_k \neq 2, 3$, and one can find a subsequence k_j such that

$$w_{k_j+1} > 2,\; n_{k_j+1} \geq 2,\; \text{ and } n_{k_j} = -2, -3, \pm 4, \dots$$

or

$$w_{k_j+1} < -2,\; n_{k_j+1} \leq -2,\; \text{ and } n_{k_j} = -3, \pm 4, \dots\,.$$

Using Proposition 6 (v), one has, in the first case

$$\frac{q_{k_j}}{q_{k_j-1}} \leq -2 + r \text{ or } \frac{q_{k_j}}{q_{k_j-1}} \geq 4 - r \;\Rightarrow\; u_{k_j+1} \in \left[\frac{1}{-2+r} + \varepsilon_{k_j}, \frac{1}{4-r} + \varepsilon_{k_j}\right]$$

and, in the second case,

$$\frac{q_{k_j}}{q_{k_j-1}} \leq -3 + r \text{ or } \frac{q_{k_j}}{q_{k_j-1}} \geq 4 - r \;\Rightarrow\; u_{k_j+1} \in \left[\frac{1}{-3+r} + \varepsilon_{k_j}, \frac{1}{4-r} + \varepsilon_{k_j}\right]\,.$$

Since $1/(-2+r) = r-1$, $1/(-3+r) = -r$, $1/(4-r) < r < 1-r$, and $0 < \varepsilon_{k_j} \to 0$, there exists an integer l such that $w_{k_l+1} > 2$ and $u_{k_l+1} \in [r-1, r]$ or $w_{k_l+1} < -2$ and $u_{k_l+1} \in [-r, 1-r]$. The geodesic with end points u_{k_l+1} and w_{k_l+1} is H-reduced and $PSL(2,\mathbb{Z})$-equivalent to γ.

We finish the proof by explaining how one can derive the reduction procedure for the case $u > w$. Let $w =< n_0, n_1, \cdots >$ and consider, as before, the sequence of real pairs $\{(u_k, w_k)\}$ $(k \geq 0)$ defined by $u_0 = u$, $w_0 = w$ and:

$$w_{k+1} = ST^{-n_k} \dots ST^{-n_1} ST^{-n_0} w\,, \quad u_{k+1} = ST^{-n_k} \dots ST^{-n_1} ST^{-n_0} u\,.$$

Since $-u < -w$, one can apply the reduction procedure described above to the geodesic $\tilde{\gamma}$ from $\tilde{u} = -u$ to $\tilde{w} = -w$. Notice that $-w =< -n_0, -n_1, \cdots >$ and the sequence of pairs $\{(\tilde{u}_k, \tilde{w}_k)\}$ $(k \geq 0)$ is defined by $\tilde{u}_0 = \tilde{u}$, $\tilde{w}_0 = \tilde{w}$ and:

$$\tilde{w}_{k+1} = ST^{n_k} \dots ST^{n_1} ST^{n_0} \tilde{w}\,, \quad \tilde{u}_{k+1} = ST^{n_k} \dots ST^{n_1} ST^{n_0} u\,.$$

Using the identity $ST^n(-w) = -ST^{-n}w$, one has $\tilde{w}_k = -w_k$ and $\tilde{u}_k = -\tilde{u}_k$. From the proof above, there exists a positive integer k such that the geodesic with end points $\tilde{u}_k$ and $\tilde{w}_k$ is H-reduced and $PSL(2,\mathbb{Z})$-equivalent to $\tilde{\gamma}$. Thus, the geodesic from $u_k = -\tilde{u}_k$ to $w_k = -\tilde{w}_k$ is also H-reduced and $PSL(2,\mathbb{Z})$-equivalent to γ. □

Remark 4. (i) The proof of Theorem 3.1 gives also the algorithm for H-reducing a geodesic γ: one has to construct inductively the sequence $\{(u_k, w_k)\}$ until $|w_k| > 2$ and $\operatorname{sgn}(w_k)u_k \in [r-1, r]$; (ii) any further application of the reduction algorithm to an H-reduced geodesic yields reduced geodesics whose H-codes are left shifts of the H-code of the first reduced one.

As in the previous sections we define the H-code of an oriented geodesic γ on $\mathcal{H}$ to be the H-code of a reduced geodesic $PSL(2,\mathbb{Z})$-equivalent to γ, and prove its $PSL(2,\mathbb{Z})$-invariance by constructing a cross-section of the geodesic flow on M, directly related to the notion of H-reduced geodesics.

Construction of the cross-section. We describe the construction of the cross section C_H for the geodesic flow on M, such that successive returns to the cross section correspond to left-shifts in the H-code. We define $C_H = P \cup Q_1 \cup Q_2$ to be a subset of the unit tangent bundle SM, where P consists of all tangent vectors with base points in the circular side of F and pointing inward such that the corresponding geodesic is H-reduced; Q_1 consists of all tangent vectors with base points on the right vertical side of F pointing inwards, such that if γ is the corresponding geodesic, then $TS(\gamma)$ is H-reduced; Q_2 consists of all tangent vectors with base points on the left vertical side of F pointing inwards, such that if γ is the corresponding geodesic, then $T^{-1}S(\gamma)$ is H-reduced. Notice that $C_H = \pi(C_h)$ where C_h is the set all unit tangent vectors with base points on the unit semi-circle $|z| = 1$ and pointing outward such that the associated geodesic on $\mathcal{H}$ is H-reduced (Figure 12).

One can show similarly to the proof of Theorem 1.2 that $C_H = P \cup Q_1 \cup Q_2$ is indeed a cross-section for the geodesic flow on M, hence every geodesic γ can be represented as a bi-infinite sequence of segments σ_i between successive returns to C_H. To each segment σ_i is associated the corresponding H-reduced geodesic γ_i, so that $\lceil \gamma_{i+1} \rfloor$ differs from $\lceil \gamma_i \rfloor$ by a left shift. Thus we associate to γ a bi-infinite coding sequence, defined up to a shift, which we call the *H-code* of γ and denote by $\langle \gamma \rangle$. The argument of Corollary 1 shows that the H-code is $PSL(2,\mathbb{Z})$-invariant.

The set of all oriented geodesics on M can be described symbolically as a countable 1-step Markov chain $X_H \subset \mathcal{N}_H^{\mathbb{Z}}$ with infinite alphabet $\mathcal{N}_H = \{n \in \mathbb{Z}, |n| \geq$

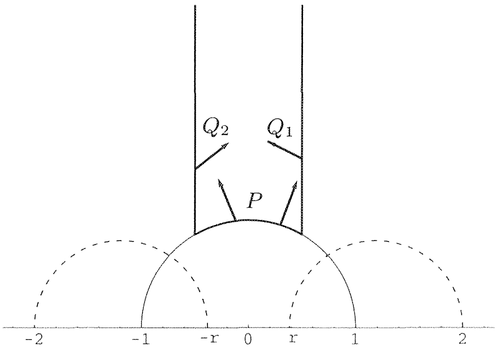


FIGURE 12. The cross section $C_H = P \cup Q_1 \cup Q_2$

2} and transition matrix H,

$$H(n,m) = \begin{cases} 0 & \text{if } |n| = 2 \text{ and } nm > 0, \\ 1 & \text{otherwise}. \end{cases} \tag{3.1}$$

Each oriented geodesic γ corresponds to its H-code, $\langle\gamma\rangle \in X_H$ and each bi-infinite sequence of nonzero integers $x \in X_H$ produces a geodesic on $\mathcal{H}$ from $u(x)$ to $w(x)$, where

$$w(x) = \langle n_1, n_2, \ldots\rangle \quad , \quad \frac{1}{u(x)} = \langle\!\langle n_0, n_{-1}, \ldots\rangle\!\rangle \, .$$

The partition of the cross-section. The infinite partition of the cross-section C_H corresponding to the H-code can be constructed as follows. We parameterize the cross-section C_h by (ϕ, θ), where $\phi \in [0, \pi]$ parameterizes the unit semicircle (counterclockwise) and $\theta \in [-\pi/2, (3\pi)/2]$ is the angle the unit vector makes with the positive horizontal axis (counterclockwise). The angle θ depends on the position ϕ and is determined by the condition that the corresponding geodesic is H-reduced.

The partition of C_h (and therefore of C_H by projection) corresponding to the arithmetic H-code ("the horizontal rectangles") and its iteration under the first return map R to the cross-section C_h ("the vertical rectangles") is shown on Figure 13. Its elements ("the horizontal rectangles") are labeled by the symbols of the alphabet $\mathbb{N}_H$, $C_h = \sqcup_{n \in \mathbb{N}_H} C_n$ and are defined by the following condition: $C_n = \{\mathbf{x} \in C_h, n_0(\mathbf{x}) = n\}$, i.e. it consists of all tangent vectors $\mathbf{x}$ in C_h such that the coding sequence $x \in X$ of the corresponding geodesic with this initial vector has its first symbol in the H-code $n_0(x) = n$. Thus, for $n \geq 2$, C_n consists of all tangent vectors $\mathbf{x} \in C_h$ such that the corresponding geodesic goes from $r - 1 < u < r$ to $n - 1/2 < w < n + 1/2$. For $n \leq -2$, C_n consists of all tangent vectors $\mathbf{x} \in C_h$ such that the corresponding geodesic goes from $-r < u < 1 - r$ to $n - 1/2 < w < n + 1/2$.

Some results of this section can be illustrated geometrically. If $n_0(\mathbf{x}) = n$ and $n_1(\mathbf{x}) = m$ for some vector $\mathbf{x} \in C_H$, then $R(C_n) \cap C_m \neq \emptyset$. Therefore, as follows from Figure 13, if $R(C_n) \cap C_m \neq \emptyset$ and $n = \pm 2$, then $nm < 0$. Moreover, all intersections are transversal, hence, according to Theorem 7.9 of [**Ad**], our partition is Markov.

When does the H-code coincide with the geometric code? The next theorem gives a sufficient condition for the geometric code and the arithmetic H-code of a geodesic γ on M to coincide.

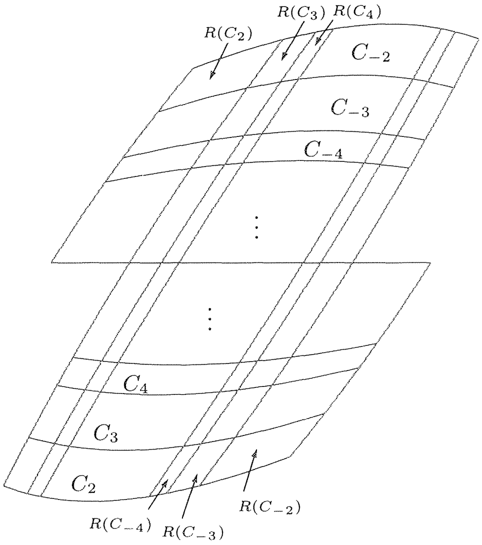


FIGURE 13. Infinite partition for the H-code and its image under the return map R

THEOREM 3.2. The geometric code and the arithmetic H-code of γ coincide if $|n_i| \geq 2$ and

$$\left|\frac{1}{n_i} + \frac{1}{n_{i+1}}\right| \leq \frac{1}{2}, \tag{3.2}$$

where $\langle\gamma\rangle = \langle n_i \rangle_{i=-\infty}^{\infty}$.

PROOF. Let $x = \{\dots, n_{-2}, n_{-1}, n_0, n_1, \dots\}$ be a sequence of integers with $|n_i| \geq 2$ and satisfying (3.2). Consider the geodesic $\gamma(x)$ on $\mathcal{H}$, from

$$u(x) = \frac{1}{(n_{-1}, n_{-2}, \dots)} = \cfrac{1}{n_{-1} - \cfrac{1}{n_{-2} - \cfrac{1}{\ddots}}} \quad \text{to} \quad w(x) = (n_0, n_1, \dots) = n_0 - \cfrac{1}{n_1 - \cfrac{1}{\ddots}}.$$

Since $x \in X_H$, the H-code of $\gamma(x)$ is $\langle\gamma(x)\rangle = \langle n_{-2}, n_{-1}, n_0, n_1, \dots\rangle$. We showed in [**KU**, Theorem 1.4], that such a geodesic $\gamma(x)$ from $u(x)$ to $w(x)$ has the geometric code $[\gamma(x)] = [\dots, n_{-1}, n_0, n_1, n_2, \dots]$. Therefore the geometric code and the arithmetic code of $\pi(\gamma(x))$ coincide (up to a shift). □

REMARK 5. There exist geodesics that are not geometrically Markov, for which the geometric code and the H-code coincide. For example, consider the closed geodesic γ given by the axis of $T^5ST^3ST^{-2}S$. Its geometric code is $[\gamma] = [5, 3, -2]$ and coincides with the H-code $\langle\gamma\rangle = \langle\overline{5, 3, -2}\rangle$. However, γ is not geometrically Markov. A natural question would be to characterize completely the class of geodesics for which the two codes coincide.

REMARK 6. One can easily notice that the H-code and G-code of a geodesic γ coincide if $n_i \geq 3$, and, the H-code, G-code and geometric code coincide for positive geodesics.

4. Symbolic representation of the geodesic flow

Geodesic flow as a special flow. Let $C_\alpha \subset SM$ $(\alpha = G, A, H)$ be a cross-section for the geodesic flow $\{\varphi^t\}$ on M constructed for one of the arithmetic codes studied in the previous sections, and X_α the corresponding set of coding sequences. Every $\mathbf{x} \in C_\alpha$ defines an oriented geodesic $\gamma(\mathbf{x})$ on M which will return to C_α infinitely often. Let $R_\alpha : C_\alpha \to C_\alpha$ be the first return map, and $f_\alpha : C_\alpha \to \mathbb{R}$ be the time of the first return on C_α defined as follows: for $\mathbf{x} \in C_\alpha$, $R_\alpha(\mathbf{x}) = \varphi^t(\mathbf{x})$, $f_\alpha(\mathbf{x}) = t$. Then $\{\varphi^t\}$ can be represented as the special flow on the space

$$C_\alpha{}^{f_\alpha} = \{(\mathbf{x}, y) : \mathbf{x} \in C_\alpha, 0 \le y \le f_\alpha(\mathbf{x})\}$$

with the ceiling function f_α by the formula $\varphi^t(\mathbf{x}, y) = (\mathbf{x}, y + t)$ with the identification $(\mathbf{x}, f_\alpha(\mathbf{x})) = (R_\alpha(\mathbf{x}), 0)$.

In the previous sections for each arithmetic code we have established a bijective map $\mathrm{Cod}_\alpha : C_\alpha \to X_\alpha$ by $\mathrm{Cod}_\alpha : \mathbf{x} \mapsto (\gamma(\mathbf{x}))$ such that the diagram

$$\begin{tikzcd}
C_\alpha \arrow[r, "\mathrm{Cod}_\alpha"] \arrow[d, "\mathrm{R}_\alpha"'] & X_\alpha \arrow[d, "\sigma_\alpha"] \\
C_\alpha \arrow[r, "\mathrm{Cod}_\alpha"] & X_\alpha
\end{tikzcd}$$

is commutative. Here σ_α is the left shift $\sigma_\alpha : X_\alpha \to X_\alpha$ defined for $x = (n_i(x))_{i=-\infty}^{\infty}$ by $(\sigma_\alpha x)_i = n_{i+1}(x)$. Thus we obtain three symbolic representations of the geodesic flow (for $\alpha = G, A, H$) on the space

$$X_\alpha{}^{f_\alpha} = \{(x, y) : x \in X_\alpha, 0 \le y \le f_\alpha(x)\}$$

given by the formula $\varphi^t(x, y) = (x, y + t)$ with the identification $(x, f_\alpha(x)) = (\sigma_\alpha x, 0)$, where $(X_\alpha, \sigma_\alpha)$ is the space of α-coding sequences, and f_α is the time of the first return to the cross-section C_α.

Calculation of the return time. Let $\alpha = G, A, H$. The ceiling function $f_\alpha(x)$ on X_α is the length of the segment between successive returns of the geodesic $\gamma(x)$ to the cross-section C_α. The following theorem was proved in [**GK**] for the G-code. The proof for the A- and H-code is the same.

THEOREM 4.1. Let $x \in X_\alpha$ and $w(x)$, $u(x)$ be the end points of the corresponding geodesic $\gamma(x)$. Then

$$f_\alpha(x) = 2\log|w(x)| + \log g(x) - \log g(\sigma_\alpha x)$$

where

$$g(x) = \frac{|w(x) - u(x)|\sqrt{w(x)^2 - 1}}{w(x)^2\sqrt{1 - u(x)^2}}.$$

References

[Ad] R. Adler, *Symbolic dynamics and Markov partitions*, Bull. Amer. Math. Soc. **35** (1998), no. 1, 1–56.

[AF1] R. Adler and L. Flatto, *Cross section maps for geodesic flows, I (The Modular surface)*, Birkhäuser, Progress in Mathematics (ed. A. Katok) (1982), 103–161.

[AF2] R. Adler and L. Flatto, *Cross section map for geodesic flow on the modular surface*, Contemp. Math. **26** (1984), 9–23.

[AF3] R. Adler and L. Flatto, *Geodesic flows, interval maps, and symbolic dynamics*, Bull. Amer. Math. Soc. **25** (1991), no. 2, 229–334.

[Arn] P. Arnoux, *Le codage des flot géodésique sur la surface modulaire*, Enseign. Math. **40** (1994), 29–48.

[Ar] E. Artin, *Ein Mechanisches System mit quasiergodischen Bahnen*, Abh. Math. Sem. Univ. Hamburg **3** (1924), 170–175.

[BS] R. Bowen and C. Series, *Markov maps associated with Fuchsian groups*, Inst. Hautes Études Sci. Publ. Math. No. 50 (1979), 153–170.

[F] D. Fried, *Reduction theory over quadratic imaginary fields*, preprint.

[GK] B. Gurevich and S. Katok, *Arithmetic coding and entropy for the positive geodesic flow on the modular surface*, Moscow Mathematical Journal **1** (2001), no. 4, 569-582.

[GL] D. J. Grabiner and J. C. Lagarias, *Cutting sequences for geodesic flow on the modular surface and continued fractions*, Monatsh. Math. **133** (2001), no. 4, 295–339.

[Hed] G. A. Hedlund, *A metrically transitive group defined by the modular group*, Amer. J. Math. **57** (1935), 668–678.

[H] A. Hurwitz, *Über eine besondere Art der Kettenbruch-Entwicklung reeler Grössen*, Acta Math. **12** (1889) 367–405.

[KH] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, 1995.

[K1] S. Katok, *Fuchsian groups*, University of Chicago Press, 1992.

[K2] S. Katok, *Coding of closed geodesics after Gauss and Morse*, Geom. Dedicata **63** (1996), 123–145.

[K3] S. Katok, *Continued fractions, hyperbolic geometry and quadratic forms*, MASS Selecta, Amer. Math. Soc., 121–160, 2003.

[KU] S. Katok and I. Ugarcovici, *Geometrically Markov geodesics on the modular surface*, Moscow Math. Journal, to appear.

[O] C. D. Olds, *Continued fractions*, New Mathematics Library **9**, MAA, 1992.

[S1] C. Series, *On coding geodesics with continued fractions*, Enseign. Math. **29** (1980), 67–76.

[S2] C. Series, *Symbolic dynamics for geodesic flows*, Acta Math. **146** (1981), 103–128.

[S3] C. Series, *The modular surface and continued fractions*, J. London Math. Soc. (2) **31** (1985), 69–80.

[S4] C. Series, *Geometrical Markov coding of geodesics on surfaces of constant negative curvature*, Ergod. Th. & Dynam. Sys. **6** (1986), 601–625.

[Z] D. Zagier, *Zetafunkionen und quadratische Körper: eine Einführung in die höhere Zahlentheorie*, Springer-Verlag, 1982.

Preconditioning and partial differential equations

Michelle Schatzman

Laboratoire de Mathématiques Appliquées de Lyon
CNRS and UCBL
69622 Villeurbanne Cedex
France
schatz@maply.univ-lyon1.fr

Abstract. The purpose of this article is to explain how some apparently simple problems from numerical linear algebra are in fact extremely difficult, so that we cannot hope to solve them effectively *in general*. However, if we build and analyze algorithms to solve them in special cases of interest for the numerical analysis of partial differential equations, we find that the theory needed to validate these methods is of the same nature as that used for pseudo-differential operators, though the operators considered probably do not enter the framework of pseudo-differential operators. This relationship between distant fields of analysis displays how the interaction between different parts of mathematics, motivated by problems of practical origin, leads to interesting questions and solutions. The article is intended for mathematicians of all backgrounds and is written so that beginning graduate students with a good background in PDE's and analysis can read it, and hopefully enjoy it also.

1. Introduction

This article contains a description of how different fields of mathematics fertilize one another in the course of solving some highly practical and apparently simple problems.

This article is presented for non specialists and does not include technical proofs. I have tried to write it in simple language. Hopefully, a graduate student with a good background in partial differential equations and analysis should understand it.

The simple problem is to solve a system of d linear equations in d unknowns, for d very large. Very large would mean for instance $d = 10^6$.

In section 2, I show that this problem requires a very large number of operations, and I emphasize the importance of choosing an appropriate algorithm in order to have any chance to succeed before the end of the universe.

Then, I move on to a more analytical point of view.

Let A be the matrix of our linear system. The condition number of A is the number $\kappa(A) = \|A\| \|A^{-1}\|$, which provides an upper bound of the relative error on the solution of a linear system in terms of the relative errors on the right hand side and the matrix. The condition number is also involved in the convergence properties of the so-called iterative methods: in an iterative method, we construct a sequence of approximations to the solution of the linear system.

For large matrices with lots of vanishing coefficients i.e. *sparse matrices*, the iterative methods are the methods of choice: they let us obtain a solution when Gaussian elimination and its analogues require too many operations to give a practical answer.

It turns out that the numerical solution of linear systems is meaningless if the condition number is too large, and on average, in floating point arithmetic, for very large matrices, the condition number is so large that for all practical purposes, the matrices might as well be considered as singular. This problem can be handled if we use longer mantissas, which is not always a feasible solution.

However, it is common practice to solve for numerical approximations of large systems of linear equations coming from the discretization of partial differential equations. This is in fact possible only because these matrices are very special and very rare, as is shown in section 3. I would expect that the reason why they are so special is that they have a particular algebraic structure: such a structure has not been described in algebraic terms and this question deserves to be studied.

The choice of good algorithms contains the question of preconditioning. If we want to solve $Ax = b$, we might as well try to solve $B^{-1}Ax = B^{-1}b$ or $AB^{-1}y = b, \quad x = By$, provided that the matrix B is a "good" approximation of A and it has nice algorithmic properties. If the condition number of $B^{-1}A$ or of AB^{-1} is small relatively to that of A, the numerical solution will be faster than that of the original problem. If in addition, we have a precise algorithm for multiplying $B^{-1}A$ by a vector, it will be also more precise.

But one has to understand in mathematical terms what one means by saying that B is a "good" approximation of A.

As an example, in section 4, I show how the construction of approximate inverses is a standard trick of the analysis of partial differential equations, by explaining the basics of parametrices of elliptic operators.

We start from the fundamental solution for an elliptic constant coefficient operator $P(D)$: it is obtained by a combination of the Fourier method and Cauchy-Kowalevskaia theorem.

How to work with a variable coefficients elliptic operator $P(x, D)$? We construct a parametrix, i.e. an asymptotic series giving an approximate inverse of a partial differential, or more generally pseudo-differential operator.

The first step of the construction gives an approximate inverse S_0 such that $\mathbb{1} - P(x, D)S_0$ and $\mathbb{1} - S_0P(x, D)$ smooth the data by one degree of differentiability.

The process can be iterated: the operator S_j is such that $\mathbb{1} - S_jP(x, D)$ and $\mathbb{1} - P(x, D)S_j$ smooth the data by j degrees of differentiability.

Eventually, one can get an asymptotic series of sum S_∞, such that $\mathbb{1} - S_\infty P(x, D)$ and $\mathbb{1} - P(x, D)S_\infty$ are infinitely smoothing operators.

The validation of this strategy relies in an essentially way on the stationary phase method and on many estimates and technical steps.

This set of ideas has filtered into numerical analysis in many ways: paraxial approximation of variable coefficient wave equations (Bamberger et al., [**3**], [**2**]), boundary element methods, which are appropriate for integral formulation of PDE's, and these are nothing but another word for Green's function, preconditioning of variable coefficient problems by wavelets (Piquemal and Liandrat, [**37**]), $\mathcal{H}$-matrices (Hackbusch and coworkers, [**4**], [**20**], [**19**], [**18**], [**22**], [**26**], [**25**], [**17**], [**23**], [**24**], [**21**]), which are very much motivated by integral equations.

This is the standard use of analysis for numerical analysis: we take a classical method from the analysis of partial differential equations, and we turn it into an applicable numerical method.

In the last part of this paper, I give another example of the importation of PDE methods into numerical analysis.

But before sketching this example, I must stress the difference between approximate inverses for theoreticians of PDE's and approximate inverses for numerical analysts: for a theoretician of partial differential equations, a good approximate inverse S of $P(x, D)$ is such that $\mathbb{1} - SP(x, D)$ has strong smoothing properties. For a numerical analyst, a good approximate inverse of A is an S for which there is a fast computational algorithm giving the matrix vector product Sx, and the condition number of SA is small with respect to the condition number of A. It may also have other desirable properties.

In section 5, I move on to a specific numerical problem: how does one approximate numerically the solutions of the heat equation in one dimension, and I show that though this problem looks quite simple, it contains a few delicate points. Spectral approximations are very popular, because their accuracy is limited only by the smoothness of the data. Unfortunately, they involve rather full and very ill-conditioned matrices, and nobody likes to solve a system having this kind of matrix. Therefore, it is tempting to precondition the problem by using another kind of discretization, namely finite elements, which are known to yield a sparse matrix.

In the elliptic case, this is a very classical process: in 1980, Orszag [**32**] suggested preconditioning by finite differences; he gave an argument for the spectral equivalence between the spectral and finite differences stiffness matrices in the case of periodic boundary conditions and a Fourier basis, and stated that this equivalence still holds in many other cases. Haldenwang et al. [**28**] give an argument for spectral equivalence between the stiffness matrices for finite differences and Chebyshev spectral approximation. In [**7**], Canuto and Quarteroni tested a large number of preconditioners for Chebyshev spectral calculations, including preconditioning by finite elements, and gave numerical estimates of the spectral radii of the different numerical methods; in [**11**] Deville and Mund test a variety of finite elements methods for the same problem and in [**12**], they extend their ideas to more general classes of orthogonal polynomials.

In the case of a time dependent equation, there are two ways to solve the problem. One is to apply the method of lines, and then precondition an operator of the form $\mathbb{1} + \Delta t A$, with A a discretization of the Laplacian. Since, most of the time, no error estimation is performed on iterative methods, this source of error will be left out. Moreover, instabilities may come from the approximate character of the resolution; this question is now studied by M. Ribot.

Another strategy is the one embodied by the Residual Smoothing Scheme (RSS), as presented in [**1**].

This is a scheme that has been floating around for some time; the name of Orszag has been mentioned in this respect, but I have been unable to find a precise reference.

In order to prove the stability of RSS for the heat equation, we have to prove the spectral equivalence of two matrices, uniformly with respect to the number of discretization points.

Section 6 describes the strategy used by Magali Ribot [**39**], [**40**] for solving this problem. The result could also have been obtained as a consequence of results by Parter [**34**], who used totally different methods.

The principle of the method of validation is based on the method of the stationary phase, that is on the techniques used for constructing parametrices of elliptic equations.

Therefore, the conclusion, given in Section 7, is that though it is not (yet) very popular, a considerable progress may be expected from importing some very theoretical techniques into numerical analysis, but this must be done in an unprejudiced way. Considering the operators of numerical analysis as some funny pseudo-differential operators would not do, because the analysis of boundary problems for pseudo-differential operators meets with many difficulties, and it may not be a good idea to go in this direction. However, some of the wisdom of pseudo-differential operators has its place in numerical analysis, This is already well-known for the construction of artificial boundary conditions, paraxial approximations, integral equations, multi-polar methods, $\mathcal{H}$ matrices. But it can be probably used in many other places, provided that the spirit is imported, but not necessarily the techniques created for specific situations.

Conversely, the matrices of the numerical analysis of partial differential equations happen to be very rare objects; most probably, if we had known how rare they are, we would not have even tried to solve problems from the numerical analysis of PDE's!

But ignorance is blissful only in innocence. Once the information is out, there is no reason not to look at the question from an algebraic point of view.

Numerical analysis should start a dialogue with contemporary algebra and ask how the matrices of numerical analysis can be characterized on algebraic grounds. There may exist algebraically characterizable classes of matrices among which the matrices of numerical analysis are very common, and for which the methods of resolution are much more efficient than in the general case, even for very large matrices.

2. The simplest numerical problem

The simplest problem of numerical linear algebra is to solve a linear system of d equations with d unknowns. Let A be a $d \times d$ matrix, real or complex, and let b be a $d \times 1$ vector. We want to solve the linear system

$$Ax = b. \tag{2.1}$$

The theory is without mysteries: there exists a unique solution to (2.1) iff A is regular.

Practically, how do we do that?

Let us first kill Cramer's formulas. The calculation of the determinant of a $d \times d$ matrix requires the calculation of $d!$ products of d numbers, hence $(d-1)d!$ multiplications, and then the addition of the $d!$ results, i.e. $d! - 1$ additions, that is for one determinant,

$$d!(d-1) + d! - 1 = (d!)d - 1.$$

We have to calculate $d+1$ such determinants, and to perform d divisions; the total operation count is

$$(d+1)((d!)d - 1) + d \sim d\big((d+1)!\big).$$

For $d = 100$, which is considered as a small matrix, Stirling's formula gives

$$101! \sim 101^{101.5} e^{-101} \sqrt{2\pi}$$

and therefore

$$100 \times 101! \sim 9.4 \times 10^{161}.$$

With a computer working at 1 gigaflops, that is (10^9 floating point operations per second), we can perform

$$10^9 \times 365 \times 86400 \sim 3.5310^{16}\text{operations per year}$$

and therefore, we will need approximately 3×10^{145} years to complete the operation. If the age of the universe is 15 billion years, this means that the solution of this smallish linear system would require 2×10^{135} times the age of the universe.

Not very realistic, which means that the chosen algorithm was stupid, and every first year student knows that.

The alternative algorithms are usually based on Gaussian elimination, with its variants: Cholesky's method, LDU and a few more.

In terms of operation counts, Gaussian elimination, even with pivoting, requires $O(2d^3/3)$ operations, so that our smallish system of 100 equations with 100 unknowns requires $2 \times 10^6/3$ operations, and on the same machine as above, it will take two thirds of a millisecond. This operation count is explained in all books of numerical analysis, for instance in [**42**].

If we consider a more common machine, a modern personal computer that might work at 10 megaflops, the Cramer method takes 2×10^{137} years and the Gauss elimination requires 0.066 second.

It is also well-known that sparsity improves the situation; however, this improvement is not good enough for really large matrices, because of *fill-in*: it is known that, in general, in the factors L and U of the LU decomposition, the portion of line between the most removed non zero element and the diagonal is filled with non zero coefficients; in fact, there may be even more non zero elements in L and U, depending on the exact distribution of the non zero terms in $A = LU$. In consequence, when a linear system is obtained from a finite element discretization, the numbering of the elements and of the vertices may play a very important rôle.

By the way, a matrix by vector multiplication requires about $2d^2$ floating point operations. A naïve analysis seems to show that one cannot do better than that number of operations for a full matrix. But this is not true if the matrix has a special structure: a very well-known example of that situation is the case of an FFT. A discrete Fourier transform is a linear transformation which can be performed most efficiently through the Fast Fourier Transform algorithm (FFT): the operation count is then $O(d \log d)$ instead of $2d^2$.

Numerical analysts are not interested in solving baby problems such as a system of 100 equations with 100 unknowns. A large problem is for instance a problem with $d = 10^6$ unknowns.

This would correspond for instance to the discretization of a stationary partial differential equation in a cube, with 100 discretization points or modes in each direction.

In this case, Gaussian elimination requires $2 \times 10^{18}/3$ floating point operations, and on the previous machine, the computing time would be of 21 years and about two months.

More reasonable but not yet fast enough.

This is the reason why the so-called iterative methods have been developed. The principle of an iterative method is the following: instead of solving by a *direct* method such as Gaussian elimination, which is supposed to give the result up to computer arithmetic error, also called round-off error, we agree to be satisfied with the construction of a sequence of approximations x_n to the solution, where we decide to stop when a certain error criterion is satisfied.

The simplest methods are well known and easy to describe, but not very efficient. For instance, Jacobi's method is defined by the data of an initial guess x^0; given x^n, we calculate the solution of

$$\sum_{j=1}^{i-1} A_{ij} x_j^n + A_{ii} x_i^{n+1} + \sum_{j=i+1}^{d} A_{ij} x_j^n = b_i, \quad 1 \le i \le d. \tag{2.2}$$

In other words, in each row of the system, we think that the off-diagonal terms are data and we just solve for the diagonal terms, which is of course algorithmically very easy. A slightly better alternative is Gauss-Seidel's method, where we think that at row i of the system, the data are the x_k^{n+1} for $k \le i - 1$ and the x_k^n for

$k \geq i+1$: we just have to solve for the term x_{ii}^{n+1} and once again, this is very easy:

$$\sum_{j=1}^{i-1} A_{ij} x_j^{n+1} + A_{ii} x_i^{n+1} + \sum_{j=i+1}^{d} A_{ij} x_j^n = b_i, \quad 1 \leq i \leq d. \tag{2.3}$$

The difference between (2.2) and (2.3) is that in the former method, we use only data from the previous step $x_1^n, \dots x_{i-1}^n$ at line i, while in the latter, we use data which have just been calculated in the present step $x_1^{n+1}, \dots, x_{i-1}^{n+1}$ and data from the previous step.

In fact, these elementary iterative methods perform badly in systems derived from the discretization of partial differential equations, because such systems are generally not strongly diagonally dominant. This is very easy to see for a finite element method and an elliptic problem of order 2, determined by the bilinear form

$$\begin{aligned} a(u,v) = \int_\Omega \Big(& \sum_{i,j=1}^{N} a_{ij}(x) \partial_i u(x)\, \partial_j v(x) \\ & + \sum_{i=1}^{N} b_i(x) \partial_i(x) u(x)\, v(x) + c(x) u(x) v(x) \Big)\, \mathrm{d}x. \end{aligned}$$

If ϕ_k is a basis function whose support does not meet the boundary of the domain Ω, we observe that

$$\sum_l a(\phi_k, \phi_l) = a(\phi_k, \sum_{l \in I(k)} \phi_l).$$

where $I(k)$ is the set of indices l such that the support of ϕ_l intersects the support of ϕ_k. On the support of ϕ_k,

$$\sum_{l \in I(k)} \phi_l = 1$$

and therefore,

$$\sum_l a(\phi_k, \phi_l) = a(\phi_k, 1) = \int c(x) \phi_k(x)\, \mathrm{d}x$$

and therefore, the sum of the coefficients along the line k is an $O(h^N)$, while we expect that each individual coefficient will be of size $O(h^{N-2})$.

However, these methods are interesting building blocks for other methods, in particular multi-grid methods.

Of course, one may well wonder why such a primitive method as (2.2) of (2.3) might give any kind of result whatsoever.

Here is a sufficient condition for the convergence of these methods: assume that the matrix A is positive definite and that it decomposes as $A = M - N$; if $M + N^*$ is Hermitian positive definite, then the iterative method defined by $Mx^{n+1} = Nx^n + b$ converges.

The main interest of iterative methods is that many matrices of numerical analysis are sparse, i.e., very few of their coefficients are non zero. Here is an idea of what sparse means, for a low order approximation of a scalar elliptic problem, there are $O(10)$ non vanishing coefficients in spatial dimension 2 and $O(50)$ in spatial dimension 3.

For a precise approximation of the system of three-dimensional elasticity, we might have several hundred non vanishing coefficients per row.

The main idea is that multiplying a sparse matrix by a vector is cheap, provided that we use the sparse structure when coding the multiplication: we just do not want to multiply anything by 0. Observe that this approach means that our linear operators are *not* written as matrices, but as algorithms. In the jargon of numerical analysis, we say that we do not *assemble* matrices, i.e. we never produce a list of their coefficients, we just produce an algorithm which enables us to perform the matrix by vector multiplication.

The standard efficient iterative methods for solving linear systems are the conjugate gradient, if A is Hermitian positive definite, or GMRES, the biconjugate gradient and many others. There are many references which describe them, and the reader is referred for instance to [**41**] for a detailed description of many classes of iterative methods.

Until now, we have not considered another numerical question: what is the size of the error we make when solving a system of linear equations? Here is the answer: assume that δA and δb are perturbations of the data A and b. Provided that $A+\delta A$ is still regular, the perturbed solution $x+\delta x$ satisfies

$$(A+\delta A)(x+\delta x)=b+\delta b. \tag{2.4}$$

If we subtract (2.1) from (2.4), we find

$$\delta x = A^{-1}\big(\delta b-\delta A x-\delta A\delta x\big). \tag{2.5}$$

We apply now the triangle inequality to (2.5), we divide by the vector norm $|x|$, assuming that it does not vanish, and we get

$$\frac{|\delta x|}{|x|}\le \|A^{-1}\|\left(\frac{|\delta b|}{|x|}+\|\delta A\|+\frac{\|\delta A\|\,|\delta x|}{|x|}\right). \tag{2.6}$$

Here $|\ |$ is an arbitrary vector norm and $\|A\|$ is the operator norm of A, i.e. the maximum of $|Ax|$ over the ball $\{|x|\le 1\}$.

We use now the obvious inequality

$$|b|\le \|A\|\,|x| \iff |x|\ge |b|\,\|A\|^{-1}$$

to simplify the right hand side of (2.6):

$$\frac{|\delta x|}{|x|}\le \|A^{-1}\|\,\|A\|\left(\frac{|\delta b|}{|b|}+\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta A\|\,|\delta x|}{\|A\|\,|x|}\right).$$

We introduce the *condition number* of A:

$$\kappa(A)=\|A\|\,\|A^{-1}\|.$$

The condition number depends on the choice of operator norm $\|\ \|$, and hence on the choice of the underlying vector norm $|\ |$. By default, $|\ |$ is the canonical hermitian vector norm.

When A is Hermitian positive definite, and $|\,|$ is the Hermitian norm, $\kappa(A)$ is the ratio of the largest to the smallest eigenvalue of A.

The error on the solution can now be estimated as follows:

$$\left(1-\frac{\|\delta A\|\kappa(A)}{\|A\|}\right)\frac{|\delta x|}{|x|}\le \kappa(A)\left(\frac{|\delta b|}{|b|}+\frac{\|\delta A\|}{\|A\|}\right). \tag{2.7}$$

Assume that $\kappa(A)\|\delta A\|/\|A\|$ is at most equal to $1/2$, and that A and b are known with a relative error of 10^{-m}. In practice, this means that we work with a mantissa of length m in floating point representation. Then, very coarsely, relation (2.7) tells us that the number of correct decimals in the mantissa of x is at least $m-\log_{10}\kappa(A)$.

Of course, this statement is sloppy: it assumes that the relative error on the coefficients is comparable to the relative error on the norm; if this is not true, it is possible to obtain more detailed statements, depending on the particular structure of the matrix under consideration.

Thus, $\kappa(A)$ is the natural object which describes the feasibility of the resolution of a system. For instance, in classical double precision arithmetic, the mantissa has 16 decimal digits. If the condition number is larger than 10^{16}, the numerical solution is probably meaningless, and in double precision, a matrix with condition number larger than 10^{16} is as good as singular

As always in the partly experimental field of numerical analysis, this statement must be taken with a grain of salt. In some special circumstances, we may control better the growth of round-off error, when we do better than applying the triangle inequality. This may happen under appropriate algebraic conditions, as has been shown by Boros *et al.* [**6**] for a special class of matrices, and also by Demmel and Koev [**9**] for a different problem. The basic idea is that the really bad operation is the addition of numbers of opposite sign and comparable magnitude: this is the operation that creates the most relative error. If we can build an algorithm that does not perform this kind of operation, there is much less degradation in the relative error.

This question winds tightly together algebraic and algorithmic preoccupations: can you obtain a well-defined result in exact arithmetic, when you restrict the set of permissible operations? Which of these exact arithmetic expressions are accessible under this type of restrictions?

There is more to the condition number.

Let us consider the simplest iterative method namely Richardson's method given by

$$M(x^{n+1} - x^n) = \alpha^n(b - Ax^n). \tag{2.8}$$

Here, M is a regular matrix: its choice is crucial for the efficiency of the method.

Let us analyze the simplest case: take M to be the identity matrix $\mathbf{1}$, assume that α does not depend on n and that A is Hermitian positive definite. We write (x, y) for the canonical Hermitian scalar product.

The error e^n is $e^n = x^n - x$, and it satisfies the recurrence relation $e^{n+1} = (\mathbf{1} - \alpha A)e^n$. Therefore, we choose the real number α so as to minimize the spectral radius of the matrix $\mathbf{1} - \alpha A$. By a straightforward analysis, we find that the optimal choice is

$$\alpha = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are respectively the smallest and largest eigenvalues of A. Therefore, the spectral radius of $\mathbf{1} - \alpha A$ is

$$\frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)} = \frac{1 - 1/\kappa(A)}{1 + 1/\kappa(A)}. \tag{2.9}$$

The number of iterations necessary to divide the error by 2 is proportional to $\kappa(A)$.

Of course, the method (2.8) is rather crude. There are two ways of improving it.

One way of improving the rate is to precondition the method, i.e. to choose a matrix M in (2.8) such that the condition number of $M^{-1}A$ is much smaller than the condition number of A.

Remember that we assumed A to be Hermitian positive definite. If M happens to be Hermitian positive definite, an analogous analysis can be performed using the vector norm $|x|_A = \left(x^\top Ax\right)^{1/2}$; this works because $M^{-1}A$ is Hermitian relative to the scalar product deduced from the norm $|\ |_A$. Now, the rate of convergence depends on the condition number of $M^{-1}A$, the formula being analogous to (2.9).

The other method is to perform a conjugate gradient algorithm, instead of a Richardson algorithm. The conjugate gradient can be written

$$\begin{aligned}
&p^0 = r^0 = b - Ax^0;\\
&\text{and while } p^k \text{ does not vanish:}\\
&q^k = Ap^k,\\
&\alpha^k = \frac{(r^k, p^k)}{(q^k, p^k)},\\
&x^{k+1} = x^k + \alpha^k p^k,\\
&r^{k+1} = r^k - \alpha^k q^k,\\
&\beta^{k+1} = \frac{(r^{k+1}, r^{k+1})}{(r^k, r^k)},\\
&p^{k+1} = r^{k+1} + \beta^{k+1} p^k.
\end{aligned} \tag{2.10}$$

The rate of convergence of algorithm (2.10) is at most

$$\frac{\sqrt{\lambda_{\max}(A)} - \sqrt{\lambda_{\min}(A)}}{\sqrt{\lambda_{\max}(A)} + \sqrt{\lambda_{\min}(A)}} = \frac{1 - 1/\sqrt{\kappa(A)}}{1 + 1/\sqrt{\kappa(A)}},$$

a result which is proved for instance in [**41**].

Therefore, the number of steps needed to divide the error by 2 is proportional to $\sqrt{\kappa(A)}$, which is much smaller than $\kappa(A)$, if $\kappa(A)$ is large.

Of course, these are upper bounds, but without more precise qualifications, that is all we have.

Moreover, there is a preconditioned version of the conjugate gradient algorithm,

$$\begin{aligned}
&r^0 = b - Ax^0, \quad p^0 = z^0 = M^{-1} r^0,\\
&\text{and while } p^k \text{ does not vanish,}\\
&q^k = Ap^k,\\
&\alpha^k = \frac{(z^k, r^k)}{(q^k, p^k)},\\
&x^{k+1} = x^k + \alpha^k p^k,\\
&r^{k+1} = r^k - \alpha^k q^k,\\
&M z^{k+1} = r^{k+1},\\
&\beta^{k+1} = \frac{(z^{k+1}, r^{k+1})}{(z^k, r^k)},\\
&p^{k+1} = z^{k+1} + \beta^{k+1} p^k.
\end{aligned} \tag{2.11}$$

If M is Hermitian positive definite and the vector norm is $(x, y)_M = x^\mathsf{T} M y$, the rate of convergence of algorithm (2.11) is

$$\frac{1 - 1/\sqrt{\kappa(M^{-1/2} A M^{-1/2})}}{1 + 1/\sqrt{\kappa(M^{-1/2} A M^{-1/2})}}$$

so that we gain a lot by preconditioning the algorithm.

In practice, one must choose the matrix M in such a way that a system with matrix M is easy to solve.

The ideal choice would be to take $M = A$; but this is impossible, since we are really looking for an algorithm which provides the inverse of A. Therefore, we look for a matrix M which approximates A while remaining algorithmically easy to invert.

Beyond solving a system with matrix M, the algorithm (2.11) uses only matrix by vector multiplications, addition of vectors and scalar product of vectors.

Let us give an order of magnitude of these condition numbers. In the case of a second order elliptic problem discretized on an approximately uniform grid in a cube, with 100 discretization points for each dimension, the number of unknowns is 10^6, while the condition number is $O(10^4)$. This condition number can be substantially degraded if the mesh contains very flat elements or if the order of magnitude of the coefficients of the problem varies widely. These orders of magnitude are important: as we shall see below, there are very few $10^6 \times 10^6$ matrices with condition number less than 10^4 or even less than 10^5.

3. How many bad matrices are there?

If we start from another side of the problem, it is legitimate to ask the following question: given a dimension d, what is the relative volume of the set of matrices in $GL_d(\mathbb{R})$ or $GL_d(\mathbb{C})$ whose condition number is larger than x?

This question makes sense, since the condition number is invariant by the transformation $A \mapsto tA$ for all $t \neq 0$. It has been studied first by Jim Demmel [**10**] and the solution given by Alan Edelman [**15**] deserves a description. First, we need the definition of the Frobenius norm of a matrix:

$$\|A\|_F = \left(\sum_{i,j=1}^{d} |A_{ij}|^2 \right)^{1/2} .$$

The Frobenius norm is very easy to understand, since it is simply the Hermitian norm of matrices, seen as vectors in d^2-dimensional space. On the other hand, the operator norm of matrices is a very complicated object. Nevertheless, the following estimate is classical:

(3.1) $$\|A\| \le \|A\|_F \le \sqrt{d}\|A\|.$$

Denote the sphere in Frobenius norm of radius 1 about 0 by $\mathbb{S}^{d^2-1}$. It is equipped with the measure deduced from Lebesgue measure in d^2-dimensional space.

Let us introduce a deformed condition number:

(3.2) $$\kappa_D(A) = \|A\|_F \|A^{-1}\|,$$

which makes the theory easier.

The choice of (3.2) comes from Eckart-Young's theorem [**13**]; this theorem stated that the Frobenius distance from a matrix A to the set of singular matrices is $1/\|A^{-1}\|$. A proof of this theorem can be found for instance in Blum et al's book [**5**].

The relative volume of the set of matrices in $\mathbb{S}^{d^2-1}$ whose condition number κ_D is larger than x is a probability $P(x,d)$:

$$P(x,d) = \frac{|\{A \in \mathbb{S}^{d^2-1} : \quad \kappa_D(A) \ge x\}|}{|\mathbb{S}^{d^2-1}|}$$

and its density is exactly known. In the real case, the density is

$$p(x,d) = \mu x^{1-d^2}(x^2-d)^{d(d+1)/2-2} {}_2F_1\left(\frac{d-1}{2}, \frac{d+2}{2}; \frac{d^2+d-2}{2}; -(x^2-d)\right),$$

where μ is given by

$$\mu = 2d\Gamma\left(\frac{d+1}{2}\right)\Gamma\left(\frac{d^2}{2}\right)\sqrt{\pi}\Gamma\left(\frac{d(d+1)}{2}-1\right).$$

Here ${}_2F_1$ is the Gauss hyper-geometric function. An equivalent for $x \gg \sqrt{d}$ and $d \gg 1$ of the distribution function, i.e. $P(x,d) = \int_{-\infty}^{x} p(y,d)\mathrm{d}y$ is known:

$$P(x,d) \sim \frac{d^{3/2}}{x}.$$

In the complex case, the density formula is much simpler:

$$p(x,d) = 2d(d^2-1)x^{1-2d^2}(x^2-d)^{d^2-2},$$

and there is an exact formula for the probability distribution:

$$P(x,d) = 1 - \left(1 - \frac{d}{x^2}\right)^{d^2-1}. \tag{3.3}$$

Moreover, for $d \gg 1$, $\kappa_D(A) \sim \sqrt{d}\kappa_2(A)/2$, which is a nice observation, since it means that the average ratio of the operator and Frobenius norms is almost in the middle of the interval defined by (3.1).

We combine all these observations and apply them to the case $d = 10^6$ and $\kappa_2(A) = 10^4$; then $x = \kappa_D(A) \sim 5 \times 10^6$. Then, the numerical application of (3.3) gives

$$P(10^4, 10^6) = 1 - \left(1 - \frac{10^6}{25 \times 10^{12}}\right)^{25\times 10^6 \times 4 \times 10^4 - 1} \sim 1 - e^{-4\times 10^4}.$$

Therefore the probability that, in the complex Frobenius sphere $\mathbb{S}^{(10^{12}-1)}$, a matrix has a condition number less than 10^4 is at most $e^{-4\times 10^4} = 10^{-17371}$.

This is such a small probability that if we had not constructed such a matrix in an *ad hoc* fashion, we would have been extremely lucky to find it just by chance. The probability of finding it is the same as the probability that random typing will get a text of 12316 characters completely right using an alphabet of 26 letters plus space. More generally, the set of matrices such that $\kappa(A)$, the ordinary condition number satisfies the inequality

$$1 \ll \kappa(A)/2 \ll d$$

has relative volume in the Frobenius sphere $\exp(-4d^3/\kappa(A)^2)$, which is very small.

To make matters worse, Edelman has also proved in [**14**] that for matrices with independent Gaussian normal coefficients, the average of the logarithm of the condition number is the logarithm of the dimension plus a bounded number, while the condition number itself is so bad that none of its moments is bounded. Since we deal here with random matrices, the moments are taken with respect to the probability measure; in particular, there is no average of the condition number over this class of random matrices.

The practical consequence is that, if we pick "randomly" a $10^6 \times 10^6$ matrix, on average, we will lose 6 decimal places when solving the corresponding linear system. Of course, once again, this is a very coarse statement, and it has to be qualified by all sorts of ifs and buts. Nevertheless, it is a rule of thumb: solving large linear system with no specific information on their structure is a hard problem.

In other words, the matrices of numerical analysis of partial differential equations are very special objects. There is no universal method that could work for large matrices. In the numerical analysis of partial differential equations, we have to use the special algebraic structure of the matrices, and therefore, it would be very interesting to characterize specifically these matrices from an algebraic viewpoint.

4. Preconditioning and parametrices

A preconditioner for a linear system of equations (2.1) is a matrix B such that $\kappa(AB^{-1})$ or $\kappa(B^{-1}A)$ is small relatively to $\kappa(A)$; moreover, there should exist an algorithm of low complexity for solving the system

$$By = c. \tag{4.1}$$

In other words, B^{-1} is an approximate inverse of A which can be realized as a low complexity algorithm.

Constructing inverses of low complexity is an idea that is very close to the construction of a parametrix for a partial differential operator.

Let us explain briefly what a parametrix is. Write $\partial_j = \partial/\partial x_j$ and $D_j = -i\partial_j$. When we have variables x and ξ, the derivation operator with respect to x will be written $D_{j,x} = -\mathrm{i}\partial/\partial x_j$ in order to avoid confusion.

Recall that the Schwartz space $\mathcal{S}(\mathbb{R}^N)$ is the space of infinitely differentiable functions which decrease fast to 0 at infinity as well as their derivatives of all order. If u belongs to this space, its Fourier transform is

$$\hat{u}(\xi) = \int_{\mathbb{R}^N} e^{-\mathrm{i}x\cdot\xi} u(x)\,\mathrm{d}x,$$

and the Fourier inversion formula is

$$u(x) = \frac{1}{(2\pi)^N}\int_{\mathbb{R}^N} \mathrm{e}^{\mathrm{i}x\cdot\xi}\hat{u}(\xi)\,\mathrm{d}\xi.$$

We have the identities

$$\widehat{D_i u} = \xi_i \hat{u}, \quad \widehat{x_i u} = -D_i \hat{u}.$$

A constant coefficient differential operator is a finite sum of the form

$$P(D) = \sum a_\alpha D^\alpha, \tag{4.2}$$

where we have used the multi-index notation

$$\alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{N}^N, \quad \partial^\alpha = \partial_1^{\alpha_1} \dots \partial_N^{\alpha_N}.$$

Provided that we define

$$\xi^\alpha = \xi_1^{\alpha_1} \dots \xi_N^{\alpha_N},$$

the action of $P(D)$ on $\hat{u} \in \mathcal{S}(\mathbb{R}^N)$ can be expressed very simply:

$$\big(\widehat{P(D)u}\big)(\xi) = \sum a_\alpha \xi^\alpha \hat{u}(\xi) = P(\xi)\hat{u}(\xi).$$

Therefore, it is tempting to write the inverse of the operator $P(D)$ by

$$P(D)^{-1}u = \frac{1}{(2\pi)^N}\int_{\mathbb{R}^N} \frac{\hat{u}(\xi)\mathrm{e}^{ix\cdot\xi}\,\mathrm{d}\xi}{P(\xi)}.$$

This formula makes sense immediately if P does not vanish over $\mathbb{R}^N$. When this is not true, a famous theorem of L. Ehrenpreis [**16**] and B. Malgrange [**29**], [**30**], [**31**] shows the existence of a fundamental solution of the operator $P(D)$, i.e. a solution of

$$P(D)u = \delta$$

where δ is the Dirac mass at $0 \in \mathbb{R}^N$. A (very) short proof of this result is given by M. E. Taylor in his book [**44**], pages 33–35. Let us pretend now that we do not know how to find fundamental solutions for constant coefficient operators, and let us take a rather particular case, where the set Z of zeros of P is compact in $\mathbb{R}^N$. So as to avoid all difficulties, we will even assume that P is elliptic; this means that if P_m is the homogeneous part of highest degree m of P, then P_m vanishes only at $0 \in \mathbb{R}^N$.

Let χ be a smooth cut-off function which is equal to 0 in a neighborhood of 0 and to 1 outside of a large ball or radius R about 0. The operator S given by

$$(Su)(x) = \frac{1}{(2\pi)^N} \int_{\mathbb{R}^N} \frac{\chi(\xi)\hat{u}(\xi)\mathrm{e}^{ix\cdot\xi}\,\mathrm{d}\xi}{P(\xi)}$$

is well defined. Moreover, we may calculate $P(D)S$ and $SP(D)$. A straightforward calculation gives

$$D_j Su = \frac{1}{(2\pi)^N} \int_{\mathbb{R}^N} \frac{\chi(\xi)\xi_j\hat{u}(\xi)\mathrm{e}^{ix\cdot\xi}\,\mathrm{d}\xi}{P(\xi)},$$

and therefore

$$u - P(D)Su = \frac{1}{(2\pi)^N} \int_{\mathbb{R}^N} (1 - \chi(\xi))\hat{u}(\xi)\mathrm{e}^{ix\cdot\xi}\,\mathrm{d}\xi.$$

If we let $\hat{\phi}(\xi) = 1 - \chi(\xi)$, we see that $\mathbf{1} - P(D)S$ is simply the convolution operation with ϕ. As $\hat{\phi}$ is a smooth function with compact support, $\mathbf{1} - P(D)S$ is an infinitely smoothing operator.

As $P(D)$ commutes with S, $\mathbf{1} - SP(D)$ is also an infinitely smoothing operator.

What good is this for solving partial differential equations?

Since ϕ is the inverse Fourier transform of a compactly supported smooth function, it is an entire function of exponential type, i.e. it satisfies the estimate

$$|\phi(x + \mathrm{i}y)| \leq C\exp(L|y|), \quad \forall y \in \mathbb{R}^N.$$

Here, we have extended ϕ to $\mathbb{C}^N$ by letting

$$\phi(x + \mathrm{i}y) = \frac{1}{(2\pi)^N} \int_{\mathbb{R}^N} \hat{\phi}(\xi)\mathrm{e}^{\mathrm{i}(x+iy)\cdot\xi}\,\mathrm{d}\xi.$$

By Cauchy-Kowalevskaia theorem, it is possible to construct a solution of

$$P(D)w = \phi$$

and even to take w entire of exponential type. Then, a fundamental solution of $P(D)$ is

$$v = w + K$$

where K is the inverse Fourier transform of $(1 - \chi(\xi))/P(\xi)$. As $(1 - \chi(\xi))/P(\xi)$ increases at most polynomially at infinity, K is a temperate distribution. Therefore, if f is a compactly supported data, $u = v * f$ is well defined, and it is a solution of $P(D)u = f$.

This is not yet very interesting. If we consider now a variable coefficient elliptic operator

$$P(x, D) = \sum_{\alpha} a_\alpha(x)D^\alpha$$

we are tempted to apply the same type of process to find a parametrix, which we assume to be of the form

$$(Su)(x) = \frac{1}{(2\pi)^N} \int_{\mathbb{R}^N} \mathrm{e}^{\mathrm{i}x\cdot\xi}k(x,\xi)\hat{u}(\xi)\,\mathrm{d}\xi.$$

We assume that for each x, $P(x, \xi)$ is elliptic, and we call $P_m(x, \xi)$ the homogeneous part of degree m in ξ. The ellipticity means that for all x, $P_m(x, \cdot)$ vanishes only at $\xi = 0$.

We want to choose k so that $P(x, D)S - \mathbb{1}$ will be an infinitely smoothing operator. We observe immediately that

$$D_{j,x}(\mathrm{e}^{ix\cdot\xi}w(x)) = \mathrm{e}^{ix\cdot\xi}(\xi_j + D_{j,x})w(x).$$

This implies the following identity

$$P(x, D)(\mathrm{e}^{ix\cdot\xi}w(x)) = \mathrm{e}^{ix\cdot\xi}\big(P(\xi + D_x)w\big)(x),$$

and we can decompose $P(x, \xi + D_x)$ into a sum of homogeneous polynomials of degree $m - j$ in ξ and of degree j in D:

$$P(x, \xi + D_x) = P_m(x, \xi) + \sum_{j=1}^{m} P_j(x, \xi, D_x).$$

Assume formally that

$$k(x, \xi) = k_0(x, \xi) + k_1(x, \xi) + k_2(x, \xi) + \dots$$

where k_j will be homogeneous in ξ of degree $n - j$.

We just apply the standard asymptotic argument, by equating the terms of the same degree of homogeneity in ξ:

$$\begin{aligned}
&P_m(x, \xi)k_0(x, \xi) = 1,\\
&P_m(x, \xi)k_1(x, \xi) + P_{m-1}(x, \xi, D_x)k_0(x, \xi) = 0,\\
&P_m(x, \xi)k_2(x, \xi) + P_{m-1}(x, \xi, D_x)k_1(x, \xi) + P_{m-1}(x, \xi, D_x)k_2(x, \xi) = 0,
\end{aligned}$$

and so on. Of course, we cannot really invert P_m because of its zero set, but we are going to pretend that we can; then

$$\begin{aligned}
&k_0(x, \xi) = 1/P_m(x, \xi)\\
&k_1(x, \xi) = -(P_{m-1}(x, \xi, D_x)K_0(x, \xi))/P_m(x, \xi), \text{ and so on,}
\end{aligned}$$

so that, at least formally, we can write all the terms k_j of the expansion. Of course, many tricks are needed to validate this expansion into a *bona fide* kernel, and this is the crux of the classical theory of pseudo-differential operators. This presentation of the beginning of the theory has followed the first section of chapter 1 of Trèves' book [**45**].

When we apply $\mathbb{1} - P(x, D)k_0(x, D)$ to a function u whose derivatives of order at most k are square integrable, we obtain a function whose derivatives of order at most $k + 1$ are square integrable; if we define indeed

$$(S_0u)(x) = \frac{1}{(2\pi)^N}\int_{\mathbb{R}^N} \mathrm{e}^{\mathrm{i}x\cdot\xi}k_0(x, \xi)\hat{u}(\xi)\,\mathrm{d}\xi,$$

then

$$\big((\mathbb{1} - P(x, D)S_0)u\big)(x) = -\frac{1}{(2\pi)^N}\int_{\mathbb{R}^N}\left(\sum_{j=1}^{m} P_j(x, \xi, D_x)k_0(x, \xi)\right)\hat{u}(\xi)\,\mathrm{d}\xi. \tag{4.3}$$

But $P_j(x, \xi, D_x)$ is of degree $m - j$ with respect to ξ and $k_0(x, \xi)$ is of degree $-m$ in ξ. Therefore, if the x derivatives do not contribute any messy terms, we have to expect that, at infinity, the kernel in the integral (4.3) contributes terms which decrease as $|\xi|^{-1}$. Therefore, it is possible to prove that if u belongs to the Sobolev space H^m, i.e. its derivatives of order at most m are square integrable, then $(\mathbb{1} - P(x, D)S_0)u$ belongs to H^{m+1}.

More generally, $\mathbb{1} - P(x, D)(k_0 + \dots + k_j)$ gains j degrees of smoothness.

I have carefully hidden the technicalities of the construction of parametrices. Let it just be known that the method of stationary phase plays here a prominent rôle.

It is not really possible to infer directly from this process a preconditioner for a given variable coefficient operator. Nevertheless, the fundamental solution obtained in the first part of the present section yields a Green function for the constant coefficient operator. Green functions are very much used when solving boundary problems in homogeneous media, since simple and double layers can be used to treat the boundary conditions.

In the case of variable coefficients, parametrices have not been used for preconditioning; however, the idea of wavelet preconditioning for variable coefficients elliptic operators has been used by Piquemal and Liandrat [**37**]. Once the technology of wavelets has been developed, theoretical ideas from PDE's become much easier to implement numerically, though the treatment of boundary conditions is not yet very satisfactory.

The important message is that the construction of approximate inverses is a classical method in analysis, and therefore, it is not surprising that it is used in numerical analysis. A striking and recent example is the theory of $\mathcal{H}$-matrices by Hackbusch and coworkers ([**4**], [**20**], [**19**], [**18**], [**22**], [**26**], [**25**], [**17**], [**23**], [**24**], [**21**]): they define a class of matrices for which there is a fast algorithm which encodes the matrix by vector multiplication. The inverse of an $\mathcal{H}$-matrix is not an $\mathcal{H}$-matrix, but an $\mathcal{H}$ matrix possesses an approximate inverse which is an $\mathcal{H}$-matrix. The finite element discretizations of elliptic problems and the matrices obtained from integral equations happen to be $\mathcal{H}$-matrices.

5. A specific problem

Most often, in numerical analysis, the error analysis does not involve the influence of preconditioners or of the criteria used to stop iterations, and there might be food for thought in this area. Even more striking is the fact that the construction of preconditioners is more art than science. I believe that science is more powerful than art, and that it must go beyond the ideas of the artist in order to define systematic strategies for tackling problems. One should try to analyze rigorously methods which have been proposed and whose efficiency is demonstrated in practice: the hope is that a scientific analysis will let us understand better the causes of and the limits to efficiency.

Let us take therefore an extremely simple problem, namely the one-dimensional heat equation:

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \quad x \in (0,1), \quad t \geq 0, \tag{5.1}$$

with Dirichlet boundary conditions and initial data u_0 at time $t = 0$.

In order to discretize (5.1), we rewrite it as a variational problem. For this purpose, we multiply equation (5.1) by a function $v(x)$, and we integrate by parts over $(0,1)$. If we choose a function v which vanishes at 0 and 1, the integrated part goes away and we are left with

$$\int_0^1 \frac{\partial u}{\partial t}(x,t)v(x)\,\mathrm{d}x + \int_0^1 \frac{\partial u}{\partial x}(x,t)\frac{\partial v}{\partial x}(x)\,\mathrm{d}x = 0. \tag{5.2}$$

Of course, we have to say a few things on functional analysis; let $H_0^1(0,1)$ be the space of square integrable functions u, with square integrable derivative, such that $u(0)$ and $u(1)$ vanish. It so happens that square integrable functions whose derivative is square integrable are almost everywhere equal to continuous functions, and therefore, this definition makes sense.

In relation (5.2), u must be continuous from $\mathbb{R}^+$ to $H_0^1(0,1)$ with square integrable derivative over $(0,1)\times\mathbb{R}^+$ and the equality must hold for all $v \in H_0^1(0,1)$.

Then, in order to discretize, we choose a finite dimensional space of functions V included in $H_0^1(0,1)$, we replace in (5.2) the time derivative by a finite difference, and we write the following formulation:

$$u^{n+1} \in V \text{ and for all } v \in V, \tag{5.3}$$
$$\int_0^1 \frac{u^{n+1}-u^n}{\Delta t} v \,\mathrm{d}x + \int_0^1 \frac{\mathrm{d}u^n}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x = 0.$$

Since V is finite dimensional, we just have to decompose this problem on a basis of V; denoting by M the mass matrix, i.e. the matrix of the bilinear form

$$V \times V \to \mathbb{R}, \tag{5.4}$$
$$(u,v) \mapsto \int_0^1 uv\,\mathrm{d}x,$$

and by K the stiffness matrix, which is the matrix of the bilinear form

$$V \times V \to \mathbb{R}, \tag{5.5}$$
$$(u,v) \mapsto \int_0^1 u'v'\,\mathrm{d}x,$$

the relation (5.3) can be rewritten

$$M\frac{U^{n+1}-U^n}{\Delta t} + KU^n = 0, \tag{5.6}$$

where U^n is the vector of coordinate of u^n at the discrete time $n\Delta t$.

In the nice cases where M happens to be diagonal, the resolution of (5.6) is trivial, since it requires only a matrix by vector multiplication, a vector addition and a multiplication by a scalar.

Unfortunately, the method (5.6) is very inefficient, because unless the following condition is satisfied:

$$\Delta t \lambda_{\max}(M^{-1}K) \le 2 \tag{5.7}$$

the numerical solution will develop exponentially increasing oscillations. The explanation of this behavior is as follows: if U^n is an eigenvector of $M^{-1}K$ corresponding to the largest eigenvalue λ of this operator, we will have the identity

$$U^{n+1} = U^n - \Delta t M^{-1}KU^n = (1-\lambda\Delta t)U^n,$$

and therefore, if (5.7) does not hold, the magnitude of U^n is multiplied by a negative number of absolute value larger than 1. Since any round-off error is susceptible of triggering the development of a component in the direction of an eigenvector corresponding to the largest eigenvalue of $M^{-1}K$, we see why the choice (5.7) is necessary.

However, the largest eigenvalue of $M^{-1}K$ diverges when the precision of the discretization tends to 0.

Let us see why this is true. We choose for instance V to be the space of continuous functions over $[0,1]$ whose restriction to each interval $[j/J,(j+1)/J]$ is a polynomial of degree at most 1; define the following basis of V:

$$\phi_j(x) = \begin{cases} 1 - J|x-j/J| & \text{if } |x-j/J| \le 1/J, \\ 0 & \text{otherwise,} \end{cases}$$

Instead of exact integration, we use the trapezium formula in (5.4); then relation (5.6) can be written

$$\frac{U^{n+1}-U^n}{\Delta t}+J^2\begin{pmatrix}2 & -1 & 0 & \dots & 0\\ -1 & 2 & -1 & \ddots & 0\\ 0 & -1 & 2 & \ddots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ 0 & 0 & 0 & \dots & 2\end{pmatrix}=0; \tag{5.8}$$

here, the matrix has $J-1$ rows and $J-1$ columns.

The reader will check that

$$v_i=\sin(i(J-1)\pi/J)$$

is an eigenvector of the matrix appearing in (5.8); its corresponding eigenvalue is $4J^2\sin^2(\pi(J-1)/2J)$, and it can be shown that it is the largest eigenvalue of this matrix; therefore, if $\Delta x=1/J\ll 1$ is the mesh size, the condition (5.7) is satisfied only if

$$\Delta t\Delta x^2\sin^2(\pi(J-1)/2J)\le 1,$$

which is equivalent to

$$\Delta t\le\Delta x^2/2,$$

up to higher order terms in Δx.

The alternative to this sorry situation is to use an implicit scheme, i.e., instead of (5.6), we use

$$M\frac{U^{n+1}-U^n}{\Delta t}+KU^{n+1}=0. \tag{5.9}$$

The price to pay is that we have to solve a system with matrix $M+\Delta tK$. For a small problem such as the one-dimensional problem considered here, it cheap and easy. But in higher dimension, this is a time-consuming process, and therefore it is important to find a fast alternative, which could also include non uniform meshes, non constant coefficients, and so on.

When the stiffness matrix has nice algebraic properties, such as the matrix written in (5.8), or its higher dimensional analogues, it is possible to apply adapted strategies, such as a fast Fourier transform. But, if the coefficients are not constant, or if the mesh is not uniform, this strategy breaks down, and we are left with a bear of a problem.

These considerations make natural the idea of the residual smoothing scheme (RSS) of Averbuch, Cohen and Israeli [**1**]; given a matrix K_1 which is algorithmically simpler than K but resembles K enough to make the process stable, replace (5.9) by

$$M\frac{U^{n+1}-U^n}{\Delta t}+KU^n+\tau K_1(U^{n+1}-U^n)=0. \tag{5.10}$$

Here τ is a positive parameter that has to be tuned in order to make the method efficient and stable: if τ is large, the precision is poor; if τ is small, the stability is destroyed.

The advantage of this method is that the only systems we have to solve have matrix $M+\tau\Delta tK_1$.

In [**38**], we have proved that if M is the identity matrix, K and K_1 are hermitian positive definite, and τ is large enough, (5.10) is stable; the number τ just has to be larger than or equal to

$$\min_{x \neq 0} \frac{x^{\mathsf{T}} K x}{2 x^{\mathsf{T}} K_1 x}.$$

If this bound is independent of the number of discretization point, RSS is unconditionally stable.

The proof is contained in [**38**], and the norm used to describe the stability is the energy norm, i.e. $\|x\|_A = (x^{\mathsf{T}} A x)^{1/2}$.

The specific question that I would like to describe now is related to the so-called spectral methods. These methods use for V a space of polynomials instead of a space of piecewise polynomial functions. For convenience, we will work now on the interval $(-1, 1)$ instead of $(0, 1)$. In simple geometries, these are very efficient methods, because they are potentially of infinite precision. In other words, the order of the discretization error is bounded only by the smoothness of the data, and for infinitely differentiable functions, the discretization error decreases faster than any power of the number of discretization points.

However, spectral methods have two substantial defects: the stiffness matrices are not sparse, and the size of the largest eigenvalue of the stiffness matrix is expected to be larger than J^4, J being the degree of the polynomial. On the other hand, when they are written under an appropriate form, spectral methods are also collocation methods. This means that it is possible to find points ξ_j, $1 \leq \xi_j \leq J - 1$ where (5.6) is equivalent to

$$\text{(5.11)} \qquad \frac{u^{n+1}(\xi_j) - u^n(\xi_j)}{\Delta t} - \frac{\partial^2 u^n}{\partial x^2}(\xi_j) = 0.$$

The reader should be reminded at this point that u^n and u^{n+1} are polynomials. In other words, at the nodes ξ_j, we just write a point-wise equality, hence the name *collocation*. Moreover, the mass matrix is diagonal. In fact, it is not necessarily a good choice to solve (5.6) under the form (5.11) and it can be argued that the situation looks better under the form (5.6), since there, M and K are positive definite matrices. Some obvious discretizations of (5.11) do not have this property.

But if we want to apply an implicit method for solving (5.9), we fall back onto the question of preconditioning.

Some nice propositions have proved effective: in the case of Dirichlet boundary conditions, we know what the ξ_j, $1 \leq j \leq J - 1$ should be: they are the zeroes of the derivative of the Legendre polynomial of degree J. Recall that the Legendre polynomials are the orthogonal polynomials over $[-1, 1]$, relatively to the weight 1. The idea developed by Orszag [**32**], Canuto and Quarteroni [**7**], Deville and Mund [**12**], [**11**] is to precondition the spectral method by a finite difference or a finite element method whose nodes are those of the spectral method; they are also the nodes of a Gauss-Lobatto-Legendre quadrature formula. This means that we choose V_1 to be a space of continuous and piecewise polynomial functions, the pieces being bounded by the ξ_j and ± 1. The corresponding stiffness matrix will be denoted by K_1.

For many years, this brand of preconditioning was used without possessing a mathematical proof of any of the properties of the method, the simplest being that K_1 and K ought to be spectrally equivalent independently of the number of discretization points.

A series of articles of Parter [**33**], [**34**], [**35**] proves this uniform spectral equivalence and many other results of interest; the method of proof depends strongly on very detailed results in the theory of orthogonal polynomials, which were deduced from results of Gatteschi; these are obtained by Sturm sequences type techniques and comparison of solutions of ordinary differential equations.

When we tried to understand the RSS for a spectral method, with a finite element preconditioning, we were unaware of the results of Parter, and we studied in detail at the asymptotics used by the specialists of spectral methods. We could not find in Szegő's bible of results of orthogonal polynomials [**43**], the asymptotics we needed, and therefore, we embarked on a very technical enterprise: give asymptotics for the Legendre polynomials and the zeroes of the derivative of Legendre polynomials, with error estimates. This is a work that Magali Ribot accomplished [**39**], [**40**].

6. Strategies and how to validate them

In order to explain better the strategies, I have to be more specific than in the previous section. Let L_N be the N-th Legendre polynomial and let $\xi_1, \dots \xi_{N-1}$ be the zeroes of L'_N. They all belong to the open interval $(-1, 1)$ and they will be arranged in increasing order. Define $\xi_0 = -1$, $\xi_N = 1$.

The basis functions will be the Lagrange basis function over $\xi_0, \dots, \xi_N$ so that they are given as

$$\psi_j(x) = \prod_{\substack{0 \le i \le N \\ i \ne j}} \frac{(x - \xi_i)}{(\xi_j - \xi_i)}, \quad 0 \le j \le J$$

Define also the numbers

$$\begin{cases} \rho_0 = \rho_N = \dfrac{2}{N(N+1)}, \\ \rho_j = \dfrac{2}{N(N+1)L_N^2(\xi_j)}, \quad 1 \le j \le N-1. \end{cases} \tag{6.1}$$

Then, the spectral mass matrix is

$$M_S = \begin{pmatrix} \rho_1 & 0 & \dots & 0 \\ 0 & \rho_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \rho_{N-1} \end{pmatrix},$$

while the coefficients of the stiffness matrix K_S are given by

$$(K_S)_{i,j} = \begin{cases} \dfrac{4}{N(N+1)L_N(\xi_i)L_N(\xi_j)(\xi_i - \xi_j)^2}, & \text{if } i \ne j, \\ \dfrac{2}{3(1-\xi_i^2)L_N(\xi_i)^2}, & \text{if } i = j. \end{cases}$$

The matrix K_S is full; in dimension 2, the spectral discretization of Laplace operator does not yield a full matrix, since the stiffness matrix will be of the form $M_S \otimes K_S + K_S \otimes M_S$. The structure of this matrix is shown in Figure 14. The number of non vanishing elements per line is equivalent to $2J$ for a $(J-1)^2 \times (J-1)^2$ matrix. For a more complicated elliptic operator, in particular one which would involve cross-derivatives, the matrix would indeed be full.

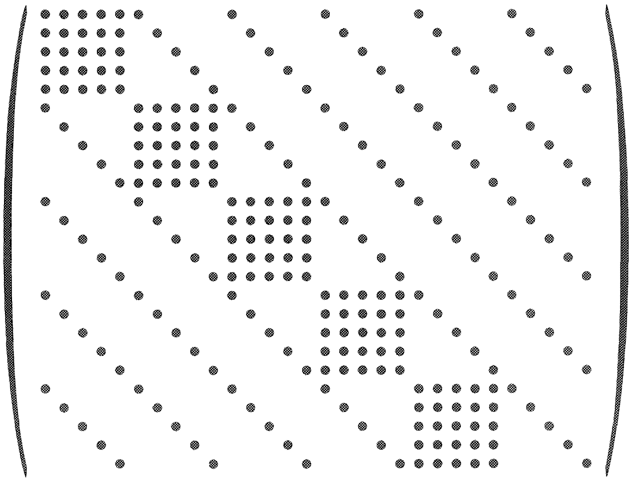


FIGURE 14. The structure of the matrix $M_S \otimes K_S + K_S \otimes M_S$ for $J = 6$. Possibly non vanishing elements are indicated by a black circle.

The mass matrix for finite elements with nodes at the ξ_j, using numerical integration is

$$M_S = \begin{pmatrix} (\xi_2+1)/2 & 0 & \dots & 0 \\ 0 & (\xi_{i+1}-\xi_{i-1})/2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & (1+\xi_{J-2})/2 \end{pmatrix},$$

and the coefficients of the finite elements stiffness matrix, which is tridiagonal, are

$$(K_F)_{ij} = \begin{cases} \dfrac{1}{\xi_i - \xi_{i-1}} + \dfrac{1}{\xi_{i+1} - \xi_i} & \text{if } i = j, \\ \dfrac{1}{\xi_{i-1} - \xi_i} & \text{if } i = j+1, \\ \dfrac{1}{\xi_i - \xi_{i+1}} & \text{if } i = j-1. \end{cases}$$

Let us describe now the RSS scheme with preconditioning by finite elements. In the method (5.6), we let $V^n = M_S^{1/2} U^n$, so that the equation in V^n becomes

$$\frac{V^{n+1} - V^n}{\Delta t} + M_S^{-1/2} K_S M_S^{1/2} V^n = 0.$$

The corresponding RSS scheme is then

$$\frac{V^{n+1} - V^n}{\Delta t} + \tau M_F^{-1/2} K_F M_F^{1/2} (V^{n+1} - V^n) + M_S^{-1/2} K_S M_S^{1/2} V^n = 0. \tag{6.2}$$

Observe that this method is not difficult to implement, since (6.2) can be rewritten in the coordinates $W^n = M_F^{-1/2} V_n$ as

$$(M_F + \Delta t K_F)(W^{n+1} - W^n) = -\Delta t M_F^{1/2} M_S^{-1/2} K_S M_S^{-1/2} M_F^{1/2} W^n.$$

As M_F and M_S are diagonal matrices, the bulk of the numerical work consists in solving systems with matrix $M_F + \Delta t K_F$, which are much faster to solve than the system with matrix $M_S + \Delta t K_S$.

Therefore, the question is to prove that $M_F^{-1/2} K_F M_F^{-1/2}$ is spectrally equivalent to $M_S^{-1/2} K_S M_S^{-1/2}$, uniformly with respect to J.

The first step, which is due to Parter and Rothman [**36**], is to show that K_F and K_S are spectrally equivalent, uniformly with respect to the number of discretization points.

Now, according to [**39**], the problem boils down to proving that

$$M_S^{1/2} M F^{-1/2} K_F M_F^{-1/2} M_S^{1/2}$$

is spectrally equivalent to K_F. Let us define the discrete H^1 norm by

$$\|U\|_{H^1} = \left(\sum_{j=0}^{J-1} \frac{|U_{j+1} - U_j|^2}{\xi_{j+1} - \xi_j} \right)^{1/2} ;$$

then it is equivalent to prove the above spectral equivalence or to prove that $M_F^{-1/2} M_S^{1/2}$ and its inverse are bounded in discrete H^1 operator norm, uniformly with respect to the discretization parameter.

The diagonal elements of $M_F^{-1} M_S$ are called σ_k and they are given by the explicit formula

$$\sigma_k = \frac{2\rho_k}{\xi_{k+1} - \xi_{k-1}}$$

with ρ_k being defined at (6.1).

Define

$$\mu_k = \frac{2 - |\xi_k| - |\xi_{k+1}|}{\xi_{k+1} - \xi_k} \left| \frac{1}{\sigma_{k+1}} - \frac{1}{\sigma_k} \right|^2 .$$

Then, another reduction performed in [**39**] implies that it suffices to bound the sum

$$\sum_{k=0}^{J-1} \mu_k \tag{6.3}$$

in order to obtain the desired result. Thus, it suffices to have precise asymptotics of the Legendre polynomials and their derivatives, in order to conclude.

The Legendre polynomials belong to the so-called family of ultra-spherical polynomials: $L_J = P_J^{(1/2)}$, and their derivatives are also ultra-spherical polynomials:

$$\frac{d}{dx} P_J^{(1/2)}(x) = P_{J-1}^{(3/2)}(x), \quad \frac{d^2}{dx^2} P_J^{(1/2)}(x) = 3P_{J-2}^{(5/2)}(x).$$

It turns out that the literature, and more precisely Szegő's book [**43**] contains asymptotics of the zeroes of ultra-spherical polynomials in the following regions:

(1) the first K zeroes can be related to the zeroes of an appropriately scaled Bessel function, and the error estimate depends on K;
(2) the zeroes indexed by j for $\alpha J \le j \le (1-\alpha)N$ can be related to the zeroes of trigonometric functions, for some $\alpha > 0$.

There was no information in Szegő's book on the intermediate region. Therefore, the option followed in [**40**] was to write the classical integral representation formula

$$P_J^{(\lambda)}(x) = \frac{2^{1-2\lambda}}{(\Gamma(\lambda)^2)} \frac{\Gamma(J+2\lambda)}{J!} \int_0^\pi \left(x + \mathrm{i}\sqrt{1-x^2}\cos\phi\right)^J \sin^{2\lambda-1}\phi \, \mathrm{d}\phi. \tag{6.4}$$

The principle of the asymptotics is to apply the method of the stationary phase to the representation (6.4). Some really technical work is required, first because the phase is *not* of the form $\mathrm{i}Jf(x,\phi)$, with a real valued f. The function f in the phase has a positive imaginary part. The second reason for the technical difficulties is that some very precise asymptotics are needed in order to conclude that the sum (6.3) is bounded: it is necessary to get three terms for each of $P_J^{(1/2)}$ and $P_J^{(3/2)}$, two terms for $P_j^{(5/2)}$ and one for $P_J^{(7/2)}$ Moreover, in order to get all these terms, one has to differentiate composite functions several times, and this is a process that can be qualified as messy, though it is possible to organize it with the help of trees. A tree formulation of Faá di Bruno's theorem can be found in [**27**], and the process was already known to Cayley [**8**].

Conversely, once one knows that it suffices to write a clean asymptotic expansion using the stationary phase method, the strategy is clear and is easy to reproduce in other cases. It could also probably be transferred to symbolic computation codes, provided that these softwares gave clean error estimate, in all the parameters of the problem.

7. Where the snake is eating its own tail

Though it is not (yet?) very popular, progress may be expected from importing some very theoretical techniques into numerical analysis, but this must be done in an unprejudiced way. Considering the operators of numerical analysis as some funny pseudo-differential operators would not do, because the analysis of boundary problems for pseudo-differential operators meets with many difficulties, and it may not be a good idea to go in this direction. However, some of the wisdom of pseudo-differential operators has its place in numerical analysis, This is already well-known for the construction of artificial boundary conditions and paraxial approximations. But there is more, in particular for constructing and analyzing preconditioners.

Conversely, the matrices of the numerical analysis of partial differential equations happen to be very rare objects, and there is probably an algebraic reason why this should be so. Therefore, numerical analysis should turn to modern algebra and ask the question of recognizing and analyzing the very special structure of the matrices of numerical analysis.

It is somewhat sobering to observe that the inclination and the culture of mathematicians play such a rôle in the choice of mathematical strategies. With Magali Ribot, we looked in the direction of the stationary phase method, because it belonged to my culture, and being the senior author, I steered the junior author in the direction I understood. But other methods did work as well, as Parter's results show.

As the strategy used for tackling a given problem is highly dependent on the background of the authors who use it, it is sensible to expand the set of tools and the culture of people who deal with applied problems. We may not be always in the situation where a high-brow and a low-brow method both work.

Beyond that, I believe that realistic numerical analysis has to leave its well-ploughed furrows, and use unusual mathematical techniques — not for their own sake, but because there may be a large number of problems, waiting for us out there, which will have very nice solutions if we incorporate another culture into the standard culture of numerical analysis. Conversely, I believe also that many mathematicians, who are not usually interested in numerical analysis, would find there some fascinating questions, provided that the translation to their language is properly performed.

In particular, there are many problems which are highly non commutative, others which seem to require a good command of algebra, and still others which mix geometry together with analysis. Their presentation would require a much longer paper to be substantiated and will be left for later.

References

[1] A. Averbuch, A. Cohen, and M. Israeli. A stable and accurate explicit scheme for parabolic evolution equations. http://www.ann.jussieu.fr/~cohen/para.ps.gz, 1998.

[2] A. Bamberger, B. Engquist, L. Halpern, and P. Joly. Higher order paraxial wave equation approximations in heterogeneous media. *SIAM J. Appl. Math.*, 48(1):129–154, 1988.

[3] A. Bamberger, B. Engquist, L. Halpern, and P. Joly. Parabolic wave equation approximations in heterogenous media. *SIAM J. Appl. Math.*, 48(1):99–128, 1988.

[4] Mario Bebendorf and Wolfgang Hackbusch. Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients. *Numer. Math.*, 95(1):1–28, 2003.

[5] Lenore Blum, Felipe Cucker, Michael Shub, and Steve Smale. *Complexity and real computation*. Springer-Verlag, New York, 1998. With a foreword by Richard M. Karp.

[6] Tibor Boros, Thomas Kailath, and Vadim Olshevsky. Pivoting and backward stability of fast algorithms for solving Cauchy linear equations. *Linear Algebra Appl.*, 343/344:63–99, 2002. Special issue on structured and infinite systems of linear equations.

[7] Claudio Canuto and Alfio Quarteroni. Preconditioned minimal residual methods for Chebyshev spectral calculations. *J. Comput. Phys.*, 60(2):315–337, 1985.

[8] Arthur Cayley. On the theory of the analytical forms called trees. *Phil. Magazine*, XIII:172–176. Mathematical Papers, Vl. 2., Nr. 152, p. 475.

[9] James Demmel and Plamen Koev. Necessary and sufficient conditions for accurate and efficient rational function evaluation and factorizations of rational matrices. In *Structured matrices in mathematics, computer science, and engineering, II (Boulder, CO, 1999)*, volume 281 of *Contemp. Math.*, pages 117–143. Amer. Math. Soc., Providence, RI, 2001.

[10] James W. Demmel. The probability that a numerical analysis problem is difficult. *Math. Comp.*, 50(182):449–480, 1988.

[11] M. Deville and E. Mund. Chebyshev pseudospectral solution of second-order elliptic equations with finite element preconditioning. *J. Comput. Phys.*, 60(3):517–533, 1985.

[12] M. O. Deville and E. H. Mund. Finite-element preconditioning for pseudospectral solutions of elliptic problems. *SIAM J. Sci. Statist. Comput.*, 11(2):311–342, 1990.

[13] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

[14] Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM J. Matrix Anal. Appl.*, 9(4):543–560, 1988.

[15] Alan Edelman. On the distribution of a scaled condition number. *Math. Comp.*, 58(197):185–190, 1992.

[16] Leon Ehrenpreis. Solution of some problems of division. I. Division by a polynomial of derivation. *Amer. J. Math.*, 76:883–903, 1954.

[17] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij. $\mathcal{H}$-matrix approximation for elliptic solution operators in cylinder domains. *East-West J. Numer. Math.*, 9(1):25–58, 2001.

[18] Ivan P. Gavrilyuk, Wolfgang Hackbusch, and Boris N. Khoromskij. $\mathcal{H}$-matrix approximation for the operator exponential with applications. *Numer. Math.*, 92(1):83–111, 2002.

[19] L. Grasedyck, W. Hackbusch, and B. N. Khoromskij. Solution of large scale algebraic matrix Riccati equations by use of hierarchical matrices. *Computing*, 70(2):121–165, 2003.

[20] Lars Grasedyck and Wolfgang Hackbusch. Construction and arithmetics of $\mathcal{H}$-matrices. *Computing*, 70(4):295–334, 2003.

[21] W. Hackbusch. A sparse matrix arithmetic based on $\mathcal{H}$-matrices. I. Introduction to $\mathcal{H}$-matrices. *Computing*, 62(2):89–108, 1999.

[22] W. Hackbusch and S. Börm. Data-sparse approximation by adaptive $\mathcal{H}^2$-matrices. *Computing*, 69(1):1–35, 2002.

[23] W. Hackbusch and B. N. Khoromskij. A sparse $\mathcal{H}$-matrix arithmetic: general complexity estimates. *J. Comput. Appl. Math.*, 125(1-2):479–501, 2000. Numerical analysis 2000, Vol. VI, Ordinary differential equations and integral equations.

[24] W. Hackbusch and B. N. Khoromskij. A sparse $\mathcal{H}$-matrix arithmetic. II. Application to multi-dimensional problems. *Computing*, 64(1):21–47, 2000.

[25] W. Hackbusch and B. N. Khoromskij. Blended kernel approximation in the $\mathcal{H}$-matrix techniques. *Numer. Linear Algebra Appl.*, 9(4):281–304, 2002.

[26] Wolfgang Hackbusch and Steffen Börm. $\mathcal{H}^2$-matrix approximation of integral operators by interpolation. *Appl. Numer. Math.*, 43(1-2):129–143, 2002. 19th Dundee Biennial Conference on Numerical Analysis (2001).

[27] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 1993. Nonstiff problems.

[28] P. Haldenwang, G. Labrosse, S. Abboudi, and M. Deville. Chebyshev 3-D spectral and 2-D pseudospectral solvers for the Helmholtz equation. *J. Comput. Phys.*, 55(1):115–128, 1984.

[29] Bernard Malgrange. Equations aux dérivées partielles à coefficients constants. I. Solution élémentaire. *C. R. Acad. Sci. Paris*, 237:1620–1622, 1953.

[30] Bernard Malgrange. Equations aux dérivées partielles à coefficients constants. II. Equations avec second membre. *C. R. Acad. Sci. Paris*, 238:196–198, 1954.

[31] Bernard Malgrange. Existence et approximation des solutions des équations aux dérivées partielles et des équations de convolution. *Ann. Inst. Fourier, Grenoble*, 6:271–355, 1955–1956.

[32] Steven A. Orszag. Spectral methods for problems in complex geometries. *J. Comput. Phys.*, 37(1):70–92, 1980.

[33] Seymour V. Parter. On the Legendre-Gauss-Lobatto points and weights. *J. Sci. Comput.*, 14(4):347–355, 1999.

[34] Seymour V. Parter. Preconditioning Legendre special collocation methods for elliptic problems. I. Finite difference operators. *SIAM J. Numer. Anal.*, 39(1):330–347 (electronic), 2001.

[35] Seymour V. Parter. Preconditioning Legendre spectral collocation methods for elliptic problems. II. Finite element operators. *SIAM J. Numer. Anal.*, 39(1):348–362 (electronic), 2001.

[36] Seymour V. Parter and Ernest E. Rothman. Preconditioning Legendre spectral collocation approximations to elliptic problems. *SIAM J. Numer. Anal.*, 32(2):333–385, 1995.

[37] Anne-Sophie Piquemal and Jacques Liandrat. A comparison between a new wavelet preconditioner for finite difference operators and some other multilevel preconditioners in 1D. *Int. J. Pure Appl. Math.*, 6(1):1–13, 2003.

[38] M. Ribot and M. Schatzman. Extrapolation of the Residual Smoothing Scheme. part of the Ph. D. Thesis of M. Ribot, Decembre 2003.

[39] Magali Ribot. Application of the Residual Smoothing Scheme to the preconditioning of spectral methods by finite elements methods. part of the Ph. D. Thesis of M. Ribot, December 2003.

[40] Magali Ribot. Asymptotic of Legendre polynomials and of their extrema. part of the Ph. D. Thesis of M. Ribot, December 2003.

[41] Yousef Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003.

[42] Michelle Schatzman. *Numerical Analysis. A mathematical introduction*. Oxford University Press, 2002.

[43] Gábor Szegő. *Orthogonal polynomials*. American Mathematical Society, Providence, R.I., fourth edition, 1975. American Mathematical Society, Colloquium Publications, Vol. XXIII.

[44] Michael E. Taylor. *Pseudodifferential operators*, volume 34 of *Princeton Mathematical Series*. Princeton University Press, Princeton, N.J., 1981.

[45] François Trèves. *Introduction to pseudodifferential and Fourier integral operators. Vol. 1*. Plenum Press, New York, 1980. Pseudodifferential operators, The University Series in Mathematics.

POSTER SESSIONS

The poster sessions were organized by Tatiana Ivanova. All participants were invited to submit an abstract and present a poster at the conference. Participants were encouraged to include information about themselves, works in progress, and future works to facilitate contact between participants with common interests.

Posters related to one of the lecture series were presented together. The posters were divided into three sessions, and the corresponding abstracts are included below according to their sessions.

POSTER SESSION I
BIOMATHEMATICS / APPLIED MATHEMATICS

Using Distances in Multidimensional Statistics

Susan Holmes

Stanford University Statistics Department
Sequoia Hall, Serra Mall, CA94305 Stanford, U.S.A.
e-mail: susan@stat.stanford.edu

Abstract. Biology now requires the use of non standard parameters generalising work done on multivariate Euclidean spaces to spaces of parameters that are not embeddable in Euclidean structures.

I will present examples of extracting useful information by using distances between non standard objects in computational biology. In particular, trees, permutations and networks.

Visualisation of distances often provides much more information that the simple distributions.

Examples include :

- Comparing Phylogenetic trees from different DNA data.
- Comparing Hierarchical clustering trees on melanoma patiens.
- Comparing protein interaction networks.
- Constructing confidence sets for non standard data.
- Comparing permutations of genes along different genomes.

These all require use of interactive multidimensional visualisation techniques.

Some of these pose interesting statistical questions on how to build probability distributions that are defined if the sample sizes increase, interesting algorithmic questions arise as the computations of distances become exponentially difficult and good approximations are needed.

Mathematics Subject Classifications (2000): 62H30, 62H25, 60B05, 92B10, 92B05.

Keywords: Phylogenetic Trees, Gene Order Data, Distance, Multivariate Analyses, CAT(0) space, Bootstrap.

Zooming into Fullerenes

Irene Sciriha and Patrick W. Fowler*

Dept of Mathematics, Faculty of Science
University of Malta, Malta
e-mail: irene.sciriha-aquilina@um.edu.mt

**School of Chemistry, University of Exeter, UK*
e-mail: PWFowler@ex.ac.uk

Abstract. Carbon does not appear only in the form of diamond and graphite. Fullerenes, a third family of allotropes of carbon (**C**) exists as large stable clusters of **C** atoms. A trivalent polyhedron P is a cubic graph which may be embedded on a convex 3-D surface and a fullerene, $\mathbf{C_n}$, is P with twelve pentagons and $n-12$ hexagons.

According to the **Hűckel molecular orbital theory** (HMOT), the spectrum consisting of the eigenvalues of the adjacency matrix A of a graph G gives a good estimate of the spread of the π- elctron energy levels when G has the same structure as the (**C**)-skeleton of a molecule, a solution to a simplified **Schrődinger's equation.** The eigenvalue zero of **A** indicates the presence of NBOs or zero energy levels which in fullerenes is relatively rare. A corresponding eigenvector, called a kernel eigenvector, describes a NBO and determines a unique subgraph called the core that characterizes the charge-rich **C** centres contributed by the NBO-electron.

The distribution of the twelve pentagons and other hexagonal faces that tesselate the surface on which a $\mathbf{C_n}$-fullerene is embedded determines the substructures that force the NBO to be occupied. Of particular interest are the nut fullerenes so called because their skeleton is a nut graph that implies equidistributivity of the charge contributed by the NBO electron.

We study the substructures in fullerenes and other trivalent polyhedra, that determine the presence of the eigenvalue zero. Together with the symmetry group of the graph, they shed new light on singular graphs and polyhedra in particular.

Mathematics Subject Classifications (2000): 05C50, 92E10, 05C25, 05C90, 20F28.

Keywords: non-bonding orbital (NBO), core, fullerenes, adjacency matrix, minimal configuration (mc), nut graph, automorphism group.

Cellular Neural Network Models in Biology and Ecology

Angela Slavova

Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Sofia 1113, Bulgaria
e-mail:slavova@math.bas.bg

Abstract. This poster deals with Cellular Neural Network (CNN) models of some parabolic differential and integro-differential equations arising in biology and ecology.

We shall present the derivation of the CNNs implementations throught spatial discretization, which suggests a methodology for converting a PDE to CNN templates and vice versa. We shall demonstrate how an autonomous CNN can serve as a unifying paradigm for active wave propagation, several well-known examples chosen from different disciplines will be modeled.

Sixty years ago Fisher showed that the propagation of a mutant gene can be modeled by a nonlinear reaction-diffusion partial differential equation (PDE):

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + qu(1-u). \tag{0.1}$$

This classic equation, also known as the "diffusional logistic" equation, has since been found to be useful in many other applications and has been widely studied. In chemical media the function $u(t,x)$ is the concentration of the reactant, D represents its diffusion coefficient, and the positive constant q specifies the rate of the chemical reaction. In media of other natures u, D, q can represent different quantities. In general, medium described by (1.1) is often refered to as a bistable medium, because it has two homogeneous stationary states, $u = 0$ and $u = 1$.

The second model we consider is a more general form of the Hodgkin-Huxley model for the propagation of the voltage pulse through a nerve axon which is referred to as the FitzHugh-Nagumo equation:

$$u_t - u_{xx} = u(u-\Theta)(1-u) - b\int_0^t u(s,x)ds, \tag{0.2}$$

$0 < x, t < 1$, $0 < \Theta < 1/2$, $b \geq 0$. The proposed equation (1.2) is nonlinear parabolic integro- differential equation, in which u_t is the first partial derivative of $u(t,x)$ with respect to t, u_{xx} is the second derivative of u with respect to x, u is a membrain potential in a nerve axon, the steady state $u = 0$ represents the resting state of the nerve.

Dynamical behavior of such models is studied using the describing function technique. Travelling wave solutions are constructed and their structure and stability are investigated for the CNN equations.

Mathematics Subject Classifications (2000): 92B20, 34K57, 34C55.

Keywords: Cellular Neural Networks, Fisher equation, FitzHugh-Nagumo equation, travelling waves.

Mathematical Virology: A novel approach to the protein stoichiometry of viral capsids.

Reidun Twarock

Centre for Mathematical Science, City University, Northampton Square, London EC1V 0HB
e-mail: r.twarock@city.ac.uk

Abstract. A vital part of infectious virus particles is the protein shell, called the viral capsid. It protects the viral genome and is formed by so-called morphological units, entities composed of usually five or six protein subunits. The derivation of

mathematical models for the location and the types of these morphological units in the viral capsids is important as this information is key for an understanding of the viral assembly process and hence for the design of anti-viral therapeutics.
For a large class of viruses this problem is solved by Caspar-Klug theory [**1**], and the classification implied by this theory is presently used in the classification and evaluation of experimental data. However, for a significant number of viruses, including for example important cases such as polyoma virus – the causative agent for cervical cancer in women – this theory does not apply. In [**2**] a new theory has been introduced that uses tiling theory, a theory investigating tessellations of surfaces in terms of a given finite set of different shapes called tiles, combined with group theory to generalize Caspar-Klug theory. The new theory [**3**] not only accommodates the open cases that are known at present, but also makes predictions about novel viral structures that have not yet been discovered. Furthermore, apart from predicting the types of the morphological units for the open cases correctly, the new theory also predicts the bonding structure between protein subunits in the capsid both for the novel and the Caspar-Klug cases, a result that could not be obtained within the framework of Caspar-Klug theory.
In this contribution we demonstrate the new theory for the example of polyoma virus based on [**2**].

References

[1] D.L.D. Caspar and A. Klug, Cold Spring Harbor Symp. Quant. Biol. 27, 1 (1962).
[2] R. Twarock, "A new view on quasi-equivalence: A tiling approach to virus capsid formation", accepted by J. Theor. Biol.
[3] R. Twarock, "Classification of viral capsids based on the tiling principle", in preparation.

Mathematics Subject Classifications (2000): 05B45, 62P10, 92C40

Keywords: tiling theory, viral capsids, protein stoichiometry

Algebraic Classification of Discrete Kinetic Models

Mirela Cristina Vinerean

Karlstad University
Universitetgatan 2, 651 88 Karlstad, Sweden
e-mail: mirela.vinerean@kau.se

Abstract. The basic equation in kinetic theory is the Boltzmann equation for time-evolution of the particle density $f = f(x, t; v, \varepsilon, ...)$, where x, t, v,ε represent the position, the time, the velocity and the internal energy of the particle in the phase space.

Discrete kinetic models (DKMs) or simply, discrete velocity models (DVMs) in the particular case when there exist no internal degrees of freedom, are models where all phase coordinates, except the space one, are discretized (i.e. the velocities are assumed to be able to take a finite number of values). In this case, the Boltzmann equation is replaced by a system of differential equations easier to analyze from the mathematical or numerical point of view.

In many interesting papers on DVMs, authors postulate from the beginning that the finite velocity space with "good" properties is given and only after this step, study the Boltzmann equation (system). Contrary to this approach, our aim is not to study the equations, but to discuss all possible choices of finite phase spaces (sets) satisfying this type of "good restrictions". Due to the velocity discretization is well-known that it is possible to have DVMs with "spurious" summational invariants (conservation laws which are not linear combination of physical invariants). Our purpose is to give a method (algorithm) for constructing normal models (without spurious invariants) and to classify all normal plane models with small number of velocities (which usually appear in applications).

In the first step we describe DKMs as algebraic systems. We introduce for this an abstract discrete model (ADM) which is defined by the matrix of reactions (same as for the concrete model). This matrix contains as rows all vector of reactions, which can be written as n-dimensional vectors $(k_1, ., k_n)$ with $k_i \in \mathbb{Z}$, describing the "jump" from a pre-reaction state to a new reaction state. The conservation laws corresponding to the many-particle system are uniquely determined by the ADM, or equivalently, by its corresponding matrix of reactions, and do not depend on the concrete realization.

We find the restrictions on ADM such that it is a realization of some concrete DM and in the next step we give a general method of constructing normal models (using the results on ADMs). Having the general algorithm, we consider in more details, the particular cases of models with mass and momentum conservation (inelastic lattice gases with pair collisions) and models with mass, momentum and energy conservation (elastic lattice gases with pair collisions).

Mathematics Subject Classifications (2000): 82C40, 76P05.

Keywords: kinetic theory, discrete kinetic (velocity) models, conservation laws.

POSTER SESSION II
PURE MATHEMATICS and POPULARIZATION OF MATHEMATICS

On Universality in Varieties of Semigroups

Marie Demlová

Department of Mathematics, Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
e-mail: demlova@math.feld.cvut.cz

Abstract. Endomorphisms of any algebra together with the operation composition form a monoid, so called *endomorphism monoid.*

We say that a class $\mathcal{K}$ of algebras is *monoid universal* if for every monoid $\mathbb{M}$ there exists an algebra $\mathbb{A} \in \mathcal{K}$ such that $\mathbb{M}$ and the endomorphism monoid of $\mathbb{A}$ are isomorphic. An important and useful generalization of a monoid universality is alg-universality (this notion plays a key role in proofs that a concrete category is monoid universal). A concrete category $\mathcal{K}$ is *alg-universal* if the category of all graphs and compatible mapping can be fully embedded into $\mathcal{K}$.

For semigroup varieties the notions of monoid universality and alg-universality coincide. All alg-universal varieties of semigroups were completely described by V. Koubek and J. Sichler. Usually, if a semigroup variety is not alg-universal then there exist trivial homomorphisms between any pair of its semigroups. One of the possibilities how to forbid trivial homomorphisms is an expansion of similarity type. A variety $\mathcal{V}$ has an *alg-universal nullary expansion* if we can add finite number of nullary operations to the type so that the new variety is alg-universal.

All band varieties that have alg-universal nullary expansion were fully described by M. Demlová and V. Koubek. We present a further result in this investigation.

For a semigroup variety $\mathcal{V}$, let $\sqrt{\mathcal{V}}$ denote the class of all semigroups $\mathbf{S} = (S, \cdot)$ such that the subsemigroup of $\mathbf{S}$ on the set $S^2 = \{s \cdot t | s, t \in S\}$ belongs to $\mathcal{V}$. Then $\sqrt{\mathcal{V}}$ is a variety.

Theorem: If $\mathcal{V}$ is the variety of left-zero semigroups or the variety of right-zero semigroups then the variety $\sqrt{\mathcal{V}}$ has alg-universal expansion by two nullary operations.

The above result is in contrast with the following proposition proved by M. Demlová and V. Koubek:

Proposition: If $\mathcal{V}$ is the variety of left-zero semigroups or the variety of right-zero semigroups then every algebra from the variety $\sqrt{\mathcal{V}}$ can be reconstructed from its endomorphism up to isomorphism.

Mathematics Subject Classifications (2000): 20M07, 20M15, 18B15.

Keywords: endomorphism, semigroup, variety, universal category, nullary operation

Regularity Results for Functionals with Non Standard Growth

Michela Eleuteri

Dipartimento di Matematica di Trento
via Sommarive 14, 38050 Povo (Trento)
e-mail: eleuteri@science.unitn.it

Abstract. The aim of this poster is to present some regularity results for scalar minimizers of functionals with non standard growth, also in the case of a minimization problem with obstacle. We deal with functionals of the following type:

$$\mathcal{F}(u, \Omega) := \int_\Omega f(x, u(x), Du(x))dx \ ,$$

where Ω is a bounded open set of $\mathbb{R}^n$, while $f : \Omega \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ is a Carathéodory function and $u \in W^{1,1}_{\text{loc}}(\Omega, \mathbb{R})$. The main assumption is that f satisfies a $p(x)-$type growth, that is

$$|z|^{p(x)} \leq f(x, u, z) \leq L(1 + |z|^{p(x)})$$

where $p(x) > 1$ is a *variable growth exponent* and it is continuous and f satisfies suitable assumptions of convexity with respect to the variable z. Such types of energies owe their importance to the fact that several models (also non variational) coming from Mathematical Phisics are built using a variable growth exponent. Under sharp assumption on the modulus of continuity $p(x)$ we prove regularity results

for the minimizers. In particular, if $p(x)$ is Hölder continuous, the gradient is Hölder continuous too. A model functional that satisfies our assumption is:

$$\int_\Omega a(x,u)|Du|^{p(x)}\,dx\ ,$$

where $a(x,u)$ is a bounded function which is far from zero and continuous.

Mathematics Subject Classifications (2000): 49N60, 49J10.

Keywords: Regularity results, non standard growth.

Generalized Polynomials in Ergodic Theory and Number Theory

Inger Johanne Håland Knutson

Dept of mathematics, Agder University College,
Serviceboks 422, NO-4604 Kristiansand, Norway
e-mail: inger.j.knutson@hia.no

Abstract. A *generalized polynomial* is a function $q : \mathbb{Z} \to \mathbb{R}$ obtained from finitely many polynomials by use of the greatest integer function, addition and multiplication, like the following examples: $[\sqrt{2}n]\sqrt{2}n$, $\Big[[\pi n^2]\sqrt[3]{2}n^3 + \sqrt{3}n\Big][en]\pi$, $[\sqrt{2}n][\sqrt{3}n] - [\sqrt{6}n^2]$. Since the family of generalized polynomials is a natural extension of the family of usual polynomials, I have been interested in investigating some results in ergodic theory involving polynomials to show that they are also true for some classes of generalized polynomials. Here are some results about polynomials that we extend:

1. If $p(t)$ is a real polynomial with at least one coefficient other than the constant term irrational, then the sequence $p(n)$, $n = 1, 2, \ldots$, is uniformly distributed modulo 1. (H.Weyl 1916).

2. A set $R \subset \mathbb{Z}$ is called a set of recurrence if given any invertible measure preserving system $(X, \mathcal{B}, \mu, T)$ and $A \in \mathcal{B}$, $\mu(A) > 0$, there exists $n \in R \setminus \{0\}$ such that $\mu(A \cap T^{-n}A) > 0$. If $p(n)$ is an integer-valued polynomial with $p(0) = 0$, then $\{p(n) \mid n \in \mathbb{Z}\}$ is a set of recurrence. (Furstenberg, Sárközy)

3. Let $(X, \mathcal{B}, \mu, T)$ be a weakly mixing measure preserving system, and let $p_1(n), \ldots, p_k(n)$ be pairwise essentially distinct integer-valued polynomials. Then for any $f_1, \ldots, f_k \in L^\infty(X, \mathcal{B}, \mu)$, one has

$$\Big\| \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} T^{p_1(n)} f_1 \cdots T^{p_k(n)} f_k - \prod_{i=1}^{k} \int f_i \Big\|_2 = 0$$

(V. Bergelson 1986)

The classes of uniformly distributed generalized polynomials and those giving rise to sets of recurrence include generalized polynomials with sufficiently independent coefficients. We also show that a multiple recurrence theorem (corresponding to 3.) is true for wide classes of generalized polynomials including oscillating generalized polynomials of, for example, the form $[[\alpha n]\beta n] - [[\beta n]\alpha n]$, where α, β and $\frac{\alpha}{\beta}$ are irrational.

Some of the work presented in the poster is joint with V. Bergelson.

Mathematics Subject Classifications (2000): 28D05, 11J71, 11B83.

Keywords: polynomials, uniform distribution, set of recurrence, multiple recurrence.

Some Practical Combinatorial Computations of the Matrix Functions and Applications

Rajae Ben Taher

Faculte des Sciences
Universite Moulay Ismail
Meknes, 50000
Maroc
e-mail: bentaher@fsmek.ac.ma

Abstract. We present some explicit formulas for A^n $(n \geq r)$ and e^{tA} for every $r \times r$ matrix A, using here some linear recurrence relations in the algebra of square matices. We extend some of this results to the general case of $f(A)$, where f is defined on the spectrum of A. New formulas are obtained for $f(A)$ are given and some well-known formulas are derived. Examples and applications are also studied.

Mathematical Subject Classifications (2000): Primary 15A99, 40A05. Secondary 40A25, 45M05, 15A18.

Keywords : algebra of matrices, linear recurrence relations, combinatorial powers of a matrix, matrix functions.

Study of Finnish Women in Mathematics

Marjo Lipponen-Salhi and Jennifer J. von Reis

Department of Mathematics, 20014 University of Turku, Finland
e-mail: marlip@utu.fi, jvonreis@fulbrightweb.org

Abstract. Finland was the first country in Europe to grant women sufferage. It is a country widely known for its progressive policies in gender equality. However, this equality is not visible in the field of mathematics. For example, in Finland there is currently no woman professor of mathematics. Also, the participation in the EWM has been alarmingly low compared to the number of women involved in mathematics. This discrepancy was the inspiration for our study.

Our study aims to gather information about the current situation of women in mathematics in Finland. We plan to raise awareness and participation in EWM, to find out the factors for success, and underlying difficulties for women in the field of math. In particular we would like to see if there are any candidates for future professorships.

The methods of our study include a questionnaire sent out to all women in mathematics departments who have at least a M.Sc. degree. Included in the questionnaire was information about the EWM, the mentoring project, and the forthcoming meeting in Luminy. We asked for statistics from each mathematics department about the number of graduate students, people in the department, visiting researchers, and degrees granted seperated by gender. We also requested information on all Ph.D. degrees granted in mathematics from the Statistics Finland. To give us a better

interpretation of the original questionnaire, we have also sent out a slightly modified version to roughly twenty male mathematicians at the University of Turku.

The Department of Women Studies at the University of Turku is interested in the project. We have had informal discussion with them and have been invited to present our results at a seminar in their department. This contact has already been fruitful in terms of ideas and inspiration.

The study is still on-going and the interpretation of the results we do have is not final. However, already there is evidence of a large interest in the EWM meetings and the mentoring program. Women who are still in a Ph.D. program have been more active in responding to our study as well as foreigners. Since the obstacles tend to appear after the completion of the Ph.D. degree, the study does not yet give much information about the real problems facing women in their careers. There was little evidence of any ambition to obtain a professorship, though this might be partly due to Finnish modesty. In the responses there is also a general feeling of doubt about the choice of mathematics and the future.

We plan to further the study with interviews with women mathematicians. The website for the project, www.math.utu.fi/EWM will be further developed. Most importantly, we are also planning a meeting for all Finnish women mathematicians through which we hope to begin to increase the participation of women in the EWM.

Funding for the project has been provided by Research Programme MaDaMe funded by the Academy of Finland. This study is also connected to a research project funded by a Fulbright Grant to Finland and the Center for International Mobility in Finland.

Keywords: Gender issues, mentors, factors for success, Finnish women in mathematics.

POSTER SESSION III: NUMERICAL METHODS

Recent Methodologies in Applied Scientific Computing

Maria Mercede Cerimele, Daniela Mansutti, Francesca Pistella and Rosa Maria Spitaleri

Istituto per le Applicazioni del Calcolo-CNR,
Viale del Policlinico 137, 00161 Roma, Italy
e-mail: spitaleri@iac.rm.cnr.it

Abstract. We present recent methodologies based on the development of models and methods for handling complex application problems.

Differential computing is a powerful tool for the investigation of physical phenomena. The numerical approach to dissect and understand a physical phenomenon includes appropriate differential modeling, an effective design of the overall computational process, the definition of accurate and robust algorithms, the implementation of advanced software tools allowing a fast evaluation of computed intermediate and final results.

We are developing models and methods looking at the computational simulation as a cyclical process which is able to accumulate knowledge cycle by cycle.

We believe that several knowledge environments are involved in simulation processes and integrated in powerful methods and tools. We focus on fundamental substeps of numerical simulations, linking developed models, algorithms and software tools.

We illustrate PDE models, computational algorithms and numerical results on:

- multigrid variational image segmentation,
- growth of a wall film resulting from spray impingement,
- solidification processes in microgravity.

Mathematics Subject Classifications (2000): 65N06, 65N22, 65N55

Keywords: differential computing, PDE modeling, finite differences, multigrid computation

Validation and Verification of a Dynamic Blood Flow Model - Using Numerical Simulations in FEMLAB

Sofie Inari Castella
in cooperation with Jacob Kirkensgaard Hansen and Ingunn Gunnarsdottir

Roskilde University
IMFUFA, Postbox 260, 4000 Roskilde, Denmark
e-mail: sic@ruc.dk

Abstract. The following project is a bachelor project from Roskilde University, Denmark.
In the project we seeked to validate and verificate a non-linear dynamic model of blood flow in arteries. Based on the concepts of validation and verification we "investigated" the model thoroughly. This includes a discussion of the assumptions that underlies the model, an alternative simulation setup in the commercial program FEMLAB in order to compare two different numerical solutions, and a comparison of simulations with scanning data from a human artery.

It is concluded that we in general succeded in verificating the model mathematically, even though we through an alternative deduction of one of the models equations found a missing term that should not be neglected under the model assumptions. Further it is concluded, that we did not succeed in producing comparible alternative simulations, which primarily is do to implementational difficulties in the commercial simulation program, FEMLAB. Finally it is concluded that we did not succeed in validating the model based on the scanning data, since data and simulations concern two quite different arteries. Allthough there are indications of qualitative similarities between data and model.

Mathematics Subject Classifications (2000): 35Q30, 35Q51, 76D05, 76M10, 91B74, 93A30.

Keywords: Blood Flow, non-linear Navier-Stokes, verification, validation, FEMLAB.

Simulation of Waves through Soft Heterogenous Tissue: Approximation of Rays by Algebraic Curves.

Cathrine Tegnander

Department of Mathematical Sciences
NTNU, NO-7491 Trondheim, Norway
e-mail:cathrine@math.ntnu.no

Abstract. We study simulation of ultrasound waves through soft tissue (breast). Most of the existing techniques for reconstruction of images by ultrasound waves are based on homogenous background. Since the breast consists of strongly heteogenous soft tissue, we study the deviation of the propagated wave, compared with the propagation in a homogenous case. Raytracing is used in this work. Our aim is to understand the importance of this deviation due to changes in the velocity. We would also like to understand whether methods based on heterogenous media (from seismic applications) might improve the existing imaging.

We model the propagation of a given pulse f (MHz) through soft (inhomogenous) tissue. With reasonable approximation, we can model our propagation by a linear wave equation.

Let t denote time (sec). We consider propagation in 2D (but this can easily be generalised to 3D) so $x \in R^2$ (m^2). v is velocity (m/sec) with $v^2 = \frac{1}{\rho\kappa}$, ρ is density (kg/m^3) and κ is compressibility (Pa^{-1}). Then, we use the linear model given by

$$\nabla \cdot \frac{1}{\rho} \nabla p = \frac{1}{\rho v^2} \frac{\delta^2 p}{\delta t^2},$$

where $p = p(x, t)$ is the pressure.

The breast is highly heterogenous, but with a small variation in velocity. We study the importance of this variation by considering the wave front. We follow the wave front in a given direction (ray tracing), and construct a 2nd order curve that approximates the ray. This gives a family of rays or algebraic curves covering the wave front.

We study a numerical example where we trace the numerical rays in some directions from an initial given point. The given pulse is highfrequency of size 10-15 MHz such that the wave length is of order 150μ-m. Then we effectuate the L_1-differences between a straight line ray and the numerical ray, and further the L_1-difference between some second order curves and the numerical ray.

Mathematics Subject Classifications (2000): 35L05, 65M06, 65M25, 65M32, 14H50.

Keywords: wave equation, ray tracing, algebraic curves, ultrasound imaging.

Modelling Valveless Pumping – with Experimental Validation

Stine Timmermann

Dept. of Mathematics and Physics
Roskilde University, P.O. 260, 4000 Roskilde, Denmark
e-mail: stinet@ruc.dk

Abstract. This abstract describes an ongoing master thesis project with planned termination in June 2004.

When a water filled torus consisting of two elastic tubes with different elasticity is compressed symmetrically and periodically in a place of asymmetry a unidirectional mean flow in the system is created. The direction and size of the mean flow depend among other things upon the frequency of compression and the difference in elasticity of the tubes.

This phenomenon is called *valveless pumping* and since its discovery in the 18th century scientists have tried to describe and understand the phenomenon. Their works mainly deal with one-dimensional models of an equivalent system consisting of two vessels in connection with a periodic compressed elastic tube.

The aim of the master thesis is to investigate the phenomenon through experiments and by making a *two-dimensional* mathematical model of the torus system. In outline: A mathematical description of the torus system consists of the Navier Stokes equations with periodic boundary conditions. The modelling difficulties lie both in the moving boundary caused by the elasticity of the tube and in the description of the pumping mechanism. The purpose of making a two-dimensional model is among other things to maintain information about the velocity profile to be compared with the experimental results.

The master thesis is a continuation of the work done by J.T. Ottesen, Journal of Mathematical Biology, vol. 46 (4): 309-332 APR 2003.

Mathematics Subject Classifications (2000): 76D05, 35Q30, 65N06, 91B74, 93A30.

Keywords: flow, periodic compression, elastic tubes, mathematical model, Navier Stokes equations.

Some Recent Results About Steady Flows of Viscoelastic Fluids

Colette Guillopé

Laboratoire d'Analyse et de Mathématiques Appliquées
CNRS and Université Paris XII-Val de Marne
61, avenue du Général de Gaulle, 94010 Créteil Cedex, France
e-mail: guillope@univ-paris12.fr

Abstract. Flows of viscoelastic liquids might be modeled by a system of coupled partial differential equations, a Navier-Stokes (or Euler) equation type for the velocity field, a transport equation for the non-Newtonian part of the stress tensor, and the usual equation describing the conservation of mass.

Recently we have been interested in slightly compressible fluids, in an attempt to understand how important the effects of small compressibility are near sharp corners, or next to irregular obstacles. In a first stage we only consider regular solutions.

A first study was done by R. Talhouk[2]: the behavior of the stress tensor is modeled by a relatively simple equation, the Oldroyd model, the flow is confined into a regular bounded domain Ω, and subject to homogeneous boundary conditions. In collaboration with R. Talhouk, we also studied a similar problem for flows around an obstacle, i.e. flows outside a bounded domain. A second study, done in collaboration with A. Hakim[3] and R. Talhouk, concerns the White-Metzner model, in which the parameters entering the constitutive equation depend nonlinearly on the rate of deformation tensor.

The case of an exterior domain, i.e. a domain which is the complement of a bounded domain of $\mathbb{R}^3$, is fairly interesting because it leads to the study of three linear problems, an Oseen problem for the linearized velocity, a transport problem for the linearized non-Newtonian stress, and a Neumann problem for a modified pressure.

Results of existence and uniqueness of regular solutions are obtained for small compressibility, small exterior forces, and for small velocity of the flow at infinity, but without any condition on the viscosity of the fluid.

Mathematics Subject Classifications (2000): 35Q35, 76A10, 76N10.

Keywords: Navier-Stokes equations, transport equation, Oseen equation, exterior domain, viscoelastic fluids.

[2] Faculté des Sciences - Section 1, Université libanaise, Beyrouth, Liban

[3] Faculté des Sciences de Guéliz, Marrakech, Maroc

About European Women in Mathematics

by Laura Tedeschini Lalli

EWM, European Women in Mathematics, was established in 1986, at the International Congress of Mathematicians in Berkeley, and held its first meeting later that year in Paris.

Our history begins with informal meetings in which we exchanged our iews of professional life, as well as our mathematics.

While later the association became official, with its seat in Helsinki, still to this date the General Meetings are the structural backbone of EWM, with their strong stress on exchanging ideas and making contacts.

The general aims of EWM are:

- to encourage women to take up and continue their studies in mathematics.
- to support women with or desiring careers in research in mathematics or mathematics related fields.
- to provide a meeting place for these women.
- to foster international scientific communication among women and men in the mathematical community.
- to cooperate with groups and organizations, in Europe and elsewhere, with similar goals.

Within our general aims, some reflections on our obstacles and/or difficulties took place along the years, yielding to actions and more focused questions and projects. We have learned with experience that each subject that has prompted our attention, inevitably led to discover obstacles, or discrimination, or pointless difficulties, acting against free and respectful communication and dissemination of mathematics, regardless of gender.

We think this is one of our contributions to the mathematical community at large, so let us see briefly some of these points.

Women academicians in mathematics across Europe. It has been very clear since the meeting in Warwick, December 1988, that when you walk into a department of Mathematics in a European University, your chances of meeting with a female mathematician vary deeply, and strogly depend on the country you are in.

In years such as these, in which Europe denounces a shortage of graduates in technical subjects, we think educational institutions should observe and meditate what happens in countries like France, India, Italy, Russia, Spain, to name just a few, where girls are not scared away from mathematics during highschool and college, on the basis of surrounding cultural expectations.

On this subject we have some statistics available through our webpage at: http://www.math.helsinki.fi/EWM/, and the video "Women in Mathematics across Cultures", produced by EWM and also available through the web. At the meeting in Luminy, a special day was organized by *Femmes et Maths* , about "Mathématiques au féminin en Méditerranée".

Age limits. Together with the stereotype of a male mathematician, goes the idea that math is a young person's task. We worked together with the Committee fo Women and Mathematics of the European Mathematical Society, at removing or soften age limits for prizes and positions, as well as at understanding what kind of career breaks are more likely for women, and can be overcome with due help.

Isolation of women in research. Unfortunately, it is still the case in many European universities that women are singularities within the mathematical departments. Frequently this has a well known inhibiting effect on us, resulting in self-consciousness or defensiveness, both particularly negative when we start our professional path. And if we are inhibited, we do not speak about mathematics, and if we do not speak about mathematics we do not learn how to speak about mathematics, and the loop traps us. The vicious circle of communication, well-known to many, creates a steady isolation which becomes sterile and depressing, as opposed to the temporary isolation which is necessary to all creative work. In fact, we think many problems arise for women in mathematical research from the different types of isolation (communication, life passages...) adding to the second, necessary one, and making it seem unbearable.

For much the same reasons, men are always welcome to the scientific pat of our meetings.

EWM issues a yearly newletter and maintains an e-mail list and a list of discussion. All the relative information can be gathered from our webpage at

http://www.math.helsinki.fi/EWM/

To join EWM, please either contact your regional coordinator (list in this volume), or contact our office in Helsinki directly from our web page.

Rome, September 2004

European Women in Maths Web-based Mentoring Scheme

By Cathy Hobbs

In August 2001 the European Union agreed to fund a project proposed by EWM to provide web-based mentoring to women in mathematical sciences in Europe. Recent reports had highlighted (yet again) the lack of women in higher positions in academia across scientific disciplines. The EU is committed to improving the human potential across Europe, and in particular, to realising the talent of the female population, so this project was funded as a step towards encouraging women to progress in their mathematical science careers. The funding from the EU for the scheme finished in August 2003 but the scheme is still operating, based at Oxford Brookes University, Oxford, UK.

Aim and scope of project

The aim of the web-based mentoring scheme is to enable new women mathematical scientists (e.g. graduate students, those considering graduate work, post-doctoral students) to find mentors amongst the mathematical science community. In this context, a mentor is someone who can listen to what their mentee is saying, provide advice on academic issues such as applying for jobs, applying for grants, when and where to publish research work, and also may act as a role-model to the mentee. Mentors may also advise on broader gender-related issues faced by women in a mainly male-dominated environment. The mentors in this scheme are volunteers, not trained counsellors, who are willing to share their own experiences with less-experienced mathematicians.

Using the web to facilitate the mentoring scheme enables women to form links with mentors across Europe. Because of the wide distribution of mentors and mentees across Europe, they mainly communicate by e-mail but they have the freedom to structure their own mentoring relationship. This may mean telephone contacts and face-to-face meetings where appropriate (eg both attending a conference together).

Similar schemes are now starting up across the world, for example that run by the American Women in Maths organisation. Our scheme links with them to provide mentors for European women, and also to provide US mentors for those considering studying in the US. We hope that schemes of this nature will contribute to the support network for women in mathematical sciences and encourage women to progress in their mathematical science careers.

Web Site

The website for this scheme was designed and implemented by a professional web designer. On the home page the basic idea of the scheme is outlined. Users can then go to sections on signing up to be a mentor, signing up to request a mentor, profiles of existing women in mathematics and information on careers, education and on mentoring generally.

The sign-up pages give further information about the role of a mentor and guidelines on the time commitment required to be part of the scheme. There are links to guidelines on being a mentor and being a mentee, as appropriate, and guidelines on electronic communication. The mentor or mentee can then fill out an online form.

The data provided goes directly to a database which is only accessible to the administrators of the site. Matching of pairs is then done, paying particular attention to the aspects the mentor and mentee have highlighted. For example, for some people the subject area is most important whereas others would feel that geographical location is more critical. The mentor/mentee pair will then be informed of the matching and provided with basic details of their partner. They are then free to conduct their mentoring relationship as they see fit.

Results and feedback

So far the website has had nearly 8000 visitors. 50 mentors have signed up and 60 mentees. There have been around 30 successful matches. Matching mentees successfully is quite difficult since there is not always a good overlap between the mentees' requirements and what the mentors have to offer. This is particularly true of research interests, which seems to be the most important matching factor. We do not always manage to find mentors with the same research interests as the mentees, which is the main reason for the number of unmatched mentees.

Evaluation was planned from the start of the project. Statistics were collected on the website and database on numbers of visitors, numbers of mentors and mentees signed up and related data. Questionnaires were e-mailed to mentors and mentees in March and August 2003. The feedback showed overwhelming support for the scheme, which seems to be meeting a deep need amongst its target audience, both mentors and mentees. The majority of respondents gave positive answers on the operation of the scheme and the results so far. Almost all respondents felt the matching process was clear and most thought the communication from the mentoring scheme helpful. All mentors and mentees were content with the choice of mentee/mentor made for them.

Respondents found the site easy to navigate, the guidelines useful, the biographies valuable, the form easy to fill in and the right number of questions asked, although one mentee would have liked the mentor to indicate whether they had a family. Most felt no need for more structured procedures, although one mentor suggested it would be useful to send out discussion tips regularly to give mentors and mentees something to help things move and another suggested a quarterly email newsletter.

Most mentees felt no need for formal training but did say they might be interested at a later stage.

All mentors had agreed frequency of contact times with their mentees and most had kept up their contact on a regular basis. Only one mentor reported a 'fizzling out' of the agreement. Contact was mainly by email (75%), others had face to face meetings.

Around two thirds of mentors thought that being a mentor had benefited them in some way and over half of the mentees felt that having a mentor had made a difference to their career plan even after such a short time period. Many reported gaining opportunities to study and work through contacts made, and an increase in confidence. This is key for many women who often lack confidence in their own abilities. Communicating with an impartial advisor is providing mentees with such opportunities. One mentee commented "I feel more confident to pursue a purely academic career, not having a single female maths lecturer at my university got me doubting those plans, I have to admit". Another stressed the important of the mentoring for those who did not have much money to travel or attend conferences and make contacts.

Athena Award

In 2003 the mentoring scheme won a UK prize for the best use of information technology in advancing the careers of women in science, engineering and technology. The Royal Society of Great Britain (the leading society of scientists in the UK) and the British Computer Society sponsored this prize, known as the Athena Award. The prize was around 4500 Euros and will be used to fund a project to collect current statistical data on the numbers of women in mathematics in Europe (to compare with the figures collected by EWM in 1993).

Further Information

The website is at

http://ewm.brookes.ac.uk

If you would like further information about the scheme or the statistics project, contact Dr Cathy Hobbs (cahobbs@brookes.ac.uk).

CWI TRACTS

1 D.H.J. Epema. *Surfaces with canonical hyperplane sections.* 1984.

2 J.J. Dijkstra. *Fake topological Hilbert spaces and characterizations of dimension in terms of negligibility.* 1984.

3 A.J. van der Schaft. *System theoretic descriptions of physical systems.* 1984.

4 J. Koene. *Minimal cost flow in processing networks, a primal approach.* 1984.

5 B. Hoogenboom. *Intertwining functions on compact Lie groups.* 1984.

6 A.P.W. Böhm. *Dataflow computation.* 1984.

7 A. Blokhuis. *Few-distance sets.* 1984.

8 M.H. van Hoorn. *Algorithms and approximations for queueing systems.* 1984.

9 C.P.J. Koymans. *Models of the lambda calculus.* 1984.

10 C.G. van der Laan, N.M. Temme. *Calculation of special functions: the gamma function, the exponential integrals and error-like functions.* 1984.

11 N.M. van Dijk. *Controlled Markov processes; time-discretization.* 1984.

12 W.H. Hundsdorfer. *The numerical solution of nonlinear stiff initial value problems: an analysis of one step methods.* 1985.

13 D. Grune. *On the design of ALEPH.* 1985.

14 J.G.F. Thiemann. *Analytic spaces and dynamic programming: a measure theoretic approach.* 1985.

15 F.J. van der Linden. *Euclidean rings with two infinite primes.* 1985.

16 R.J.P. Groothuizen. *Mixed elliptic-hyperbolic partial differential operators: a case-study in Fourier integral operators.* 1985.

17 H.M.M. ten Eikelder. *Symmetries for dynamical and Hamiltonian systems.* 1985.

18 A.D.M. Kester. *Some large deviation results in statistics.* 1985.

19 T.M.V. Janssen. *Foundations and applications of Montague grammar, part 1: Philosophy, framework, computer science.* 1986.

20 B.F. Schriever. *Order dependence.* 1986.

21 D.P. van der Vecht. *Inequalities for stopped Brownian motion.* 1986.

22 J.C.S.P. van der Woude. *Topological dynamix.* 1986.

23 A.F. Monna. *Methods, concepts and ideas in mathematics: aspects of an evolution.* 1986.

24 J.C.M. Baeten. *Filters and ultrafilters over definable subsets of admissible ordinals.* 1986.

25 A.W.J. Kolen. *Tree network and planar rectilinear location theory.* 1986.

26 A.H. Veen. *The misconstrued semicolon: Reconciling imperative languages and dataflow machines.* 1986.

27 A.J.M. van Engelen. *Homogeneous zero-dimensional absolute Borel sets.* 1986.

28 T.M.V. Janssen. *Foundations and applications of Montague grammar, part 2: Applications to natural language.* 1986.

29 H.L. Trentelman. *Almost invariant subspaces and high gain feedback.* 1986.

30 A.G. de Kok. *Production-inventory control models: approximations and algorithms.* 1987.

31 E.E.M. van Berkum. *Optimal paired comparison designs for factorial experiments.* 1987.

32 J.H.J. Einmahl. *Multivariate empirical processes.* 1987.

33 O.J. Vrieze. *Stochastic games with finite state and action spaces.* 1987.

34 P.H.M. Kersten. *Infinitesimal symmetries: a computational approach.* 1987.

35 M.L. Eaton. *Lectures on topics in probability inequalities.* 1987.

36 A.H.P. van der Burgh, R.M.M. Mattheij (editors). *Proceedings of the first international conference on industrial and applied mathematics (ICIAM 87).* 1987.

37 L. Stougie. *Design and analysis of algorithms for stochastic integer programming.* 1987.

38 J.B.G. Frenk. *On Banach algebras, renewal measures and regenerative processes.* 1987.

39 H.J.M. Peters, O.J. Vrieze (eds.). *Surveys in game theory and related topics.* 1987.

40 J.L. Geluk, L. de Haan. *Regular variation, extensions and Tauberian theorems.* 1987.

41 Sape J. Mullender (ed.). *The Amoeba distributed operating system: Selected papers 1984-1987.* 1987.

42 P.R.J. Asveld, A. Nijholt (eds.). *Essays on concepts, formalisms, and tools.* 1987.

43 H.L. Bodlaender. *Distributed computing: structure and complexity.* 1987.

44 A.W. van der Vaart. *Statistical estimation in large parameter spaces.* 1988.

45 S.A. van de Geer. *Regression analysis and empirical processes.* 1988.

46 S.P. Spekreijse. *Multigrid solution of the steady Euler equations.* 1988.

47 J.B. Dijkstra. *Analysis of means in some nonstandard situations.* 1988.

48 F.C. Drost. *Asymptotics for generalized chi-square goodness-of-fit tests.* 1988.

49 F.W. Wubs. *Numerical solution of the shallow-water equations.* 1988.

50 F. de Kerf. *Asymptotic analysis of a class of perturbed Korteweg-de Vries initial value problems.* 1988.

51 P.J.M. van Laarhoven. *Theoretical and computational aspects of simulated annealing.* 1988.

52 P.M. van Loon. *Continuous decoupling transformations for linear boundary value problems.* 1988.

53 K.C.P. Machielsen. *Numerical solution of optimal control problems with state constraints by sequential quadratic programming in function space.* 1988.

54 L.C.R.J. Willenborg. *Computational aspects of survey data processing.* 1988.

55 G.J. van der Steen. *A program generator for recognition, parsing and transduction with syntactic patterns.* 1988.

56 J.C. Ebergen. *Translating programs into delay-insensitive circuits.* 1989.

57 S.M. Verduyn Lunel. *Exponential type calculus for linear delay equations.* 1989.

58 M.C.M. de Gunst. *A random model for plant cell population growth.* 1989.

59 D. van Dulst. *Characterizations of Banach spaces not containing l^1.* 1989.

60 H.E. de Swart. *Vacillation and predictability properties of low-order atmospheric spectral models.* 1989.

61 P. de Jong. *Central limit theorems for generalized multilinear forms.* 1989.

62 V.J. de Jong. *A specification system for statistical software.* 1989.

63 B. Hanzon. *Identifiability, recursive identification and spaces of linear dynamical systems, part I.* 1989.

64 B. Hanzon. *Identifiability, recursive identification and spaces of linear dynamical systems, part II.* 1989.

65 B.M.M. de Weger. *Algorithms for diophantine equations.* 1989.

66 A. Jung. *Cartesian closed categories of domains.* 1989.

67 J.W. Polderman. *Adaptive control & identification: Conflict or conflux?.* 1989.

68 H.J. Woerdeman. *Matrix and operator extensions.* 1989.

69 B.G. Hansen. *Monotonicity properties of infinitely divisible distributions.* 1989.

70 J.K. Lenstra, H.C. Tijms, A. Volgenant (eds.). *Twenty-five years of operations research in the Netherlands: Papers dedicated to Gijs de Leve.* 1990.

71 P.J.C. Spreij. *Counting process systems. Identification and stochastic realization.* 1990.

72 J.F. Kaashoek. *Modeling one dimensional pattern formation by anti-diffusion.* 1990.

73 A.M.H. Gerards. *Graphs and polyhedra. Binary spaces and cutting planes.* 1990.

74 B. Koren. *Multigrid and defect correction for the steady Navier-Stokes equations. Application to aerodynamics.* 1991.

75 M.W.P. Savelsbergh. *Computer aided routing.* 1992.

76 O.E. Flippo. *Stability, duality and decomposition in general mathematical programming.* 1991.

77 A.J. van Es. *Aspects of nonparametric density estimation.* 1991.

78 G.A.P. Kindervater. *Exercises in parallel combinatorial computing.* 1992.

79 J.J. Lodder. *Towards a symmetrical theory of generalized functions.* 1991.

80 S.A. Smulders. *Control of freeway traffic flow.* 1996.

81 P.H.M. America, J.J.M.M. Rutten. *A parallel object-oriented language: design and semantic foundations.* 1992.

82 F. Thuijsman. *Optimality and equilibria in stochastic games.* 1992.

83 R.J. Kooman. *Convergence properties of recurrence sequences.* 1992.

84 A.M. Cohen (ed.). *Computational aspects of Lie group representations and related topics. Proceedings of the 1990 Computational Algebra Seminar at CWI, Amsterdam.* 1991.

85 V. de Valk. *One-dependent processes.* 1994.

86 J.A. Baars, J.A.M. de Groot. *On topological and linear equivalence of certain function spaces.* 1992.

87 A.F. Monna. *The way of mathematics and mathematicians.* 1992.

88 E.D. de Goede. *Numerical methods for the three-dimensional shallow water equations.* 1993.

89 M. Zwaan. *Moment problems in Hilbert space with applications to magnetic resonance imaging.* 1993.

90 C. Vuik. *The solution of a one-dimensional Stefan problem.* 1993.

91 E.R. Verheul. *Multimedians in metric and normed spaces.* 1993.

92 J.L.M. Maubach. *Iterative methods for non-linear partial differential equations.* 1994.

93 A.W. Ambergen. *Statistical uncertainties in posterior probabilities.* 1993.

94 P.A. Zegeling. *Moving-grid methods for time-dependent partial differential equations.* 1993.

95 M.J.C. van Pul. *Statistical analysis of software reliability models.* 1993.

96 J.K. Scholma. *A Lie algebraic study of some integrable systems associated with root systems.* 1993.

97 J.L. van den Berg. *Sojourn times in feedback and processor sharing queues.* 1993.

98 A.J. Koning. *Stochastic integrals and goodness-of-fit tests.* 1993.

99 B.P. Sommeijer. *Parallelism in the numerical integration of initial value problems.* 1993.

100 J. Molenaar. *Multigrid methods for semiconductor device simulation.* 1993.

101 H.J.C. Huijberts. *Dynamic feedback in nonlinear synthesis problems.* 1994.

102 J.A.M. van der Weide. *Stochastic processes and point processes of excursions.* 1994.

103 P.W. Hemker, P. Wesseling (eds.). *Contributions to multigrid.* 1994.

104 I.J.B.F. Adan. *A compensation approach for queueing problems.* 1994.

105 O.J. Boxma, G.M. Koole (eds.). *Performance evaluation of parallel and distributed systems - solution methods. Part 1.* 1994.

106 O.J. Boxma, G.M. Koole (eds.). *Performance evaluation of parallel and distributed systems - solution methods. Part 2.* 1994.

107 R.A. Trompert. *Local uniform grid refinement for time-dependent partial differential equations.* 1995.

108 M.N.M. van Lieshout. *Stochastic geometry models in image analysis and spatial statistics.* 1995.

109 R.J. van Glabbeek. *Comparative concurrency semantics and refinement of actions.* 1996.

110 W. Vervaat, H. Holwerda (ed.). *Probability and lattices.* 1997.

111 I. Helsloot. *Covariant formal group theory and some applications.* 1995.

112 R.N. Bol. *Loop checking in logic programming.* 1995.

113 G.J.M. Koole. *Stochastic scheduling and dynamic programming.* 1995.

114 M.J. van der Laan. *Efficient and inefficient estimation in semiparametric models.* 1995.

115 S.C. Borst. *Polling models.* 1996.

116 G.D. Otten. *Statistical test limits in quality control.* 1996.

117 K.G. Langendoen. *Graph reduction on shared-memory multiprocessors.* 1996.

118 W.C.A. Maas. *Nonlinear $\mathcal{H}_\infty$ control: the singular case.* 1996.

119 A. Di Bucchianico. *Probabilistic and analytical aspects of the umbral calculus*. 1997.

120 M. van Loon. *Numerical methods in smog prediction*. 1997.

121 B.J. Wijers. *Nonparametric estimation for a windowed line-segment process*. 1997.

122 W.K. Klein Haneveld, O.J. Vrieze, L.C.M. Kallenberg (editors). *Ten years LNMB – Ph.D. research and graduate courses of the Dutch Network of Operations Research*. 1997.

123 R.W. van der Hofstad. *One-dimensional random polymers*. 1998.

124 W.J.H. Stortelder. *Parameter estimation in nonlinear dynamical systems*. 1998.

125 M.H. Wegkamp. *Entropy methods in statistical estimation*. 1998.

126 K. Aardal, J.K. Lenstra, F. Maffioli, D.B. Shmoys (eds.) *Selected publications of Eugene L. Lawler*. 1999.

127 E. Belitser. *Minimax estimation in regression and random censorship models*. 2000.

128 Y. Nishiyama. *Entropy methods for martingales*. 2000.

129 J.A. van Hamel. *Algebraic cycles and topology of real algebraic varieties*. 2000.

130 P.J. Oonincx. *Mathematical signal analysis: wavelets, Wigner distribution and a seismic application*. 2000.

131 M. Ruzhansky. *Regularity theory of Fourier integral operators with complex phases and singularities of affine fibrations*. 2001.

132 J.V. Stokman. *Multivariable orthogonal polynomials and quantum Grassmannians*. 2001.

133 N.R. Bruin. *Chabauty methods and covering techniques applied to generalised Fermat equations*. 2002.

134 E.H. van Brummelen. *Numerical methods for steady viscous free-surface flows*. 2003.

135 K. Dajani, J. von Reis (eds.). *European Women in Mathematics – Marseille 2003. Proceedings of the 11th conference of EWM*. 2005.

MATHEMATICAL CENTRE TRACTS

1 T. van der Walt. *Fixed and almost fixed points.* 1963.

2 A.R. Bloemena. *Sampling from a graph.* 1964.

3 G. de Leve. *Generalized Markovian decision processes, part I: model and method.* 1964.

4 G. de Leve. *Generalized Markovian decision processes, part II: probabilistic background.* 1964.

5 G. de Leve, H.C. Tijms, P.J. Weeda. *Generalized Markovian decision processes, applications.* 1970.

6 M.A. Maurice. *Compact ordered spaces.* 1964.

7 W.R. van Zwet. *Convex transformations of random variables.* 1964.

8 J.A. Zonneveld. *Automatic numerical integration.* 1964.

9 P.C. Baayen. *Universal morphisms.* 1964.

10 E.M. de Jager. *Applications of distributions in mathematical physics.* 1964.

11 A.B. Paalman-de Miranda. *Topological semigroups.* 1964.

12 J.A.Th.M. van Berckel, H. Brandt Corstius, R.J. Mokken, A. van Wijngaarden. *Formal properties of newspaper Dutch.* 1965.

13 H.A. Lauwerier. *Asymptotic expansions.* 1966, out of print: replaced by MCT 54.

14 H.A. Lauwerier. *Calculus of variations in mathematical physics.* 1966.

15 R. Doornbos. *Slippage tests.* 1966.

16 J.W. de Bakker. *Formal definition of programming languages with an application to the definition of ALGOL 60.* 1967.

17 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 1.* 1968.

18 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 2.* 1968.

19 J. van der Slot. *Some properties related to compactness.* 1968.

20 P.J. van der Houwen. *Finite difference methods for solving partial differential equations.* 1968.

21 E. Wattel. *The compactness operator in set theory and topology.* 1968.

22 T.J. Dekker. *ALGOL 60 procedures in numerical algebra, part 1.* 1968.

23 T.J. Dekker, W. Hoffmann. *ALGOL 60 procedures in numerical algebra, part 2.* 1968.

24 J.W. de Bakker. *Recursive procedures.* 1971.

25 E.R. Paërl. *Representations of the Lorentz group and projective geometry.* 1969.

26 European Meeting 1968. *Selected statistical papers, part I.* 1968.

27 European Meeting 1968. *Selected statistical papers, part II.* 1968.

28 J. Oosterhoff. *Combination of one-sided statistical tests.* 1969.

29 J. Verhoeff. *Error detecting decimal codes.* 1969.

30 H. Brandt Corstius. *Exercises in computational linguistics.* 1970.

31 W. Molenaar. *Approximations to the Poisson, binomial and hypergeometric distribution functions.* 1970.

32 L. de Haan. *On regular variation and its application to the weak convergence of sample extremes.* 1970.

33 F.W. Steutel. *Preservations of infinite divisibility under mixing and related topics.* 1970.

34 I. Juhász, A. Verbeek, N.S. Kroonenberg. *Cardinal functions in topology.* 1971.

35 M.H. van Emden. *An analysis of complexity.* 1971.

36 J. Grasman. *On the birth of boundary layers.* 1971.

37 J.W. de Bakker, G.A. Blaauw, A.J.W. Duijvestijn, E.W. Dijkstra, P.J. van der Houwen, G.A.M. Kamsteeg-Kemper, F.E.J. Kruseman Aretz, W.L. van der Poel, J.P. Schaap-Kruseman, M.V. Wilkes, G. Zoutendijk. *MC-25 Informatica Symposium.* 1971.

38 W.A. Verloren van Themaat. *Automatic analysis of Dutch compound words.* 1972.

39 H. Bavinck. *Jacobi series and approximation.* 1972.

40 H.C. Tijms. *Analysis of (s,S) inventory models.* 1972.

41 A. Verbeek. *Superextensions of topological spaces.* 1972.

42 W. Vervaat. *Success epochs in Bernoulli trials (with applications in number theory).* 1972.

43 F.H. Ruymgaart. *Asymptotic theory of rank tests for independence.* 1973.

44 H. Bart. *Meromorphic operator valued functions.* 1973.

45 A.A. Balkema. *Monotone transformations and limit laws.* 1973.

46 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 1: the language.* 1973.

47 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 2: the compiler.* 1973.

48 F.E.J. Kruseman Aretz, P.J.W. ten Hagen, H.L. Oudshoorn. *An ALGOL 60 compiler in ALGOL 60, text of the MC-compiler for the EL-X8.* 1973.

49 H. Kok. *Connected orderable spaces.* 1974.

50 A. van Wijngaarden, B.J. Mailloux, J.E.L. Peck, C.H.A. Koster, M. Sintzoff, C.H. Lindsey, L.G.L.T. Meertens, R.G. Fisker (eds.). *Revised report on the algorithmic language ALGOL 68.* 1976.

51 A. Hordijk. *Dynamic programming and Markov potential theory.* 1974.

52 P.C. Baayen (ed.). *Topological structures.* 1974.

53 M.J. Faber. *Metrizability in generalized ordered spaces.* 1974.

54 H.A. Lauwerier. *Asymptotic analysis, part 1.* 1974.

55 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 1: theory of designs, finite geometry and coding theory.* 1974.

56 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 2: graph theory, foundations, partitions and combinatorial geometry.* 1974.

57 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 3: combinatorial group theory.* 1974.

58 W. Albers. *Asymptotic expansions and the deficiency concept in statistics.* 1975.

59 J.L. Mijnheer. *Sample path properties of stable processes.* 1975.

60 F. Göbel. *Queueing models involving buffers.* 1975.

63 J.W. de Bakker (ed.). *Foundations of computer science.* 1975.

64 W.J. de Schipper. *Symmetric closed categories.* 1975.

65 J. de Vries. *Topological transformation groups, 1: a categorical approach.* 1975.

66 H.G.J. Pijls. *Logically convex algebras in spectral theory and eigenfunction expansions.* 1976.

68 P.P.N. de Groen. *Singularly perturbed differential operators of second order.* 1976.

69 J.K. Lenstra. *Sequencing by enumerative methods.* 1977.

70 W.P. de Roever, Jr. *Recursive program schemes: semantics and proof theory.* 1976.

71 J.A.E.E. van Nunen. *Contracting Markov decision processes.* 1976.

72 J.K.M. Jansen. *Simple periodic and non-periodic Lamé functions and their applications in the theory of conical waveguides.* 1977.

73 D.M.R. Leivant. *Absoluteness of intuitionistic logic.* 1979.

74 H.J.J. te Riele. *A theoretical and computational study of generalized aliquot sequences.* 1976.

75 A.E. Brouwer. *Treelike spaces and related connected topological spaces.* 1977.

76 M. Rem. *Associons and the closure statements.* 1976.

77 W.C.M. Kallenberg. *Asymptotic optimality of likelihood ratio tests in exponential families.* 1978.

78 E. de Jonge, A.C.M. van Rooij. *Introduction to Riesz spaces.* 1977.

79 M.C.A. van Zuijlen. *Empirical distributions and rank statistics.* 1977.

80 P.W. Hemker. *A numerical study of stiff two-point boundary problems.* 1977.

81 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 1.* 1976.

82 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 2.* 1976.

83 L.S. van Benthem Jutting. *Checking Landau's "Grundlagen" in the AUTOMATH system.* 1979.

84 H.L.L. Busard. *The translation of the elements of Euclid from the Arabic into Latin by Hermann of Carinthia (?), books vii-xii.* 1977.

85 J. van Mill. *Supercompactness and Wallmann spaces.* 1977.

86 S.G. van der Meulen, M. Veldhorst. *Torrix I, a programming system for operations on vectors and matrices over arbitrary fields and of variable size.* 1978.

88 A. Schrijver. *Matroids and linking systems.* 1977.

89 J.W. de Roever. *Complex Fourier transformation and analytic functionals with unbounded carriers.* 1978.

90 L.P.J. Groenewegen. *Characterization of optimal strategies in dynamic games.* 1981.

91 J.M. Geysel. *Transcendence in fields of positive characteristic.* 1979.

92 P.J. Weeda. *Finite generalized Markov programming.* 1979.

93 H.C. Tijms, J. Wessels (eds.). *Markov decision theory.* 1977.

94 A. Bijlsma. *Simultaneous approximations in transcendental number theory.* 1978.

95 K.M. van Hee. *Bayesian control of Markov chains.* 1978.

96 P.M.B. Vitányi. *Lindenmayer systems: structure, languages, and growth functions.* 1980.

97 A. Federgruen. *Markovian control problems; functional equations and algorithms.* 1984.

98 R. Geel. *Singular perturbations of hyperbolic type.* 1978.

99 J.K. Lenstra, A.H.G. Rinnooy Kan, P. van Emde Boas (eds.). *Interfaces between computer science and operations research.* 1978.

100 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 1.* 1979.

101 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 2.* 1979.

102 D. van Dulst. *Reflexive and superreflexive Banach spaces.* 1978.

103 K. van Harn. *Classifying infinitely divisible distributions by functional equations.* 1978.

104 J.M. van Wouwe. *GO-spaces and generalizations of metrizability.* 1979.

105 R. Helmers. *Edgeworth expansions for linear combinations of order statistics.* 1982.

106 A. Schrijver (ed.). *Packing and covering in combinatorics.* 1979.

107 C. den Heijer. *The numerical solution of nonlinear operator equations by imbedding methods.* 1979.

108 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 1.* 1979.

109 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 2.* 1979.

110 J.C. van Vliet. *ALGOL 68 transput, part I: historical review and discussion of the implementation model.* 1979.

111 J.C. van Vliet. *ALGOL 68 transput, part II: an implementation model.* 1979.

112 H.C.P. Berbee. *Random walks with stationary increments and renewai theory.* 1979.

113 T.A.B. Snijders. *Asymptotic optimality theory for testing problems with restricted alternatives.* 1979.

114 A.J.E.M. Janssen. *Application of the Wigner distribution to harmonic analysis of generalized stochastic processes.* 1979.

115 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 1.* 1979.

116 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 2.* 1979.

117 P.J.M. Kallenberg. *Branching processes with continuous state space.* 1979.

118 P. Groeneboom. *Large deviations and asymptotic efficiencies.* 1980.

119 F.J. Peters. *Sparse matrices and substructures, with a novel implementation of finite element algorithms.* 1980.

120 W.P.M. de Ruyter. *On the asymptotic analysis of large-scale ocean circulation.* 1980.

121 W.H. Haemers. *Eigenvalue techniques in design and graph theory.* 1980.

122 J.C.P. Bus. *Numerical solution of systems of nonlinear equations.* 1980.

123 I. Yuhász. *Cardinal functions in topology - ten years later.* 1980.

124 R.D. Gill. *Censoring and stochastic integrals.* 1980.

125 R. Eising. *2-D systems, an algebraic approach.* 1980.

126 G. van der Hoek. *Reduction methods in nonlinear programming.* 1980.

127 J.W. Klop. *Combinatory reduction systems.* 1980.

128 A.J.J. Talman. *Variable dimension fixed point algorithms and triangulations.* 1980.

129 G. van der Laan. *Simplicial fixed point algorithms.* 1980.

130 P.J.W. ten Hagen, T. Hagen, P. Klint, H. Noot, H.J. Sint, A.H. Veen. *ILP: intermediate language for pictures.* 1980.

131 R.J.R. Back. *Correctness preserving program refinements: proof theory and applications.* 1980.

132 H.M. Mulder. *The interval function of a graph.* 1980.

133 C.A.J. Klaassen. *Statistical performance of location estimators.* 1981.

134 J.C. van Vliet, H. Wupper (eds.). *Proceedings international conference on ALGOL 68.* 1981.

135 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part I.* 1981.

136 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part II.* 1981.

137 J. Telgen. *Redundancy and linear programs.* 1981.

138 H.A. Lauwerier. *Mathematical models of epidemics.* 1981.

139 J. van der Wal. *Stochastic dynamic programming, successive approximations and nearly optimal strategies for Markov decision processes and Markov games.* 1981.

140 J.H. van Geldrop. *A mathematical theory of pure exchange economies without the no-critical-point hypothesis.* 1981.

141 G.E. Welters. *Abel-Jacobi isogenies for certain types of Fano threefolds.* 1981.

142 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 1.* 1981.

143 J.M. Schumacher. *Dynamic feedback in finite- and infinite-dimensional linear systems.* 1981.

144 P. Eijgenraam. *The solution of initial value problems using interval arithmetic; formulation and analysis of an algorithm.* 1981.

145 A.J. Brentjes. *Multi-dimensional continued fraction algorithms.* 1981.

146 C.V.M. van der Mee. *Semigroup and factorization methods in transport theory.* 1981.

147 H.H. Tigelaar. *Identification and informative sample size.* 1982.

148 L.C.M. Kallenberg. *Linear programming and finite Markovian control problems.* 1983.

149 C.B. Huijsmans, M.A. Kaashoek, W.A.J. Luxemburg, W.K. Vietsch (eds.). *From A to Z, proceedings of a symposium in honour of A.C. Zaanen.* 1982.

150 M. Veldhorst. *An analysis of sparse matrix storage schemes.* 1982.

151 R.J.M.M. Does. *Higher order asymptotics for simple linear rank statistics.* 1982.

152 G.F. van der Hoeven. *Projections of lawless sequencies.* 1982.

153 J.P.C. Blanc. *Application of the theory of boundary value problems in the analysis of a queueing model with paired services.* 1982.

154 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part I.* 1982.

155 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part II.* 1982.

156 P.M.G. Apers. *Query processing and data allocation in distributed database systems.* 1983.

157 H.A.W.M. Kneppers. *The covariant classification of two-dimensional smooth commutative formal groups over an algebraically closed field of positive characteristic.* 1983.

158 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 1.* 1983.

159 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 2.* 1983.

160 A. Rezus. *Abstract AUTOMATH.* 1983.

161 G.F. Helminck. *Eisenstein series on the metaplectic group, an algebraic approach.* 1983.

162 J.J. Dik. *Tests for preference.* 1983.

163 H. Schippers. *Multiple grid methods for equations of the second kind with applications in fluid mechanics.* 1983.

164 F.A. van der Duyn Schouten. *Markov decision processes with continuous time parameter.* 1983.

165 P.C.T. van der Hoeven. *On point processes.* 1983.

166 H.B.M. Jonkers. *Abstraction, specification and implementation techniques, with an application to garbage collection.* 1983.

167 W.H.M. Zijm. *Nonnegative matrices in dynamic programming.* 1983.

168 J.H. Evertse. *Upper bounds for the numbers of solutions of diophantine equations.* 1983.

169 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 2.* 1983.