Fractals in Engineering

Jacques Lévy-Véhel and Evelyne Lutton (Eds.)

Fractals in Engineering

New Trends in Theory and Applications

With 106 Figures

Springer

Jacques Lévy-Véhel
Evelyne Lutton

INRIA
Rocquencourt
Domaine de Voluceau-Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France

British Library Cataloguing in Publication Data
Fractals in engineering : new trends in theory and
applications
1. Engineering mathematics 2. Fractals
I. Lévy-Véhel, Jacques, 1960- II. Lutton, Evelyne, 1962-
620'.001514742
ISBN-10: 1846280478

Library of Congress Control Number: 2005927902

ISBN-10: 1-84628-047-8 e-ISBN: 1-84628-048-6 Printed on acid-free paper
ISBN-13: 978-1-84628-047-4

Printed in Germany

9 8 7 6 5 4 3 2 1

Springer Science+Business Media
springeronline.com

Foreword

This volume is a sequel to the books *Fractals: Theory and Applications in Engineering* (Springer-Verlag, 1999) and *Fractals in Engineering. From Theory to Industrial Applications* (Springer-Verlag, 1997), presenting some of the most recent advances in the field. It is a fascinating exercise to follow the progress of knowledge in this interdisciplinary area, as witnessed by these three volumes.

First, confirming previous trends observed in 1997 and 1999, applied mathematical research on fractals has now reached a mature level, where beautiful theories are developed in direct contact with engineering concerns. The four papers in the *Mathematical Aspects* section constitute valuable additions to the set of tools needed by the engineer: Synthetic pictures modelling and rendering in computer graphics (*Theory and Applications of Fractal Tops*, by Michael Barnsley), curve approximation and "fractal B-splines" (*Splines, Fractal Functions, and Besov and Triebel-Lizorkin Spaces*, by Peter Massopust), deep understanding of the Hölderian properties of certain stochastic processes useful in a large number of applications (*Hölderian random functions*, by Antoine Ayache *et al.*), and study of the invariant measure of a coupled discrete dynamical system (*Fractal Stationary Density in Coupled Maps*, by Jürgen Jost *et al.*).

The second section of the book describes novel physical applications as well as recent progress on more classical ones. The paper *A Network of Fractal Force Chains and Their Effect in Granular Materials under Compression* by Luis E. Vallejo *et al.* offers an explanation to the well-known experimental fact that granular material develop fractal fragments as a result of compres-

sion. In *Percolation and permeability of three dimensional fracture networks with a power law size distribution*, V.V. Mourzenko *et al.* provide a new and interesting addition to the large body of work devoted to fractal analysis of percolation in fracture networks. They perform a thorough numerical study of percolation in polydisperse fracture networks, allowing to define an appropriate percolation parameter and to develop two heuristic analytical models. A new and very promising application of fractal analysis to acoustics in the frame of urban structures is developed by Philippe Woloszyn in *Acoustic diffraction patterns from fractal to urban structures: Applications to the Sierpinski triangle and to a neoclassical urban facade*. Rolf Bader develops another application to acoustics, proposing an interesting *Turbulent $k - \epsilon$ model of flute-like musical instrument sound production.*

The section on *Chemical Engineering* features two papers. In *A simple discrete stochastic model for laser-induced jet-chemical etching*, Alejandro Mora *et al.* describe a discrete stochastic model for the description of laser-induced wet-chemical etching. This model enables one to describe the aspect of the surface depending on the velocity of the laser beam. A deep study of fluid mixing in two dimensions is made in *Invariant structures and multifractal measures in 2d mixing systems* by Massimiliano Giona *al.*, through a connection between geometric invariant structures and the spatial distribution of periodic points.

Fractal modelling of financial time series has a long and rich history. The section on *Finance* focuses on the specific question of long range dependence, with two papers. In *Long range dependence in financial markets*, Rama Cont discusses the relevance of this property in financial modelling, and highlights possible economic mechanisms accounting for its presence in financial time series. Pierre Bertrand derives in *Financial Modelling by Multiscale Fractional Brownian Motion* the price of a European option for this model of stock prices.

Application of fractal analysis to Internet traffic, which is the topic of the fifth section, started in the 1990's, and an extremely large number of studies have been devoted to this topic in recent years. The paper *Limiting Fractal Random Processes in Heavy-Tailed Systems* by Ingemar Kaj investigates the asymptotic behavior of stochastic processes build through aggregation of independent subsystems and simultaneous time rescaling. This behavior depends considerably on the relative speed of aggregation degree and rescaling. Although primarily of interest in telecommunications, these results extend in higher dimensions (e.g. spatial Poisson point processes). The concept of crossing tree previously introduced by the authors for estimating the Hurst index of self-similar processes is used as a tool for *A non-parametric test for self-similarity and stationarity in network traffic*, by Owen Jones *et al.*.

The last section deals with applications in image processing. In *Continuous evolution of functions and measures toward fixed points of contraction mappings*, Jerry Bona *et al.* study a class of evolution equations associated with contraction mappings on a Banach space of functions. This enables one to perform continuous, fractal-like, "touch-up" operations on images. Fahima

Nekka *et al.* use the autocorrelation function, the regularization dimension as well as the Hausdorff measure spectrum function to analyze textures in *Various Mathematical Approaches to Extract Information from Textures of Increasing Complexities*. The celebrated inverse problem of fractal coding is the topic of *Fractal Inverse Problem: Approximation Formulation and Differential Methods* by Eric Guérin *et al.* Using an analytical approach, they obtain interesting results both in one and two dimensions.

While it is obviously impossible to cover the wealth of all applications of fractal analysis in engineering sciences in a single volume, this book does provide an overview of some of the more prominent recent advances, which should be of interest to anyone willing to keep up with the fast pace of development in this field.

We would like to thank all the authors who have contributed to this book. Thanks also to Nathalie Gaudechoux for her Latex skills. Finally, we are grateful to INRIA and our publisher Springer-Verlag for their support.

Jacques Lévy Véhel,
Evelyne Lutton.

Contents

List of Contributors

Pierre Adler
IPGP
Tour 24
4 Place Jussieu
75252 Paris Cedex 05, France
adler@ipgp.jussieu.fr

Alessandra Adrover
Dipartimento di Ingegneria Chimica
Facoltà di Ingegneria
Università di Roma "La Sapienza"
via Eudossiana 18
00184 Roma, Italy
alex@giona.ing.uniroma1.it

Antoine Ayache
Laboratoire P. Painlevé,
CNRS UMR 8524, Université Lille 1,
59650 Villeneuve d'Ascq cedex,
France
Antoine.Ayache@math.univ-lille1.fr

Rolf Bader
University of Hamburg
Institute of Musicology
Neue Rabenstr. 13
20354 Hamburg, Germany
R_Bader@t-online.de

Michael Barnsley
Australian National University
Canberra, ACT 0200, Australia
mbarnsley@aol.com

Pierre Bertrand
Laboratoire de Mathématiques -
UMR CNRS 6620,
Université Blaise Pascal (Clermont-
Ferrand II),
24 Avenue des Landais, 63117
Aubière Cedex, France.
Pierre.Bertrand@math.univ-bpclermont.fr

Jerry L. Bona
Department of Mathematics,
Statistics and Computer Science
The University of Illinois at Chicago
Chicago, Illinois, USA 60607-7045
bona@math.uic.edu

Stefano Cerbelli
Dipartimento di Ingegneria Chimica
Facoltà di Ingegneria
Università di Roma "La Sapienza"
via Eudossiana 18
00184 Roma, Italy
stefano@giona.ing.uniroma1.it

Zamri Chik
Department of Civil and
Environmental Engineering
University of Pittsburgh
Pittsburgh, PA 15261, U.S.A.
irzamri@yahoo.com

Rama Cont
Centre de Mathématiques appliquées,
Ecole Polytechnique, France
Rama.Cont@polytechnique.fr

Massimiliano Giona
Dipartimento di Ingegneria Chimica
Facoltà di Ingegneria
Università di Roma "La Sapienza"
via Eudossiana 18
00184 Roma, Italy
max@giona.ing.uniroma1.it

Eric Guérin
LIRIS
Université Claude Bernard
Bâtiment Nautibus
43, Bd du 11 Novembre
69622 Villeurbanne Cedex, France
eric.guerin@liris.cnrs.fr

Maria Haase
Institut für Höchstleistungsrechnen
(IHR)
University of Stuttgart
70569 Stuttgart, Germany
mh@ica.uni-stuttgart.de

Philippe Heinrich
Laboratoire P. Painlevé
CNRS UMR 8524
Université Lille 1
59650 Villeneuve d'Ascq cedex,
France
Philippe.Heinrich@math.univ-lille1.fr

Owen Dafydd Jones
Department of Mathematics &
Statistics
The University of Melbourne
Parkville, VIC, 3010, Australia
o.d.jones@ms.unimelb.edu.au

Jürgen Jost
Max Planck Institute for
Mathematics in the Sciences
Inselstrasse 22-26
D-04103 Leipzig, Germany
jjost@mis.mpg.de

Ingemar Kaj
Dept. of Mathematics
Uppsala University
Box 480
SE 751 06 Uppsala, Sweden
ikaj@math.uu.se

Kiran M. Kolwankar
Max Planck Institute for
Mathematics in the Sciences
Inselstrasse 22-26
D-04103 Leipzig, Germany
Kiran.Kolwankar@mis.mpg.de

Bernd Lehle
vFlow Engineering GmbH
70499 Stuttgart, Germany

Jun Li
Faculté de Pharmacie and Centre de
Recherches Mathématiques
Université de Montréal, C.P. 6128
Succ. Centre-ville
Montréal (Québec),
Canada H3C 3J7
li@crm.umontreal.ca

Sebastian Lobo-Guerrero
Department of Civil and Environ-
mental Engineering
University of Pittsburgh
Pittsburgh, PA 15261, U.S.A.
sel2@pitt.edu

Laurence Marsalle
Laboratoire P. Painlevé
CNRS UMR 8524
Université Lille 1
59650 Villeneuve d'Ascq cedex,
France
Laurence.Marsalle@math.univ-lille1.fr

Peter Massopust
Engineering and Research Development
Tuboscope Pipeline Services
2835 Holmes Road
Houston, TX 77051, USA
pmassopust@varco.com

Alejandro Mora
Institut für Höchstleistungsrechnen (IHR)
University of Stuttgart
70569 Stuttgart, Germany
ica2am@csv.ica.uni-stuttgart.de

V.V. Mourzenko
LCD, SP2MI
BP 179
86960 Futuroscope Cedex, France
mourzenko@lcd.ensma.fr

Fahima Nekka
Faculté de Pharmacie and Centre de Recherches Mathématiques
Université de Montréal C.P. 6128
Succ. Centre-ville
Montréal (Québec),
Canada H3C 3J7
fahima.nekka@umontreal.ca

Peter J. Plath
Institut für Angewandte und Physikalische Chemie
Chemische Synergetik
University of Bremen
28334, Germany

Thomas Rabbow
Institut für Angewandte und Physikalische Chemie
Chemische Synergetik
University of Bremen
28334, Germany

Yuan Shen
Dept of Statistics
University of Warwick
Coventry, CV4 7AL,
United Kingdom
shen@stats.warwick.ac.uk

Charles Suquet
Laboratoire P. Painlevé
CNRS UMR 8524
Université Lille 1
59650 Villeneuve d'Ascq cedex,
France
Charles.Suquet@math.univ-lille1.fr

Jean-François Thovert
LCD
SP2MI
BP 179
86960 Futuroscope Cedex, France
thovert@lcd.ensma.fr

Eric Tosan
LIRIS
Université Claude Bernard
Bâtiment Nautibus
43, Bd du 11 Novembre
69622 Villeurbanne Cedex, France
eric.tosan@liris.cnrs.fr

Luis E. Vallejo
Department of Civil and Environmental Engineering
University of Pittsburgh
Pittsburgh, PA 15261, U.S.A.
vallejo@engrng.pitt.edu

Edward R. Vrscay
Department of Applied Mathematics
University of Waterloo
Waterloo
Ontario, Canada N2L 3G1
ervrscay@uwaterloo.ca

Philippe Woloszyn
Acoustic dept. Cerma Lab.
UMR CNRS 1563
E.A.N.
rue Massenet, BP 81931
F-44319 Nantes Cedex 3, France
philippe.woloszyn@cerma.archi.fr

1

MATHEMATICAL ASPECTS

Theory and Applications of Fractal Tops

Michael Barnsley

Australian National University, Canberra
mbarnsley@aol.com

Summary. We consider an iterated function system (IFS) of one-to-one contractive maps on a compact metric space. We define the **top** of an IFS; define an associated symbolic dynamical system; present and explain a fast algorithm for computing the top; describe an example in one dimension with a rich history going back to work of A.Rényi [*Representations for Real Numbers and Their Ergodic Properties*, Acta Math. Acad. Sci. Hung.,**8** (1957), pp. 477-493]; and we show how tops may be used to help to model and render synthetic pictures in applications in computer graphics.

1 Introduction

It is well-known that an iterated function system (IFS) of 1-1 contractive maps, mapping a compact metric space into itself, possesses a set attractor and various invariant measures. But it also possesses another type of invariant object which we call a *top*. One application of tops is to modelling and rendering new families of synthetic pictures in computer graphics. Another application is to information theory and data compression. Tops are mathematically fascinating because they have a rich symbolic dynamics structure, they support intricate Markov chains, and they provide examples of IFS with place-dependent probabilities in a regime where not much research has taken place.

In this paper we define the top of an IFS; define an associated symbolic dynamical system; present and explain a fast algorithm for computing the top; describe an example in one dimension with a rich history going back to work of A.Rényi [11]; and we show how tops may be used to help to model and render synthetic pictures in applications in computer graphics.

This is a short version of a paper, [5], which includes proofs and more detail. This work was supported by the Australian Research Council.

The author thanks John Hutchinson for many useful discussions and much help with this work. The author thanks Louisa Barnsley for editorial help and for producing the graphics. The author thanks a referee for helpful comments.

2 The Top of an IFS

Let an iterated function system (IFS) be denoted

$$\mathcal{W} := \{\mathbb{X}; w_0, ..., w_{N-1}\}. \tag{1}$$

This consists of a finite of sequence of one-to-one contraction mappings

$$w_n : \mathbb{X} \to \mathbb{X}, n = 0, 2, ..., N-1 \tag{2}$$

acting on the compact metric space

$$(\mathbb{X}, d) \tag{3}$$

with metric d so that for some

$$0 \leq l < 1 \tag{4}$$

we have

$$d(w_n(x), w_n(y)) \leq l \cdot d(x, y) \tag{5}$$

for all $x, y \in \mathbb{X}$.

Let A denote the attractor of the IFS, that is $A \subset \mathbb{X}$ is the unique non-empty compact set such that

$$A = \bigcup_n w_n(A).$$

Let the associated code space be denoted by $\Omega = \Omega_{\{0,1,...,N-1\}}$. This is the space of infinite sequences of symbols $\{\sigma_i\}_{i=1}^{\infty}$ belonging to the alphabet $\{0, 1, ..., N-1\}$ with the discrete product topology. We will also write $\sigma = \sigma_1\sigma_2\sigma_3... \in \Omega$ to denote a typical element of Ω, and we will write ω_k to denote the k^{th} element of the sequence $\omega \in \Omega$. We order the elements of Ω according to

$$\sigma < \omega \text{ iff } \sigma_k < \omega_k$$

where k is the least index for which $\sigma_k \neq \omega_k$.

Let

$$\phi : \Omega \to A$$

denote the associated *continuous* addressing map *from* code space *onto* the attractor of the IFS. We note that the set of addresses of a point $x \in A$, defined to be $\phi^{-1}(x)$, is compact and so possesses a unique largest element. We denote this value by $\tau(x)$. That is, $\tau : A \to \Omega$ is defined by

$$\tau(x) = \max\{\sigma \in \Omega : \phi(\sigma) = x\}.$$

We call τ the *tops function* of the IFS. We also call $G_\tau := \{(x, \tau(x)) : x \in A\}$ the graph of the top of the IFS or simply *the top* of the IFS.

The top of an IFS may be described as follows: consider the *lifted* IFS

$$\{\mathbb{X}\times\Omega : W_0, W_1, ..., W_{N-1}\}$$

where

$$W_n(x,\sigma) = (w_n(x), n\sigma)$$

where, for the avoidance of any doubt, $n\sigma := \omega$ where $\omega_1 = n$ and $\omega_{n+1} = \sigma_n$ for $n = 0, 1, ..., N-1$. (The metric d_Ω on Ω is defined, for all $\omega, \sigma \in \Omega$, by $d_\Omega(\omega,\sigma) = 0$ when $\omega = \sigma$, and otherwise $d_\Omega(\omega,\sigma) = \frac{1}{2^k}$ where k is the least integer for which $\sigma_k \neq \omega_k$.) Let the unique attractor of this IFS be denoted by $\widehat{A}$. Then the projection of $\widehat{A}$ on the $\mathbb{X}$-direction is $\mathbb{X}$, and in the Ω-direction it is Ω. The top of the original IFS is related to $\widehat{A}$ according to:

$$G_\tau = \{(x,\sigma) \in \widehat{A} : (x,\omega) \in \widehat{A} \Longrightarrow \omega \leq \sigma\}.$$

This latter formulation is useful because we can use the chaos game algorithm (also called a Markov Chain Monte Carlo (MCMC) algorithm or a random iteration algorithm) to compute approximations to $\widehat{A}$ and hence to G_τ. According to this method we select a sequence of symbols

$$\sigma_1\sigma_2\sigma_3... \in \{1, 2, ..., N\}^\infty$$

with probability $p_n > 0$ for the choice $\sigma_k = n$, independent of all of the other choices. We also select $X_0 \in \mathbb{X}$ and let

$$X_{n+1} = W_{\sigma_{n+1}}(X_n) \text{ for } n = 0, 1, 2, ... \ .$$

Then, almost always,

$$\lim_{k\to\infty} \overline{\{X_n : n = k, k+1, ...\}} = \widehat{A}$$

where $\overline{S}$ denotes the closure of the set S. (The value of the limit is intersection of the decreasing sequence of compact sets whose limit is being taken.) This algorithm provides in many cases a simple efficient fast method to compute approximations to the attractor of an IFS, for example when $\mathbb{X} = \square$, a compact subset of $\mathbb{R}^2$. By keeping track of points which, for each approximate value of $x \in \mathbb{X}$, have the greatest code space value, we can compute approximations to G_τ. We illustrate this approach in the following example which we continue in Section 6.

Example 1. Consider the IFS

$$\{[0,1] \subset \mathbb{R}; w_0(x) = \alpha x, w_2(x) = \alpha x + (1-\alpha)\} \tag{6}$$

We are interested in the case where

$$\frac{1}{2} < \alpha < 1,$$

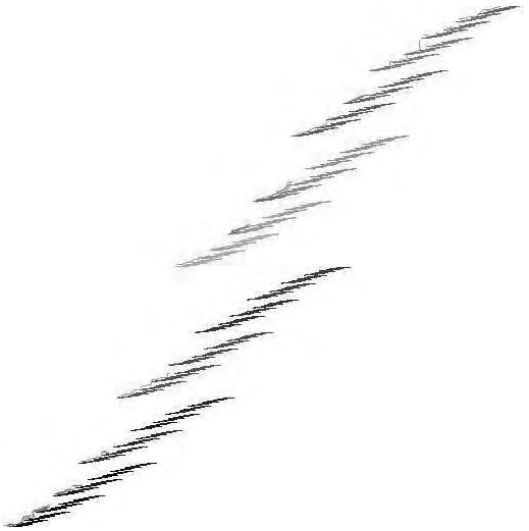


Fig. 1. The attractor of the IFS in Equation 7. This represents the attractor $\widehat{A}$ of the lifted IFS corresponding to Equation 6. The top of the IFS is indicated in red. The visible part of the "x-axis" represents the real interval $[0, 1]$ and the visible part of the "y-axis" represents the code space Ω between the points 000000000.... and 111111111.....

which we refer to as "overlapping" because $w_0([0, 1]) \cap w_1([0, 1])$ contains a non-empty open set. In Figure 1 we show the attractor $\widehat{A}$ of the associated lifted IFS, and upon this attractor we have indicated the top with some red squiggles. Figure 1 was computed using random iteration: we have represented points in code space by their binary expansions which are interpreted as points in $[0, 1]$. Since the invariant measure of both the IFS and the lifted IFS contain no atoms, the information lost by this representation is irrelevant to pictures. Accordingly, the actual IFS used to compute Figure 1 is

$$\{[0,1]\times[0,1] \subset \mathbb{R}; W_0(x,y) = (\alpha x, \frac{1}{2}y), W_2(x,y) = (\alpha x + (1-\alpha), \frac{1}{2}y + \frac{1}{2})\} \quad (7)$$

with $\alpha = \frac{2}{3}$.

3 Application of Tops to Computer Graphics

Here we introduce the application of tops to computer graphics. There is a great deal more to say about this, but to serve as motivation as well as to provide an excellent method for graphing fractal tops, we explain the basic idea here.

A **picture function** is a mapping $\mathfrak{P} : D_{\mathfrak{P}} \subset \mathbb{R}^2 \to \mathfrak{C}$ where $\mathfrak{C}$ is a colour space, for example $\mathfrak{C} = [0, 255]^3 \subset \mathbb{R}^3$. The domain $D_{\mathfrak{P}}$ is typically a rectangular subset of $\mathbb{R}^2$: we often take

$$D_{\mathfrak{P}} = \square := \{(x, y) \in \mathbb{R}^2 : 0 \leq x, y \leq 1\}.$$

The domain of a picture function is an important part of its definition; for example a segment of a picture may be used to define a picture function. A picture in the usual sense may then be thought of as the graph of a picture function. But we will use the concepts of picture, picture function, and graph of a picture function interchangeably. We do not discuss here the important questions of how such functions arise in image science, for example, nor about the relationship between such abstract objects and real world pictures. Here we assume that given a picture function, we have some process by which we can render it to make pictures which may be printed, viewed on computer screens, etc. This is far from a simple matter in general.

Let two IFS's

$$\mathcal{W} := \{\square; w_0, ..., w_{N-1}\} \text{ and } \widetilde{\mathcal{W}} := \{\square; \widetilde{w}_0, ..., \widetilde{w}_{N-1}\}$$

and a picture function

$$\widetilde{\mathfrak{P}} : \square \to \mathfrak{C}$$

be given. Let A denote the attractor of the IFS $\mathcal{W}$ and let $\widetilde{A}$ denote the attractor of the IFS $\widetilde{\mathcal{W}}$. Let

$$\tau : A \to \Omega$$

denote the tops function for $\mathcal{W}$. Let

$$\widetilde{\phi} : \Omega \to \widetilde{A} \subset \square$$

denote the addressing function for the IFS $\widetilde{\mathcal{W}}$. Then we define a new picture function

$$\mathfrak{P} : A \to \mathfrak{C}$$

by

$$\mathfrak{P} = \widetilde{\mathfrak{P}} \circ \widetilde{\phi} \circ \tau.$$

This is the **unique picture function** defined by the IFS's $\mathcal{W}$, $\widetilde{\mathcal{W}}$, and the picture $\widetilde{\mathfrak{P}}$. We say that it has been **produced by tops + colour stealing**. We think in this way: colours are "stolen" from the picture $\widetilde{\mathfrak{P}}$ to "paint" code space; that is, we make a code space picture, that is the function $\widetilde{\mathfrak{P}} \circ \widetilde{\phi} : \Omega \to \mathfrak{C}$, which we then use together with top of $\mathcal{W}$ to paint the attractor A.

Notice the following points. (i) Picture functions have properties that are determined by their source; digital pictures of natural scenes such as clouds and sky, fields of flowers and grasses, seascapes, thick foliage, etc. all have their own distinctive palettes, relationships between colour and position, "continuity" and "discontinuity" properties, and so on. (ii) Addressing functions are

Fig. 2. Colours were stolen from this picture to produce Figure 4 and the right-hand image in Figure 3.

continuous. (iii) Tops functions have their own special properties; for example they are continuous when the associated IFS is totally disconnected, and they contain the geometry of the underlying IFS attractor A plus much more, and so may have certain self-similarities and, assuming the IFS are built from low information content transformations such as similitudes, possess their own harmonies. Thus, the picture functions produced by tops plus colour stealing may define pictures which are interesting to look at, carrying a natural palette, possessing certain continuities and discontinuities, and also certain self-similarities. There is much more one could say here.

The stolen picture $\widetilde{\mathfrak{P}} \circ \widetilde{\phi} \circ \tau$ may be computed by random iteration, by coupling the lifted IFS associated with $\mathcal{W}$ to the IFS $\widetilde{\mathcal{W}}$. This is the method used until recently, and has been described in [3] and in [4]. Recently we have discovered a much faster algorithm for computing the entire stolen picture at a given resolution. It is based on a symbolic dynamical system associated with the top of $\mathcal{W}$. We describe this new method in Section 5.

In Figure 3 we illustrate the attractor, and invariant measure, and a picture defined by tops + colour stealing, all for the IFS of projective transformations

$$\mathcal{W} = \{\square; w_n(x,y) = (\frac{a_n x + b_n y + c_n}{g_n x + h_n y + j_n}, \frac{d_n x + e_n y + f_n}{g_n x + h_n y + j_n}), n = 0,1,2,3\} \quad (8)$$

where the coefficients are given by

n	a_n	b_n	c_n	d_n	e_n	f_n	g_n	h_n	j_n
0	1.901	−0.072	0.186	0.015	1.69	0.028	0.563	−0.201	2.005
1	0.002	−0.044	0.075	0.003	−0.044	0.104	0.002	−0.088	0.154
2	0.965	−0.352	0.058	1.314	−0.065	−0.191	1.348	−0.307	0.075
3	−0.325	−0.0581	−0.029	−1.229	−0.001	0.199	−1.281	0.243	−0.058

The picture from which the colours were stolen is shown in Figure 2.

A close-up on the tops picture is illustrated in Figure 4.

4 The Tops Dynamical System

We show how the IFS leads to a natural dynamical system $T : G_\tau \to G_\tau$. The notation is the same as above. We also need the shift mapping

Fig. 3. From left to right, the attractor, an invariant measure, and a picture of the top made by colour stealing, for the IFS in Equation 8. Figure 4 shows a zoom on the picture of the top.

Fig. 4. Close-up on the fractal top in Figure 3. Details of the structure are revealed by colour-stealing.

$$S : \Omega \to \Omega$$

defined by

$$S\sigma_1\sigma_2\sigma_3... = \sigma_2\sigma_3...$$

for all $\sigma = \sigma_1\sigma_2\sigma_3.. \in \Omega$.

Lemma 1. *Let* $(x, \sigma) \in G_\tau$. *Then* $(w_{\sigma_1}^{-1}(x), S\sigma) \in G_\tau$.

Proof. See [5].

Lemma 2. *Let* $(x, \sigma) \in G_\tau$. *Then there is* $(y, \omega) \in G_\tau$ *such that* $(w_{\sigma_1}^{-1}(y), S\omega) = (x, \sigma) \in G_\tau$.

Proof. See [5].

It follows that the mapping

$$T : G_\tau \to G_\tau \text{ defined by } T(x, \sigma) = (w_{\sigma_1}^{-1}(x), S\sigma)$$

is well-defined, and onto. It can be treated as a dynamical system which we refer to as $\{G_\tau, T\}$. As such we may explore its invariant sets, invariant measures, other types of invariants such as entropies and information dimensions, and its ergodic properties, using "standard" terminology and machinery.

We can we project $\{G_\tau, T\}$ onto the Ω-direction, as follows: let

$$\Omega_\gamma := \{\sigma \in \Omega : (x, \sigma) \in G_\tau \text{ for some } x \in \mathbb{X}\}$$

Then Ω_γ is a shift invariant subspace of Ω, that is $S : \Omega_\gamma \to \Omega_\gamma$ with

$$S(\Omega_\gamma) = \Omega_\gamma,$$

and we see that $\{\Omega_\gamma, S\}$ is a **symbolic dynamical system**, see for example [10].

Indeed, $\{\Omega_\gamma, S\}$ is the symbolic dynamical system corresponding to a partition of the domain of yet a third dynamical system $\{A, \widetilde{T}\}$ corresponding to a mapping $\widetilde{T} : A \to A$ which is obtained by projecting $\{G_\tau, T\}$ onto A. This system is defined by

$$\widetilde{T}(x) = \begin{cases} w_{N-1}^{-1}(x) \ if & x \in D_{N-1} := w_{N-1}(A), \\ w_{N-2}^{-1}(x) \ if & x \in D_{N-2} := w_{N-2}(A) \setminus w_{N-1}(A) \\ \quad . \quad . & \quad . \\ w_0^{-1}(x) \ if & x \in D_0 := w_0(A) \setminus \bigcup_{n=1}^{N-1} w_n(A) \end{cases} \tag{9}$$

for all $x \in A$ and we have

$$\widetilde{T}(A) = A.$$

We call $\{A, \widetilde{T}\}$ the **tops dynamical system** (TDS) associated with the IFS.

Fig. 5. Illustrates the domains D_0, D_1, D_2, D_3 for the tops dynamical system associated with the IFS in Equation 8. This was the IFS used in Figures 3 and 4. Once this "picture" has been computed it is easy to compute the tops function. Just follow orbits of the tops dynamical system!

$\{\Omega_\gamma, S\}$ is the symbolic dynamical system obtained by starting from the tops dynamical system $\{A, T\}$ and partitioning A into the disjoint sets $D_0, D_1, ..., D_{N-1}$ defined in Equation 9, where

$$A = \bigcup_{n=0}^{N-1} D_n \text{ and } D_i \cap D_j = \varnothing \text{ for } i \neq j.$$

An example of such a partition is illustrated in Figure 5. We refer to $\{\Omega_\gamma, S\}$ as the **symbolic tops dynamical system** associated with the original dynamical system.

Theorem 1. *The tops dynamical system $\{A, \widetilde{T}\}$ and the symbolic dynamical system $\{\Omega_\gamma, S\}$ are conjugate. The identification between them is provided by the tops function $\tau : A \to \Omega_\gamma$. That is,*

$$\widetilde{T}(x) = \phi \circ S \circ \tau(x)$$

for all $x \in A$, and μ is an invariant measure for $\{A, \widetilde{T}\}$ iff $\tau \circ \mu$ is an invariant measure for $\{\Omega_\gamma, S\}$.

Proof. Follows directly from everything we have said above.

5 Symbolic Dynamics Algorithm for Computing Tops

Corollary 1. *The value of $\tau(x)$ may be computed by following the orbit of $x \in A$ as follows. Let $x_1 = x$ and $x_{n+1} = \widetilde{T}(x_n)$ for $n = 1, 2, ...$, so that the orbit of x is $\{x_n\}_{n=1}^{\infty}$. Then $\tau(x) = \sigma$ where $\sigma_n \in \{0, 1, ..., N-1\}$ is the unique index such that $x_n \in D_n$ for all $n = 1, 2, ...$*

Corollary 1 provides us with a delightful algorithm for computing approximations to the top of an IFS in cases in which we are particularly interested, namely when the IFS acts in $\mathbb{R}^2$ as in Equation 8.

Here we describe briefly one of many possible variants of the algorithm, with concentration on the key idea.

(i) Fix a resolution $L \times M$. Set up a rectangular array of little rectangular boxes, "pixels", corresponding to a rectangular region in $\mathbb{R}^2$ which contains the attractor of the IFS. Each box "contains" either a null value or a floating point number $x_{l,m} \in \mathbb{R}^2$and a finite string of indexes $\omega_{l,m} = \sigma^1_{l,m}\sigma^2_{l,m}...\sigma^{L_{m,n}}_{l,m}$ where $L_{l,m}$ denotes the length of the string and each $\sigma^k_{l,m} \in \{0, 1, ..., N-1\}$. The little boxes are initialized to null values.

(ii) Run the standard random iteration algorithm applied to the IFS with appropriate choices of probabilities, for sufficiently many steps to ensure that it has "settled" on the attractor, then record in all of those boxes, which taken together correspond to a discretized version of the attractor, a representative floating point value of $x \in A$ and the highest value, so far encountered, of the map index corresponding to that x-value. This requires that one runs the algorithm for sufficiently many steps that each box is visited many times and that, each time a map with a higher index value than the one recorded in a visited little box, the index value at the box is replaced by the higher index. [The result will be a discretized "picture" of the attractor, defined by the boxes with non null entries, partitioned into the domains $D_0, D_1, .., D_{N-1}$ with a high resolution value of $x \in A$ for each pixel. We will use these high resolution values to correct at each step the approximate orbits of $\widetilde{T} : A \to A$ which otherwise would rapidly loose precision and "leave" the attractor.]

(iii) Choose a little rectangular box, indexed by say l_1, m_1, which does not have a null value. (If the value of $\tau(x_{l_1,m_1})$, namely the string ω_{l_1,m_1}, is already recorded in the little rectangular box to sufficient precision, that is $L_{l_1,m_1} = L$, say go to another little rectangular box until one is found for which $L_{l_1,m_1} = 1$.) Keep track of $l_1, m_1, \sigma_{l_1,m_1}$. Compute $w_{\sigma_{l_1,m_1}}(x_{l_1,m_1})$ then discretize and identify the little box l_2, m_2 to which it belongs. If $L_{l_2,m_2} = L$ then set

$$\omega_{l_1,m_1} = \sigma^1_{l_1,m_1}\sigma^1_{l_2,m_2}\sigma^2_{l_2,m_2}...\sigma^{L-1}_{l_2,m_2}$$

and go to (iv). If $l_2, m_2 = l_1, m_1$ set

$$\omega_{l_1,m_1} = \sigma^1_{l_1,m_1}\sigma^1_{l_1,m_1}\sigma^1_{l_1,m_1}...\sigma^1_{l_1,m_1}$$

and go to (iv). Otherwise, keep track of $l_1, m_1, \sigma_{l_1,m_1}; l_2, m_2, \sigma_{l_2,m_2}$ and repeat the iterative step now starting at l_2, m_2 and computing $w_{\sigma_{l_2,m_2}}(x_{l_2,m_2})$.

Continue in this manner until *either* one lands in a box for which the string value is already of length L, in which case one back-tracks along the orbit one has been following, filling in all the string values up to length L, *or* until the sequence of visited boxes first includes the address of one box twice; i.e. a discretized periodic orbit is encountered. The strings of all of the points on the periodic orbit can now be filled-out to length L, and then the strings of all of the points leading to the periodic cycle can be deduced and entered into their boxes.

(iv) Select systematically a new little rectangular box and repeat step (iii), and continue until all the strings $\omega_{l,m}$ have length L. Our final approximation is

$$\tau(x_{l,m}) = \omega_{l,m}.$$

This algorithm includes "pixel-chaining" as described in [9] and is very efficient because only one point lands in each pixel during stages (iii) and (iv).

6 Analysis of Example 1

6.1 Invariant Measures and Random Iteration on Fractal Tops

We are interested in developing algorithms which are able to compute directly the graph G_τ of the tops function by some form of random iteration in which, at every step, the new points remain on G_τ. For while the just-described algorithm is delightful for computing approximations to the whole of G_τ it appears to be cumbersome and to have large memory requirements if very high resolution approximations to a part of G_τ are required, as when one "zooms in on" a fractal. But the standard chaos game algorithm has huge benefits in this regard and we would like to be able to repeat the process here. (The variant of the random iteration algorithm mentioned earlier, where one works on the whole of $\widehat{A}$ and keeps track of "highest values" is clearly inefficient in overlapping cases where the measure of the overlapping region is not negligible. Orbits of points may spend time wandering around deep within $\widehat{A}$ rarely visiting the top!)

But this is not the only motivation for studying stochastic processes and invariant measures on tops. Such processes have very interesting connections to information theory and data compression. In the end one would like to come back, full-circle, to obtain insights into image compression by understanding these processes. This topic in turn relates to IFS's with place-dependent probabilities and to studies concerning when such IFS's possess unique invariant measures.

Here we extend our discussion of Example 1 in some detail, to show the sort of thing we mean. We show that this example is related to a dynamical system studied by A. Renyi [11] connected to information theory. The IFS in

this example also shows up in the context of Bernoulli convolutions, where the absolute continuity or otherwise of its invariant measures is discussed. We present a theorem concerning this example which is, hopefully, new.

Much of what we say may be generalized extensively.

6.2 The TIFS and Markov Process for Example 1

We continue Example 1. The tops IFS (TIFS) associated with the IFS in Equation 6 is the "IFS" made from the following two functions, one of which has as its domain a set that is not all of $[0,1]$:

$$\widetilde{w}_0(x) = \alpha x \text{ for all } x \in [0, \frac{1-\alpha}{\alpha}); \tag{10}$$
$$\widetilde{w}_1(x) = \alpha x + (1-\alpha) \text{ for all } x \in [0,1].$$

We are interested in invariant measures for the following type of Markov process. This relates to a "chaos game" with place-dependent probabilities. Define a Markov transition probability by

$$\widetilde{P}(x,B) = \widetilde{p}_0(x)\chi_B(\widetilde{w}_0(x)) + \widetilde{p}_1(x)\chi_B(\widetilde{w}_1(x)). \tag{11}$$

Here

$$\widetilde{p}_0(x) = \begin{cases} p_0 \ for \ 0 \le x < \frac{1}{\alpha} - 1 \\ 0 \ for \ \frac{1}{\alpha} - 1 < x \le 1 \end{cases} \tag{12}$$

and

$$\widetilde{p}_1(x) = \begin{cases} p_1 \ for \ 0 \le x < \frac{1}{\alpha} - 1 \\ 1 \ for \ \frac{1}{\alpha} - 1 < x \le 1 \end{cases} \tag{13}$$

where $p_0, p_1 > 0$ and $p_0 + p_1 = 1$. $\widetilde{P}(x,B)$ is the probability of transfer from $x \in [0,1]$ into the Borel set B. Intuitively, pick a number $i \in \{0,1\}$ according to the distribution $\widetilde{p}_i(x)$ and transfer from x to $\widetilde{w}_i(x)$. What types of measures may be generated by such random orbits? When are such measures invariant for the Markov process? A Borel measure $\widetilde{\mu}$ on $[0,1]$ is said to be invariant for the Markov process iff

$$\widetilde{\mu}(B) = \int \widetilde{P}(x,B) d\widetilde{\mu}(x)$$

for all Borel sets $B \subset [0,1]$.

6.3 The TDS and Trapping Region for Example 1

The tops dynamical system (TDS) associated with this TIFS in Equation 6 is obtained by "inverting" the TIFS and is readily found to be defined as follows:

$$D_0 = [0, 1-\alpha), D_1 = [1-\alpha, 1]$$

and

$$\widetilde{T}(x) = \begin{cases} \frac{1}{\alpha}x & for\ x \in D_0 \\ \frac{1}{\alpha}x - (1 - \frac{1}{\alpha}) & for\ x \in D_1 \end{cases}. \tag{14}$$

Notice that orbits under $\widetilde{T}$ of points in $(\frac{1-\alpha}{\alpha}, 1)$ eventually arrive in the interval $R_{trapping} = [0, \frac{1-\alpha}{\alpha}]$ and that once there, they never leave again. We refer to $R_{trapping}$ as the trapping region. Notice too that $\widetilde{T}(1) = 1$ and that $x = 1$ is a repulsive fixed point of the dynamical system. That is, the tops dynamical system $\widetilde{T} : [0, 1] \to [0, 1]$ admits the invariant measure $\widetilde{\mu} = \delta_1(x)$. Note that this atomic measure is also invariant for the Markov process with transition probability $\widetilde{P}(x, B)$. But the following theorem tells us that there are no other possible atoms than one at $x = 1$ for invariant probability measures for the process.

6.4 Non-atomic invariant measures for Example 1

The proof of the following result typifies an argument which may be used in many cases to establish the existence of non-atomic invariant measures for TIFS.

Theorem 2. *Let $\widetilde{\mu}$ be an invariant probability measure for the Markov process described by Equation 11. Then $\widetilde{\mu}(\{a\}) = 0$ for all $a \in [0, 1)$.*

Proof. See [5].

6.5 The TIFS and TDS for Example 1 in the trapping region

From this point forward in this section we concentrate on the behaviour of the TDS and the TIFS in the trapping region. We study invariant probability measures and ergodic properties for the TIFS and TDS restricted to the trapping region. In view of Theorem 2, we restrict attention to invariant probability measures which do not contain atoms. This will allow us to modify the "trapped" TDS/TIFS on any countable sets of points without altering potential invariant probability measures off a set of measure zero.

We make the change of variable $x' = \frac{1-\alpha}{\alpha}x = g(x)$ to rescale the behaviour of the TDS restricted to $R_{trapping}$ to produce an equivalent dynamical system acting on $[0, 1]$: that is $\widetilde{\widetilde{T}} : [0, 1] \to [0, 1]$ is defined by $\widetilde{\widetilde{T}} = g \circ \widetilde{T} \circ g^{-1}$. The result is

$$\widetilde{\widetilde{D}}_0 = [0, \alpha), \widetilde{\widetilde{D}}_1 = [\alpha, 1]$$

and

$$\widetilde{\widetilde{T}}(x) = \begin{cases} \frac{1}{\alpha}x & for\ x \in \widetilde{\widetilde{D}}_0 \\ \frac{1}{\alpha}x - 1 & for\ x \in \widetilde{\widetilde{D}}_1 \end{cases}. \tag{15}$$

Notice that

$$\widetilde{\widetilde{T}}(x) = (\beta x) \tag{16}$$

where (y) denotes the fractional part of the real number y and

$$\beta = \frac{1}{\alpha}.$$

By using the symbol β we connect our problem to standard notation for a much studied dynamical system, Equation 16, see for example [11], [10]. The focus of all of the work, with which we are familiar, connected to the dynamical system in Equation 15, concerns invariant measures which are absolutely continuous with respect to Lebesgue measure and which maximize topological entropy. This is not our focus. We are interested in the existence and computation of invariant measures for the associated IFS with place-dependent, typically piecewise constant, probabilities.

Next we invert the trapped TDS to produce a corresponding restricted TIFS. This is defined with the aid of the two functions

$$\begin{aligned} \widetilde{\widetilde{w}}_0(x) &= \alpha x \text{ for all } x \in [0,1); \\ \widetilde{\widetilde{w}}_1(x) &= \alpha x + \alpha \text{ for all } x \in [0, \frac{1-\alpha}{\alpha}]. \end{aligned} \tag{17}$$

The Markov process in the trapped region corresponds to the transition probability

$$\widetilde{\widetilde{P}}(x,B) = \widetilde{\widetilde{p}}_0(x)\chi_B(\widetilde{\widetilde{w}}_0(x)) + \widetilde{\widetilde{p}}_1(x)\chi_B(\widetilde{\widetilde{w}}_1(x)). \tag{18}$$

with the place-dependent probabilities

$$\widetilde{\widetilde{p}}_0(x) = \begin{cases} p_0 \ for \ 0 \le x \le \frac{1}{\alpha} - 1 \\ 1 \ for \ \frac{1}{\alpha} - 1 < x \le 1 \end{cases} \tag{19}$$

and

$$\widetilde{\widetilde{p}}_1(x) = \begin{cases} p_1 \ for \ 0 \le x \le \frac{1}{\alpha} - 1 \\ 0 \ for \ \frac{1}{\alpha} - 1 < x \le 1 \end{cases} \tag{20}$$

where $p_0, p_1 > 0$ and $p_0 + p_1 = 1$. The reader may find it interesting to compare the trapped TIFS in Equations 17, 19 and 20 (*system II*) to the original TIFS in Equations 10, 12 and 13 (*system I*). The two systems look very similar. But the trapped system admits no invariant probability measure which contains an atom.

Notice that if $\widetilde{\widetilde{\mu}}$ is an invariant probability measure for the trapped system (*system II*) then

$$\widetilde{\mu} = g \circ \widetilde{\widetilde{\mu}}.$$

is an invariant measure for the original system (*system I).*

6.6 Existence and Unicity of Invariant Measures for the trapped IFS

The following theorem asserts that there are two basically different situations which can occur.

Theorem 3. *Let* $\frac{1}{2} < \alpha < 1$. *Let* $\widetilde{\widetilde{w}}_0 : [0,1] \to [0,1]$ *be defined by* $\widetilde{\widetilde{w}}_0(x) = \alpha x$. *Let* $\widetilde{\widetilde{w}}_1 : [0, \frac{1}{\alpha} - 1] \to [0,1]$ *be defined by* $\widetilde{\widetilde{w}}_1(x) = \alpha x + \alpha$. *Let a Markov transition probability be given by Equation 18, with probabilities given by Equations 19 and 20. Then the Markov process possesses at least one and at most two linearly independent invariant probability measures. In particular, there exists at most two invariant probability measures which are also ergodic invariant measures for the tops dynamical system* $\widetilde{\widetilde{T}} : [0,1] \to [0,1]$ *defined by*

$$\widetilde{\widetilde{T}}(x) = (\frac{1}{\alpha}x) \text{ for all } x \in [0,1],$$

where (y) *denotes the fractional part of the real number* y. *If* α *possesses certain arithmetic properties, namely that* $\beta = \frac{1}{\alpha}$ *is a "*β*-number", then the Markov process possesses a unique invariant probability measure which is ergodic for* $\widetilde{\widetilde{T}}$.

Proof. See [5]. This relies on analysis of a corresponding code space system which we describe next.

6.7 Corresponding Code Space Systems

We can convert the system in Theorem 3 to an equivalent symbolic dynamical system with the aid of the maps

$$\widehat{w}_0(x,\sigma) = (\alpha x, 0\sigma) \text{ for all } (x,\sigma) \in [0,1) \times \Omega \tag{21}$$
$$\widehat{w}_1(x,\sigma) = (\alpha x + \alpha, 1\sigma) \text{ for all } (x,\sigma) \in [0, \frac{1-\alpha}{\alpha}] \times \Omega$$

We call the system described by Equation 21 "the lifted trapped TIFS".This system possesses a maximal invariant set $G_{\widetilde{\tau}} \subset [0,1) \times \Omega$ which obeys

$$G_{\widetilde{\tau}} = \widehat{w}_0(G_{\widetilde{\tau}}) \cup \widehat{w}_1(G_{\widetilde{\tau}})$$

$G_{\widetilde{\tau}}$ is the graph of a one-to-one function $\widetilde{\tau} : [0,1] \to \Omega$. The projection of this system onto the $[0,1]$-direction gives us back the original system, the one in Theorem 3.

Projection in the Ω-direction provides us with a symbolic dynamical system $S : \Omega_\gamma \to \Omega_\gamma$ where $\widetilde{\tau}([0,1]) = \Omega_\gamma$, and $S(\Omega_\gamma) = \Omega_\gamma$. $\{\Omega, S\}$ is the usual shift dynamical system on code space and the symbolic dynamical system $\{\Omega_\gamma, S\}$ is obtained by restricting the domain of S to Ω_γ. We could choose to write $S : \Omega_\gamma \to \Omega_\gamma$ as $S|_{\Omega_\gamma} : \Omega_\gamma \to \Omega_\gamma$ and $\{\Omega_\gamma, S|_{\Omega_\gamma}\} = \{\Omega_\gamma, S\}$, but do not. Note that we can compute $\widetilde{\tau}(x)$ by following the orbit of $x \in [0,1]$ under the trapped TDS (Equation 15).

Thus we obtain the *symbolic IFS with place-dependent probabilities*

$$\{\Omega_\gamma; s_0|_{\Omega_\gamma}, s_1|_{\Omega_\gamma}; p_0(\sigma), p_1(\sigma)\}$$

where

$$s_0|_{\Omega_\gamma}(\sigma) = 0\sigma \text{ for all } \sigma \in \Omega_\gamma \tag{22}$$
$$s_1|_{\Omega_\gamma}(\sigma) = 1\sigma \text{ for all } \sigma \in \Omega_\gamma,$$

and the probabilities are given by $p_i(\sigma) = \widetilde{\widetilde{p}}_i(\widetilde{\tau}^{-1}(\sigma))$, that is

$$p_0(\sigma) = \begin{cases} p_0 \text{ for all } \sigma \in \Omega_\gamma \text{ with } 0 \le \sigma \le \gamma \\ 1 \text{ for all } \sigma \in \Omega_\gamma \text{ with } \gamma < \sigma \le 1 \end{cases} \tag{23}$$

and

$$p_1(\sigma) = \begin{cases} p_1 \text{ for all } \sigma \in \Omega_\gamma \text{ with } 0 \le \sigma \le \gamma \\ 0 \text{ for all } \sigma \in \Omega_\gamma \text{ with } \gamma < \sigma \le 1 \end{cases} \tag{24}$$

where $\gamma = \widetilde{\tau}(\frac{1-\alpha}{\alpha})$. This system is equivalent to the one in Theorem 3. It is the restriction to Ω_γ of the IFS $\{\Omega; s_0, s_1; p_0(\sigma), p_1(\sigma)\}$ where

$$s_0(\sigma) = 0\sigma \text{ for all } \sigma \in \Omega \tag{25}$$
$$s_1(\sigma) = 1\sigma \text{ for all } \sigma \in \Omega$$

and the probabilities $p_i : \Omega \to [0,1]$ are defined in the same way as the $p_i : \Omega_\gamma \to [0,1]$ in Equations 23 and 24 with Ω_γ replaced by Ω.

The system described by Equation 25 is an extension of the system described by Equation 22. Notice the following relationship between these two systems: *Ω_γ is the unique set attractor of the IFS* $\{\Omega; s_0, s_1|_{[0,\gamma]}\}$; that is Ω_γ is the unique nonempty compact subset of Ω such that

$$\Omega_\gamma = s_0(\Omega) \cup s_1|_{[0,\gamma]}(\Omega) = s_0(\Omega) \cup s_1(\Omega \cap [0,\gamma]).$$

By ignoring a countable set of points we can embed the code space in Ω in $[0,1]$ to make pictures of, and simplify the visualization of, the *symbolic IFS's* in Equations 22 and 25. For example the symbolic system in Equation 22 may be represented by the following pretty IFS with place-dependent probabilities

$$\{[0,1]; s_0 : [0,1] \to [0,1], s_1 : [0,\gamma] \to [0,1]; p_0(x), p_1(x)\}$$

where

$$s_0(x) = \frac{1}{2}x \text{ for all } x \in [0,1] \tag{26}$$
$$s_1(x) = \frac{1}{2}x + \frac{1}{2} \text{ for all } x \in [0,1]$$

where

$$p_0(x) = \begin{cases} p_0 \text{ for all } 0 \le x \le \gamma \\ 1 \text{ for all } \gamma < x \le 1 \end{cases}$$

and

$$p_1(x) = \begin{cases} p_1 \text{ for all } 0 \leq x \leq \gamma \\ 0 \text{ for all } \gamma < x \leq 1 \end{cases}$$

Note that we use the symbols $s_0(\cdot)$ and $s_1(\cdot)$ to denote the maps and $p_0(\cdot)$ to denote the probabilities for embedded system.

Our Markov process, represented on Ω_γ, involves applying the maps $s_0|_{\Omega_\gamma}, s_1|_{\Omega_\gamma}$ with probabilities p_0, p_1 respectively when $\sigma \in \Omega_\gamma$ with $\sigma \leq \gamma$, and applying the map $s_0|_{\Omega_\gamma}$ with probability one when $\sigma \in \Omega_\gamma$ with $\sigma > \gamma$. This process equivalent to the Markov process on $[0, 1]$ corresponding to the transition probability $\widetilde{\widetilde{P}}(x, B)$ in Equation 18 in Theorem 3.

But we can also consider the corresponding process on Ω which involves applying the maps s_0, s_1 with probabilities p_0, p_1 respectively when $\sigma \in \Omega$ with $\sigma \leq \gamma$, and applying the map s_0 with probability one when $\sigma \in \Omega$ with $\sigma > \gamma$. This process extends, to all of Ω, the domain of the original symbolic process. We will find it very useful to consider this latter process. This is because, when we change α, the maps and the space upon which they act remain unaltered; only the set of values of $\sigma \in \Omega$ for which $p_1(\sigma) = 0$ changes. In the original system, in Theorem 3, the slopes of the maps, the location where a probability becomes zero, and the set of allowed codes, all change with α.

We find that the structure of the system in Theorem 3 depends fundamentally on whether or not γ, the address of the point $\frac{1-\alpha}{\alpha}$, which may be computed by following the orbit of $\frac{1-\alpha}{\alpha}$ under the trapped tops dynamical system $\widetilde{\widetilde{T}} : [0, 1] \to [0, 1]$, terminates in an endless string of zeros or not. That is, on whether or not

$$\frac{r}{2^{k-1}} = \frac{1-\alpha}{\alpha} = \beta - 1$$

for some pair of positive integers k and r.

We work with the closure of the symbolic system. We are interested in those invariant measures which correspond to orbits of the following type of random iteration. When the current point lies in the region $0000... \leq \sigma \leq \gamma$ there is a nonzero probability that s_0 may be applied to the current point and there is a nonzero probability that s_1 may be applied to the current point. When the current point lies in the region $\gamma < \sigma < 1$ the map s_0 is applied with probability one. Theorem 2 implies that this Markov process possesses no invariant measures which include point masses.

Notice that if γ terminates in $\overline{0}$ or $\overline{1}$ then $S : \Omega_\gamma \to \Omega_\gamma$ is open, that is it maps open sets to open sets. This implies that the inverse branches of S are continuous in the product topology.

Another approach to the proof of Theorem 3 in the case $\gamma = x....xx1\overline{0}$ uses the formulation, and a theorem, of W. Parry [10] which involves intrinsic Markov chains. Our framework is broader since we work on all of Ω, which has a further implication: in the "recurrent" case, we obtain a stronger result than I. Werner, [12], regarding convergence of orbits produced by random iteration in the case of certain graph-directed IFS. This also has implications which are

entirely new, so far as we know, for the "not open" case, which corresponds for example to the case where γ is irrational.

References

1. M. F. Barnsley, *Fractals Everywhere*, Academic Press, New York, NY, 1988.
2. M. F. Barnsley and S. Demko, *Iterated Function Systems and the Global Construction of Fractals,* R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. **399** (1985), pp. 243-275.
3. M. F. Barnsley and L. F. Barnsley, *Fractal Transformations*, in "The Colours of Infinity: The Beauty and Power of Fractals", by Ian Stewart et al., Published by Clear Books, London, (2004), pp.66-81.
4. M. F. Barnsley, *Ergodic Theory, Fractal Tops and Colour Stealing,* Lecture Notes*,* Australian National University, *2004.*
5. M. F. Barnsley, *Theory and Application of Fractal Tops (Full Version),* Preprint, Australian National University, 2005.
6. M.F.Barnsley, J.E.Hutchinson, O. Stenflo *A Fractal Valued Random Iteration Algorithm and Fractal Hierarchy* (2003) to appear in Fractals journal.
7. J. Elton, *An Ergodic Theorem for Iterated Maps,* Ergodic Theory Dynam. Systems, 7 (1987), pp. 481-488.
8. J. E. Hutchinson, *Fractals and Self-Similarity*, Indiana. Univ. Math. J., 30 (1981), pp. 713-749.
9. N. Lu, *Fractal Imaging*, Academic Press, (1997).
10. W. Parry, *Symbolic Dynamics and Transformations of the Unit Interval*, Trans. Amer. Math. Soc., **122** (1966),368-378.
11. A.Rényi, *Representations for Real Numbers and Their Ergodic Properties*, Acta Math. Acad. Sci. Hung., **8** (1957), pp. 477-493.
12. I. Werner, *Ergodic Theorem for Contractive Markov Systems,* Nonlinearity **17** (2004) 2303-2313

Splines, Fractal Functions, and Besov and Triebel-Lizorkin Spaces

Peter Massopust

Engineering and Research Development, Tuboscope Pipeline Services, Houston, USA
pmassopust@varco.com

Summary. In this paper, some of the relationships between splines and fractal functions are considered. In particular, it is shown that fractal analogs of B-splines can be defined. Moreover, conditions on fractal functions are derived that guarantee their inclusion in approximation spaces, such as Besov and Triebel-Lizorkin spaces. These conditions generalize earlier results obtained in [8].

1 Brief Summary of Spline and Fractal Function Theory

This section provides a brief summary of the theory of splines and reacquaints the reader with the construction of fractal functions. For more detailed presentations of either subject, the reader is referred to [1, 2, 5, 7].

1.1 Splines and B-Splines

A *spline function* or, for short, a *spline* is a piecewise polynomial function joined together on subintervals with certain continuity or smoothness conditions. More precisely, let $X = \{a = x_0 < x_1 < \cdots < x_n < x_{n+1} = b\}$ be sequence of real numbers and $k \in \mathbb{N}$. A *spline of degree* $k-1$ or *order* k on $[a, b]$ is a function $s : [a, b] \to \mathbb{R}$ such that

1. $\forall i = 1, \ldots, n+1 : s\big|_{[x_{i-1}, x_i)} \in \mathbb{P}^{k-1}$, the linear space of all real polynomials of degree $< k$;
2. $s \in C^{k-2}[a, b]$.

It can be shown [2] that the set $S_{X,k}$ of all spline functions s of order k forms a real vector space of dimension $n + k$.

B-splines are special local basis functions for $S_{X,k}$. They provide a computationally efficient and numerically stable framework for the evaluation of and approximation by splines.

Let $t_1 \leq t_2 \leq \cdots \leq t_{n+k}$ be a nondecreasing sequence of real numbers. The sequence $\boldsymbol{t} = (t_1, t_2, \ldots, t_{n+k})$ is termed a *knot vector*. Some of the *knots*

t_i may be repeated, i.e., it may happen that $t_i = t_{i+1} = \dots t_{i+\nu_i}$ for some $\nu_i \in \mathbb{N}$.

Now suppose that $k, n \in \mathbb{N}$. The *B-spline $B_{i,k,\boldsymbol{t}}$ of order k (degree $k-1$)* is recursively defined as follows.

$$B_{i,1,\boldsymbol{t}} := \chi_{[t_i, t_{i+1})} \tag{1}$$

(Here $\chi : \mathbb{R} \to \{0, 1\}$ denotes the characteristic function of a subset of $\mathbb{R}$.) and

$$B_{i,k,\boldsymbol{t}}(x) := \frac{x - t_i}{t_{i+k-1} - t_i} B_{i,k-1,\boldsymbol{t}}(x) + \frac{t_{i+k} - x}{t_{i+k} - t_{i+1}} B_{i+1,k-1,\boldsymbol{t}}(x) \tag{2}$$

for $i = 1, \dots, n$ and $k \geq 2$. Should it happen that $t_{i+k-1} = t_i$ or $t_{i+k} = t_{i+1}$ for one of the coefficients in the above recursive definition of $B_{i,k,\boldsymbol{t}}$, then the entire coefficient is set equal to zero. Figure 1 below shows examples of a constant B_1, linear B_2, and cubic B-spline B_4.

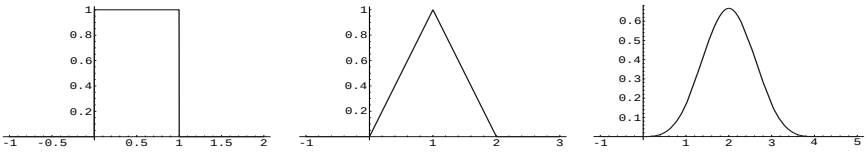


Fig. 1. A constant, linear, and cubic B-spline

Some of the main properties of B-splines are:

- $B_{i,k,\boldsymbol{t}}(x) > 0$ for $x \in (t_i, t_{i+k})$.
- $B_{i,k,\boldsymbol{t}}\big|_{[t_i, t_{i+1})} \in \mathbb{P}^{k-1}$ and a piecewise polynomial on each interval $[t_i, t_{i+k})$.
- $B_{i,k,\boldsymbol{t}}(x) = 0$ for $x \in (-\infty, t_1) \cup (t_{n+k}, \infty)$.
- If ν_i is the multiplicity of knot t_i, i.e., $t_{i-1} \neq t_i = t_{i+1} = \dots t_{i+\nu_i - 1} \neq t_{i+\nu_i}$, then $B_{i,k,\boldsymbol{t}} \in C^{k-1-\nu_i}$ in a neighborhood of t_i. Loosely speaking,

  degree = smoothness at knot + multiplicity of knot

- $\sum_{i=1}^{n} B_{i,k,\boldsymbol{t}}(x) = 1$, $\forall x \in [t_k, t_{n+1}]$.

Now reconsider the spline space $S_{X,k}$ consisting of splines of order k. The elements of $S_{X,k}$ have prescribed degrees of smoothness ν_i at each of the knots x_i, $i = 1, \dots, n$, whereas the smoothness of B-splines is controlled by the multiplicity of knots coinciding with x_i. This observation allows that for any space $S_{X,k}$ one can find a knot vector $\boldsymbol{t} = (t_j)_{1 \leq j \leq m+k}$ whose associated B-splines form a basis of $S_{X,k}$.

To this end, let $\boldsymbol{t} = (t_i)_{1 \leq i \leq m+k}$ be defined by

$$t_1 = \cdots = t_k := a = x_0; \quad t_{k+\sum_{\ell=1}^{j}\nu_\ell+i} := x_j,\ i = 0, 1, \ldots, \nu_j - 1;\ j = 1, \ldots, n; \tag{3}$$

$$t_{k+\sum_{j=1}^{n}\nu_j+i} := b = x_{n+1},\ i = 1, \ldots, k;\ m = k + \textstyle\sum_{j=1}^{n}\nu_j.$$

Then the functions $B_{j,k,\boldsymbol{t}}$, $j = 1, \ldots, m$, are linearly independent and elements of $S_{X,k}$. Thus, they form a basis of $S_{X,k}$.

Theorem 1 (Representation Theorem for Splines). *Every spline $s \in S_{X,k}$ defined on the interval $[a, b]$ has a unique expansion in terms of B-splines of order k:*

$$s(x) = \sum_{j=1}^{m} c_j\, B_{j,k,\boldsymbol{t}}(x). \tag{4}$$

Equivalently, if $\mathcal{S}_{k,\boldsymbol{t}} := \operatorname{span}\{B_{j,k,\boldsymbol{t}}\}$ then $S_{X,k} = \mathcal{S}_{k,\boldsymbol{t}}$.

Definition 1. *A B-spline whose knots are equally spaced, i.e., $\forall j : t_j - t_{j-1} =$* constant *is termed a* cardinal spline.

1.2 Fractal Functions

There are several ways to construct fractal functions [1, 5], but one of the most general constructions is via fixed points of so-called *Read-Bajraktarević operators* $\mathscr{T}$ [7].

Theorem 2. *Let Ω be a compact subset of $\mathbb{R}^n$ and $1 < N \in \mathbb{N}$. Suppose that $u_i : \Omega \to \Omega$ are contractive homeomorphisms and $v_i : \Omega \times \mathbb{R} \to \Omega$ is uniformly Lipschitz in the second argument with Lipschitz constant λ_i, $i = 1, \ldots, N$. Let*

$$\mathscr{T}(f) := \sum_{i=1}^{N} \left[v_i(u_i^{-1}(\cdot), f \circ u_i^{-1}(\cdot))\right] \chi_{u_i(\Omega)} \tag{5}$$

Then, if $\max\{|\lambda_i|\} < 1$, the operator $\mathscr{T}$ is contractive on $L^\infty(\Omega)$ and its unique fixed point $\mathfrak{F} : \Omega \to \mathbb{R}$ satisfies

$$\mathfrak{F} = \sum_{i=1}^{N} \left[v_i(u_i^{-1}(\cdots), \mathfrak{F} \circ u_i^{-1}(\cdot))\right] \chi_{u_i(\Omega)} \tag{6}$$

Definition 2. *The unique fixed point $\mathfrak{F}$ of the operator $\mathscr{T}$ defined in 5 is called an* ($\mathbb{R}$-valued) fractal function.

In this paper, a subclass of fractal functions is considered. More precisely, for all $i = 1, \ldots, N$, let the contractive mappings u_i and the Lipschitz mappings v_i be given by

$$u_i \in \operatorname{Aff}(n) = \operatorname{GL}(n, \mathbb{R}) \ltimes \mathbb{R}^n \quad \text{and} \quad v_i(\cdot, *) := p_i^M(\cdot) + \lambda_i(*), \tag{7}$$

where Aff (n) denotes the affine group on $\mathbb{R}^n$, which is the semi-direct product of the general linear group of nonsingular matrices in $\mathbb{R}^{n\times n}$ and the additive group of translations in $\mathbb{R}^n$ ($\cong$ to $\mathbb{R}^n$). The functions $p_i^M \in \mathbb{P}^M$ are real polynomials of degree at most $M \in \mathbb{N}$ in Ω and the scalars $|\lambda_i| < 1$, but are otherwise free parameters.

To simplify notation, extend $\mathfrak{F}$ to all of $\mathbb{R}$ by setting it identically equal to zero off Ω. Analogously, set

$$P_M := \begin{cases} \sum_{i=1}^{N} p_i^M \circ u_i^{-1}\, \chi_{u_i(\Omega)}, & \text{on } \Omega \\ 0, & \text{on } \mathbb{R}\setminus\Omega. \end{cases}$$

The Read-Baractarević operator now has the form

$$f \mapsto P_M + \sum_{i=1}^{N} \lambda_i\, f \circ u_i^{-1} \tag{8}$$

and the fixed point equation for $\mathfrak{F}$ becomes

$$\mathfrak{F} = P_M + \sum_{i=1}^{N} \lambda_i\, \mathfrak{F} \circ u_i^{-1}. \tag{9}$$

The graphs of fractal functions may be fractal sets with non-integral Hausdorff-Besicovitch dimension, but there also exist fixed points $\mathfrak{F}$ of $\mathscr{T}$ that are of class C^m, $m \in \mathbb{N}$ [6,8].

A fractal function $\mathfrak{F}$ generated in the above fashion, clearly depends on the set of polynomials $\boldsymbol{p}_M := (p_1^M, \ldots, p_N^M) \in \prod_{i=1}^{N} \mathbb{P}^M$ and the set of parameters $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_N) \in (-1,1)^N$. Indeed, the mapping

$$\prod_{i=1}^{N} \mathbb{P}^M \ni \boldsymbol{p}_M \mapsto \mathfrak{F}_{\boldsymbol{p}_M} \in L^\infty \tag{10}$$

is a linear isomorphism [4,7,8].

2 Fractal Analogs of B-Splines

B-splines provide a family of approximation function of increasing smoothness (by increasing the support). This becomes quite transparent when cardinal B-splines are considered. In this case, let the knot vector be the sequence of integers whose terms have knot multiplicity one. Then the B-splines $B_{i,k,\mathbb{Z}}$ are translates of $B_{0,k,\mathbb{Z}}$: $B_{i,k,\mathbb{Z}}(x) = B_{0,k,\mathbb{Z}}(x-i)$, and we may drop the subscripts 0 and $\mathbb{Z}$. *The* kth-order B-spline B_k is then supported on $[0,k]$ and is an element of C^{k-1}. Moreover, the kth-order B-spline can be computed as the k-fold, $1 < k \in \mathbb{N}$, convolution of the characteristic function $\chi_{[0,1)}$:

$$B_k = *^k \chi_{[0,1)} = B_{k-1} * \chi_{[0,1)} = \int_0^1 B_{k-1}(\cdot - t)\, dt. \tag{11}$$

Now fix integers $k, N > 1$ and let $\Omega := [0, k]$. For $i \in \{1, \ldots, k\}$, define contractions $u_i : \Omega \to \Omega$ by $u_i = k^{-1}(\cdot) + i - 1$. Furthermore, determine polynomials $p_i^k \in \mathbb{P}^k$ such that

$$\begin{aligned} &p_1^k(0) = 0 \quad \text{and} \quad \forall i = 1, \ldots, k : \ (i, B_k(i)) \in \operatorname{graph} p_i^k \\ &\forall i = 1, \ldots, k-1 \, \forall \varkappa = 0, \ldots, k-2 : \ D^\varkappa(p_i^k)(i+) = D^\varkappa(p_{i+1}^k)(i-) = B_k(i). \end{aligned} \tag{12}$$

Here D denotes the differentiation operator. Using these two sets of mappings, one defines a Read-Bajraktarević operator

$$\mathscr{T}(f) := \sum_{i=1}^{k} \left[p_i^k(k(\cdot - i + 1)) + \lambda_i \, f(k(\cdot - i + 1)) \right] \chi_{[i-1,i)} \tag{13}$$

for a set of parameters λ_i with $\max |\lambda_i| < 1$. By the uniqueness of the interpolation problem, the first term in the above sum equals B_k and thus the fixed point $\mathfrak{B}_k$ of $\mathscr{T}$ can be written as

$$\mathfrak{B}_k(x) = B_k(x) + \sum_{i=1}^{k} \lambda_i \, \mathfrak{B}_k(m(x - i + 1)) \chi_{[i-1,i)}(x) \tag{14}$$

It is important to realize that $\mathfrak{B}_k$ depends not only on k (and thus the polynomials p_i^k) but also the set $\{\lambda_i\}$. The size of the modulus of these parameters determines to which function space $\mathfrak{B}_k$ belongs and whether $\operatorname{graph} \mathfrak{B}_k$ has nonintegral Hausdorff-Besicovitch dimension. For simplicity, this latter dependence will be suppressed.

Recall that any fractal function $\mathfrak{F}$, as defined in Section 1.2, can be iteratively generated employing the following process. Let $f_0 \equiv 0$ be the zero function in some function space $\mathcal{F}$. Define $f_m := \mathscr{T} f_{m-1}$, $m \in \mathbb{N}$. Then f_m converges to $\mathfrak{F}$ in the topology of $\mathcal{F}$; in most cases this is some locally convex or norm topology. Notice that choosing $f_0 \equiv 0$, gives $f_1 = P_M$. In the above situation, the fixed point $\mathfrak{B}_k$ is built upon the B-spline B_k in an iterative fashion by adding vertically scaled copies of B_k to $B_k \circ u_i$. If $\max |\lambda_i| = 0$, then obviously $\mathfrak{B}_k = B_k$. One may think of the fractal function $\mathfrak{B}_k$ as a family of parameterized B-splines (of increasing complexity as $\max |\lambda_i| \uparrow 1-$). Varying the moduli of the parameters λ_i creates, for instance, symmetric and differentiable fractal functions. Approximation-theoretic properties of these fractal analogs of B-splines will be investigated elsewhere.

By Theorem 1, every spline of order k is a linear combination of B_k and it is natural to ask whether a fractal function $\mathfrak{F}_{\boldsymbol{p}}$ defined on a compact domain $\Omega \subset \mathbb{R}$ generated by polynomials p_i^k whose restriction to Ω is a spline s of order k is a linear combination of fractal functions related to those generated by B-splines B_k.

To this end, suppose that $\mathfrak{F}_{\boldsymbol{p}}$ is generated by polynomials p_i^k whose restriction to $\Omega := [0, N]$, $1 < N \in \mathbb{N}$, constitutes a spline s of order k. By the representation theorem for splines 1, there exists a finite sequence $c_1, \ldots, c_m$ of real numbers and an associated finite sequence $B_{1,k}, \ldots, B_{m,k}$ of B-splines of order k with knots on the integers such that $s = \sum_{j=1}^{m} c_j B_{j,k}$. Let $\boldsymbol{p}_k := (p_1^k, \ldots, p_N^k)$, then by the properties of B-splines

$$s\big|_{u_i(\Omega)} = p_i^k = \sum_{j=i}^{i+k-1} c_j \, B_{j,k}\big|_{u_i(\Omega)}. \tag{15}$$

Hence,

$$\begin{aligned} \boldsymbol{p}_k &= (p_1^k, \ldots, p_N^k) \\ &= \left(\sum_{j=1}^{k} c_j B_{j,k}\big|_{u_1(\Omega)}, \ldots, \sum_{j=1}^{k} c_{N-k+j} B_{N-k+j}\big|_{u_N(\Omega)} \right) \end{aligned} \tag{16}$$

Employing the linear isomorphism 10 and invoking the uniqueness of the fixed point of a Read-Bajraktarević operator, gives

$$\begin{aligned} \mathfrak{F}_{\boldsymbol{p}} &= \mathfrak{F}_{\sum_{j=1}^{k}(c_j B_{j,k}|_{u_1(\Omega)}, \ldots, \sum_{j=1}^{k} c_{N-k+j} B_{N-k+j}|_{u_N(\Omega)})} \\ &= \sum_{j=1}^{k} \mathfrak{F}_{(c_j B_{j,k}|_{u_1(\Omega)}, \ldots, \sum_{j=1}^{k} c_{N-k+j} B_{N-k+j}|_{u_N(\Omega)})}, \end{aligned} \tag{17}$$

which is the analog of the representation of splines in terms of B-splines.

Figure 2 below depicts two fractal analogs of B-splines. The function on the right is differentiable on its domain and resembles the usual third order B-spline.

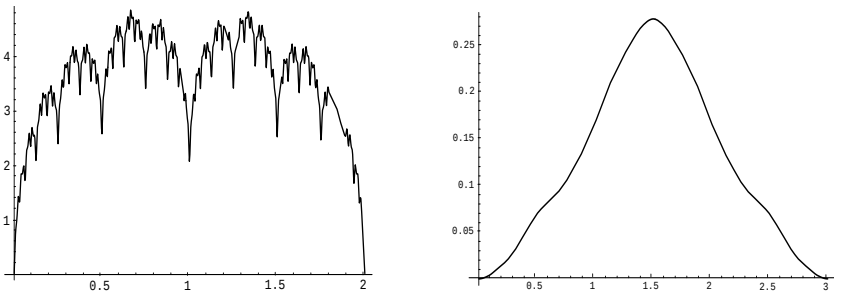


Fig. 2. Fractal analogs of B-splines: $\mathfrak{B}_2$ (left) and $\mathfrak{B}_3$ (right)

3 Besov and Triebel-Lizorkin Spaces

The theory of Besov and Triebel-Lizorkin spaces is very rich and has numerous applications to partial differential equations and approximation theory, including finite elements, splines and wavelets. Originally, these spaces were developed to close the gaps in the ladders of smoothness spaces such as the Hölder spaces C^s, $s \in \mathbb{R}_0^+$, and the classical Sobolev spaces $W^m(L^n)$, $m, n \in \mathbb{Z}_0^+$. This section provides a very rudimentary introduction to these spaces and the interested reader is referred to [9] and the references therein.

Let X be a real or complex vector space. A mapping $\|\cdot\| : X \to \mathbb{R}_0^+$ is called a *quasi-norm* if it satisfies all the usual conditions of a norm except for the triangle inequality, which is replaced by

$$\|x + y\| \le c\,(\|x\| + \|y\|) \tag{18}$$

for a constant $c \ge 1$. If $c = 1$, then $\|\cdot\|$ is a norm. A complete quasi-normed space is called a *quasi-Banach space.*

Recall that the M-th order difference operator Δ_h^M of step size $h \in \mathbb{R}^n$ is defined by

$$(\Delta_h^M f)(x) := \sum_{m=0}^{M} (-1)^{M-m} \binom{M}{m} f(x + m\,h). \tag{19}$$

Definition 3. *Let $0 < p, q \le \infty$, let $\sigma_p := \left(\dfrac{1}{\min\{p,1\}} - 1\right)$, and suppose that $s > \sigma_p$. Suppose $M \in \mathbb{N}$ is such that $M > s \ge M - 1$. Then a function $f \in L^p$ belongs to the Besov space $B_q^s(L^p)$ iff*

$$\|f\|_{B_q^s(L^p)} := \begin{cases} \|f\|_{L^p} + \left(\displaystyle\int_{\mathbb{R}^n} |h|^{-sq} \|\Delta_h^M f\|_{L^p}^q \frac{dh}{|h|^n}\right)^{\frac{1}{q}} < \infty, & 0 < q < \infty \\ \|f\|_{L^p} + \displaystyle\sup_{0 \ne h \in \mathbb{R}^n} |h|^{-s} \|\Delta_h^M f\|_{L^p} < \infty, & q = \infty. \end{cases} \tag{20}$$

Note that $B_q^s(L^p)$ is a Banach space for $1 \le p, q \le \infty$; otherwise $B_q^s(L^p)$ is a quasi-Banach space.

Definition 4. *Let $0 < p, q \le \infty$ and suppose $s > \dfrac{n}{\min\{p,q\}}$. If $M \in \mathbb{N}$ is such that $M > s \ge M - 1$, then a function $f \in L^p$ is said to belong to the Triebel-Lizorkin space $F_q^s(L^p)$ iff*

$$\|f\|_{F_q^s(L^p)} := \begin{cases} \|f\|_{L^p} + \left\| \left(\displaystyle\int_{\mathbb{R}^n} |h|^{-sq} \, |\Delta_h^M f(\cdot)|^q \frac{dh}{|h|^n} \right)^{\frac{1}{q}} \right\|_{L^p}, & 0 < q < \infty \\ \| \displaystyle\sup_{0 \ne h \in \mathbb{R}^n} |h|^{-s} \, |\Delta_h^M f(\cdot)| \, \|_{L^p}, & q = \infty. \end{cases} \tag{21}$$

As above, $F_q^s(L^p)$ is a Banach space for $1 \leq p, q \leq \infty$; otherwise a quasi-Banach space.

To show the versatility of Besov and Triebel-Lizorkin spaces, some commonly known function spaces are expressed as special cases of these function spaces.

Hölder spaces:

$$\text{For } s > 0 \text{ and } s \notin \mathbb{N},\ C^s = B_\infty^s(L^\infty). \tag{22}$$

Sobolev spaces:

$$\text{For } 1 < p < \infty \text{ and } m \in \mathbb{N},\ W^m(L^p) = F_2^m(L^p). \tag{23}$$

Slodeckij spaces:

$$\text{For } 1 \leq p < \infty \text{ and } s > 0,\ W^s(L^p) = B_p^s(L^p) = F_p^s(L^p). \tag{24}$$

Bessel potential spaces:

$$\text{For } 1 < p < \infty \text{ and } s > 0\text{: } H^s(L^p) = F_2^s(L^p). \tag{25}$$

For $s \in \mathbb{R}$, $H^s(L^p) = \{f \in \mathscr{S}' : \|\mathscr{F}^{-1}[(1+|\xi|^2)^{s/2}\,\mathscr{F}f(\xi)](\cdot)\|_{L^p} < \infty\}$
(Here $\mathscr{F}$ denotes the Fourier transform on the Schwartz space $\mathscr{S}'$.)

4 Fractal Functions of Besov and Triebel-Lizorkin Type

In this section, we state conditions on a fractal function defined by mappings of type 7 to belong to $B_q^s(L^p)$ and $F_q^s(L^p)$. These conditions generalize earlier results [8] for Sobolev and Hölder spaces.

For this purpose, it is further assumed that the linear part of $u_i \in \mathrm{Aff}\,(n)$ is a similarity (transformation), i.e., a mapping $S_i : \Omega \to \Omega$, such that $|S_i(\omega) - S_i(\omega')| = \gamma_i|\omega - \omega'|$, for all $\omega, \omega' \in \Omega$ and a constant $-1 < \gamma_i < 1$.

Theorem 3. *Suppose the assumptions of Definition 3 are satisfied. Let Ω be a compact subset of $\mathbb{R}^n$ and let $\mathfrak{F} : \Omega \to \mathbb{R}$ be a fractal function. If*

$$\begin{cases} \displaystyle\sum_i |\lambda_i|^q \gamma_i^{q\left(\frac{1}{p}-s\right)+1-n} < 1, & 0 < q < \infty \\[2ex] \displaystyle\sum_i |\lambda_i|\, \gamma_i^{\frac{1}{p}-s} < 1, & q = \infty, \end{cases} \tag{26}$$

then $\mathfrak{F} \in B_q^s(L^p)$

Proof. It suffices to show that the Read-Bajractarević operator $\mathscr{T}$ in 8 with $p_i \in \mathbb{P}^{M-1}$ is contractive on $B_q^s(L^p)$ if condition 26 holds. To this end, let $f, g \in B_q^s(L^p)$. Then

$$\|\mathscr{T}f - \mathscr{T}g\|_{B_q^s(L^p)} = \left\| \sum_i (f-g) \circ u_i^{-1} \chi_{u_i(\Omega)} \right\|_{B_q^s(L^p)} \tag{27}$$

and it is enough to show that there exists a constant $0 < c < 1$ such that, with $\phi := f - g$,

$$\left\| \sum_i \phi \circ u_i^{-1} \chi_{u_i(\Omega)} \right\|_{B_q^s(L^p)} \leq c \, \|\phi\|_{B_q^s(L^p)} \tag{28}$$

if 26 holds. First, notice that for $x \in u_i(\Omega)$

$$\begin{aligned} \|\phi\|_{L^p}^p &= \int_{u_i(\Omega)} |\phi|^p \, dx = \int_{u_i(\Omega)} |\lambda_i \phi \circ u_i^{-1}|^p \, dx \\ &= |\lambda_i|^p \int_\Omega |f|^p \, du(x) = |\lambda_i|^p \, \gamma_i \, \|\phi\|_{L^p}^p. \end{aligned}$$

Thus, as is known, $\sum_i \gamma_i \, |\lambda_i|^p < 1$ implies a fractal function is in L^p.

Now, for $x \in u_i(\Omega)$ and $0 < q < \infty$,

$$\begin{aligned} &\int_{\mathbb{R}^n} |h|^{-sq} \left[\int_{u_i(\Omega)} |\Delta_h^M \phi(x)|^p \, dx \right]^{\frac{q}{p}} \frac{dh}{|h|^n} \\ &\quad = \int_{\mathbb{R}^n} |h|^{-sq} \left[\int_{u_i(\Omega)} |\lambda_i \Delta_h^M \phi \circ u_i^{-1}(x)|^p \, dx \right]^{\frac{q}{p}} \frac{dh}{|h|^n} \\ &\quad = \int_{\mathbb{R}^n} |h|^{-sq} \left[|\lambda_i|^p \int_{u_i(\Omega)} \left| \sum_{m=0}^M (-1)^{M-m} \binom{M}{m} (\phi \circ u_i^{-1})(x + mh) \right|^p dx \right]^{\frac{q}{p}} \frac{dh}{|h|^n} \\ &\quad = |\lambda_i|^q \int_{\mathbb{R}^n} |h|^{-sq} \left[\int_{u_i(\Omega)} \left| \sum_{m=0}^M (-1)^{M-m} \binom{M}{m} \phi(u_i^{-1}(x) + m(S_i^{-1} h)) \right| dx \right]^{\frac{q}{p}} \frac{dh}{|h|^n} \\ &\quad = |\lambda_i|^q \int_{\mathbb{R}^n} |h|^{-sq} \left[\int_{u_i(\Omega)} \left| (\Delta_{S_i^{-1} h}^M \phi)(u_i^{-1}(x)) \right|^p dx \right]^{\frac{q}{p}} \frac{dh}{|h|^n} \\ &\quad = |\lambda_i|^q \int_{\mathbb{R}^n} |h|^{-sq} \gamma_i^{\frac{q}{p}} \, \|\Delta_{S_i^{-1} h}^M \phi\|_{L^p}^q \frac{dh}{|h|^n} = |\lambda_i|^q \, \gamma_i^{\frac{q}{p}} \int_{\mathbb{R}^n} |S_i h|^{-sq} \, \|\Delta_h^M \phi\|_{L^p}^q \frac{d(S_i h)}{|S_i h|^n} \\ &\quad = |\lambda_i|^q \, \gamma_i^{\frac{q}{p} + 1 - sq - n} \int_{\mathbb{R}^n} |h|^{-sq} \, \|\Delta_h^M \phi\|_{L^p}^q \frac{dh}{|h|^n}, \end{aligned}$$

from which for $q < \infty$ the statement follows by setting

$$c := \max \left\{ \sum_i \gamma_i \, |\lambda_i|^p, \sum_i |\lambda_i|^q \, \gamma_i^{q\left(\frac{1}{p} - s\right) + 1 - n} \right\}.$$

Similarly, one has for $x \in u_i(\Omega)$ and $q = \infty$

$$|h|^{-s}\left(\int_{u_i(\Omega)}|\Delta_h^M\phi|^p dx\right)^{\frac{1}{p}} = |h|^{-s}\left(\int_{u_i(\Omega)}|\lambda_i\Delta_h^M\phi\circ u_i^{-1}(x)|^p dx\right)^{\frac{1}{p}}$$
$$= |h|^{-s}|\lambda_i|\left(\int_{u_i(\Omega)}|\Delta^M_{S_i^{-1}h}\phi(u_i^{-1}(x))|^p dx\right)^{\frac{1}{p}} = |\lambda_i|\,\gamma_i^{\frac{1}{p}}\,\|\Delta^M_{S_i^{-1}h}\phi\|_{L^p}$$
$$= |\lambda_i|\,\gamma_i^{\frac{1}{p}-s}\,\|\Delta_h^M\phi\|_{L^p}.$$

Setting

$$c := \max\left\{\sum_i \gamma_i\,|\lambda_i|^p, \sum_i |\lambda_i|\,\gamma_i^{\frac{1}{p}-s}\right\}$$

concludes the proof.

In the special case that was considered in [8], one obtains from 26 for $n=1$, $\gamma_i = 1/N$, $s\in\mathbb{N}$, and $q=p<\infty$ the condition given in [8], namely

$$\sum_i |\lambda_i|^p N^{ps-1} < 1 \quad\Longrightarrow\quad \mathfrak{F}\in W^s(L^p). \tag{29}$$

For $p=q=\infty$, one obtains the known requirement for a fractal function to be in C^s:

$$\sum_i |\lambda_i|\gamma_i^{-s} < 1. \tag{30}$$

Note that this special case also exemplifies 22 and 23.

Theorem 4. *Assume that the requirements in Definition 4 hold. Let Ω be a compact subset of $\mathbb{R}^n$ and let $\mathfrak{F}:\Omega\to\mathbb{R}$ be a fractal function. If*

$$\begin{cases} \sum_i |\lambda_i|^p\,\gamma_i^{(1-n)\frac{p}{q}-sp+1} < 1, & 0<p,q<\infty \\ \sum_i |\lambda_i|^p\,\gamma_i^{1-sp} < 1, & 0<p<\infty,\ q=\infty \\ \sum_i |\lambda_i|\,\gamma_i^{-s} < 1, & p=q=\infty, \end{cases} \tag{31}$$

then $\mathfrak{F}\in F_q^s(L^p)$.

Proof. The proof mimicks that of Theorem 3; again, one needs to show that

$$\left\|\sum_i \phi\circ u_i^{-1}\chi_{u_i(\Omega)}\right\|_{F_q^s(L^p)} \le c\,\|\phi\|_{F_q^s(L^p)} \tag{32}$$

for some constant $0<c<1$. To this end, consider $x\in u_i(\Omega)$. Then

$$\int_{u_i(\Omega)} \left(\left[\int_{\mathbb{R}^n} |h|^{-sq} \, |\Delta_h^M \phi(x)|^q \, \frac{dh}{|h|^n} \right]^{\frac{1}{q}} \right)^p dx$$

$$= \int_{u_i(\Omega)} \left(\left[\int_{\mathbb{R}^n} |h|^{-sq} \, |\lambda_i \Delta_h^M \phi \circ u_i^{-1}(x)|^q \, \frac{dh}{|h|^n} \right]^{\frac{1}{q}} \right)^p dx$$

$$= |\lambda_i|^p \int_{u_i(\Omega)} \left(\left[\int_{\mathbb{R}^n} |h|^{-sq} \, |(\Delta^M_{S_i^{-1}h} \phi)(u_i^{-1}(x))|^q \, \frac{dh}{|h|^n} \right]^{\frac{1}{q}} \right)^p dx$$

$$= |\lambda_i|^p \int_{u_i(\Omega)} \gamma_i^{(-sq+1-n)\frac{p}{q}} \left(\int_{\mathbb{R}^n} |h|^{-sq} \, |\Delta_h^M \phi(u_i^{-1}(x))|^q \, \frac{dh}{|h|^n} \right)^{\frac{p}{q}} dx$$

$$= |\lambda_i|^p \, \gamma_i^{(-sq+1-n)\frac{p}{q}+1} \left\| \left(|h|^{-sq} \, |\Delta_h^M \phi(\cdot)|^q \, \frac{dh}{|h|^n} \right)^{\frac{1}{q}} \right\|_{L^p},$$

giving the statement for $p, q < \infty$.

If $0 < p < \infty$ and $q = \infty$, one has for $x \in u_i(\Omega)$

$$\int_{\mathbb{R}^n} \left(\sup |h|^{-s} \, |\Delta_h^M \phi(x)| \right)^p dx = \int_{u_i(\Omega)} \left(\sup |h|^{-s} |\lambda_i \, \Delta_h^M \phi \circ u_i^{-1}(x)| \right)^p dx$$

$$= |\lambda_i|^p \int_{u_i(\Omega)} \left(\sup |h|^{-s} \, |\Delta^M_{S_i^{-1}h} \phi(u_i^{-1}(x))| \right)^p dx$$

$$= |\lambda_i|^p \, \gamma_i^{-sp} \int_{u_i(\Omega)} \left(\sup |h|^{-s} \, |\Delta_h^M \phi(u_i^{-1}(x))| \right)^p dx$$

$$= |\lambda_i|^p \, \gamma_i^{1-sp} \, \| \sup |h|^{-s} \, |\Delta_h^M \phi(\cdot)\|_{L^p}^p,$$

which implies the result in this case.

Lastly, if $p = q = \infty$ and $x \in u_i(\Omega)$ again, one obtains

$$\sup_{x \in u_i(\Omega)} \sup_{0 \neq h \in \mathbb{R}^n} |h|^{-s} \, |\Delta_h^M \phi(x)| = \sup \sup |h|^{-s} \, |\lambda_i \, \Delta_h^M \phi \circ u_i^{-1}(x)|$$

$$= |\lambda_i| \, \sup \sup |h^{-s}| \, |\Delta^M_{S_i^{-1}h} \phi(u_i^{-1}(x))| = |\lambda_i| \, \gamma_i^{-s} \sup \sup \, |h^{-s}| \, |\Delta_h^M \phi(u_i^{-1}(x))|$$

$$\leq |\lambda_i| \, \gamma_i^{-s} \sup_{x \in \mathbb{R}^n} \sup_{0 \neq h \in \mathbb{R}^n} |h|^{-s} \, |\Delta_h^M \phi(x)|,$$

which concludes this case and finishes the proof.

References

1. Barnsley M F (1986) Constr. Approx. 2:303–329.
2. de Boor C (2001) A Practical Guide to Splines. Revised edition, Springer Verlag, New York
3. Dahmen W (1999) Travaux Mathématiques 15 – 76.

4. Geronimo J, Hardin D, Massopust P (1995) Fractal surfaces, multiresolution analyses and wavelet transforms. In: O Y-L, Toet A, Foster D, Heijmans H, Meer P (eds) Shape in Picture – Mathematical description of shape in grey-level images. NATO ASI Series Vol. 126. Springer-Verlag, Berlin Heidelberg New York
5. Hutchinson J (1981) Indiana Univ. J. Math. 30:713–747.
6. Massopust P (1993) Zeitschrift für Analysis u. i. Anwend. 12:201–210.
7. Massopust P (1995) Fractal Functions, Fractal Surfaces, and Wavelets. Academic Press, New York
8. Massopust P (1997) Chaos, Solitons & Fractals 8(2):171–190
9. Triebel H (1983) Theory of Function Spaces. Birkhäuer, Basel

Hölderian random functions

Antoine Ayache, Philippe Heinrich, Laurence Marsalle, and Charles Suquet

Laboratoire P. Painlevé, CNRS UMR 8524, Université Lille 1, 59655 Villeneuve d'Ascq cedex, France. Antoine.Ayache@math.univ-lille1.fr

Summary. Hölder regularity which plays a key rôle in fractal geometry raises an increasing interest in probability and statistics. In this paper we discuss various aspects of local and global regularity for stochastic processes and random fields. As a main result we show the invariability of the pointwise Hölder exponent of a continuous and nowhere differentiable random field which has stationary increments and satisfies a zero-one law. We also survey some recent uses of Hölder spaces in limit theorems for stochastic processes and statistics.

1 Introduction

The concept of Hölder regularity is quite important in fractal geometry, signal and image processing, finance, statistics and telecommunications [8]. Hölder exponents have been used frequently to measure the roughness of a curve or of a surface [10]; applications in signal and image processing are numerous and include interpolation, segmentation [26] and denoising [27]. They are closely related to other fractal indices such as fractal dimensions, self-similarity parameters and multifractal spectra (see e.g. [12,18,44,45]). On the other hand, the Hölder spaces provide a functional framework for limit theorems in the theory of stochastic processes. The use of Hölder topologies leads to more precise results than the classical framework of continuous functions spaces.

This paper discuss both uses of Hölder regularity in the study of stochastic processes.

1.1 Hölder exponents

When looking for random fields modeling some roughness, it is quite natural to investigate the pointwise Hölder regularity of various extensions of the well known Brownian motion.

Recall that $\{B_H(t), t \in \mathbb{R}^d\}$, the fractional Brownian motion (fBm) of Hurst parameter $H \in (0,1)$ is the real-valued, self-similar and stationary

increments continuous Gaussian field defined for every $t \in \mathbb{R}^d$ as the Wiener integral

$$B_H(t) = \int_{\mathbb{R}^d} \frac{e^{it.\xi} - 1}{|\xi|^{H+d/2}} \, d\widehat{W}(\xi), \tag{1}$$

where $d\widehat{W}$ is a complex-valued white noise. This field was first introduced by Kolmogorov [20] for generating Gaussian "spirals" in a Hilbert space. Later, the seminal article of Mandelbrot and Van Ness [30] emphasized its relevance for the modeling of natural phenomena (hydrology, finance,...) and thus greatly contributed to make it popular. Since then, this field turns out to be a very powerful tool in modeling. The monograph of Doukhan, Oppenheim and Taqqu [11] offers a systematic treatment of fBm, as well as an overview of different areas of applications.

The field $\{B_H(t), t \in \mathbb{R}^d\}$ is a natural generalization of the Wiener process ($\{B_{1/2}(t), t \in \mathbb{R}\}$ is a Wiener process) and shares many nice properties with it. However, one of the main advantages of fBm with respect to Wiener process is that its increments are correlated and can even display long range dependence. Still, fBm is not always a realistic model. Indeed, its pointwise Hölder exponent remains constant all along its trajectory which can be a serious drawback in several applications (see for example [2–4, 28]). Generally speaking, a multifractional field is a field with continuous trajectories that extends fBm and whose pointwise Hölder exponent is allowed to change from one point to another. Recall that $\{\alpha_X(t), t \in T\}$ the pointwise Hölder exponent of a continuous and nowhere differentiable field $\{X(t), t \in T\}$ is defined for every $t \in T$ as

$$\alpha_X(t) = \sup\Big\{\alpha;\ \limsup_{h \to 0} \frac{|X(t+h) - X(t)|}{|h|^\alpha} = 0\Big\}. \tag{2}$$

A paradigmatic example of a multifractional field is multifractional Brownian motion (mBm). It was introduced independently in [28] and in [7] but the denomination multifractional Brownian motion is due to Lévy Véhel. MBm is obtained by substituting to the Hurst parameter in the harmonizable representation (1) of fBm a continuous function $t \mapsto H(t)$ with values in $(0, 1)$. When the function $H(\cdot)$ is smooth enough (typically when it is a C^1 function), the pointwise Hölder exponent of mBm satisfies for any $t \in T$, almost surely $\alpha_{\mathrm{mBm}}(t) = H(t)$, which means that it can change from one point to another.

In [5] it has been proved that when the increments of any order of mBm are stationary the function $H(\cdot)$ is constant (i.e. mBm reduces to an fBm). Moreover, no example of a multifractional field with stationary increments has been constructed yet. *This is why it seems natural to wonder whether there exists a continuous, nowhere differentiable and stationary increments field* $\{X(t), t \in T\}$ *whose pointwise Hölder exponent changes from one point to another.* In Section 2 we show that the answer is negative when we impose in addition to $\{X(t), t \in T\}$ to satisfy a zero-one law.

1.2 Hölder spaces as a functional framework

In many situations some uniform control on the regularity is needed. For instance let us consider the following statistical problem. Having observed random variables $X_1, \ldots, X_n$, we need to test the null hypothesis that $X_1, \ldots, X_n$ have the same expectation μ_0 against the alternative of a change from μ_0 to μ_1 between the unknown instants k^* and m^* with going back to μ_0 after m^*. This is known as the *epidemic model*. It is quite natural, see [41] for a step by step explanation, to use here the weighted test statistics

$$\mathrm{UI}(n,a) := \max_{1\le i<j\le n} \frac{\big|S(j)-S(i)-S(n)(t_j-t_i)\big|}{|t_j-t_i|^a}$$

where $S(n) := \sum_{1\le i\le n} X_i$, $t_i := i/n$ and $0 < a < 1/2$. The asymptotic distribution of $\mathrm{UI}(n,a)$ follows from the weak convergence in the Hölder space $\mathcal{H}^a$ of a partial sums process ξ_n (a precise definition of Hölder spaces are given in Section 3). The practical interest of the exponent a here lies in the sensitivity of the test. Detecting the shortest epidemics requires to take the biggest possible a and this leads to investigate weak convergence of ξ_n in $\mathcal{H}^a$.

In Section 4 we survey some recent advances in the asymptotic theory of sequences of stochastic processes considered as random elements in some Hölder space $\mathcal{H}$. The issues addressed may be classified along the following two main directions.

A) Classical limit theorems for normalized sums $b_n^{-1}S_n$ of independent random elements X_i in $\mathcal{H}$: laws of large number, central limit theorems, see e.g. [31], [32], [35], [40].
B) Weak convergence in $\mathcal{H}$ of sequences of random elements ξ_n of the form

$$t \longmapsto \xi_n(t) = G_n(X_1, \ldots, X_n, t),$$

where $X_1, \ldots, X_n$ is usually a sample of i.i.d. random variables or random elements in some Banach space and G_n a function smooth enough to ensure the membership in $\mathcal{H}$ of ξ_n .

Problem A) is directly connected to the Probability Theory in Banach Spaces. It is well known in this area that the limit theorems for a sequence of random elements $b_n^{-1}S_n$ in some separable Banach space $\mathbb{B}$ involve the geometry of $\mathbb{B}$. For instance, if the X_i's are i.i.d. with null expectation, then the square integrability of $\|X_1\|$ gives the asymptotic normality of $n^{-1/2}S_n$ when $\mathbb{B}$ is of type 2 (e.g. $\mathbb{B}$ is a Hilbert space or a L^p space with $2 \le p < \infty$). But if $\mathbb{B} = c_0$, the classical space of sequences converging to 0, we can find a *bounded* random element X_1 in c_0 which does not satisfy the CLT, i.e. the corresponding sequence $n^{-1/2}S_n$ is not asymptotically Gaussian. This makes hopeless characterizing the CLT in a general $\mathbb{B}$ in terms of integrability properties of X_1 only. Because all the Hölder spaces $\mathcal{H}$ under consideration here contain a subspace isomorphic to c_0 and are concrete function spaces,

they provide an interesting framework to study the asymptotic behavior of $b_n^{-1} S_n$ in a context where the geometry of the Banach space is "bad".

Problem B) is more oriented to statistical applications. Indeed the weak convergence $\xi_n \xrightarrow{\mathcal{D}} \xi$ in some function space E means

$$\mathbf{E}\, g(\xi_n) \xrightarrow[n\to\infty]{} \mathbf{E}\, g(\xi), \tag{3}$$

for every continuous and bounded function $g : E \to \mathbb{R}$. By the continuous mappingtheorem, this implies that for every functional $f : E \to \mathbb{R}$, continuous with respect to the strong topology of E,

$$f(\xi_n) \xrightarrow[n\to\infty]{} f(\xi), \quad \text{in distribution.} \tag{4}$$

The classical functional frameworks for such convergence $\xi_n \xrightarrow{\mathcal{D}} \xi$ are the Skorohod space when ξ_n has jumps and some space $\mathcal{C}$ of continuous functions when ξ_n has continuous paths. The interest in replacing, whenever possible, $\mathcal{C}$ by $\mathcal{H}$ is that this strenghtening of the topology on the paths space enlarges the set of continuous functionals f. Usually, the random functions ξ_n share more smoothness than their weak limit ξ. For instance in the Hölderian version of the invariance principle for partial sums processes, the paths of ξ_n are random polygonal lines, while ξ is a Brownian motion. In such cases the global smoothness of ξ put a natural bound in the choice of the "best" space $\mathcal{H}$. The example of the Brownian motion W shows here that the classical ladder of Hölder spaces $\mathcal{H}^a$ is not rich enough. Indeed $\mathcal{H}^a$ is the space of functions x whose increments $x(t+h) - x(t)$, $h \geq 0$ are $O(h^a)$ uniformly in t. Due to Lévy's result on W's modulus of uniform continuity [25, Th. 52,2], it seems desirable to consider also the spaces of functions x such that $x(t+h) - x(t) = O\big(h^{1/2} \ln^b(1/h)\big)$. In more generality, this leads to introduce a ladder of Hölder spaces $\mathcal{H}^\rho$, where membership of x in $\mathcal{H}^\rho$ is equivalent to the uniform estimate $x(t+h) - x(t) = O(\rho(h))$, for some weight function ρ.

This raises a third problem which in some sense is also preliminary to Problem A):

C) Given a stochastic process $X = \{X(t),\, t \in T\}$, find conditions in terms of its finite dimensional distributions so that X admits a version with paths in the Hölder space $\mathcal{H}^\rho$.

2 Critical Exponents

2.1 Zero-One laws and Exponents

Let $X = \{X(t), t \in T\}$ be a real process with, say, separable and metric time set T. We can view X as a random element in $\mathbb{R}^T$ endowed with product σ-field $\mathcal{B}(\mathbb{R})^{\otimes T}$ where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-field of $\mathbb{R}$. The kind of zero-one law we shall focus on can be stated as follows:

Definition 1. *We will say that* X satisfies a zero-one law *if for each measurable linear subspace* V *of* $\mathbb{R}^T$,

$$\mathbf{P}\,(X \in V) = 0 \text{ or } 1. \tag{5}$$

It is known that Gaussian, stable and some infinitely divisible (without Gaussian component) processes satisfy (5). Their associated finite-order chaos processes do as well. We refer the reader to the paper [43] by Rosinski and Samorodnitsky and the references therein. One classical application of such a zero-one law is to establish regularity properties of paths. For instance, neglecting measurability questions at first glance, V could be:

1. the space of bounded functions on T,
2. the space of continuous functions on T, if T is compact,
3. the space of uniformly continuous functions on T,
4. the space of Lipschitz functions on T,
5. the space of a-Hölderian functions on T,
6. the space of absolutely continuous functions on T, if T is an interval in $\mathbb{R}$.

If for some countable dense subset S of T, we have

$$\mathbf{P}\big(\forall t \in T,\, \exists (s_n)_{n\geq 1} \subset S,\, s_n \to t,\, X(s_n) \to X(t)\big) = 1,$$

then the measurability of $\{X \in V\}$, for the V's displayed above, can be established provided the underlying probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is complete (as can be always assumed). Indeed, the events $\{X \in V\}$ may then be expressed, up to negligeable sets, as ones involving only the restriction of X to S. The following result illustrates how useful this zero-one law can be.

Theorem 1. *Assume that* T *is an open subset of* $\mathbb{R}^d$ *(*$d \in \mathbb{N}$*). Let* $X = \{X(t),\, t \in T\}$ *be a continuous, nowhere differentiable process satisfying the zero-one law (5). Then, for all* $t \in T$*, the pointwise Hölder exponent of* X *at* t *is almost surely deterministic. In other words, for all* $t \in T$*, there exists a number* $H(t) \in [0,1]$ *such that*

$$\mathbf{P}\big(\alpha_X(t) = H(t)\big) = 1.$$

Proof. This result has already been established by Ayache and Taqqu for Gaussian processes, see [6]. Their proof is based on the same key-argument (zero-one law), but the one we give here uses it more directly and explicitely.

We set $S = \mathbb{Q}^d$. Let t be some arbitrary point of the open set T and choose $\eta > 0$ such that the ball $B(t,\eta)$ be in T. Since X has almost all continuous and nowhere differentiable paths on T, we know that $\alpha_X(t,\omega)$ belongs to $[0,1]$ for almost all $\omega \in \Omega$. We can thus define

$$u_*(t) := \sup\big\{u \in \mathbb{R}; \mathbf{P}(\alpha_X(t) \leq u) = 0\big\},$$
$$u^*(t) := \inf\big\{u \in \mathbb{R}; \mathbf{P}(\alpha_X(t) \leq u) = 1\big\}.$$

By definition, the interval $[u_*(t), u^*(t)]$ is the support of the distribution function of $\alpha_X(t)$. To prove that $\alpha_X(t)$ is almost surely deterministic, we only need to check that $u^*(t) \leq u_*(t)$. Let $u < u^*(t)$, we thus have $\mathbf{P}(\alpha_X(t) > u) > 0$. On the event $\{\alpha_X(t) > u\}$, we have $\limsup_{h\to 0} |h|^{-u}|X(t+h) - X(t)| = 0$ which implies that $h \mapsto h^{-u}(X(t+h) - X(t))$ is bounded on the countable bounded subset $\{h \in S;\, |h| < \eta\}$. It follows, by inclusion, that

$$0 < \mathbf{P}\big(\alpha_X(t) > u\big) \leq \mathbf{P}\left(\sup_{\substack{|h|<\eta \\ h\in S}} \frac{|X(t+h) - X(t)|}{|h|^u} < \infty\right).$$

We shall prove that this last probability is equal to 1, using the zero-one law (5). To this end, note that the event

$$\left\{\sup_{\substack{|h|<\eta \\ h\in S}} \frac{|X(t+h) - X(t)|}{|h|^u} < \infty\right\}$$

can clearly be written $\{X \in V\}$ for some linear subspace V, which is $\mathcal{B}(\mathbb{R})^{\otimes T}$-measurable since it involves only countable many projections $x \mapsto x(t)$ from $\mathbb{R}^T$ to $\mathbb{R}$. The zero-one law ensures consequently that

$$\mathbf{P}\left(\sup_{|h|<\eta} \frac{|X(t+h) - X(t)|}{|h|^u} < \infty\right) = 1,$$

where we skipped the restriction $h \in S$ in the supremum thanks to the continuity of X. But now, a simple inclusion of events yields

$$\mathbf{P}\left(\limsup_{h\to 0} \frac{|X(t+h) - X(t)|}{|h|^u} < \infty\right) = 1,$$

which can be read as $\mathbf{P}\left(u \leq \alpha_X(t)\right) = 1$ or, equivalently, as $\mathbf{P}(\alpha_X(t) < u) = 0$. This means that $u \leq u_*(t)$ which gives $u^*(t) \leq u_*(t)$, since u is arbitrary in $(-\infty, u^*(t))$. Besides, by definition $u_*(t) \leq u^*(t)$, whence $u_*(t) = u^*(t)$. We call $H(t)$ this common value. We just have proved that the distribution function of $\alpha_X(t)$ jumps from 0 to 1 at $H(t)$ in other words $\mathbf{P}(\alpha_X(t) = H(t)) = 1$.

Remark 1. Other exponents may be defined to characterize the regularity of a function, for instance the *local Hölder exponent* at time t

$$\tilde{\alpha}_X(t) = \sup\left\{\alpha; \exists \eta > 0 \sup_{u,v\in B(t,\eta)} \frac{|X(u) - X(v)|}{|u - v|^\alpha} < \infty\right\},$$

where $B(t, \eta)$ denotes the open ball centered at t and of radius η, and the *global Hölder exponent* on a compact set $K \subset T$

$$\beta_X = \sup\left\{\beta;\ \sup_{u,v\in K}\frac{|X(u)-X(v)|}{|u-v|^\beta}<\infty\right\}.$$

When X is a continuous nowhere differentiable process, satisfying a zero-one law, the same property as in Theorem 1 holds. More precisely, for all compact subset K of T, β_X is almost surely deterministic, and for all $t\in T$, $\tilde{\alpha}_X(t)$ is almost surely deterministic. The proof of both results is the same as for Theorem 1, the only change concerns the measurable subspace of $\mathbb{R}^T$ involved in the zero-one law. In the case of $\tilde{\alpha}_X(t)$, we use

$$\tilde{V} := \bigcup_{n\in\mathbb{N}^*}\left\{x\in\mathbb{R}^T;\ \sup_{u,v\in B(t,1/n)\cap S}\frac{|x(u)-x(v)|}{|u-v|^\alpha}<\infty\right\},$$

where S is a countable dense subset of T, and in the case of β_X, we define

$$W := \left\{x\in\mathbb{R}^T;\ \sup_{u,v\in K\cap S}\frac{|x(u)-x(v)|}{|u-v|^\alpha}<\infty\right\}.$$

2.2 Processes with Stationary Increments

Throughout all this paragraph, T, the set of times, can be taken to be equal to any non-empty open subset of $\mathbb{R}^d$ ($d\in\mathbb{N}$).

Let $X=\{X(t), t\in T\}$ denote a continuous nowhere differentiable process, for which a zero-one law holds (see (5)). Thanks to Subsection 2.1, we know that the pointwise Hölder exponent of X at t is deterministic, but depends on t. Now, we assume besides that X has stationary increments. This means that $(X(s_2)-X(s_1),\ldots,X(s_n)-X(s_1))$ and $(X(s_2+t)-X(s_1+t),\ldots,X(s_n+t)-X(s_1+t))$ are identically distributed for any $s_1,\ldots,s_n,s_1+t,\ldots,s_n+t\in T$ and any integer $n\geq 2$. Since the pointwise Hölder exponent is defined by means of increments, we can show that it doesn't depend anymore on t.

Theorem 2. *Let $X=\{X(t),t\in T\}$ be a continuous nowhere differentiable process, with stationary increments. We assume that a zero-one law holds for X. Then there exists $H\in[0,1]$ such that for all $t\in T$*

$$\mathbf{P}\left(\alpha_X(t)=H\right)=1.$$

Proof. The scheme of the proof is the following: since T is a non-empty set, it contains at least one element, that we will denote 0 for the sake of simplicity. Using an equivalent definition of the pointwise Hölder exponent, we prove that $\alpha_X(t)$ and $\alpha_X(0)$ have the same law, for all $t\in T$. Thanks to Theorem 1, we know that there exists $H=H(0)\in[0,1]$ such that the law of $\alpha_X(0)$ is a Dirac mass at point H. Consequently, for all $t\in T$, the law of $\alpha_X(t)$ is a Dirac mass at point H.

As in the proof of Theorem 1, S denotes a countable dense subset of T. Let t be a fixed point of T. It can be easily shown that the pointwise Hölder exponent of X at t is given by:

$$\alpha_X(t) = \liminf_{h \to 0} \frac{\log |X(t+h) - X(t)|}{\log |h|},$$

with the usual convention that $\log 0 = -\infty$. This definition reads as

$$\begin{aligned} \alpha_X(t) &= \sup_{R>0} \inf_{|h|<R} \frac{\log |X(t+h) - X(t)|}{\log |h|} \\ &= \sup_{n \in \mathbb{N}} \inf_{\substack{|h|<1/n \\ h \in S}} \frac{\log |X(t+h) - X(t)|}{\log |h|}, \end{aligned}$$

the last equality coming from the monotonicity of the infimum with respect to R and from the continuity of the paths of X. To obtain the identity in law between $\alpha_X(t)$ and $\alpha_X(0)$, we introduce an increasing sequence $(S_k)_{k \geq 1}$ of *finite* sets such that $\cup_{k \geq 1} S_k = S$. Then, note that for $u \in \mathbb{R}$

$$\{\alpha_X(t) \leq u\} = \bigcap_{n \in \mathbb{N}}{\downarrow} \bigcap_{m \in \mathbb{N}}{\downarrow} \bigcup_{k \in \mathbb{N}}{\uparrow} \left\{ \min_{\substack{|h|<1/n \\ h \in S_k}} \frac{\log |X(t+h) - X(t)|}{\log |h|} < u + \frac{1}{m} \right\},$$

so that, for every $t \in T$ and $u \in \mathbb{R}$, by sequential monotonic continuity of $\mathbf{P}$,

$$\mathbf{P}\left(\alpha_X(t) \leq u\right) = \lim_n \lim_m \lim_k \mathbf{P} \left(\min_{\substack{|h|<1/n \\ h \in S_k}} \frac{\log |X(t+h) - X(t)|}{\log |h|} < u + \frac{1}{m} \right).$$

The event mentioned in the last probability involves a *finite* number of increments of X, all of them based on point t. The stationarity of the increments implies that we can replace them with the analogue increments, based on point 0. It follows that for all $u \in \mathbb{R}$

$$\mathbf{P}(\alpha_X(0) \leq u) = \mathbf{P}(\alpha_X(t) \leq u),$$

which means that $\alpha_X(t)$ and $\alpha_X(0)$ have the same law, for all $t \in T$. We already know that the pointwise Hölder exponent of X at point 0 is deterministic, so that there exists $H \in [0,1]$ such that $\mathbf{P}(\alpha_X(0) = H) = 1$. The equality in law between $\alpha_X(t)$ and $\alpha_X(0)$ thus leads to $\mathbf{P}(\alpha_X(t) = H) = 1$, for all $t \in T$.

3 Hölder spaces

Let us introduce the Hölder spaces by an informal description of the most familiar case. For fixed $0 < a < 1$, $\mathcal{H}^a$ is the set of functions $x : [0,1] \to \mathbb{R}$ such that $|x(t) - x(s)| \leq K|t-s|^a$ for some constant K depending only on x and a. The best constant K in this uniform estimate defines a semi-norm on the vector space $\mathcal{H}^a$. By adding $|x(0)|$ to this semi-norm we obtain a norm $\|x\|_a$

which makes $\mathcal{H}^a$ a non separable Banach space. Clearly if $0 < a < b < 1$, $\mathcal{H}^b$ is topologically embedded in $\mathcal{H}^a$ and all these Hölder spaces are topologically embedded in the classical Banach space $\mathcal{C}$ of continuous functions $[0,1] \to \mathbb{R}$.

To remedy the non separability drawback of $\mathcal{H}^a$, one introduces its subspace $\mathcal{H}^{a,o}$ of functions x such that $|x(t) - x(s)| = o(|t-s|^a)$ uniformly. This subspace is *closed* (hence also a Banach space for the same norm $\|x\|_a$) and *separable*.

One interesting feature of the spaces $\mathcal{H}^{a,o}$ is the existence of a basis of triangular functions, see [9]. It is convenient to write this basis as a triangular array of functions, indexed by the dyadic numbers. Let us denote by D_j the set of dyadic numbers in $[0,1]$ of level j, i.e.

$$\mathrm{D}_0 = \{0,1\}, \qquad \mathrm{D}_j = \left\{(2l-1)2^{-j};\ 1 \le l \le 2^{j-1}\right\}, \quad j \ge 1.$$

Write for $r \in \mathrm{D}_j$, $j \ge 0$,

$$r^- := r - 2^{-j}, \quad r^+ := r + 2^{-j}.$$

For $r \in \mathrm{D}_j$, $j \ge 1$, the triangular Faber-Schauder functions Λ_r are continuous, piecewise affine with support $[r^-, r^+]$ and taking the value 1 at r:

$$\Lambda_r(t) = \begin{cases} 2^j(t - r^-) & \text{if } t \in (r^-, r]; \\ 2^j(r^+ - t) & \text{if } t \in (r, r^+]; \\ 0 & \text{else.} \end{cases}$$

When $j = 0$, we just take the restriction to $[0,1]$ in the above formula, so

$$\Lambda_0(t) = 1 - t, \quad \Lambda_1(t) = t, \quad t \in [0,1].$$

The sequence $\{\Lambda_r; r \in \mathrm{D}_j,\ j \ge 0\}$ is a Schauder basis of $\mathcal{C}$. Each $x \in \mathcal{C}$ has a unique expansion

$$x = \sum_{j=0}^{\infty} \sum_{r \in \mathrm{D}_j} \lambda_r(x) \Lambda_r, \tag{6}$$

with uniform convergence on $[0,1]$. The Schauder scalar coefficients $\lambda_r(x)$ are given by

$$\lambda_r(x) = x(r) - \frac{x(r^+) + x(r^-)}{2}, \quad r \in \mathrm{D}_j,\ j \ge 1, \tag{7}$$

and in the special case $j = 0$, by $\lambda_0(x) = x(0)$ and $\lambda_1(x) = x(1)$. The partial sum $\sum_{j=0}^{n}$ in the series (6) gives the linear interpolation of x by a polygonal line between the dyadic points of level at most n.

Ciesielski [9] proved that $\{\Lambda_r; r \in \mathrm{D}_j,\ j \ge 0\}$ is also a Schauder basis of each space $\mathcal{H}^{a,o}$ (hence the convergence (6) holds in the $\mathcal{H}^a$ topology when $x \in \mathcal{H}^{a,o}$) and that the norm $\|x\|_a$ is equivalent to the following sequence norm :

$$\|x\|_a^{\mathrm{seq}} := \sup_{j \ge 0} 2^{ja} \max_{r \in \mathrm{D}_j} |\lambda_r(x)|.$$

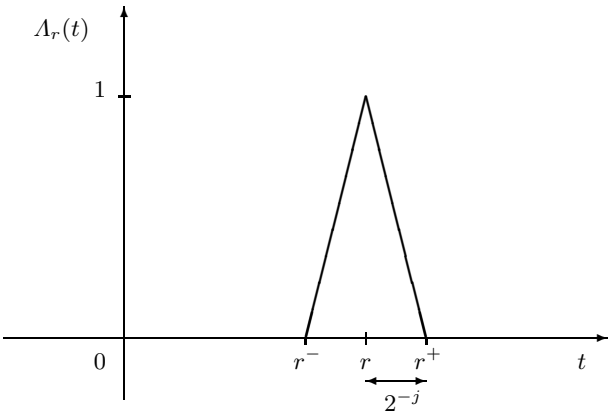


Fig. 1. The Faber-Schauder triangular function Λ_r

This equivalence of norms provides a very convenient discretization procedure to deal with Hölder spaces and is extended in Račkauskas and Suquet [34] to the more general setting of Hölder spaces of Banach space valued functions x, with a modulus of continuity controlled by some weight function ρ.

Let $(\mathbb{B}, \|\ \|)$ be a separable Banach space. We write $\mathcal{C}_{\mathbb{B}}$ for the Banach space of continuous functions $x : [0,1] \to \mathbb{B}$ endowed with the supremum norm $\|x\|_\infty := \sup\{\|x(t)\|\,;\ t \in [0,1]\}$. Let ρ be a real valued non decreasing function on $[0,1]$, null and right continuous at 0. Put

$$\omega_\rho(x,\delta) := \sup_{\substack{s,t\in[0,1],\\ 0<t-s<\delta}} \frac{\|x(t)-x(s)\|}{\rho(t-s)}.$$

Denote by $\mathcal{H}^\rho_{\mathbb{B}}$ the set of continuous functions $x : [0,1] \to \mathbb{B}$ such that $\omega_\rho(x,1) < \infty$. The set $\mathcal{H}^\rho_{\mathbb{B}}$ is a Banach space when endowed with the norm

$$\|x\|_\rho := \|x(0)\| + \omega_\rho(x,1).$$

Define

$$\mathcal{H}^{\rho,o}_{\mathbb{B}} = \{x \in \mathcal{C}_{\mathbb{B}} :\ \lim_{\delta\to 0} \omega_\rho(x,\delta) = 0\}.$$

Then $\mathcal{H}^{\rho,o}_{\mathbb{B}}$ is a closed separable subspace of $\mathcal{H}^\rho_{\mathbb{B}}$. We shall abbreviate $\mathcal{C}_{\mathbb{R}}$, $\mathcal{H}^\rho_{\mathbb{R}}$ and $\mathcal{H}^{\rho,o}_{\mathbb{R}}$ in $\mathcal{C}$, $\mathcal{H}^\rho$ and $\mathcal{H}^{\rho,o}$ correspondingly. Our main examples of Hölder spaces use as weight function $\rho = \rho_{a,b}$, $0 < a < 1$, $b \in \mathbb{R}$ defined by:

$$\rho_{a,b}(h) := h^a \ln^b(c/h), \quad 0 < h \le 1,$$

for a suitable constant c. For $\rho = \rho_{a,b}$, we shall write $\mathcal{H}^{a,b}_{\mathbb{B}}$ and $\mathcal{H}^{a,b,o}_{\mathbb{B}}$ for $\mathcal{H}^\rho_{\mathbb{B}}$ and $\mathcal{H}^{\rho,o}_{\mathbb{B}}$ respectively and we abbreviate $\mathcal{H}^{a,0,o}_{\mathbb{B}}$ in $\mathcal{H}^{a,o}_{\mathbb{B}}$. As above, the subscript $\mathbb{B}$ will be omitted when $\mathbb{B} = \mathbb{R}$.

In what follows, we assume that the weight function ρ satisfies the following technical conditions where c_1, c_2 and c_3 are positive constants:

$$\rho(0) = 0,\ \rho(\delta) > 0,\ 0 < \delta \leq 1; \tag{8}$$

$$\rho \text{ is non decreasing on } [0,1]; \tag{9}$$

$$\rho(2\delta) \leq c_1 \rho(\delta), \quad 0 \leq \delta \leq 1/2; \tag{10}$$

$$\int_0^\delta \frac{\rho(u)}{u}\, du \leq c_2 \rho(\delta), \quad 0 < \delta \leq 1; \tag{11}$$

$$\delta \int_\delta^1 \frac{\rho(u)}{u^2}\, du \leq c_3 \rho(\delta), \quad 0 < \delta \leq 1. \tag{12}$$

For instance, elementary computations show that the functions $\rho_{a,b}$ satisfy Conditions (8) to (12), for a suitable choice of the constant c, namely $c \geq \exp(b/a)$ if $b > 0$ and $c > \exp(-b/(1-a))$ if $b < 0$. For any ρ satisfying (8) to (12), we have the equivalence of norms :

$$\|x\|_\rho \sim \|x\|_\rho^{\text{seq}} := \sup_{j \geq 0} \frac{1}{\rho(2^{-j})} \max_{r \in \mathrm{D}_j} \|\lambda_r(x)\|\,,$$

where the $\mathbb{B}$-valued coefficients $\lambda_r(x)$ are still defined by (7).

The space $E = \mathcal{H}_{\mathbb{B}}^{\rho,o}$ may be used as the topological framework for limit theorems. Among various continuous functionals f for which the convergence (4) holds, let us mention the norms $f_1(x) = \|x\|_\rho$ and $f_2(x) = \|x\|_\rho^{\text{seq}}$, which are closely connected to the test statistics proposed below for the detection of epidemic changes. Other examples of Hölder continuous functionals and operators like p-variation, fractional derivatives are given in Hamadouche [15].

4 Random elements in Hölder spaces

4.1 Processes with a version in Hölder space

Let $\mathbb{B}$ be a Banach space. We consider a given $\mathbb{B}$-valued stochastic process $\xi = \{\xi(t),\ t \in T\}$, continuous in probability and discuss the problem of existence of a version of ξ with almost all paths in $\mathcal{H}_{\mathbb{B}}^{\rho,o}$. For simplicity we restrict this presentation to the case $T = [0,1]$. The results presented here are proved in [34], in the more general case of $\mathbb{B}$-valued random fields. The main analytic tool in this problem is the following generalization of the Faber-Schauder decomposition.

Proposition 1. *For a $\mathbb{B}$-valued array $\nu = (\nu_r;\ j \geq 0, r \in D_j)$, consider the following conditions.*

(a) $\displaystyle\sum_{j=0}^{\infty} \max_{r \in D_j} \|\nu_r\| < \infty.$

(b) $\sup_{j\geq 0} \dfrac{1}{\rho(2^{-j})} \max_{r\in D_j} \|\nu_r\| < \infty.$

(c) $\lim_{J\to\infty} \sup_{j>J} \dfrac{1}{\rho(2^{-j})} \max_{r\in D_j} \|\nu_r\| = 0.$

Define the sequence $(y_J)_{J\geq 0}$ *of continuous piecewise affine functions by*

$$y_J := \sum_{j=0}^{J} \sum_{r\in D_j} \nu_r \Lambda_r.$$

Then (a) implies the convergence in $\mathcal{C}_{\mathbb{B}}$ *of* y_J *to some function* y*. Condition (b) gives the same convergence plus the membership in* $\mathcal{H}^{\rho}_{\mathbb{B}}$ *for* y*. Condition (c) gives the convergence of* y_J *to* y *in* $\mathcal{H}^{\rho,o}_{\mathbb{B}}$*.*

Corollary 1. *For any function* $x : [0,1] \to \mathbb{B}$*, define the* $\mathbb{B}$*-valued array* $\nu = \nu(x) := (\lambda_r(x);\ j \geq 0, r \in D_j)$*. Then* x *coincides at the dyadic points of* $[0,1]$ *with some function* y *which is in* $\mathcal{C}_{\mathbb{B}}$ *under (a), in* $\mathcal{H}^{\rho}_{\mathbb{B}}$ *under (b) and in* $\mathcal{H}^{\rho,o}_{\mathbb{B}}$ *under (c).*

From Corollary 1 and continuity in probability of ξ, it is easily seen that the problem of existence of $\mathcal{H}^{\rho,o}_{\mathbb{B}}$-versions of ξ reduces to the control of the $\lambda_r(\xi)$'s which are dyadic second differences of ξ. It is convenient here to define the second differences of ξ by

$$\Delta^2_h \xi(t) := \xi(t+h) + \xi(t-h) - 2\xi(t), \quad t \in T,\ t \pm h \in T.$$

This leads to the general following result.

Theorem 3. *Let* $\xi = \{\xi(t),\ t \in T\}$ *be a* $\mathbb{B}$*-valued stochastic process, continuous in probability. Assume there exist a function* $\sigma : [0,1] \to \mathbb{R}^+$*,* $\sigma(0) = 0$ *and a function* $\Psi : (0,\infty] \to \mathbb{R}^+$*,* $\Psi(\infty) = 0$ *such that for all real numbers* $z > 0$*,* $t \in T$*,* $t \pm h \in T$*,*

$$P\big(\big\|\Delta^2_h \xi(t)\big\| > z\sigma(|h|)\big) \leq \Psi(z). \tag{13}$$

Put for $0 < u < \infty$*,*

$$R(u) = R(\Psi, \sigma, \rho, u) := \sum_{j=0}^{\infty} 2^{jd} \Psi\Big(u \frac{\rho}{\sigma}(2^{-j})\Big).$$

If $R(u_0)$ *is finite for some* $0 < u_0 < \infty$*, then* ξ *has a version in* $\mathcal{H}^{\rho}_{B}$*. If* $R(u)$ *is finite for every* $0 < u < \infty$*, then* ξ *has a version in* $\mathcal{H}^{\rho,o}_{B}$*.*

When Ψ *is non increasing and* σ *non decreasing, the same conclusions hold replacing* $R(u)$ *by*

$$I(u) := \int_0^1 \Psi\Big(u \frac{\rho}{\sigma}(s)\Big) \frac{\mathrm{d}s}{s^2}.$$

We only give here an application to the case of Gaussian processes. For other applications and examples we refer to [34].

Corollary 2. *Assume that the Gaussian $\mathbb{B}$-valued stochastic process $\xi = \{\xi(t), t \in T\}$ is continuous in probability and satisfies for each $t \in T$, $t \pm h \in T$,*

$$\mathbf{E} \left\| \Delta_h^2 \xi(t) \right\|^2 \leq \sigma^2(|h|). \tag{14}$$

(i) If $\liminf_{j\to\infty} \dfrac{\rho(2^{-j})}{j^{1/2}\sigma(2^{-j})} > 0$, *then ξ admits a version in $\mathcal{H}_B^\rho$.*

(ii) If $\lim_{j\to\infty} \dfrac{\rho(2^{-j})}{j^{1/2}\sigma(2^{-j})} = \infty$, *then ξ has a version in $\mathcal{H}_B^{\rho,o}$.*

Example 1. Let $\mathbb{B}$ be a separable Banach space and Y a centered Gaussian random element in $\mathbb{B}$ with distribution μ. A $\mathbb{B}$-valued Brownian motion with parameter μ is a Gaussian process ξ indexed by $[0,1]$, with independent increments such that $\xi(t) - \xi(s)$ has the same distribution as $|t-s|^{1/2} Y$. Hence (14) holds with $\sigma(h) = h^{1/2} \mathbf{E}^{1/2} \|Y\|_B^2$ $(h \geq 0)$. Choosing the weight function $\rho(h) = \sqrt{h \ln(\mathrm{e}/h)}$, we see that

$$\lim_{j\to\infty} \frac{\rho(2^{-j})}{j^{1/2}\sigma(2^{-j})} = \frac{1}{\mathbf{E}^{1/2} \|Y\|^2} > 0.$$

Hence by Corollary 2 (ii), the $\mathbb{B}$-valued Brownian motion ξ has a version in $\mathcal{H}_{\mathbb{B}}^\rho$. This result is optimal because of Lévy's theorem on the modulus of uniform continuity of the standard Brownian motion.

4.2 Tightness

To deal with convergence in distribution of stochastic processes considered as random elements in $\mathcal{H}_{\mathbb{B}}^{\rho,o}$, a key tool is the following tightness criterion established in Račkauskas and Suquet [40].

Theorem 4. *Suppose the Banach space $\mathbb{B}$ is separable. Then the sequence $(\xi_n)_{n\geq 1}$ of random elements in $\mathcal{H}_{\mathbb{B}}^{\rho,o}$ is tight if and only if it satisfies the two following conditions.*

(i) For each dyadic *$t \in [0,1]$, the sequence of $\mathbb{B}$-valued random variables $(\xi_n(t))_{n\geq 1}$ is tight on $\mathbb{B}$.*

(ii) For each positive ε,

$$\lim_{J\to\infty} \sup_{n\geq 1} \mathbf{P}\left(\sup_{j>J} \frac{1}{\rho(2^{-j})} \max_{r\in \mathrm{D}_j} \|\lambda_r(\xi_n)\| > \varepsilon \right) = 0.$$

4.3 Partial sums processes

Let $(X_k)_{k\geq 1}$ be a sequence of i.i.d. random elements in the separable Banach space $\mathbb{B}$. Set $S_0 := 0$, $S_k := X_1 + \cdots + X_k$, for $k = 1, 2, \ldots$ and consider the partial sums processes

$$\xi_n(t) = S_{[nt]} + (nt - [nt])X_{[nt]+1}, \quad t \in [0,1].$$

When $\mathbb{B} = \mathbb{R}$, Donsker-Prohorov invariance principle states, that if $\mathbf{E}\, X_1 = 0$ and $\mathbf{E}\, X_1^2 = \sigma^2 < \infty$, then

$$n^{-1/2}\sigma^{-1}\xi_n \xrightarrow{\mathcal{D}} W, \tag{15}$$

in $\mathcal{C}[0,1]$, where $\{W(t), t \in \mathbb{R}\}$ is a standard Wiener process. The necessity of $\mathbf{E}\, X_1^2 < \infty$ is clear here, since (15) implies the CLT for $n^{-1/2}S_n = n^{-1/2}\xi_n(1)$.

Lamperti [22] was the first who considered the convergence (15) with respect to some Hölderian topology. He proved that if $0 < a < 1/2$ and $\mathbf{E}\,|X_1|^p < \infty$, where $p > p(a) := 1/(1/2 - a)$, then (15) takes place in $\mathcal{H}^{a,o}$. This result was derived again by Kerkyacharian and Roynette [19] by another method based on Ciesielski [9] analysis of Hölder spaces by triangular functions. Further generalizations were given by Erickson [13] (partial sums processes indexed by $[0,1]^d$), Hamadouche [15] (weakly dependent sequence (X_n)), Račkauskas and Suquet [37] (Banach space valued X_i's and Hölder spaces built on the weight $\rho(h) = h^a \ln^b(1/h)$). The following result is proved in Račkauskas and Suquet [38].

Theorem 5. *Let $0 < a < 1/2$ and $p(a) = 1/(1/2 - a)$. Then*

$$n^{-1/2}\sigma^{-1}\xi_n \xrightarrow[n\to\infty]{\mathcal{D}} W \quad \textit{in the space } \mathcal{H}^{a,o}$$

if and only if $\mathbf{E}\, X_1 = 0$ and

$$\lim_{t\to\infty} t^{p(a)}\mathbf{P}(|X_1| \geq t) = 0. \tag{16}$$

Condition (16) yields the existence of moments $\mathbf{E}\,|X_1|^p$ for any $0 \leq p < p(a)$. If a approaches $1/2$ then $p(a) \to \infty$. Hence, stronger invariance principle requires higher moments.

The description of more general results requires some background on Gaussian random elements and central limit theorem in Banach spaces. Let $\mathbb{B}'$ be the topological dual of $\mathbb{B}$. For a random element X in $\mathbb{B}$ such that for every $f \in \mathbb{B}'$, $\mathbf{E}\, f(X) = 0$ and $\mathbf{E}\, f^2(X) < \infty$, the covariance operator $Q = Q(X)$ is the linear bounded operator from $\mathbb{B}'$ to $\mathbb{B}$ defined by $Qf = \mathbf{E}\, f(X)X,\ f \in \mathbb{B}'$. A random element $X \in \mathbb{B}$ (or covariance operator Q) is said to be *pregaussian* if there exists a mean zero Gaussian random element $Y \in \mathbb{B}$ with the same covariance operator as X, i.e. for all $f, g \in \mathbb{B}'$, $\mathbf{E}\, f(X)g(X) = \mathbf{E}\, f(Y)g(Y)$. Since the distribution of a centered Gaussian random element is defined by its

covariance structure, we denote by Y_Q a zero mean Gaussian random element with covariance operator Q.

For any pregaussian covariance Q there exists a $\mathbb{B}$-valued Brownian motion W_Q with parameter Q, a centered Gaussian process indexed by $[0,1]$ with independent increments such that $W_Q(t) - W_Q(s)$ has the same distribution as $|t-s|^{1/2} Y_Q$.

We say that X_1 satisfies the *central limit theorem in* $\mathbb{B}$, which we denote by $X_1 \in \mathrm{CLT}(\mathbb{B})$, if $n^{-1/2} S_n$ converges in distribution in $\mathbb{B}$. This implies that $\mathbf{E}\, X_1 = 0$ and X_1 is pregaussian. It is well known, e.g. Ledoux and Talagrand [24] that the central limit theorem for X_1 cannot be characterized in general in terms of integrability of X_1 and involves the geometry of the Banach space $\mathbb{B}$.

We say that X_1 satisfies the *functional central limit theorem in* $\mathbb{B}$, which we denote by $X_1 \in \mathrm{FCLT}(\mathbb{B})$, if $n^{-1/2}\xi_n$ converges in distribution in $\mathcal{C}_{\mathbb{B}}$. Kuelbs [21] extended the classical Donsker-Prohorov invariance principle to the case of $\mathbb{B}$-valued partial sums by proving that $n^{-1/2}\xi_n$ converges in distribution in $\mathcal{C}_{\mathbb{B}}$ to some Brownian motion W if and only if $X_1 \in \mathrm{CLT}(\mathbb{B})$ (in short $X_1 \in \mathrm{CLT}(\mathbb{B})$ if and only if $X_1 \in \mathrm{FCLT}(\mathbb{B})$). Of course in Kuelbs theorem, the parameter Q of W is the covariance operator of X_1.

The convergence in distribution of $n^{-1/2}\xi_n$ in $\mathcal{H}_{\mathbb{B}}^{\rho,o}$, which we denote by $X_1 \in \mathrm{FCLT}(\mathbb{B}, \rho)$, is clearly stronger than $X_1 \in \mathrm{FCLT}(\mathbb{B})$.

An obvious preliminary requirement for the FCLT in $\mathcal{H}_{\mathbb{B}}^{\rho,o}$ is that the $\mathbb{B}$-valued Brownian motion has a version in $\mathcal{H}_{\mathbb{B}}^{\rho,o}$. From this point of view, the critical ρ is $\rho_c(h) = \sqrt{h \ln(\mathrm{e}/h)}$ due to Lévy's Theorem on the modulus of uniform continuity of the Brownian motion. So our interest will be restricted to functions ρ generating a weaker Hölder topology than ρ_c. More precisely, we consider the following class $\mathcal{R}$ of functions ρ.

Definition 2. *We denote by $\mathcal{R}$ the class of functions ρ satisfying*

i) for some $0 < a \leq 1/2$, and some function L which is normalized slowly varying at infinity,

$$\rho(h) = h^a L(1/h), \quad 0 < h \leq 1, \tag{17}$$

ii) $\theta(t) = t^{1/2}\rho(1/t)$ is C^1 on $[1, \infty)$,
iii) for some $b > 1/2$ and some $a > 0$, $\theta(t) \ln^{-b}(t)$ is non decreasing on $[a, \infty)$.

The following result is proved in Račkauskas and Suquet [36].

Theorem 6. *Let $\rho \in \mathcal{R}$. Then $X_1 \in \mathrm{FCLT}(\mathbb{B}, \rho)$ if and only if $X_1 \in \mathrm{CLT}(\mathbb{B})$ and for every $A > 0$,*

$$\lim_{t \to \infty} t\mathbf{P}\big(\|X_1\| \geq A t^{1/2} \rho(1/t)\big) = 0. \tag{18}$$

If $\rho \in \mathcal{R}$ with $a < 1/2$ in (17) then it suffices to check (18) for $A = 1$ only. Of course the special case $\mathbb{B} = \mathbb{R}$ and $\rho(h) = h^a$ gives back Theorem 5.

In the case where $\rho(h) = h^{1/2}\ln^b(c/h)$ with $b > 1/2$, Condition (18) is equivalent to $\mathbf{E}\exp\left(\gamma\|X_1\|^{1/b}\right) < \infty$, for each $\gamma > 0$. Let us note, that for the spaces $\mathbb{B} = L_p(0,1)$, $2 \le p < \infty$, as well as for each finite dimensional space, Condition (18) yields $X_1 \in \mathrm{CLT}(\mathbb{B})$. On the other hand it is well known that for some Banach spaces existence of moments of any order does not guarantee central limit theorem. It is also worth noticing that like in Kuelbs FCLT, all the influence of the geometry of the Banach space $\mathbb{B}$ is absorbed by the condition $X_1 \in \mathrm{CLT}(\mathbb{B})$.

It would be useful to extend the Hölderian FCLT to the case of dependent X_i's. A first step was done by Hamadouche [15] in the special case where $\mathbb{B} = \mathbb{R}$ and under weak dependence (association and α-mixing). The result presented in Račkauskas and Suquet [37] provides a very general approach for $\mathbb{B}$-valued X_i's and any dependence structure, subject to obtaining a good estimate of the partial sums. Laukaitis and Račkauskas [23] obtained Hölderian FCLT for a polygonal line process based on residual partial sums of a stationary Hilbert space valued autoregression (ARH(1)) and applied it to the problem of testing stability of ARH(1) model under different types of alternatives.

4.4 Adaptive self-normalized partial sums processes

In order to relax moment assumptions like (18) in the FCLT for i.i.d. mean zero random variables X_i, Račkauskas and Suquet [33] consider the so called *adaptive self-normalized* partial sums processes. *Self-normalized* means that the classical normalization by $\sqrt{n}$ is replaced by

$$V_n = (X_1^2 + \cdots + X_n^2)^{1/2}.$$

Adaptive means that the vertices of the corresponding random polygonal line have their abscissas at the random points V_k^2/V_n^2 $(0 \le k \le n)$ instead of the deterministic equispaced points k/n. By this construction the slope of each line adapts itself to the value of the corresponding random variable.

By ζ_n (respect. ξ_n) we denote the random polygonal partial sums process defined on $[0,1]$ by linear interpolation between the vertices $(V_k^2/V_n^2, S_k)$, $k = 0, 1, \ldots, n$ (respect. $(k/n, S_k)$, $k = 0, 1, \ldots, n$). For the special case $k = 0$, we put $S_0 = 0, V_0 = 0$. By convention the random functions $V_n^{-1}\xi_n$ and $V_n^{-1}\zeta_n$ are defined to be the null function on the event $\{V_n = 0\}$. Figure 2 displays the polygonal lines $n^{-1/2}\xi_n$ and $V_n^{-1}\zeta_n$ built on a simulated sample of size $n = 800$ of the symmetric distribution given by $\mathbf{P}(|X_1| > t) = 0.5\,t^{-2.2}\mathbf{1}_{[1,\infty)}(t)$. For these simulated paths we have $\left\|n^{-1/2}\xi_n\right\|_{0.49} \simeq 24.75$, while $\left\|V_n^{-1}\zeta_n\right\|_{0.49} \simeq 3.05$. This picture shows how adaptive partition of the time interval improves slopes of polygonal line process.

Membership of X_1 in the domain of attraction of the normal distribution (DAN) means that there exists a sequence $b_n \uparrow \infty$ such that

$$b_n^{-1}S_n \xrightarrow{\mathcal{D}} N(0,1).$$

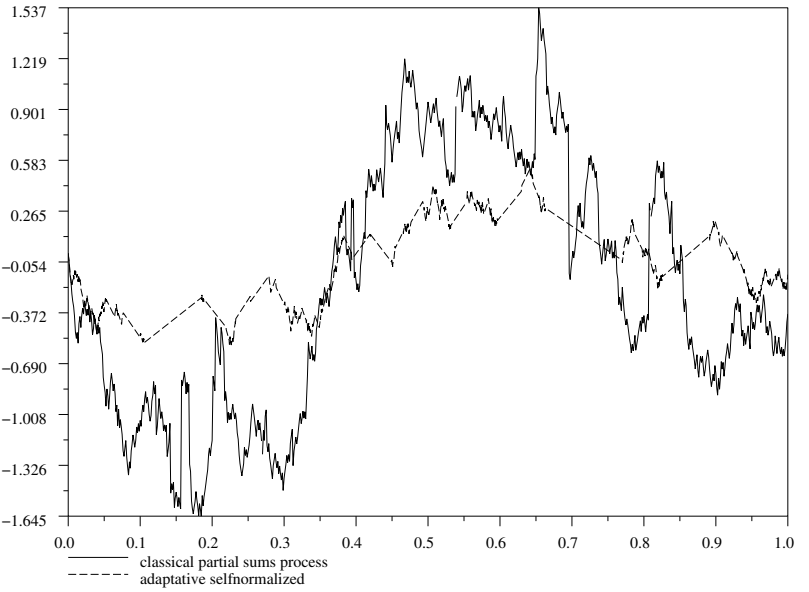


Fig. 2. Partial sums processes $n^{-1/2}\xi_n$ and $V_n^{-1}\zeta_n$

The following result is proved in Račkauskas and Suquet [33].

Theorem 7. *Assume that ρ satisfies Conditions (8) to (12) and*

$$\lim_{j\to\infty} \frac{2^j \rho^2(2^{-j})}{j} = \infty. \tag{19}$$

If X_1 is symmetric then

$$V_n^{-1}\zeta_n \xrightarrow{\mathcal{D}} W, \quad in \quad \mathcal{H}^{\rho,o}$$

if and only if $X_1 \in DAN$.

When tested with $\rho(h) = h^{1/2}\ln^b(c/h)$, (19) reduces to $j^{2b-1} \to \infty$. Due to the inclusions of Hölder spaces, this shows that Theorem 7 gives the best result possible in the scale of the separable Hölder spaces $\mathcal{H}^{a,b,o}$. Moreover, no high order moments are needed except the condition $X_1 \in DAN$ which due to well known O'Briens result is equivalent to

$$V_n^{-1} \max_{1\le k\le n} |X_k| \xrightarrow{P} 0.$$

It seems worth noticing here, that without adaptive construction of the polygonal process, the existence of moments of order bigger than 2 is necessary for

Hölder weak convergence. Indeed, if the process $V_n^{-1}\xi_n$ converges weakly to W in $\mathcal{H}^{a,o}$ for some $a > 0$, then its maximal slope $n^{-1/2}V_n^{-1}\max_{1\le k\le n}|X_k|$ converges to zero in probability. This on its turn yields $V_n^{-1}\max_{1\le k\le n}|X_k| \to 0$ almost surely, and according to Maller and Resnick (1984), $\mathbf{E}\,X_1^2 < \infty$. Hence $n^{-1}V_n^2$ converges almost surely to $\mathbf{E}\,X_1^2$ by the strong law of large numbers. Therefore $n^{-1/2}\xi_n$ converges weakly to W in $\mathcal{H}^{a,o}$ and by Theorem 5 the moment restriction (16) is necessary.

Naturally it is very desirable to remove the symmetry assumption in Theorem 7. Although the problem remains open, we can propose the following partial result in this direction (for more on this problem see Račkauskas and Suquet [33]).

Theorem 8. *If for some $\varepsilon > 0$, $\mathbf{E}\,|X_1|^{2+\varepsilon} < \infty$, then for any $b > 1/2$, $V_n^{-1}\zeta_n$ converges weakly to W in the space $\mathcal{H}^{1/2,b,o}$.*

Some extensions of this result for the non i.i.d. case are given in Račkauskas and Suquet [39].

4.5 Empirical processes

In asymptotic statistics, the empirical distribution function F_n of an i.i.d. sample $X_1, \ldots, X_n$ plays a central rôle. When the distribution function F of the X_i's is continuous, the transformation $U_i := F(X_i)$ reduces the study of the asymptotical behavior of F_n to the case of uniform $[0,1]$ distributed random variables U_i. The corresponding uniform empirical process is defined by

$$\xi_n(t) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(\mathbf{1}_{\{U_i\le t\}} - t\big), \quad t \in [0,1].$$

It is well known that ξ_n converges weakly in the Skorokhod space to the Brownian bridge B. From this convergence follows the weak $\mathcal{C}[0,1]$ convergence to B of ξ_n^{pg}, the polygonal smoothing of ξ_n. This polygonal smoothing is simply the empirical process associated to the polygonal cumulative distribution function. More precisely, let us denote by $U_{n:i}$ the order statistics of the sample $U_1, \ldots, U_n$

$$0 = U_{n:0} \le U_{n:1} \le \cdots \le U_{n:n} \le U_{n:n+1} = 1.$$

We define $\xi_n^{\mathrm{pg}} = (\xi_n^{\mathrm{pg}}(t),\, t \in [0,1])$ as the random polygonal line with vertices $(U_{n:k}, \xi_n(U_{n:k}))$, $k = 0, 1, \ldots, n+1$.

Investigating the weak Hölder convergence of ξ_n^{pg}, Hamadouche [14] proved the following result.

Theorem 9. *The sequence $(\xi_n^{\mathrm{pg}})_{n\ge1}$ converges weakly to the Brownian bridge B in every $\mathcal{H}^{a,o}$ for $0 < a < 1/4$. Moreover $(\xi_n^{\mathrm{pg}})_{n\ge1}$ is not tight in $\mathcal{H}^{a,o}$ for $a \ge 1/4$.*

At first sight this result looks somewhat surprising because the limiting process B has a version in every $\mathcal{H}^{a,o}$ for $a < 1/2$ and the paths of ξ_n^{pg} are Lipschitz functions. It illustrates the fact that polygonal smoothing of the empirical distribution function is in some sense too violent. With a convolution smoothing it is possible to achieve the convergence in any $\mathcal{H}^{a,o}$ for $a < 1/2$ (see [16] and the references therein).

Another important stochastic process connected to the empirical distribution function is the so-called uniform quantile process. Put for notational convenience

$$u_{n:i} = \mathbf{E}\, U_{n:i} = \frac{i}{n+1}, \quad i = 0, 1, \ldots, n+1.$$

The (discontinuous) uniform quantile process χ_n is given by

$$\chi_n(t) := \sqrt{n}\Big(\sum_{i=1}^{n+1} U_{n:i} \mathbf{1}_{]u_{n:i-1}, u_{n:i}]}(t) - t \Big), \quad t \in [0,1]. \tag{20}$$

We associate to χ_n the polygonal uniform quantile process χ_n^{pg} which is affine on each $[u_{n:i-1}, u_{n:i}]$, $i = 1, \ldots, n+1$ and such that

$$\chi_n^{\mathrm{pg}}(u_{n:i}) = \sqrt{n}(U_{n:i} - u_{n:i}), \quad i = 0, 1, \ldots, n+1. \tag{21}$$

Using the Hölderian FCLT (Theorem 6), Hamadouche and Suquet [17] obtained the following optimal result.

Theorem 10. *Let $\rho(h) = h^{1/2} L(1/h)$ be a weight function in the class $\mathcal{R}$. Then χ_n^{pg} converges weakly in $\mathcal{H}^{\rho,o}$ to the Brownian bridge if and only if*

$$\lim_{t \to \infty} \frac{L(t)}{\ln t} = \infty. \tag{22}$$

A third process related to empirical process is the empirical characteristic function $\mathfrak{c}_n$. Functional limit theorems for $\mathfrak{c}_n$ in Hölderian framework are investigated in [35] in the multivariate case. For simplicity we shall describe the results in the univariate case only. Let X be a real valued random variable and $(X_k)_{k \geq 1}$ a sequence of independent copies of X. Define respectively the empirical characteristic function $\mathfrak{c}_n$ and the characteristic function $\mathfrak{c}$ by

$$\mathfrak{c}_n(t) := n^{-1} \sum_{k=1}^{n} \exp(itX_k), \quad \mathfrak{c}(t) := \mathbf{E} \exp(itX), \quad t \in \mathbb{R}.$$

Here the paths of $\mathfrak{c}_n$ are smooth enough to allow membership in any $\mathcal{H}^{\rho,o}$, so we do not need any smoothing. Clearly $\mathfrak{c}_n$ appears as the sum of i.i.d. random elements in $\mathcal{H}^{\rho,o}$, so that the almost sure convergence in $\mathcal{H}^{\rho,o}$ of $\mathfrak{c}_n$ reduces to some strong law of large numbers in $\mathcal{H}^{\rho,o}$, while the weak $\mathcal{H}^{\rho,o}$ convergence of $n^{1/2}(\mathfrak{c}_n - \mathfrak{c})$ is just a central limit theorem for the random element $\xi : t \mapsto \exp(itX)$. The Hölder functions considered here (as elements of the spaces $\mathcal{H}^{\rho,o}$) can be defined on any compact interval of $\mathbb{R}$, but we shall keep $T = [0,1]$ for simplicity.

Theorem 11. *Assume that the weight function ρ belongs to the class $\mathcal{R}$. Then the convergence*

$$\sup_{\substack{t,s\in T,\\ s\neq t}} \frac{|\mathfrak{c}_n(t)-\mathfrak{c}(t)|}{\rho(|t-s|)} \xrightarrow[n\to\infty]{\text{a.s.}} 0 \tag{23}$$

holds if and only if

$$\mathbf{E}\,\rho^*(|X|)<\infty, \tag{24}$$

where

$$\rho^*(h) := \frac{1}{\rho\big(\min(1;1/h)\big)}, \quad 0<h<\infty. \tag{25}$$

In the special case where $\rho(h)=h^a$ for some $0<a<1$, Condition (24) writes $\mathbf{E}\,|X|^a<\infty$.

We refer to [35] for a discussion of the rate of convergence in (23), based on some Marcinkiewicz-Zygmund strong law of large numbers in $\mathcal{H}^{\rho,o}$.

Now consider the empirical characteristic process

$$Y_n(t)=\sqrt{n}\big(\mathfrak{c}_n(t)-\mathfrak{c}(t)\big), \quad t\in T.$$

By the multidimensional CLT, the finite dimensional distributions of (Y_n) converge to those of a complex Gaussian process Y with zero mean and covariance $\mathbf{E}\,Y(t)\overline{Y(s)}=\mathfrak{c}(t-s)-\mathfrak{c}(t)\mathfrak{c}(-s)$, $s,t\in T$.

Theorem 12. *If the distribution of X satisfies*

$$\sum_{j=1}^{\infty}\frac{\sqrt{j}}{\rho(2^{-j})}\mathbf{E}^{\,1/2}|\sin(2^{-j}X)|^4<\infty, \tag{26}$$

then (Y_n) converges in distribution to Y in the space $\mathcal{H}^{\rho,o}$.

Roughly speaking, Condition (26) may be interpreted as the square integrability of the random element $\xi: t\mapsto \exp(itX)$ in a norm a bit stronger than $\|\xi\|_\rho$, see [35]. This is not surprising because the bad geometric properties of $\mathcal{H}^{\rho,o}$ do not allow to deduce the CLT for ξ from the square integrability of $\|\xi\|_\rho$.

4.6 Detection of epidemic changes

Hölderian invariance principles like Theorems 6 and 8 have statistical applications to detection of epidemic change. This question is investigated in [41] for the case of real valued observations and in [42] for the case of Banach space valued (in fact functional) observations. We just sketch here the method.

The epidemic model may be described as follows. Having observed a sample $X_1, X_2, \ldots, X_n$ of random variables, we want to test the standard null hypothesis of constant mean

(H_0): *$X_1, \ldots, X_n$ all have the same mean denoted by μ_0,*

against the epidemic alternative

(H_A): *there are integers $1 < k^* < m^* < n$ and a constant $\mu_1 \neq \mu_0$ such that $\mathbf{E}\, X_i = \mu_0 + (\mu_1 - \mu_0)\mathbf{1}_{\{k^* < i \leq m^*\}}$, $i = 1, 2, \ldots, n$.*

To simplify notation put

$$\varrho(h) := \rho\big(h(1-h)\big), \quad 0 \leq h \leq 1.$$

For $\rho \in \mathcal{R}$, define with $t_k := k/n$, $0 \leq k \leq n$, $S(t) := \sum_{1 \leq k \leq t} X_k$,

$$\mathrm{UI}(n, \rho) = \max_{1 \leq i < j \leq n} \frac{\big|S(j) - S(i) - S(n)(t_j - t_i)\big|}{\varrho(t_j - t_i)}$$

$$\mathrm{DI}(n, \rho) = \max_{1 \leq j \leq \log n} \frac{1}{\rho(2^{-j})} \max_{r \in \mathrm{D}_j} \Big| S(nr) - \frac{1}{2}\big(S(nr^+) + S(nr^-)\big)\Big|.$$

These test statistics may be viewed as some discrete Hölder norms of the partial sums process built on the X_k's. Their relevance will be clear from the next result. In what follows, we naturally assume that the numbers of observations k^*, $m^* - k^*$, $n - m^*$ before, during and after the epidemic go to infinity with n. Write $l^* := m^* - k^*$ for the length of the epidemic.

Theorem 13. *Let $\rho \in \mathcal{R}$. Assume under (H_A) that the X_i's are independent and $\sigma_0^2 := \sup_{k \geq 1} \operatorname{Var} X_k$ is finite. If*

$$\lim_{n \to \infty} n^{1/2} \frac{h_n}{\rho(h_n)} = \infty, \quad \textit{where} \quad h_n := \frac{l^*}{n}\Big(1 - \frac{l^*}{n}\Big), \tag{27}$$

then

$$n^{-1/2}\mathrm{UI}(n, \rho) \xrightarrow[n \to \infty]{\mathrm{P}} \infty, \quad \textit{and} \quad n^{-1/2}\mathrm{DI}(n, \rho) \xrightarrow[n \to \infty]{\mathrm{P}} \infty.$$

When $\rho(h) = h^a$, (27) leads to detect *short epidemics* such that $l^* = o(n)$ and $l^* n^{-\delta} \to \infty$, where $\delta = (1 - 2a)(2 - 2a)^{-1}$. Symmetrically one can detect *long epidemics* such that $n - l^* = o(n)$ and $(n - l^*)n^{-\delta} \to \infty$. When $\rho(h) = h^{1/2} \ln^b(c/h)$ with $b > 1/2$, (27) is satisfied provided that $h_n = n^{-1} \ln^\gamma n$, with $\gamma > 2b$. This leads to detect short epidemics such that $l^* = o(n)$ and $l^* \ln^{-\gamma} n \to \infty$ as well as of long ones verifying $n - l^* = o(n)$ and $(n - l^*) \ln^{-\gamma} n \to \infty$.

Let $W = \{W(t), t \in [0,1]\}$ be a standard Wiener process and $B = \{B(t), t \in [0,1]\}$ the corresponding Brownian bridge $B(t) = W(t) - tW(1)$, $t \in [0,1]$. Consider for ρ in $\mathcal{R}$, the following random variables

$$\mathrm{UI}(\rho) := \sup_{0 < t - s < 1} \frac{|B(t) - B(s)|}{\varrho(t - s)} \tag{28}$$

and

$$\mathrm{DI}(\rho) = \sup_{j\geq 1} \frac{1}{\rho(2^{-j})} \max_{r\in \mathrm{D}_j} \Big| W(r) - \frac{1}{2}W(r^+) - \frac{1}{2}W(r^-) \Big| = \|B\|_\rho^{\mathrm{seq}}. \qquad (29)$$

These variables serve as limiting for uniform increment (UI) and dyadic increment (DI) statistics respectively. No analytical form seems to be known for the distribution function of $\mathrm{UI}(\rho)$, whereas the distribution of $\mathrm{DI}(\rho)$ is completely specified in terms of the *error function* $\operatorname{erf} x = 2\pi^{-1/2} \int_0^x \exp(-s^2)\mathrm{d}s$.

Theorem 14. *Let* $c = \limsup_{j\to\infty} j^{1/2}/\theta(2^j)$, *where* $\theta(t) = t^{1/2}\rho(1/t)$.

i) If $c = \infty$ *then* $\mathrm{DI}(\rho) = \infty$ *almost surely.*
ii) If $0 \leq c < \infty$, *then* $\mathrm{DI}(\rho)$ *is almost surely finite and its distribution function is given by*

$$\mathbf{P}\big(\mathrm{DI}(\rho) \leq x\big) = \prod_{j=1}^{\infty} \big\{ \operatorname{erf}\big(\theta(2^j)x\big) \big\}^{2^{j-1}}, \quad x > 0. \qquad (30)$$

The distribution function of $\mathrm{DI}(\rho)$ *is continuous with support* $[c\sqrt{\ln 2}, \infty)$.

The infinite product in (30) converges very fast and in practice one need to compute only four or five factors.

Theorem 15. *Under* (H_0) *with i.i.d* X_k*'s, assume that* $\rho \in \mathcal{R}$ *and for every* $A > 0$,

$$\lim_{t\to\infty} t\,\mathbf{P}\big(|X_1| > At^{1/2}\rho(1/t)\big) = 0.$$

Then

$$\sigma^{-1}n^{-1/2}\mathrm{UI}(n,\rho) \xrightarrow[n\to\infty]{\mathcal{D}} \mathrm{UI}(\rho) \quad \textit{and} \quad \sigma^{-1}n^{-1/2}\mathrm{DI}(n,\rho) \xrightarrow[n\to\infty]{\mathcal{D}} \mathrm{DI}(\rho),$$

where $\sigma^2 = \operatorname{Var} X_1$ *and* $\mathrm{UI}(\rho)$, $\mathrm{DI}(\rho)$ *are defined by (28) and (29).*

Of course when the variance σ^2 is unknown the results remain valid if σ^2 is substituted by its standard estimator $\widehat{\sigma}^2$.

References

1. Ayache A (2001) Du mouvement brownien fractionnaire au mouvement brownien multifractionnaire. Technique et Science Informatiques. 20 9:1133–1152
2. Ayache A, Lévy Véhel J (1999) Generalized multifractional Brownian motion: definition and preliminary results. In Dekking M, Lévy Véhel J, Lutton E, Tricot C (eds) Fractals: Theory and Applications in Engineering. Springer 17–32
3. Ayache A, Lévy Véhel J (2000) The Generalized multifractional Brownian motion. Statistical Inference for Stochastic Processes 3:7–18

4. Ayache A, Lévy Véhel J (2004) Identification of the pointwise Hölder exponent of Generalized Multifractional Brownian Motion. Stochastic Processes and their Applications 111:119–156
5. Ayache A, Léger S (2000) The multifractional Brownian sheet. To appear in Ann Mat Blaise Pascal
6. Ayache A, Taqqu MS (2003) Multifractional Processes with Random Exponent. Preprint to appear in Publicacions Matemàtiques
7. Benassi A, Jaffard S, Roux D (1997) Elliptic Gaussian random processes. Rev Mat Iberoamericana 13 1:19–89
8. Barral J, Lévy Véhel J (2004) Multifractal analysis of a class of additive processes with correlated nonstationary increments. Electronic Journal of Probability 9:508–543
9. Ciesielski Z (1960) On the isomorphisms of the spaces H_α and m. Bull Acad Pol Sci Ser Sci Math Phys 8:217–222
10. Davies S, Hall P (1999) Fractal analysis of surface roughness by using spatial data. J R Statist B 61 1:3–37
11. Doukhan P, Oppenheim G, Taqqu MS, (eds) (2002) Theory and Applications of Long-range Dependence. Birkhäuser, Boston
12. Falconer K (1990) Fractal Geometry: Mathematical Fondations and Applications. John Wiley and Sons, New York
13. Erickson RV (1981) Lipschitz smoothness and convergence with applications to the central limit theorem for summation processes. Annals of Probability 9:831–851
14. Hamadouche D (1998) Weak convergence of smoothed empirical process in Hölder spaces. Stat Probab Letters 36:393–400
15. Hamadouche D (2000) Invariance principles in Hölder spaces. Portugaliae Mathematica 57:127–151
16. Hamadouche D, Suquet Ch (1999) Weak Hölder convergence of stochastic processes with application to the perturbed empirical process. Applicationes Math 26:63–83
17. Hamadouche D, Suquet Ch (2004) Smoothed quantile processes in Hölder spaces. Pub IRMA Lille (Preprint) 62 IV
18. Jaffard S, Meyer Y, Ryan RD (2001) Wavelets Tools for Science & Technology. SIAM Philadelphia.
19. Kerkyacharian G, Roynette B (1991) Une démonstration simple des théorèmes de Kolmogorov, Donsker et Ito-Nisio. Comptes Rendus de l'Académie des Sciences Paris, Série I 312:877–882
20. Kolmogorov AN (1940) Wienersche Spiralen und einige andere interessante Kurven im Hilbertschen raum. Comptes Rendus (Doklady) de l'Académie des Sciences de l' URSS (NS) 26:115–118
21. Kuelbs J (1973) The invariance principle for Banach space valued random variables. Journal of Multivariate Analysis 3:161–172
22. Lamperti J (1962) On convergence of stochastic processes. Transactions of the American Mathematical Society 104:430–435
23. Laukaitis A, Račkauskas A (2002) Testing changes in Hilbert space autoregressive models. Lietuvos Matematikos Rinkinys 42:434–447
24. Ledoux M, Talagrand M (1991) Probability in Banach Spaces. Springer-Verlag, Berlin Heidelberg.
25. Lévy P (1937) Théorie de l'addition des variables aléatoires. Gauthier-Villars, Paris, Second Edition (1954)

26. Lévy Véhel J (1998) Introduction to the Multifractal Analysis of Images. In Fisher Y (ed) Fractal Image Encoding and Analysis. Springer
27. Lévy Véhel J (2002) Signal enhancement based on Hölder regularity analysis. IMA Volumes in Mathematics and its Applications 132:197–209
28. Peltier RF, Lévy Véhel J (1995) Multifractional Brownian Motion: definition and preliminary results. Rapport de recherche de l'INRIA, No 2645
29. Lifshits MA (1995) Gaussian Random Functions. Kluwer Academic Publishers, Dordrecht Boston London
30. Mandelbrot BB, Van Ness JW (1968) Fractional Brownian motions, fractional noises and applications. SIAM Review 10:422–437
31. Račkauskas A, Suquet Ch (1999) Central limit theorem in Hölder spaces. Probability and Mathematical Statistics 19:155–174
32. Račkauskas A, Suquet Ch (1999) Random fields and central limit theorem in some generalized Hölder spaces. In: Grigelionis B et al (eds) Prob Theory and Math Statist. Proceedings of the 7th Vilnius Conference (1998) TEV Vilnius and VSP Utrecht 599–616
33. Račkauskas A, Suquet Ch (2001) Invariance principles for adaptive self-normalized partial sums processes. Stochastic Processes and their Applications 95:63–81
34. Račkauskas A, Suquet Ch (2001) Hölder versions of Banach spaces valued random fields. Georgian Mathematical Journal 8 2:347–362
35. Račkauskas A, Suquet Ch (2002) Hölder convergences of multivariate empirical characteristic functions. Mathematical Methods of Statistics vol. 11:3:341–357
36. Račkauskas A, Suquet Ch (2004) Necessary and sufficient condition for the Hölderian functional central limit theorem. Journal of Theoretical Probability 17 1:221–243
37. Račkauskas A, Suquet Ch (2002) On the Hölderian functional central limit theorem for i.i.d. random elements in Banach space. In: Berkes I, Csáki E, Csörgő M (eds) Limit Theorems in Probability and Statistics. Balatonlelle (1999) (János Bolyai Mathematical Society, Budapest 2:485–498
38. Račkauskas A, Suquet Ch (2003) Necessary and sufficient condition for the Lamperti invariance principle. Theory of Probability and Mathematical Statistics 68:115–124
39. Račkauskas A, Suquet Ch (2003) Invariance principle under self-normalization for nonidentically distributed random variables. Acta Applicandae Mathematicae 79 1-2:83–103
40. Račkauskas A, Suquet Ch (2004) Central limit theorems in Hölder topologies for Banach space valued random fields. Theory of Probability and its Applications 49 1:109–125
41. Račkauskas A, Suquet Ch (2004) Hölder norm test statistics for epidemic change. Journal of Statistical Planning and Inference 126 2:495–520
42. Račkauskas A, Suquet Ch (2003) Testing epidemic change of infinite dimensional parameters. Preprint to appear in Statistical Inference for Stochastic Processes
43. Rosinski J, Samorodnitsky G (1996) Symmetrization and concentration inequalities for multilinear forms with applications to zero-one laws for Lévy chaos. Ann Probab 24 1:422–437
44. Samorodnitsky G, Taqqu MS (1994) Stable non-Gaussian random processes. Chapman & Hall.
45. Tricot C (1993) Curves and Fractal dimensions. Springer-Verlag.

Fractal Stationary Density in Coupled Maps

Jürgen Jost and Kiran M. Kolwankar

Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22-26, D-04103 Leipzig, Germany jjost@mis.mpg.de, Kiran.Kolwankar@mis.mpg.de

Summary. We study the invariant measure or the stationary density of a coupled discrete dynamical system as a function of the coupling parameter ϵ ($0 < \epsilon < 1/4$). The dynamical system considered is chaotic and unsynchronized for this range of parameter values. We find that the stationary density, restricted on the synchronization manifold, is a fractal function. We find the lower bound on the fractal dimension of the graph of this function and show that it changes continuously with the coupling parameter.

1 Introduction

Two or more coupled rhythms can under certain conditions synchronize, that is a definite relationship can develop between these rhythms. This phenomena of synchronization [1, 2] in coupled dynamical systems has acquired immense importance in recent years since it appears in natural phenomena as well as in engineering applications. What is even more interesting is the fact that even chaotic oscillations can synchronize. This observation is being utilized in secure communication [3]. Studying synchronization is also important for neural information processing [4–6].

These developments have lead to the investigation of coupled dynamical systems. The dynamical systems considered can either be continuous or discrete in time. Different types of couplings have been considered, like continuously coupled or pulsed coupled. And many coupling topologies arise in nature as well as in human applications, like random, all-to-all, scale-free, small-world, nearest neighbour lattice, etc. Also the coupling strengths can vary from element to element. We plan to study a system of two coupled maps which forms a basic building block of all these systems and allows us to separate the complexity due to coupling topologies from that coming from the chaotic nature of the dynamical system considered. As a next step, one can then consider various coupling topologies. We have demonstrated in [7] that the result of such a system of two coupled maps can be used to derive

the result for a globally coupled network of N maps. Here we are concerned with the phenomena of complete synchronization, that is, the dynamics of two systems becomes completely identical after the coupling parameter crosses a certain critical value.

The model

We consider the following coupled map system

$$X_{n+1} = AF(X_n) := S(X_n) \tag{1}$$

where $X = (x, y)^T$ is a 2-dim column vector, A is a 2×2 coupling matrix and F is a map from $\Omega = [0, 1] \times [0, 1]$ onto itself. In the present paper we take F to be the extension of the tent map $f_t : [0, 1] \to [0, 1]$,

$$f_t(x) = \begin{cases} 2x & 0 \le x \le 1/2 \\ 2 - 2x & 1/2 \le x \le 1 \end{cases}, \tag{2}$$

to two variables and we choose

$$A = \begin{pmatrix} 1 - \epsilon & \epsilon \\ \epsilon & 1 - \epsilon \end{pmatrix} \tag{3}$$

where $0 < \epsilon < 1$ is the coupling strength. This type of coupling has been called contractive or dissipative. See [1] for the physical motivation behind considering such a system. Furthermore, the row sums of A are equal to one which guarantees the existence of a synchronized solution.

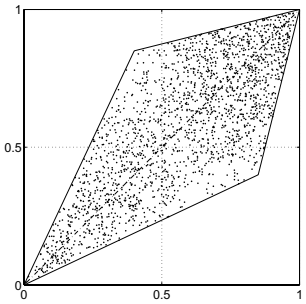


Fig. 1. Asymptotic distribution of 2000 points starting from an uniform distribution over the whole phase space Ω. $\epsilon = 0.2$.

Before proceeding we give a mathematical definition of synchronization we are interested in.

Definition 1. *A discrete dynamical system $S : \Omega \to \Omega$ is said to completely synchronize if as n tends to infinity $|S^n(x) - S^n(y)|$ tends to zero for all $(x, y) \in \Omega$.*

Using the linear stability analysis it can be shown [8] that this coupled map system synchronizes when $1/4 < \epsilon < 3/4$. The same result was also obtained by studying the evolution of the support of the invariant measure [7] which is a global result as opposed to the linear stability analysis which is carried out near the synchronized solution. This was done by showing that if $\epsilon < 1/4$ we obtain a quadrilateral of nonzero area as the support of the invariant measure. We depict this area in Fig 1 along with a distribution of points obtained from uniform distribution of initial conditions. And when ϵ crosses the value $1/4$ this quadrilateral shrinks to the line $x = y$. In this sense the synchronization transition is a discontinuous transition. But this figure is misleading as it does

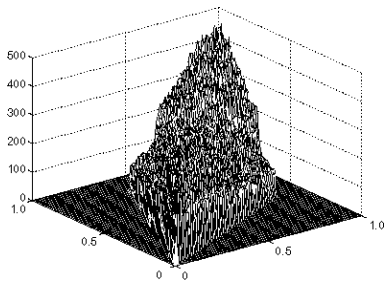


Fig. 2. The stationary density: a histogram of the distribution of points of Fig. 1. Here the number of points is 10^6 and the mesh size is 100x100. The horizontal plane depicts Ω and the vertical axis gives the number of points lying in the mesh square.

not tell us anything about the density of points. As we show in Fig. 2, if we plot the histogram then we see a quite irregular structure. In Fig. 3 we plot the section of this invariant density along the line $x = y$. We see clearly that it is an irregular function. Studying this invariant density is the object of this paper. In particular we show that this density is indeed a fractal function on the synchronization manifold, i.e., its section along line $x = y$, and the fractal dimension of the graph of this function depends on the coupling parameter.

The paper is organized as follows. In section 2 we give a brief introduction to invariant measures recalling some definitions and results required for completeness. We also outline a method to find the stationary density. We then move on to our main result in the section 3 after introducing basic concepts needed to characterize the irregularity of fractal functions. Section 4 concludes by pointing out some future directions.

2 Invariant measure

Invariant measures or the stationary densities [9] provide a useful way to study the asymptotic behavior of dynamical systems. One starts with a distribution

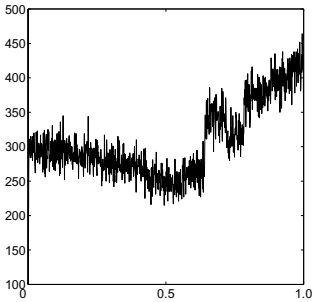


Fig. 3. The section of the stationary density along the line $x = y$. In this graph the total number of points used is 10^8 and the size of the mesh used is 1000x1000.

of initial conditions and studies its evolution as time goes to infinity. It is interesting to note that even if the dynamical system is chaotic a well behaved limit can exist which can then be used to study various average properties. In this section we give a brief introduction to invariant measures and a way to find one. We begin with the definitions (the norms used are L^1 norms throughout):

Definition 2. *A measure μ is said to be invariant under a transformation S if $\mu(S^{-1}(A)) = \mu(A)$ for any measurable subset A of Ω.*

Definition 3. *Let (X, A, μ) be a measure space and the set $D(X, A, \mu)$ be defined by $D(X, A, \mu) = \{f \in L^1(X, A, \mu) : f \geq 0$ and $||f|| = 1\}$. Any function $f \in D(X, A, \mu)$ is called a density.*

Definition 4. *Let (X, A, μ) be a measure space. A linear operator $P : L^1 \to L^1$ satisfying*
(a) $Pf \geq 0$ for $f \geq 0$, $f \in L^1$; and
(b) $||Pf|| = ||f||$, for $f \geq 0$, $f \in L^1$
is called a Markov operator.

A Markov operator satisfies a *contractive property*, viz., $||Pf|| \leq ||f||$. And this property implies the *stability property* of iterates of Markov operators, viz., $||Pf - Pg|| \leq ||f - g||$. We shall be interested in the fixed points of Markov operators.

Definition 5. *Let (X, A, μ) be a measure space and P be a Markov operator. Any $f \in D$ that satisfies $Pf = f$ is called a stationary density of P.*

A stationary density is the Radon-Nikodym derivative of an invariant measure with respect to μ.

2.1 The Frobenius-Perron operator

There exist various methods to find invariant measures. One approach is to use the so called Frobenius-Perron operator. This operator when applied to $\rho_n(x)$, the density at the nth time step, yields the density at the $(n+1)$th time step. Since all the points at the $(n+1)$th step in some set D come from the points in the set $S^{-1}(D)$ we have the following equality defining the Frobenius-Perron operator P:

$$\int\int_D P\rho(x)\mu(dx) = \int\int_{S^{-1}(D)} \rho(x)\mu(dx) \tag{4}$$

The Frobenius-Perron operator is a Markov operator.

2.2 The invariant measure of the tent map

If in (4) we choose our discrete dynamical system to be the one dimensional tent map defined in (2) and $D = [0, x]$ then the equation (4) reduces to

$$P\rho(x) = \rho(x/2) + \rho(1 - x/2). \tag{5}$$

We are interested in the fixed point solutions; this leads to

$$\rho(x) = \rho(x/2) + \rho(1 - x/2). \tag{6}$$

A solution of this functional equation is $\rho(x) = 1$. Of course, this is a trivial solution. $\delta(x)$ and $\delta(x-2/3)$ also solve this equation but we are not interested in such singular solutions since they do not span the phase space.

3 Stationary density of the coupled tent map

We now turn to coupled maps. In this section we study the stationary density on the synchronization manifold, i.e., the line $x = y$. We show that the density is a fractal function with its Hölder exponent related to the coupling parameter ϵ.

We use the following definition of Hölder continuity and its relation to the box dimension [10]:

Definition 6. *A function* $f : [0,1] \to R$ *is in* $C^{\alpha}_{x_0}$, *for* $0 < \alpha < 1$ *and* $x_0 \in [0,1]$, *if for all* $x \in [0,1]$

$$|f(x) - f(x_0)| \leq c|x - x_0|^{\alpha}. \tag{7}$$

A *pointwise Hölder exponent* $\alpha_p(x_0)$ at x_0 is the supremum of the αs for which the inequality (7) holds. We also use the relation between the Hölder continuity and the box dimension of the graph a function, $\dim_{\mathrm{B}}\mathrm{graph} f$.

Proposition 1. *If for $f : [0,1] \to R$, the pointwise Hölder exponent is α for all $x \in [0,1]$ and c in (7) is uniform then* $\dim_{\mathrm{B}}\mathrm{graph}f = 2 - \alpha$.

Now we are ready to state and prove our main result, viz, the estimate of the box dimension of the graph of the stationary density.

Theorem 1. *Let $\rho(x,y)$ be the stationary density of the coupled dynamical system (1) and let $\rho^D(x)$ be its restriction on the line $x = y$. If $\rho(x,y)$ is bounded then* $\dim_{\mathrm{B}}\mathrm{graph}\rho^D \geq d$ *where* $d = 2 + \ln(1-2\epsilon)/\ln 2$, *with* $0 < \epsilon \leq 1/4$.

Proof: We use the Frobenius-Perron operator defined in equation (4). We choose $D = [0,x] \times [0,y]$ and get

$$P\rho(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}\int\int_{S^{-1}(D)} \rho(x',y')dx'dy' \tag{8}$$

Our S in equation (1) is not invertible. In fact, it has 4 preimages. Let us denote them by S_i^{-1}, $i = 1, ..., 4$. If $X \in \Omega$, since f is symmetric, we get

$$P\rho(X) = J^{-1}(X)\sum_{i=1}^{4}\rho(S_i^{-1}(X)) \tag{9}$$

where $J^{-1}(X) = |dS^{-1}(X)/dX|$. The fixed point of this operator is given by the following functional equation for the density.

$$\begin{aligned}\rho(x,y) = \frac{1}{4|1-2\epsilon|}[&\rho(\beta x/2 - \gamma y/2, -\gamma x/2 + \beta y/2)\\ &+\rho(1-\beta x/2 + \gamma y/2, -\gamma x/2 + \beta y/2)\\ &+\rho(\beta x/2 - \gamma y/2, 1+\gamma x/2 - \beta y/2)\\ &+\rho(1-\beta x/2 + \gamma y/2, 1+\gamma x/2 - \beta y/2)]\end{aligned} \tag{10}$$

where $\gamma = \epsilon/1-2\epsilon$ and $\beta = 1+\gamma$. Since we know that a point belonging to Ω does not leave Ω, all the arguments of ρ on the right hand side of the above equation should be between 0 and 1. This gives us four lines which bound an area $0 \leq \beta x/2 - \gamma y/2 \leq 1$ and $0 \leq -\gamma x/2 + \beta y/2 \leq 1$. Lets denote this area by Γ. The support of the invariant measure should be contained in $\Gamma \cap \Omega$.

We also remark that for $0 \leq \epsilon < 1/4$, the discrete dynamical system S that we have considered is everywhere expanding and this implies that the stationary density exists [9].

The above equation can be written as, for $0 \leq \epsilon < 1/4$,

$$\rho(x,y) = \frac{1}{(1-2\epsilon)}\rho_{SS}(\beta x/2 - \gamma y/2, -\gamma x/2 + \beta y/2) \tag{11}$$

where

$$\rho_{SS}(x,y) = \frac{\rho(x,y) + \rho(1-x,y) + \rho(x,1-y) + \rho(1-x,1-y)}{4} \tag{12}$$

is the part of ρ that is symmetric around $1/2$ in both arguments. Now if we substitute $x = y$ in equation (11) we see that the arguments on both sides of the equation belong to the diagonal. As a result we obtain a functional equation

$$\rho^D(x) = \frac{1}{(1-2\epsilon)}\rho_S^D(x/2) \tag{13}$$

where we use a shorthand notation $\rho_S^D(x) = \rho_{SS}^D(x,x)$. With the change of variable $z = x/2$ and a decomposition of $\rho^D(x)$ as $\rho^D(x) = \rho_S^D(x) + \rho_A^D(x)$, the "symmetric" and "antisymmetric" part where again $\rho_A^D(x)$ is a shorthand notation for $\rho_{AS}^D(x,x) + \rho_{SA}^D(x,x) + \rho_{AA}^D(x,x)$, we arrive at

$$\rho_S^D(z) = (1-2\epsilon)\rho_S^D(2z) + g(z) \tag{14}$$

where $g(z) = (1-2\epsilon)\rho_A^D(2z)$. Its solution can be written down as

$$\rho_S^D(z) = \sum_{k=0}^{\infty}(1-2\epsilon)^k g(2^k z). \tag{15}$$

This is a Weierstrass function and if $g(z) \in C^\beta$ where $\beta > -\ln(1-2\epsilon)/\ln 2$ then the calculation in [10] for $g(z) = \sin(z)$ can be carried over and it can be shown that the pointwise Hölder exponent of this function is $-\ln(1-2\epsilon)/\ln 2$ everywhere implying that $\dim_{\mathrm{B}}\mathrm{graph}\rho_S^D = d$. And if $g(z)$ is not smooth enough then it can only increase the box dimension of $\rho^D(x)$, hence the result. □

4 Concluding Remarks

We have studied the stationary density of two coupled tent maps as a function of the coupling parameter. We find that even though the density of the individual tent map is smooth it becomes very irregular as soon as a small coupling is introduced in the sense that the pointwise Hölder exponent is small everywhere. And the density smoothness as the coupling is increased. It becomes smooth that is the Hölder exponent becomes one at the value of ϵ where the synchronization transition takes place. We have thus elucidated a new aspect of synchronization in coupled dynamical systems, beyond the standard aspects of linear or global stability of synchronized solutions.

It is a curious fact that the Hölder exponent becomes one exactly at the critical value of the coupling parameter, i.e., $\epsilon_c = 1/4$. It would be important to understand if there is any underlying principle behind this observation, that is, one valid also for other maps with varying coupling matrices.

It is interesting to note that fractal probability densities have arisen in a completely different scenario, namely the random walk problem with shrinking step lengths [11].

One should also characterize the stationary density away from the synchronization manifold. It is expected to have a more complex multifractal character [12]. The effect of different network topologies on the stationary density is another interesting topic.

One of us (KMK) would like to thank the Alexander-von-Humboldt-Stiftung for financial support.

References

1. A. Pikovsky, M. Rosenblum, and J. Kurths. *Synchronization - A Universal Concept in Nonlinear Science*. Cambridge University Press, 2001.
2. L. M. Pecora, T. L. Carroll, G. A. Johnson, D. J. Mar, and J. F. Heagy. Fundamentals of synchronization in chaotic systems, concepts, and applications. *Chaos*, 7:520, 1997.
3. G. D. VanWiggeren and R. Roy. Communication with chaotic lasers. *Science*, 279:1198, 1998.
4. D. Hansel and H. Sompolinsky. Synchronization and computation in a chaotic neural network. *Phys. Rev. Lett.*, 68:718, 1992.
5. G. Buzsáki and A. Draguhn. Neuronal oscillations in cortical networks. *Science*, 304:1926, 2004.
6. R. W. Friedrich, C. J. Habermann, and G. Laurent. Multiplexing using synchrony in the zebrafish olfactory bulb. *Nature Neuroscience*, 7:862, 2004.
7. J. Jost and K. M. Kolwankar. Global analysis of synchronization in coupled maps, 2004.
8. J. Jost and M. P. Joy. Spectral properties and synchronization in coupled map lattices. *Phys. Rev. E*, 65(1):016201, 2002.
9. A. Lasota and M. C. Mackey. *Chaos, Fractals and Noise*. Springer, 1994.
10. K. Falconer. *Fractal Geometry - Mathematical Foundations and Applications*. John Wiley, 1990.
11. P. L. Krapivsky and S. Redner. Random walk with shrinking steps. *Am. J. Phys.*, 72:591, 2004.
12. S. Jaffard. Multifractal formalism for functions part ii: Self-similar functions. *SIAM J. Math. Anal.*, 28:971, 1997.

2

PHYSICS

A Network of Fractal Force Chains and Their Effect in Granular Materials under Compression

Luis E. Vallejo, Sebastian Lobo-Guerrero, and Zamri Chik

Department of Civil and Environmental Engineering
University of Pittsburgh, Pittsburgh, PA 15261- U. S. A.
vallejo@engrng.pitt.edu,sel2@pitt.edu, irzamri@yahoo.com

Summary. Granular materials forming part of civil engineering structures such as rockfill dams and the granular base in pavement systems are subjected to large compressive stresses resulting from gravity and traffic loads respectively. As a result of these compressive stresses, the granular materials break into pieces of different sizes. The size distribution of the broken granular material has been found to be fractal in nature. However, there is no explanation to date about the mechanisms that cause the granular materials to develop a fractal size distribution. In the present study, a compression test designed to crush granular materials is presented. The tests used 5 mm glass beads and a plexiglass cylinder having an internal diameter equal to 5 cm. As a result of compression in the cylinder, the glass beads broke into pieces that had a fractal size distribution. The compression test was numerically simulated using the Discrete Element Method (DEM). The DEM simulation indicated that the particles developed a network of force chains in order to resist the compressive stress. These force chains did not have a uniform intensity but was found to vary widely through out the sample. Also, the distribution of the force chains in the sample did not involve all the grains but only a selective number of them. Thus, the force chains did not cover the whole sample. Using the box method, it was determined that the distribution of the force chains in the sample was fractal in nature. Also, the intensity of the force chains in the sample was found to be fractal in nature. Thus, the fractal nature of the intensity of the force chains and their distribution were found to be the main reason why granular material develop fractal fragments as a result of compression.

1 Introduction

Granular materials forming part of civil engineering structures such as rockfill dams and the granular base in pavement systems are subjected to sustained compressive stresses resulting from gravity and traffic loads respectively. As a result of these compressive stresses, the granular materials break into pieces of different sizes. Thus, the original engineering properties with which the rockfill

dam or the base of a pavement structure was designed with (i.e., hydraulic conductivity, shear strength, elastic moduli) will change during its engineering life. Changes in the original engineering properties could affect the stability of the structures and could make them unsafe. The size distribution of the broken granular material has been found to be fractal in nature. However, there is no explanation to date about the mechanisms that cause the granular materials to develop a fractal size distribution. In the present study, a test designed to crush granular materials and its simulation using the Discrete Element Method (DEM) are developed to understand the mechanisms involved with the development of a fractal fragmentation by granular materials.

1.1 The Crushing of Granular Materials

Granular materials form part of engineering structures such the base of flexible pavements, highway embankments, and foundations. The granular materials forming part of these structures are subjected during their engineering lives to either static or dynamic loads. As a result of these loads, particle breakage occurs [1–4]. According to Coop [3] and Lee and Farhoomand [1], particle breakage or crushing seems to be a general feature for all granular materials. Grain crushing is influenced by grain angularity, grain size, uniformity of gradation, low particle strength, high porosity, and by the stress level and anisotropy [2].

When a granular mass is subjected to a compressive load, the particles resist the load through a series of contacts between the grains (Fig. 1). The particles with highly loaded contacts are usually aligned in chains. These chains form a network of forces in the granular material that resist the applied load [5–10]. Crushing starts when these highly loaded particles fail and break into smaller pieces that move into the voids of the original material. This migration causes the settlement of a granular assembly. Also, on crushing, fines are produced and the grain size distribution curve becomes less steep. Consequently, with continuing crushing, the granular material becomes less permeable and more resistant to crushing. Grain size distribution is a suitable measure of the extent of crushing [2].

Lade et al. [2] found that if a uniform granular material is crushed, the resulting grain size distribution approaches that of a well graded soil for very large compressive loads. Bolton [4], and McDowell et al. [11] established that the grain size distribution of a granular assembly that has been crushed under large compressive loads is a fractal distribution. A well graded particle distribution or a fractal distribution represents a granular structure that is made of grains of all sizes including the original unbroken grains. There is no explanation to date, however, about the mechanisms that cause the granular materials to develop the fractal size distribution. A uniaxial compression test designed to break glass beads is presented next. This test will be numerically simulated and analyzed using the Discrete Element Method in order to clarify

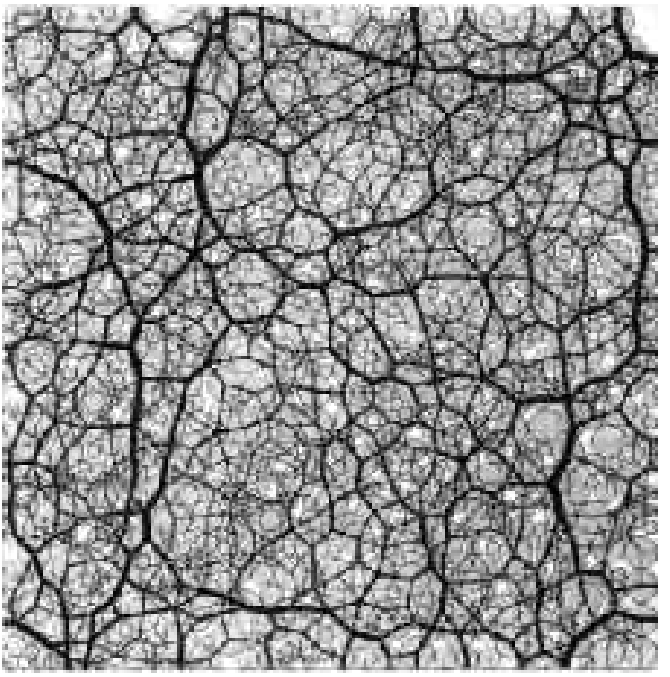


Fig. 1. Force chains in two dimensional static assembly of discs in a horizontal container with three fixed walls and one piston, on which a fixed vertical force is applied [6].

the mechanisms involved with the development of a fractal fragmentation by the glass beads.

2 Laboratory Test

2.1 Compression Test

A uniaxial compression experiment was carried out in a plexi-glass tube with a 5 cm interior diameter and filled with 5.0 mm diameter glass beads having specific gravity of 2.5 . The beads filled the tube up to a depth equal to 10 cm (Fig. 2). For meaningful test results (no wall effects), it is necessary to maintain a ratio of sample diameter to the maximum particle size of approximately 6:1 or greater (in the tests it was 10:1). The tube is set up right with a steel plug at the bottom on which the beads rest and a 2.0 kg piston head pressing against the top of the beads. The objective of the exercise is to impound a crushing condition to the grains and to investigate the crushing characteristics of the perfectly spherical beads.

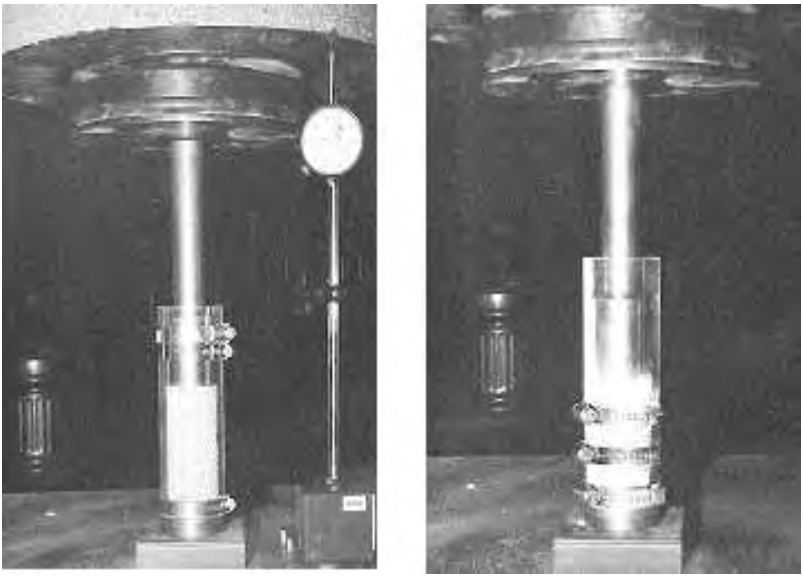


Fig. 2. Uniaxial compression test of the glass beads before and after application of uniaxial load.

The glass beads were loaded to three different uniaxial compressive stresses. These stresses were equal to 10.0 MPa, 11.8 MPa and 23.8 MPa. Due to the position of the individual grains and the consequent manner by which the particle-to-particle contact points developed and changed through out the tests, varying contact forces were experienced by individual beads in the specimen. These varying contact forces developed chains of different intensities inside the sample as outlined in Fig. 1.

As the uniaxial compressive stress was applied to the glass beads, they experience sporadic explosions. These small explosions occurred all over the particulate system without any particular pattern. With the crushing of a spheres a sudden addition of voids was created since the broken spheres were substituted solid spheres in the system. With crushing, the point-to-point contact between the spheres was reduced as the amount of fragments increased. How a sample of glass beads looked at the end of a uniaxial compression test can be seen in Fig. 3.

After the completion of the uniaxial compression test, the glass beads were removed from the tubes and a sieve analysis was performed on the broken sample. The results of the sieve analysis produced particulate size distributions that were used to determine the fragmentation fractal dimension, D_F, of the samples. The results of the sieve analysis is shown in Fig. 4. An analysis of Fig. 4 indicates that the samples upon crushing developed a grain size distribution that was well graded or fractal in nature. The samples experienced

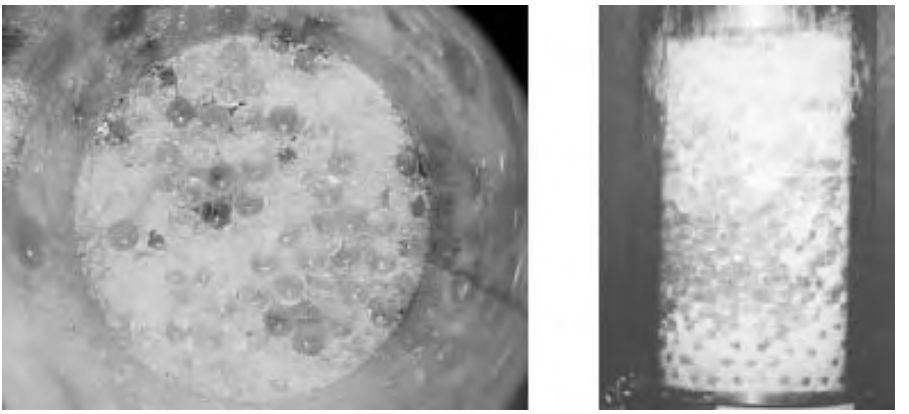


Fig. 3. The glass beads as seen from the top and the side of the plexiglass tube at the end of the uniaxial compression test.

a higher level of fragmentation with an increase in the value of the uniaxial compressive stress.

2.2 Fragmentation Fractal Dimension of the Grain Size Distribution

Grain size distribution of naturally occurring soils have been found by Tyler and Wheatcraft [12] and Hyslip and Vallejo [13] to be fractal. According Tyler and Wheatcraft, the distribution of grains by size in a natural soil can be obtained using the following equation:

$$N(R > r) \;=\; kr^{-D_F} \tag{1}$$

where $N(R > r)$ is the total number of particles with linear dimension R (radius of the particle) which is greater than a given size r; k is a proportionality constant; and D_F is the fractal dimension of the size distribution of grains. As a result of compression, the size distribution in a granular soil will change. Changes in the size distribution of the grains will be reflected in the values of D_F. Thus, grain fragmentation in soils subjected to compressive stresses can be evaluated by the changes in their fragmentation fractal dimension, D_F.

To apply the number-based relationship expressed by Eq. (1), is very time consuming. Another relationship that uses the results of a standard sieve analysis test was developed by Tyler and Wheatcraft [12] to calculate the fragmentation fractal dimension, D_F, of natural soils. This relationship is:

$$\frac{M(R < r)}{M_T} = \left(\frac{r}{r_L}\right)^{3-D_F} \tag{2}$$

where $M(R < r)$ is the cumulative mass (weight) of particles with size R smaller (finer) than a given comparative size r; M_T is the total mass (weight)

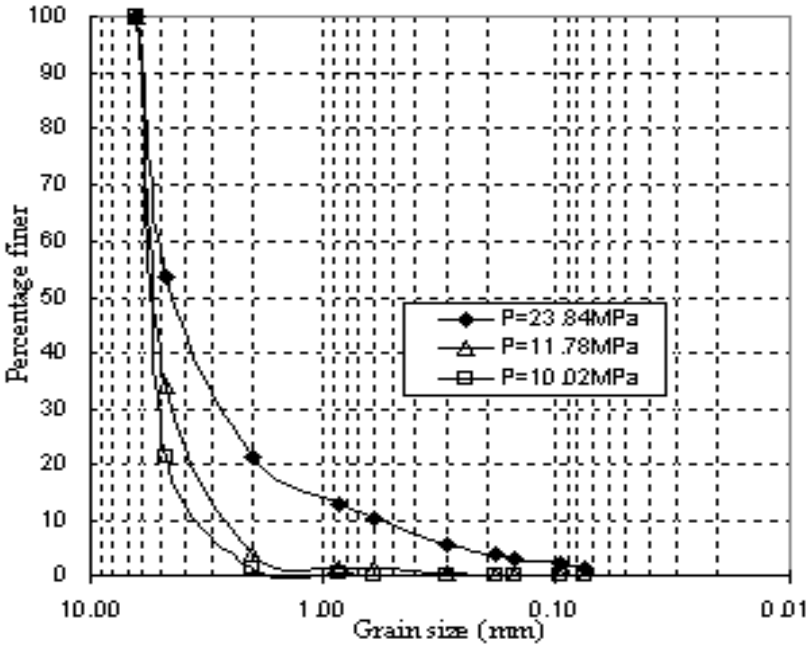


Fig. 4. Grain size distribution of the glass beads after the uniaxial compression tests.

of particles; r is the sieve size opening; r_L is the maximum particle size as defined by the largest sieve size opening used in the sieve analysis; and D_F is the fragmentation fractal dimension. The results of a sieve analysis tests using Eq. (2) can be plotted on log-log paper. The slope, m, of the best fitting line through data obtained using Eq. (2) and the fractal dimension, D_F, are related as follows:

$$D_F = 3 - m \tag{3}$$

Eqs. (2) and (3) will be used to obtain the fractal dimension of the size distribution of glass beads subjected to crushing in the uniaxial compression tests (Fig. 4). The fragmentation fractal dimension, D_F, for the grain size distributions shown in Fig. 4 are shown in Figs. 5 and 6.

An analysis of Figs. 5 and 6 indicates that the fragmentation fractal dimension, D_F, increased with the uniaxial compressive stress induced to the glass beads. The fragmentation fractal dimension measures the degree of crushing of the glass beads. The greater the fragmentation fractal dimension, the greater is the level of breaking of the beads.

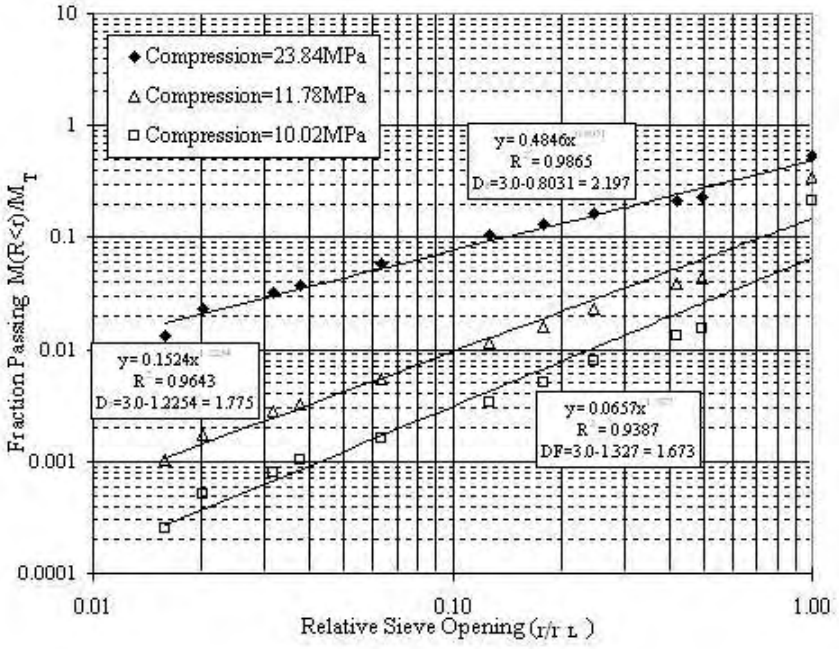


Fig. 5. The fragmentation fractal dimensions, D_F, for the grain size distributions shown in Fig. 4.

3 An Explanation for the Development of a Fractal Size Distribution in the Samples

Next, an explanation why the glass beads developed a fractal size distribution is carried out using the Discrete Element Method. The PFC^{2D} code developed by Itasca [14] was used in order to gain an understanding how the glass beads interact and distribute the reacting forces between grains that are used to resist the uniaxial compressive loads

3.1 Model of Grains Under Uniaxial Compression

A rectangular box with one fixed base wall, two fixed lateral loads, and a movable top wall was developed. The box width was equal to 0.05 m, the height of the wall was equal to 0.10 m, and the thickness of the wall was equal to 0.01 m. 950 uniform disks with a diameter equal to 2 mm were randomly generated inside the box. The density of the disks was equal to 2,500 kg/m^3, and the friction coefficient between the disks and the wall and the disks was set to be equal to 0.7. With the sample ready, the top wall of the box was allowed to move with a constant velocity of 7×10^{-7} meters/step. The top wall

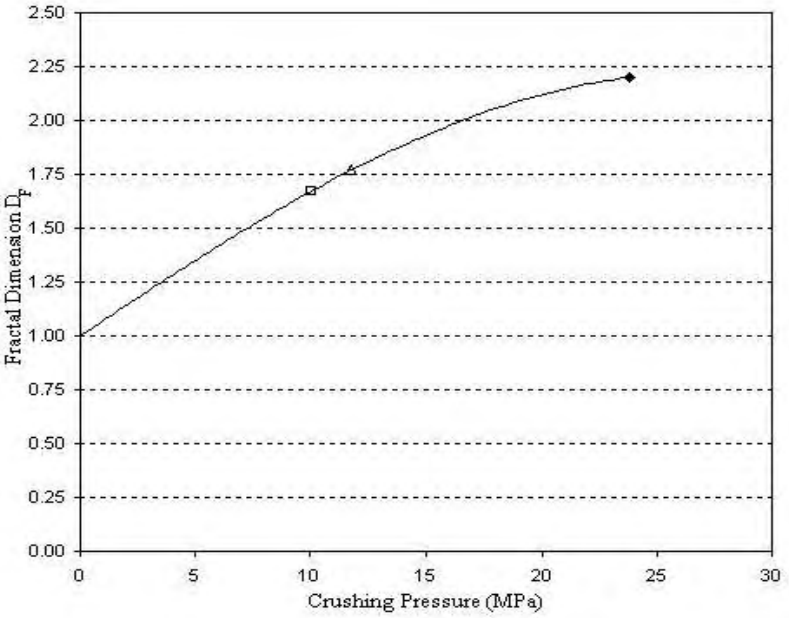


Fig. 6. Fragmentation fractal dimension versus crushing pressure.

was stopped when the top wall received a reaction force of 2×10^5 Newtons. As a result of a uniaxial compressive load, the disks developed a network of force chains that were randomly distributed trough the sample and had also varying intensities (given by the thickness of the force chains). The force chains developed by the sample of disks is shown in Fig. 7. The force chains shown in Fig. 7 are similar to the ones shown in Fig. 1.

The maximum contact force developed by two disks inside the sample was found to be equal to 1.98×10^4 Newtons which is one order of magnitude smaller than the compressive load applied to the disks.

The number of contacts (coordination number) for each disk was also obtained using the PFC^{2D} code. The most frequent coordination number for the disks was found to be slightly greater than 4 (Fig. 8).

Using the PFC^{2D} code, the particles that were heavily, moderately, and slightly loaded in the particle system shown in Fig. 7 were identified. To do this, the particles that were loaded with a contact force greater than 2/3 of the maximum force recorded in the disks (1.98×10^4 Newtons) were colored black (heavily loaded), the disks subjected to a contact force equal or smaller than 2/3 of the maximum force but greater than 1/3 of this force were colored gray (moderately loaded), and the disks subjected to a contact force equal or smaller than 1/3 of the maximum force was colored white (slightly loaded). The result of this process is shown in Fig. 9.

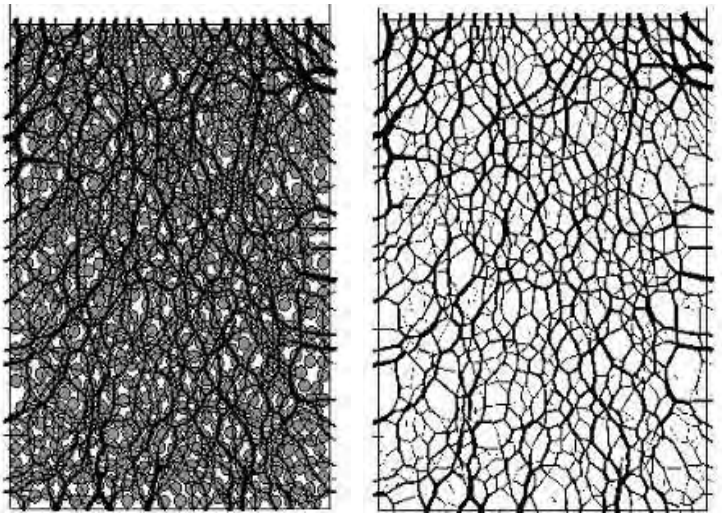


Fig. 7. Force chains developed by the uniform disks under a uniaxial compression.

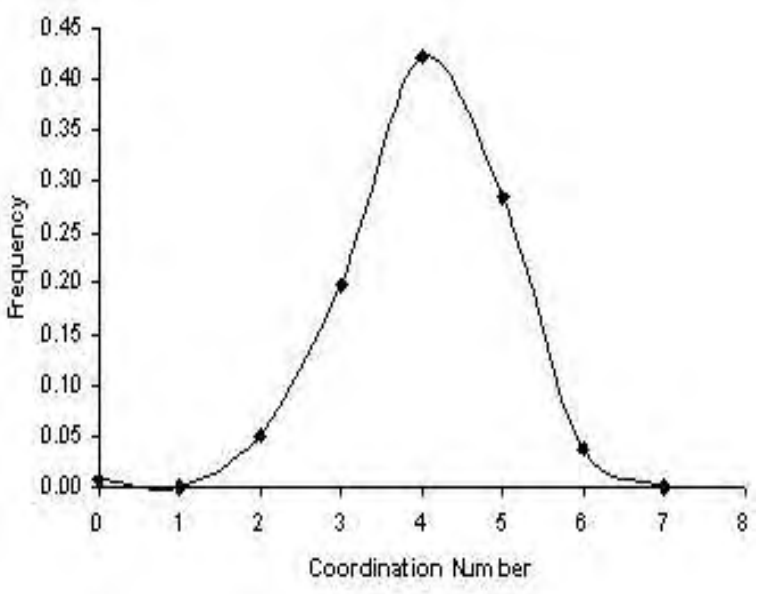


Fig. 8. Coordination number for the disks in Fig. 7.

3.2 Fractal Analysis of the Network of Force Chains

An analysis of the force chains developed by the disks indicates that these force chains are distributed like a network in the granular system and the

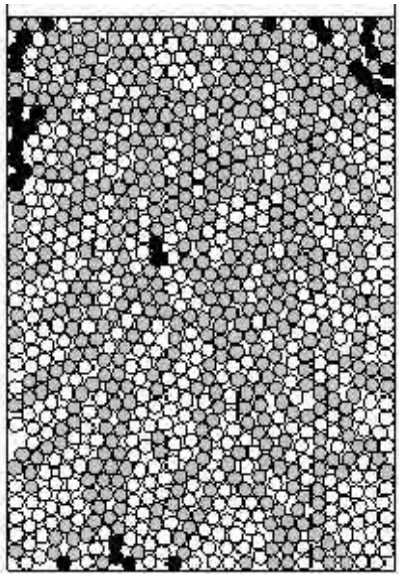


Fig. 9. Distribution of how particles are loaded by the uniaxial compressive load.

network is made of branches of varying degrees of intensity (the thicker the force chain in Fig. 7, the larger is the force). Using the box method from fractal theory, the fractal dimension of the network was obtained. The box method is an established method that has been used to measure the distribution of fracture networks in the earth crust [15,16]. The box method uses grids made of squares of different sizes that are placed on top of the force networks (Fig. 7). If one plots in a log-log paper the number of boxes intercepted by the force chains versus the size of the squares, one obtains the fractal dimension of the distribution of the force chains in Fig. 7. This has been done in Fig. 10.

An analysis of Fig. 10 indicates that the distribution of the force chains in Fig. 7 is indeed fractal. The fractal dimension of the distribution of forces in the granular system is very high and equal to 1.7431.

The intensity of the forces in the chains shown in Fig. 7 was also obtained using the PFC^{2D} code. This code has a subroutine that creates a contact pointer that goes contact by contact calculating and classifying the resultant force in each contact. At the end of the run of this subroutine, the number of the contacts with their respective contact forces are obtained. Using this information, a log-log plot of the number of contacts, N, with a force, R, greater than certain value r is plotted against the contact force, r. The result of this analysis is shown in Fig. 11.

An analysis of the results shown in Fig. 11 indicates that the intensity of force chains in the network shown in Fig. 7 follows an inverse power-law distribution. A power law hints that a system is organizing itself into a system

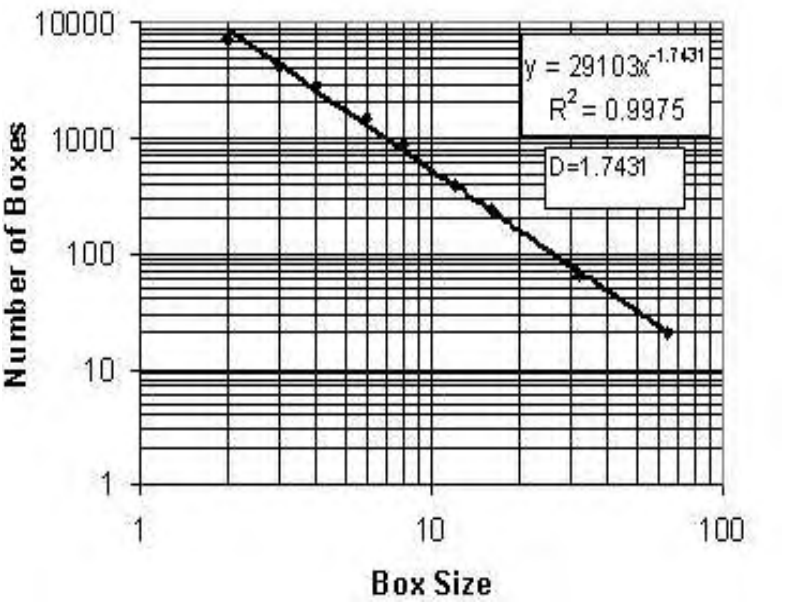


Fig. 10. The fractal dimension of the distribution of force chains in Fig. 7.

that is fractal in nature [17]. In this case with respect to the intensities of the force chains that produce a fractal granular material. The exponent in the inverse power-law distribution of forces shown in Fig. 11 was found to be equal to 4.5068. Since the granular material being produced by to the force chain network shown in Fig. 7 is a fractal one, it is safe to assume that the exponent of the power-law distribution shown in Fig. 11 represents the fractal dimension of the distribution of the force intensities shown in Fig. 7.

3.3 Discussion of the Results

When granular materials are subjected to compressive or shear loads, the materials resist the loads through a series of contacts between the grains. The particles with highly loaded contacts are usually aligned in chains. These chains form a network of forces in the granular material that resist the applied loads [5–10]. These force chains fluctuate in intensity (strength) in the granular materials. Several theories have been advanced by Liu et al [7], Coppersmith, et al [8], and Cruz Hidalgo et al. [10] to explain force fluctuations in granular materials. Howell et. al [9], and Cruz Hidalgo et al. [10] present a summary of these theories. In this study, an analysis of grain crushing and its causes is presented. The causes seem to be the network of force chains developed by the granular materials in response to applied compressive loads.

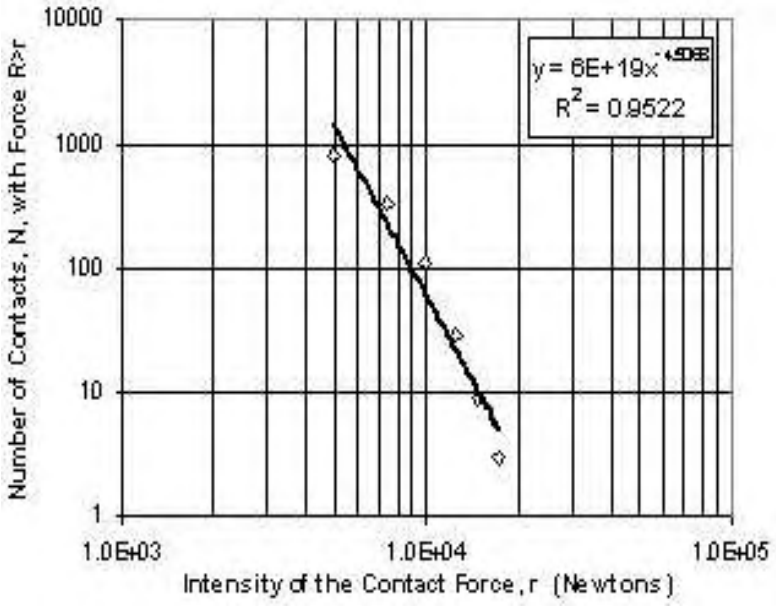


Fig. 11. Plot to obtain the fractal dimension of the force intensity distribution.

The relationship between grain crushing and force chain networks was not considered by previous studies on force chain fluctuations [7–10].

An analysis of the results of the present study shown in Figs. 7 and 9 and Figs. 10 and 11 indicates that the distribution of the network of force chains as well as their intensities are fractal in nature. Because the distribution of the force chains does not cover all the grains, some of the grains are subjected to the force chains while some of the grains are not. In fact if one looks at Figs. 7, some of the idle grains (the ones that not carry any load), can be removed without affecting the stability of the granular system. Also, these idle grains will form part of the original unbroken grains in the fractal distributions after crushing as was found in the laboratory experiments (Fig. 3).

The intensity of the force chains was also found to be fractal (Fig. 11). Thus, some of the particles will be subjected to complete crushing (the black particles with the high loads as shown in Fig. 9). Some of the particles will be partially broken under the moderate loads (the gray particles in Fig. 9). An some of the particles will resist the load without breaking (the white particles in Fig. 9). Thus, the end result of the compression process will be a system of granular material that is fractal on their size distribution as shown by the laboratory experiments (Figs. 2 to 6).

4 Conclusions

Laboratory compression tests on uniform granular materials and their simulation using the Discrete Element Method indicated the following:

(1) The effect of a compressive load exerted on a uniform system of granular material was to break the particles into a new system that had a size distribution that was fractal in nature.

(2) The cause for the fractal size distribution of the broken grains was a network of force chains that were fractal in their intensity and distribution.

5 Acknowledgments

The work described herein was supported by a Grant CMS 0301815 to the University of Pittsburgh from the National Science Foundation, Washington D.C. This support is gratefully acknowledged.

References

1. Lee, K.L., and Farhoomand, J. (1967). "Compressibility and crushing of granular soils in anisotropic triaxial compression". *Canadian Geotechnical J.*, Vol. 4, No. 1, pp. 68-86.
2. Lade, P.V., Yamamuro, J.A., and Bopp, P.A. (1996). "Significance of particle crushing in granular materials". *J. of Geotechnical Eng., ASCE*, Vol. 122, No. 4, pp. 309-316.
3. Coop, M.R. (1999). "The influence of particle breakage and state on the behavior of sand". *Proceedings of the Int. Worshop on Soil Crushability, Yamaguchi, Japan*, pp. 19-57.
4. Bolton, M.D. (1999). "The role of micro-mechanics in soil mechanics". *Proceedings of the Int. Workshop on Soil Crushability, Yamaguchi, Japan*, pp. 58-82.
5. Cundall, P.A., and Strack, O.D.L. (1979). "A discrete numerical model for granular Assemblies". *Geotechnique*, Vol. 29, No. 1, pp. 47-65.
6. Radjai, F. (1995). "Dynamique des Rotations et Frottement Collectif dans les Sys- temes Granulaires". Ph.D. Thesis, Universite de Paris-Sud XI, Orsay.
7. Liu, C.H., Nagel, D., Shecter, D., Coppersmith, S.N., Majumdar, S., Narayan, O., and Witten, T.A. (1995). "Force fluctuations in bead packs". *Science*, Vol. 269, pp. 513- 515.
8. Coppersmith, S.N., Liu, C.H., Majumdar, S., Narayan, O., and Witten, T.A. (1996). "Model for force fluctiations in bead packs". *Physical Review E*, Vol. 53, No. 5, pp. 4673-4685.
9. Howell, D.W., Behringer, R.P. (1999). "Fluctuations in granular media". *Chaos*, Vo. 9, No. 3, pp. 559-572.
10. Cruz Hidalgo, R., Grosse, C.U., Kun, F., Reinhardt, H.W., and Herrmann, H.S.(2002). "Evolution of percolating force chains in compressed granular media". *Physical Review Letters*, Vol. 89, No. 20, pp. 205501-1 - 205501-4.

11. McDowell,G.R., Bolton, M.D., and Robertson, D. (1996). "The fractal crushing of granular materials". *Int. J. of Mechanics and Physics of Solids*, Vol. 44, No. 12, pp. 2079-2102.
12. Tyler, S.W., and Wheatcraft, S.W. (1992). "Fractal scaling of soil particle-size distribution analysis and limitations". *Soil Science Society of America Journal*, Vol. 56, No. 2, pp. 47-67.
13. Hyslip, J.P., and Vallejo, L.E. (1997). "Fractal analysis of the roughness and size distribution of granular materials". *Engineering Geology*, Vol. 48, No. 3-4, pp. 231-244.
14. Itasca Consulting Group (2002). "Particle Flow Code in Two Dimensions, PFC^{2D}", Version 3.0. Minneapolis, Minnesota.
15. Watanabe, K., and Takahashi, H. (1995). "Fractal geometry characterization of geothermal reservoir fracture networks". *Journal of Geophysical Research*, Vo. 100, No. B1, pp. 521-528.
16. Acuna, J., and Yortsos, Y.C. (1997). "Application of fractal geometry to the study of networks of fractures and their pressure transient". *Water Resources Research*, Vol. 31, No. 3, pp. 527-540.
17. Strogatz, S. (2003). "Sinc: The Emerging Science of Spontaneous Order". Theia-Hyperion Press, New York, pp.338.

Percolation and permeability of three dimensional fracture networks with a power law size distribution

V.V. Mourzenko[1], Jean-François Thovert[1], and Pierre M. Adler[2]

[1] LCD, SP2MI, BP 179, 86960 Futuroscope Cedex, France
mourzenko@lcd.ensma.fr, thovert@lcd.ensma.fr
[2] IPGP, tour 24, 4 Place Jussieu, 75252 Paris Cedex 05, France
adler@ipgp.jussieu.fr

Summary. The percolation and permeability of fracture networks is investigated numerically by using a three dimensional model of plane polygons randomly located and oriented in space with sizes following a power law distribution. The influence of the range and exponent of the size distribution, of the fracture shapes and of the exponent of the individual fracture conductivity is examined. A dimensionless fracture density, which involves a shape factor, proves to be an adequate percolation parameter. In these terms, the critical density is nearly invariant, over a wide range of shape and size distribution parameters. The permeability is determined by solving the flow equations after triangulating the fracture networks. Eventually, a general expression is proposed, which is the product of the volumetric surface area, weighted by the individual fracture conductivities, and of a fairly universal function of the dimensionless network density, which accounts for the influences of the fracture shape and size distributions. Two analytical formulas are proposed which successfully fit the numerical data.

1 Introduction

Fractures and fracture networks determine the permeability of many natural rocks, and their behavior generated interest in various fields [1,4,31]. Percolation is a crucial issue for the transport properties of random systems, and it has been the topic of many studies in the past, starting with lattice systems [35]. However, percolation in fracture networks is a part of a more general continuum percolation problem. The definition of an appropriate density parameter plays a central role, and it is complicated here by two features, namely that real fracture networks are polydisperse, with a fracture size distribution which can be described in many cases by a power law [5, 6, 17, 27, 32, 36], and that the real fractures may have different and generally unknown shapes.

Various models have been considered, including 2d systems of line segments [28, 29] and 3d systems of polydisperse discs [7] or ellipses [10]. The

problem of the dependence of continuum percolation thresholds on polydispersity has also been addressed in [3, 8, 9, 19]. In general, an invariant percolation parameter can be devised, which involves the second (in 2d) or third (in 3d) moment of the object size distribution. Huseby *et al.* [15] used the concept of excluded volume to obtain percolation thresholds independent of the fracture shape for 3d monodisperse networks.

In previous contributions, we determined the geometrical properties [15] and permeability [16] of monodisperse fracture networks, and showed that the excluded volume introduced by [2] plays a crucial role; a dimensionless fracture density ρ' was defined which controls percolation and permeability whatever the fracture shape.

Therefore, the main objective of this paper is to extend the results of [15,16] to 3d fracture networks with a power law size distribution. To the best of our knowledge (see the recent review [4]), there is no such previous study on 3d permeability, although some 2d situations have been addressed [11,26].

In the present study, we consider networks of plane polygonal fractures, uniformly distributed in space with sizes following a power law distribution. The domain size is supposed to significantly exceed the size of the largest fractures in the networks, so that it is possible to determine an effective upscaled permeability. Note that the situation where fractures larger than the domain size exist is considered elsewhere [23]. Under such circumstances, percolation is governed by different mechanisms, dominated by the probability of occurence of a single fracture which crosses the whole domain.

The present contribution is organized as follows. The geometrical model of polydisperse fracture networks is described in Section 2, and some notations are introduced. In particular, a dimensionless density is defined, which will play a crucial role in the descriptions of percolation and flow properties. Percolation is addressed in Section 3. A unique value of the critical density is obtained, independent of the fracture shape and size distributions, except for a small correction for very elongated shapes. Section 4 addresses network permeability. The flow equations and the method of solution are briefly discussed. Then, a simple and very general formula for the network permeability is presented. It involves a dimensionless function of the network density, for which two analytical models are proposed. A summary of the main findings in Section 5 ends up this paper.

2 General considerations

Consider three-dimensional networks made of plane polygonal fractures, randomly oriented and located in space with a volumetric number density ρ. Each fracture is characterized by its surface area A, its perimeter P, and some measure R of its size, which in the following is the radius of its circumscribed disk.

According to various observations of fractured rocks ([1] and references herein), many real probability densities of fracture sizes follow a power law

$$n(R) = \alpha R^{-a} \tag{1}$$

where $n(R)\mathrm{d}R$ is the number of fractures with radius in the range $[R, R+\mathrm{d}R]$ and α is a normalization coefficient. In practice, the scaling exponent a ranges from 1.8 to 4.5 [5], and R varies in a large interval which can span five orders of magnitude, but it is limited by the size R_M of the largest fractures in the system and by the size R_m of the microcracks.

The percolation properties of such networks are investigated here in finite cubic domains of size L^3, with $R_m \ll R_M \ll L$. The first task to be performed, and may be conceptually the essential one, is the derivation of an adequate definition of the network density.

Note first that two definitions of a dimensionless and intrinsic network density appear possible. One is volumetric, quantified by the average number of fractures in a reference volume; the other is topological, defined as the average number of connections per fracture with other fractures in the network, *i.e.*, a mean coordination number. These two quantities are proportional one to another for given fracture shape and size distributions, but their ratio strongly depends on these statistical parameters if the reference volume is not properly defined. However, the two definitions are nicely reconciled by the concept of excluded volume, which was introduced in the context of continuum percolation by [2].

For a pair of objects F_1 and F_2, the excluded volume $V_{\mathrm{ex},12}$ is the volume around F_1 which would be excluded for the center of F_2 if the objects were impenetrable. It can be shown [1] that if the objects are two-dimensional, with areas A_i, perimeters P_i $(i = 1, 2)$, random orientations and convex contours, the mean excluded volume is

$$V_{\mathrm{ex},12} = \frac{1}{4}\left(A_1 P_2 + A_2 P_1\right) \tag{2}$$

For a set of identical polygons, this reduces to $V_{\mathrm{ex}} = AP/2$. An anisotropic orientation distribution can easily be accounted for, as shown in [1]. For a population of objects with different shapes, a global mean excluded volume $\langle V_{\mathrm{ex}} \rangle$ can be obtained by averaging (2) over the population distribution.

For networks of fractures with identical sizes, but possibly different shapes, we may use $\langle V_{\mathrm{ex}} \rangle$ to define the dimensionless fracture density ρ'

$$\rho' = \rho \, \langle V_{\mathrm{ex}} \rangle \tag{3}$$

It can be interpreted as a volumetric density, since it is the number of fractures per volume $\langle V_{\mathrm{ex}} \rangle$; however, it also represents the mean number of intersections per fracture with other fractures in the network, and as such, it is a direct measure of the connectivity. Therefore, the definition (3) incorporates both the volumetric and topologic aspects mentioned above.

This definition proved very successful in unifying the critical densities of networks of fractures with different shapes. For regular polygons with 3 to 20 vertices, as well as for rectangles with aspect ratio two, [15] obtained a nearly

constant percolation threshold $\rho'_c \approx 2.3$. It was also shown that many other geometrical features, such as the volumetric density of blocks or the cyclomatic number, as well as the permeability [16] only depend on the density ρ'.

However, it will be shown below that global connectivity (percolation) is no longer controled solely by the local one (mean coordination), in the case of size polydispersity, and the definition of the percolation parameter has to be generalized. Since shape effects are well accounted for by $\langle V_{\rm ex}\rangle$, it is useful to define the dimensionless shape factor $\langle v_{\rm ex}\rangle$, for a set of fractures with identical of different shapes,

$$\langle v_{\rm ex}\rangle = \frac{\langle V_{\rm ex}\rangle}{\langle R\rangle\langle R^2\rangle} \tag{4}$$

It can then be used to define two dimensionless densities, with different weightings of the fracture sizes

$$\rho'_{21} = \rho\ \langle v_{\rm ex}\rangle\ \langle R^2\rangle\langle R\rangle = \rho\langle V_{\rm ex}\rangle \tag{5a}$$

$$\rho'_3 = \rho\ \langle v_{\rm ex}\rangle\ \langle R^3\rangle \tag{5b}$$

The subscripts are reminders of the statistical moments of R involved in each definition. The first one, ρ'_{21} is the generalization of ρ' for monodisperse networks, since it can be shown that it is still equal to the mean number of intersections per fracture [1]. Both ρ'_{21} and ρ'_3 reduce of course to ρ' in case of equal-sized fractures.

The generation and analysis of the percolation properties of fracture networks are similar to those presented by [15]. The fractures are embedded in a cubic cell of size L; $N_{fr} = \rho L^3$ is the number of fractures in the unit cell. Centers of fractures are uniformly distributed in space, and their normal vectors are uniformly distributed on the unit sphere. An example is shown in Fig. 1.

Only large unit cells $L \geq 4R_M$ are considered here. Then, L is a natural homogenization scale over which the macroscopic permeability of a fracture network can be defined. Any scaling behavior or hydraulic properties of fracture systems which can arise when $L \leq 4R_M$ is out of the scope of this paper. The case of L of the order of or smaller than R_M is addressed in [23].

Two types of fracture systems were used, for the percolation and permeability calculations, with or in most cases without spatial periodicity. In the latter case, fracture centers were generated within the unit cell as well as outside it, and all the fractures which intersect at least one of the six faces of the cell were retained. In all cases, a network is said to percolate if a continuous path exists along, say, the $x-$direction. Since, the two types of settings yield identical results, this point is not discussed any further in this paper.

The connectivity and percolation properties of the fracture networks are determined by the exponent a and by the two dimensionless ratios

$$R'_{\rm m} = \frac{R_{\rm m}}{R_{\rm M}}, \quad L' = \frac{L}{R_{\rm M}} \tag{6}$$

In order to eliminate the influence of the lower cut-off R_m, R_m/R_M is kept as small as numerically affordable.

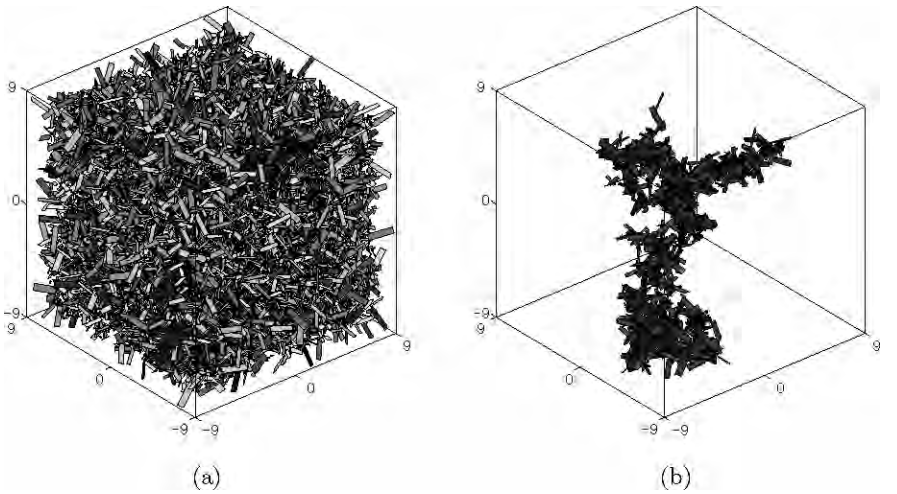


Fig. 1. Example of polydisperse network of rectangular fractures (a), with $L' = 16$, $a = 1.5$, $R'_{\rm m} = 0.1$, which contains $N_{fr} = 65900$ fractures ($\rho'_{21} = 1.06$, $\rho'_3 = 2.06$). This network contains a spanning cluster which is shown in (b).

3 Percolation

For given values of the parameters, the probability Π of having a percolating cluster which spans the cell along the $x-$ direction is derived from N_r random realizations of the system; then, the value ρ'_c for which $\Pi = 0.5$ is estimated. Π and ρ'_c depend on five or four parameters as summarized by the formulae

$$\Pi(R'_{\rm m}, L', a, \mathcal{S}, \rho'), \quad \rho'_c(R'_{\rm m}, L', a, \mathcal{S}) \tag{7}$$

where $\mathcal{S}$ denotes the fracture shape and ρ' denotes any one of the dimensionless densities defined in (5) or simply ρ. For brevity, they will be often written as $\Pi(L', \rho')$ and $\rho'_c(L')$.

In the limit of large L', the fracture networks are expected to follow the standard percolation theory with the percolation threshold $\rho'_c(\infty)$ [35]

$$\rho'_c(L') - \rho'_c(\infty) \propto L'^{-1/\nu} \tag{8}$$

where ν is the critical exponent. In our estimations of $\rho'_c(L')$, the data for $\Pi(L', \rho')$ were fitted by a two-parameter error function of the form (Fig. 2)

$$\Pi(L', \rho') = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\rho'} \frac{1}{\Delta_L} \exp\left[-\frac{(\xi - \rho'_c(L'))^2}{2(\Delta_L)^2}\right] \, {\rm d}\xi \tag{9}$$

where Δ_L is the width of the transition region of $\Pi(L', \rho')$ which follows a scaling relation in the limit of large L'

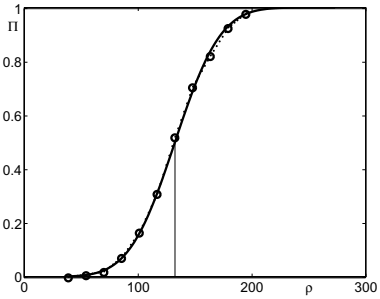


Fig. 2. The percolation probability Π versus the density ρ for networks of regular hexagons. Data are for $a = 1.5$, $L' = 4$ and $R'_{\rm m} = 0.01$. (•) corresponds to the numerical data, (——) to a fit by Eq. (9). The vertical line shows the estimated threshold $\rho_c(L)$.

$$\Delta_L \propto L'^{-1/\nu} \tag{10}$$

In practice, $\Pi(L', \rho')$ was evaluated from sets of 500 realizations, for about 10 values of the network density, evenly distributed in a range where Π varies from 0.05 to 0.95. Since there is a correspondance between ρ'_{21} and ρ'_3, for given values of $\mathcal{S}$, a and R_m, the same data sets can be used to determine $\rho'_{21c}(L)$ and $\rho'_{3c}(L)$. The 95% confidence interval is estimated to be about ± 0.04 in terms of $\rho'_{3c}(L)$.

The influence on ρ'_c of the four parameters in Eq. (7) was systematically studied. The results of this investigation will be presented elsewhere in full details [23]. We only state here the main result, which is that in the range $1.5 \leq a \leq 4$, $R_m \ll L$ and for (almost) any fracture shape, ρ'_c depends only on the domain size, and that in the limit of infinite domains, a unique value of $\rho'_c(\infty)$ applies in all cases. The independence on the various parameters is illustrated in the following examples.

In the example of Fig. 3, L' and a are kept constant, but the range of size and the fracture shapes varied. The networks contain hexagons, squares or triangles, or mixtures of hexagons with triangles or rectangles with a four to one aspect ratio. The upper set of curves shows that ρ'_c is indeed independent of R_m and $\mathcal{S}$. Note that the rightmost points are actually monodisperse networks. For comparison, the thresholds $\rho_c\langle R^3\rangle$, which do not include the shape factor $\langle v_{\rm ex}\rangle$ (see Eq. 5b), are also shown in the same figure and they are clearly much more scattered. It is the incorporation of $\langle v_{\rm ex}\rangle$ in the definition of ρ'_3 which unifies the results for the different shapes.

Conversely, the fracture shape (hexagonal) and the range of size (R'_m=0.1) are kept constant in the example of Fig. 4, whereas the exponent a and the domain size L are varied. It is seen that ρ'_c does not vary when a ranges from 1.5 to 4. However, a definite dependence on the domain size is observed, which corresponds to the well known finite size effects.

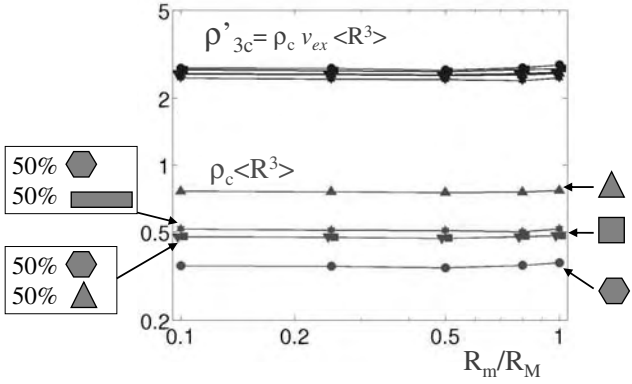


Fig. 3. The percolation threshold ρ'_{3c} for networks of fractures with various shapes and various size ranges. Data are for $L' = 6$ and $a = 1.5$. The upper set of curves is ρ'_{3c} and the lower one correspond to $\rho_c\langle R^3\rangle$.

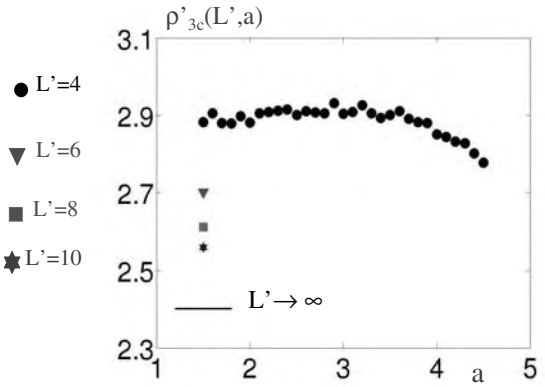


Fig. 4. The percolation threshold $\rho'_{3c}(L', a)$ for networks of hexagonal fractures with $R'_m = 0.1$, versus the exponent a, for various domain sizes L. The lower line is the extrapolation of the data for $a = 1.5$ when L' tends to infinity.

The data for increasing L' can be extrapolated for infinite systems by use of a classical technique. The combination of (8) and (10) shows that $\rho'_c(L) - \rho'_c(\infty)$ is proportional to the width Δ_L of the percolation transition zone. Hence, $\rho'_c(\infty)$ can be read on the vertical axis of the plot of $\rho'_c(L)$ versus Δ_L, which is shown in Fig. 5. The data for many cases, including various fracture shapes in monodisperse and polydisperse networks, are gathered in Fig. 5a. In all cases, the extrapolated values $\rho'_{3c}(\infty)$ fall in the narrow range

$$\rho'_{3c}(R'_{\mathrm{m}}, a, \mathcal{S}, L' \to \infty) = \rho'_{3c}(\infty) \approx 2.4 \pm 0.1 \tag{11}$$

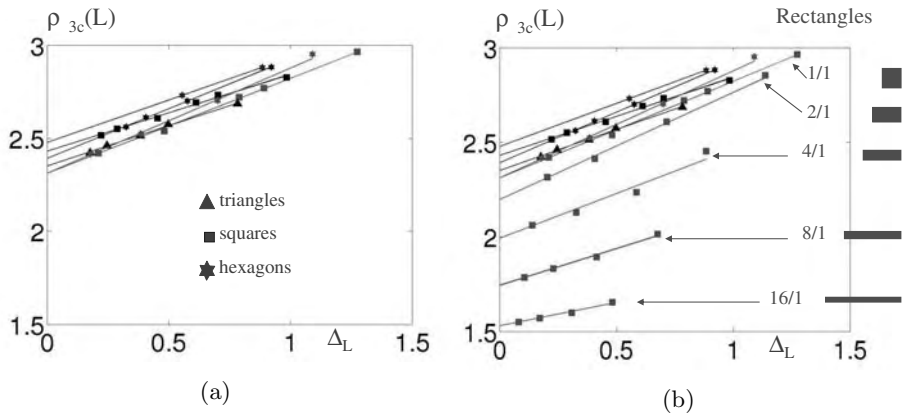


Fig. 5. The percolation threshold $\rho'_{3c}(L')$ for mono- or polydisperse networks of fractures with various shapes, versus the width Δ_L of the percolation transition. In (a), the fractures are hexagons, squares or triangles. $\rho'_{3c}(\infty)$ is the extrapolation for $\Delta_L \to 0$, which falls in the range of Eq. 11. Data for monodisperse networks of rectangles with aspect ratios from 1 to 16 are added in (b).

This applies for a variety of shapes, as well as for mixtures of fractures with different shapes (see Fig. 3).

However, when the polygons become elongated, the percolation threshold varies with the aspect ratio. Data for rectangles with aspect ratios up to 16 are shown in Fig. 5b. It appears that the threshold decreases significantly when the ratio of the rectangle length h to the rectangle width w increases.

This can be taken into account by using the shape factor $\eta = 4R/P$ of the fractures. This ratio is minimum for disks, with $\eta = 2/\pi \approx 0.637$, and it increases up to one when the shape deviates from circularity. It turns out that a quadratic correction in terms of η is very successful for the representation of the data for very different and irregular fracture shapes.

All the thresholds obtained in cells with $L' = 6$ and mono- or polydisperse size distributions with $a = 1.5$ or 2 and $R_{\rm m} = 0.1$ are plotted in Fig. 6 as functions of η. This includes networks of hexagons, squares, triangles, mixtures of hexagons with rectangles or triangles, and rectangles with h/w up to 16. The data are well fitted by the expression

$$\rho'_{3c}(L') = 2.69 \left[1 - 4\left(\eta - \frac{2}{\pi}\right)^2\right] \qquad (L' = 6) \tag{12}$$

The extrapolated data for infinite systems are also presented in Fig. 6, in comparison with the corrected version of Eq. (11),

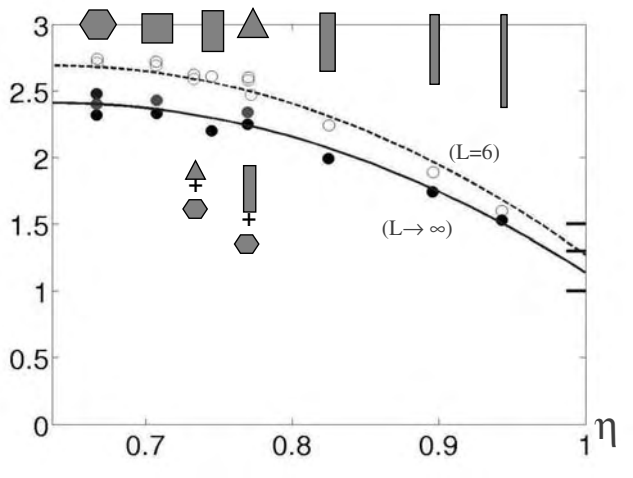


Fig. 6. The percolation thresholds $\rho'_{3c}(L'=6)$ and $\rho'_{3c}(\infty)$ for a variety of fracture shape and size distributions, in comparison with the expressions (12,13). The marks on the right are the predictions of [12, 13, 30] for infinitely elongated objects. The fracture shapes are indicated by the icons above or below the data points.

$$\rho'_{3c}(\infty) = 2.41 \left[1 - 4\left(\eta - \frac{2}{\pi}\right)^2\right] \tag{13}$$

In both cases, the deviations never exceed ± 0.1. The corrective term becomes significant, *i.e.*, larger than the error bar in (11), when $\eta > 3/4$, which corresponds for rectangles to aspect ratios larger than 2.

It can be noted that Eq. (13) predicts a threshold value 1.14 when h/w tends to infinity (*i.e.*, when $\eta \to 1$), which is in the range of the predictions 1.5 for prolate ellipsoids [13], 1 for capped cylinders [12] and 1.3 for elongated prisms [30], in the limit of infinite slenderness.

4 Permeability

At a local scale characterized by a typical aperture b, the flow of a Newtonian fluid within a fracture is governed by the Stokes equation. If b is assumed to be much smaller than the typical lateral extent $2R$ of the fracture, the flow at a scale intermediate between b and $2R$ is governed by the Darcy equation

$$\mathbf{q} = -\frac{1}{\mu}\sigma\ \overline{\nabla p} \tag{14}$$

where $\mathbf{q}$ is the locally averaged flow rate per unit width $\left[L^2T^{-1}\right]$, μ the fluid viscosity, $\overline{\nabla p}$ the pressure gradient, and σ $\left[L^3\right]$ the fracture conductivity coefficient. The mass conservation equation becomes

$$\nabla_S \cdot \mathbf{q} = 0 \tag{15}$$

where ∇_S is the two-dimensional gradient operator in the mean fracture plane.

The conductivity σ is taken to be constant over each fracture. However, we also consider the situation where σ is correlated with the fracture dimension, according to the power law

$$\sigma \propto R'^{\beta} \tag{16}$$

with $0 \leq \beta \leq 6$. The value of the exponent β depends on the physical origin of the fracture system as well as on its history [18, 20–22] but this question is out of the scope of this study

In most calculations, for non-periodic networks, prescribed pressures were applied over some inlet and outlet planes, while a no flux condition was applied over the other faces of the unit cell. The flow calculations are performed along the three directions $\alpha = x, y, z$. The corresponding permeability coefficients K_α are deduced from the flow rate Q_α via Darcy's law. A slightly different procedure was applied for periodic networks, where spatial periodicity is imposed for the flow field and the pressure gradient (see [1]). For the isotropic networks considered here, the statistical average of K_α does not depend on the direction, and it is denoted simply K in the following.

The numerical method of solution is described in details in [16]. It involves two steps. First, the fracture network is discretized; an unstructured triangulation of the fractures is built, which coincides with the fracture randomly located intersection lines. The mesh is characterized by the prescribed maximum edge length δ_M, which is set equal to $R_M/4$ in most cases. An hexagonal fracture with $R = R_M$ contains typically about 100 nodes and 140 scalene triangles. Small fractures with R of the order of δ_M or smaller contain at least 4 triangles. Let us give an example of the mesh sizes used in this study; a network with $a = 2.5$, $R'_m = 0.1$, $L' = 4$, and $\rho'_3 = 12$ contains approximately 4000 fractures with 100000 triangles and 50000 nodes. The pressure p is determined at each point of the triangular mesh, by solving the linear equations which result from a finite volume formulation.

For each set of model parameters, the flow simulations have been performed on $N_r = 25$ networks. The macroscopic permeabilities given in the following are always averages over these N_r realizations and over the three directions x, y and z. Non percolating networks with zero macroscopic permeability are also taken into account in the statistical averaging.

It is useful for the forthcoming discussion to first recall a theoretical result relative to networks of infinite plane channels with an arbitrary orientation distribution [25, 34]. If the surface area per unit volume for the fractures normal to $\mathbf{n}$ is $S(\mathbf{n})$, the permeability tensor is given by

$$\mathbf{K}_{Sn} = \sigma \int S(\mathbf{n}) \ (\mathbf{I} - \mathbf{nn}) \ \mathrm{d}\mathbf{n} \tag{17}$$

This result can be straightforwardly extended to the case when the fractures have different conductivities. When applied to isotropic networks, it yields

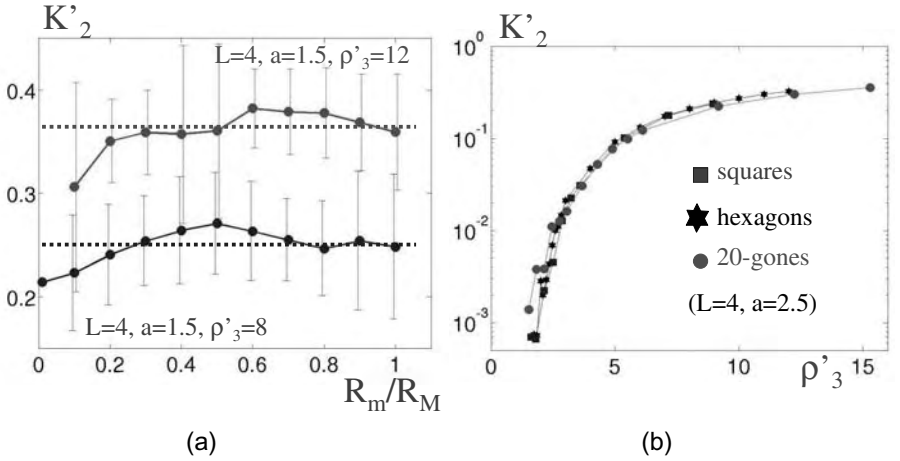


Fig. 7. The permeability of networks of hexagonal fractures, with $L' = 4$, $a = 1.5$, $\beta = 0$ and ρ'_3=8 or 12, versus R'_m (a). The error bars correspond to the statistical standard deviations. The permeability of networks of fractures with various shapes, for $L' = 4$, $a = 2.5$, $\beta = 0$ and R'_m=0.1 versus ρ'_3 (b).

$$K_{Sn} = \frac{2}{3}\rho\langle\sigma A\rangle, \quad \langle\sigma A\rangle = \int_{R_m}^{R_M} \sigma(R)A(R)n(R)\mathrm{d}R \tag{18}$$

This provides a reasonable scale for the permeabilities of the random fracture networks that we consider; therefore, we define the dimensionless permeability

$$K'_2 = \frac{K}{\rho\langle\sigma A\rangle} \tag{19}$$

The subscript 2 corresponds to the statistical moment of R used in the normalization (see Eq. 5).

As in the previous Section about percolation, we do not report here the details of the systematic investigation of all the parameters, which can be found elsewhere [24]. However, the main findings can be summarized in the simple statement that except for finite size effects when the density is near the percolation threshold, K'_2 is a function of ρ'_3 only. Again, let us illustrate this with a few examples.

The influence of the ratio R_m/R_M is tested in Fig. 7a, in two cases; it gives rise to variations of K'_2 without organisation and of an amplitude smaller than the statistical error bars. Fig. 7b shows data for various fracture shapes, which are very close together over a wide range of densities ρ'_3=2 to 16. More elongated shapes have also been tested in monodisperse networks. For instance,

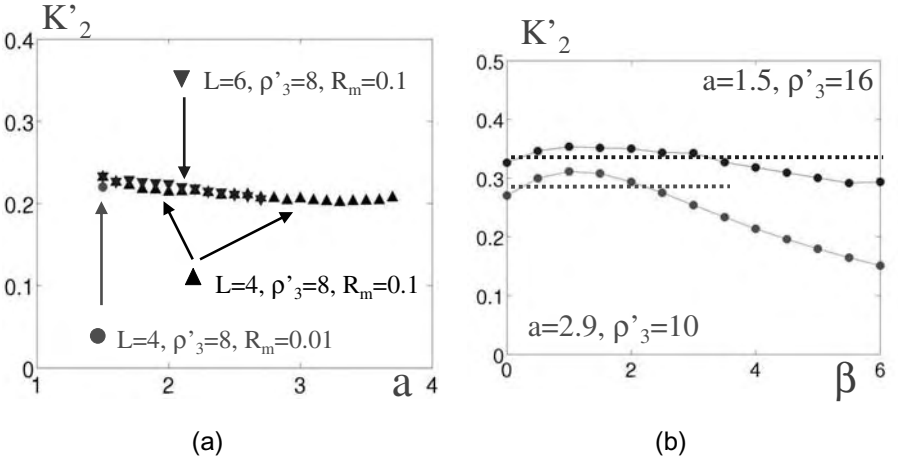


Fig. 8. The permeability of networks of hexagonal fractures, with ρ'_3=8, $\beta = 0$ and $L' = 4$ or 6,versus a (a). The permeability of networks of hexagonal fractures as a function of β, for (a=1.5, ρ'_3=10) and (a=2.9, ρ'_3=16) (b).

when ρ'_3=8 and $L' = 8$, the values of K'_2 for squares and rectangles with an aspect ratio 8 differ by less than 8%.

The effect of the exponents a and β is shown in Figs. 8a and 8b, respectively, in a variety of cases. K'_2 remains fairly constant, except when both exponents are large. In this situation, the population of fractures is dominated by the smallest ones, but their conductivities are vanishingly small. Hence, topological and flow percolations do not rely on the same paths, and a transition to a different behavior can indeed be expected.

In summary, a unified description of the macroscopic permeability K of polydisperse fracture networks with intermediate and large densities can be proposed as

$$K = \rho \langle \sigma A_p \rangle K'_2(\rho'_3) \tag{20}$$

which is a direct extension of the corresponding result of [16] about monodisperse fracture networks. The extensive term $\rho\langle\sigma A_p\rangle$ represents the volumetric area of fractures, weighted by the individual fracture conductivities. The dimensionless function K'_2 accounts for the network connectivity and it is fairly universal, as shown by the summary of our data in Fig. 9. K'_2 is a function of the same dimensionless density ρ'_3 that controls the network percolation, and incorporates the influence of the fracture shape and size distributions.

Most calculations in this work have been conducted for $a < 3$, and the model proves successful in this range even for non vanishing R_m. All the data for $0.01 \leq R'_m \leq 1$, $1.5 \leq a \leq 2.9$, $\beta = 0$ and various fracture shapes are

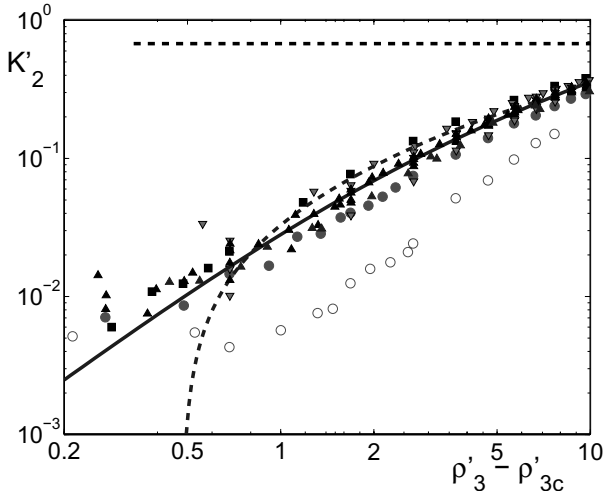


Fig. 9. The normalized permeability K_2' versus $\rho_3' - \rho_{3c}'$. Data for monodisperse hexagons (□) are from [16], with L'=4 to 10. Other symbols are for polydisperse networks of fractures with various shapes, with $R_m' = 0.1$, $L' = 4$, $a = 1.5$, 2.0, 2.5 or 2.9 and $\beta = 0$ (Δ), β =1.5 or 3 (∇) and β =6 (○). The data standing outside the set of points well described by Eq. (20) are for a=2.9 and $\beta = 6$ (open circles). The straight broken line is the Snow equation Eq. (18). The solid curve is Eq. (22). The broken curve is the prediction of Eq. (21).

represented by Eq. (20) within at most 20%, as soon as $L/R_M \geq 4$ and $\rho_3' \geq 4$. Exponents larger than 3 have not been systematically investigated but one case is considered in Fig. 8a; K_2' does not vary when a is increased up to 3.7. Finally, when $a > 4$, the volumetric surface area and the network connectivity are both controled by R_m and large fractures are uncommon. Hence it may be conjectured that such networks behave essentially like monodisperse networks of fractures with size R_m, and therefore that the model (20) is still applicable. A similar statement is possible for $a < 1$, since small fractures are then not numerous enough to play a significant role. Therefore, Eq. (20) would be a quite general and useful result, providing a reasonable estimate of the network permeability for any range and exponent of the fracture size distribution.

In the case of varying fracture conductivity σ, Eq. (20) holds for $a \leq 3$ and for moderate values of the exponent β in the scaling relation Eq. (16). It breaks down, however, when β increases up to 6.

For small densities, significant size effects are observed, which have not been systematically investigated. Still, we can formulate the general statement that they decrease when the exponent a increases, because the role played by small fractures becomes more important. For large densities, K_2' slowly tends to the value 2/3 predicted by Snow's model Eq. (18).

In view of the universality of Eq. (20), and of its high practical interest, it may be desirable to model it by an analytical formula, easier to use than the tabulated data in Fig. 9. Two such heuristic models are proposed here. The first one assumes, following [14], that the ratio $1/(1 - K/K_{Sn})$ increases linearly with ρ'_3 when ρ'_3 tends to infinity. A fit of all the data for $\beta = 0$ yields

$$K'_2 \approx \frac{2}{3}\left[1 - \frac{1}{0.1(\rho'_3 + 6.6)}\right] \tag{21}$$

The second one, which generalizes a model of [16], reads

$$K'_2 = \frac{0.1}{\rho'_3}\left(\rho'_3 - \rho'_c\right)^{1.6} \tag{22}$$

The two analytical formulas are plotted in Fig. 9. They are equally successful in representing all the numerical data for intermediate and large network densities, except for very large conductivity exponents β. Together with (20), they provide a general, simple and fairly accurate estimate of the permeability of polydisperse fracture networks.

5 Conclusion

Systematic 3d numerical simulations of percolation and flow in polydisperse fracture networks have been conducted, on the natural scale for the determination of upscaled properties, *i.e.*, on domain whose size exceeds the dimension of all the fractures. An appropriate percolation parameter ρ'_3 could be defined, which involves a shape factor $\langle v_{\mathrm{ex}} \rangle$ and the statistical moment $\langle R^3 \rangle$ of the fracture size. In these terms, the critical density is nearly constant over a wide range of shape and size distributions. A second order correction was also devised for the case of very elongated fracture shapes.

Regarding the permeability, the main findings can be summarized by the general expression (20), which involves a dimensional measure of the volumetric area of fractures and a universal function of the percolation parameter ρ'_3, for which two heuristic analytical models have been proposed.

Acknowledgements

Most computations were performed at CINES (subsidized by the MENESR). This work was also partly supported by the European contract Saltrans EVK1-Ct-2000-00062. These supports are gratefully acknowledged.

References

1. Adler P.M. and Thovert J.-F. (1999) Fractures and fracture networks, Kluwer Academic Publishers, Dordrecht

2. Balberg I., C.H. Anderson, S. Alexander, and N. Wagner (1984) Phys. Rev. E, **30**, 3933-3943
3. Berkowitz B. (1995) Math. Geol., **27**, 467-484
4. Berkowitz B. (2002) Adv. Wat. Resour., **25**, 861-884
5. Bonnet E., O. Bour, N. E. Odling, P. Davy, I. Main, P. Cowie, and B. Berkowitz (2001) Rev. Geophys., **39**, 347-383
6. Castaing C., M. A. Halawani, F. Gervais, J. P. Chiles, A. Genter, B. Bourgine, G. Ouillon, J. M. Brosse, P. Martin, A. Genna, and D. Janjou (1996) Tectonophysics, **261**, 291
7. Charlaix E., E. Guyon, and N. River (1984) Sol. State Comm., **50**, 999-1002
8. Consiglio R., D. R. Baker, G. Paul, and H. E. Stanley (2003) Physica A, **319**, 49-55
9. Dhar D. (1997) Physica A, **242**, 341-346
10. de Dreuzy J.-R., P. Davy, and O. Bour (2000) Phys. Rev. E, **62**, 5948-5952
11. de Dreuzy J.-R., P. Davy, and O. Bour (2002) Water Resour. Res., **38**, 1276, doi:10.1029/2001WR001009
12. Florian R., and Z. Neda (2001) oai:arXiv.org:cond-mat/0110067
13. Garboczi E. J., K. A. Snyder, J. F. Douglas, and M. F. Thorpe (1995) Phys. Rev. E, **52**, 819-828
14. Hestir K. and J. C. S. Long (1990) J. Geophys. Res., **95**, B13, 21,565
15. Huseby O., J.-F. Thovert, and P.M.Adler (1997) J. Phys. A, **30**, 1415-1444
16. Koudina N., Gonzalez Garcia R., Thovert J.-F. and Adler P.M. (1998) Phys. Rev. E, **57**, 4466-4479
17. Line C.E.R., D. B. Snyders and R. W. Hobbs (1997) J. Structural Geology, **19**, 687
18. Madadi M. and M. Sahimi (2003) Phys. Rev. E, **67** (2): Art. No 026309
19. Mecke K. R., and A. Seyfried (2002) Europhys. Lett., **58**, 28-34
20. Meheust Y. and J. Schmittbuhl (2000) Geophys. Res. Lett., **27**, 2989
21. Meheust Y. and J. Schmittbuhl (2003) J. Pure Appl. Geophys., **160**, 1023
22. Mourzenko V., J.-F. Thovert and P.M.Adler (2001) Transp. Porous Media, **45**, 89
23. Mourzenko V., J.-F. Thovert and P.M.Adler (2004) in preparation
24. Mourzenko V., J.-F. Thovert and P.M.Adler (2004) Phys. Rev. E, **69**, 066307
25. Oda M., Y. Hatsuyama and Y. Ohnishi (1987) J. Geophys. Res., **92**, B8, 8037
26. Odling N.E. (1993) In: *Hydrogeology of Hard Rocks, Memories of the XXIVth Congress of IAH, Oslo, 1993*, edited by Sheila and David Banks, p 290.
27. Odling N. E. (1997) J. Struct. Geol., **19**, 1257-1271
28. Robinson P. C. (1983) J. Phys. A, **16**, 605-614
29. Robinson P. C. (1984) J. Phys. A, **17**, 2823-2830
30. Saar M. O., and M. Manga (2002) Phys. Rev. E, **65**, 056131.
31. Sahimi M. (1995) Flow and Transport in Porous Media and Fractured Rock, VCH, New York
32. Scholz C.H., N. H. Dawers, J.-Z. Yu, M. H. Anders and P. A. Cowie (1993) J. Geophys. Res., **98**, 21951
33. Sisavath S., V. Mourzenko, P. Genthon, J.-F. Thovert, and P.M.Adler (2004) Geophys. International, **157**, 917-934
34. Snow D.T. (1969) Water Resour. Res., **5**, 1273
35. Stauffer D., and A.Aharony (1992) Introduction to Percolation Theory, Taylor and Francis, Bristol, PA
36. Yielding G., J. Walsh and J. Watterson (1992) First Break, **10**, 449

Acoustic diffraction patterns from fractal to urban structures: applications to the Sierpinski triangle and to a neoclassical urban facade

Philippe Woloszyn

Acoustic dept., Cerma Lab., UMR CNRS 1563, E.A.N., rue Massenet, BP 81931
F-44319 Nantes Cedex 3, France
philippe.woloszyn@cerma.archi.fr

Summary. The concept of fractal geometry, introduced by B. Mandelbrot has been explored in diverse areas of science, including acoustics [7] . The first part of this work relates the properties of far-field Fraunhofer diffraction region in wave acoustics for characterizing reflection on a self-similar structure. Therefore, the computation of the spatial Fourier transform of the Sierpinski triangle leads to its scattering intensity distribution, which describes its acoustical interference behavior. As a major application of this method, an urban facade scatter densitometry will be compared to acoustic measurements of the first reflections of its surface.The good agreement between computation and measurement allows to validate the spatial Fourier transform of the facade as an indicator of acoustic scattering.

1 Introduction

This work aims to define and validate a diffraction pattern computation model for various geometries, fractal (Sierpinski triangle) and regular (urban facades) [9] . For that, twice analytical and numerical approaches of the far-field diffraction, based on electromagnetism diffraction analogy [8], will be proceeded. In an acoustic point of view, the building doesn't reflect sound in a purely specular way, because of the irregularity of its surface (decorative elements, windows, balconies,...), which dimensions are comparable to sound wavelengths. Applying their approaches to an urban facade, both Bragg's law and Fraunhofer's model will describe the field conditions for constructive and desctructive interferences, through calculation of single and multiple scatter.

2 Diffraction on a plane

Diffraction patterns are most obvious when the incoming waves are coherent. That means the phase of the sinusoidal pattern of the sound field is deter-

ministic. Moreover, wave coherence has two domains: spatial and temporal. Consequently, the field phase information at some points in space and/or time determines the phase at other points in the same space and/or time. For spatially coherent sound, the phase difference of the measured field at two different points in space at the same time separated by a vector distance which is constant for all times. If the phase difference measurement at the same location and at two different times is the same for all points in space, the sound is temporally coherent.

2.1 Interferences production on an indent

Diffraction of a wave by a periodic structure is a consequence of phase differences that result from constructive and destructive interferences. This phenomenon can be occured when the waves pass through a periodic structure, if the repeat distance is similar to the wavelengths.

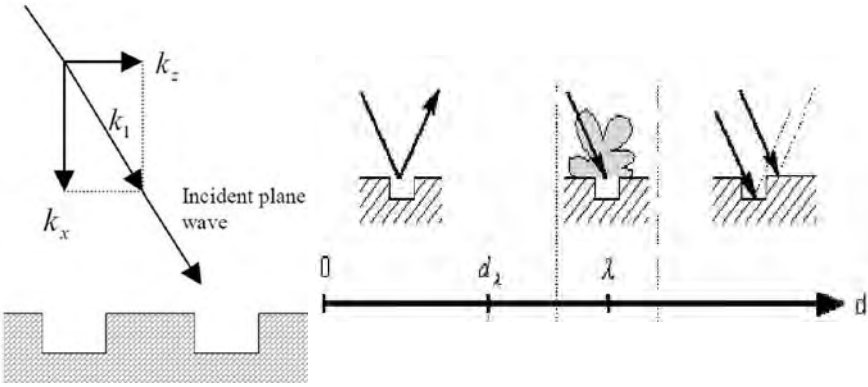


Fig. 1. Incoming wave vector on a regularly indented plane.

If sound from a single source is reflecting into two indents at different heights, Rayleighs principle indicates that the sound will spread and recombine into different wavevectors [12] .

2.2 Fraunhofer diffraction patterns

The mathematical relation between the shape and the wavesize output, relative the spatial structure input, is a Fourier transform [1]. The diffraction pattern is then defined as the Fourier transform of the structure angular density. So, the scattered intensity can be expressed as a function of the structure factor of the reflection geometry, as:

$$I(r) = \rho(S_n(r))(F_x(r)) \tag{1}$$

where ρ is the scattering density, $S_n(r)$, represents the structure factor of the structure, and $F_x(r)$, its form factor. In case of an iterative structure such as fractals, the form factor $F_x(r)$ is defined by the Fourier transform of the initiator $H_1(r)$, and the structure factor term S_n reveals the iterative procedure of the analyzed structure [5] . The structure factor takes into account the inter-plane scattering contribution, that means constructive ans destructive interference processes. Furthermore, it can be interpreted as the Fourier transform of the diffractance function $g(r)$. As a consequence, at large distances from the sound source, the far field or Fraunhofer diffraction region involves the diffracted sound pattern as a variable proportional to the reciprocal of structure dimensions. Then, for the same energy input pattern, the diffraction pattern in the Fraunhofer region has the following angular distribution form, involving the incoming sound wavelength λ scattering at angle α:

$$g(r) \propto \int_i f(\alpha) exp[-\frac{2i\pi\alpha r}{\lambda}]d\alpha \tag{2}$$

The incident plane wave scattering effect is the creation of far-field secondary waves. These waves, arising from each point on the structure, travel in all directions to rise with complex distributions of amplitude and phase. The complex amplitude at this point is obtained by adding the individual contributions from primary and secondary sources of amplitude $g(r)$, concerning the different path lengths of the travel to the reception point. So, the contribution at a distance r is specified through the following intensity distribution, which is a major parameter of the Fraunhofer diffraction pattern:

$$I(r) = |\int_i g(r) exp(-i\nu r)dr|^2 = |S_n|^2 \prod_i \frac{sin2\pi\lambda r}{2\pi\lambda r} \tag{3}$$

where ν is the spatial period of the indented structure. Previous equations 2 and 3 show identity between the measured intensity $I(r)$ and the square of the modulus of the structure diffractance Fourier transform. The structure factor S_n is the Fourier transform of the scatterers of equal strength on all points at distance r from the diffraction plane [2] .

3 Application to the Sierpinski triangle

We will apply now the Fraunhofer diffusion formalism to a known two-dimensional structure, the Sierpinski triangle, in order to explore its behaviour under a wave sollicitation. For that, an iterative term will be introduced into its structure factor.

3.1 Iterative structure factor

As the structure factor for the Fourier Transform of a level n multiplicative signal can be written as either a result of n periodic functions $H_i(r)(i =$

$1, 2, ..., n)$ with frequencies $n_i = 1/\nu_i$, or a result of n scaled replicas of the structure factor of the first level $H_1(r)$ [13], equation 4 yields:

$$S_n(r) = \prod_{i=1}^{n} H_i(r) = \prod_{i=1}^{n} H_1(r/\nu_i) \propto \prod_{i=1}^{n} T_n^D(r) H_0 \tag{4}$$

This expression of the structure factor involves the following fractal translation-dilation operator applied to the initiator H_0 as:

$$T_n^D(r) = (1 + \sum_{i=1}^{n} T_{-\overrightarrow{k_i}}) C_{1/s} \tag{5}$$

where the Sierpinski triangle is generated by the following contraction factor s and translation vector $\overrightarrow{k}$:

$$\begin{aligned} s = 1/2, \{\overrightarrow{k}\} = \{&(1/4, 0), (3/4, 0), (0, 1/4), (0, 3/4), \\ &(1/4, 3/4), (3/4, 1/4), (3/4, 3/4)\} \end{aligned} \tag{6}$$

Following the general calculation theorem of the fractal dimension D for iteratively constructed fractals given by [4] and [14], fractal dimension D is the solution of the following equation:

$$\sum_{i=1}^{k-1} s_i^D = 1 \Longrightarrow D = \frac{ln(k-1)}{ln(1/s)} \tag{7}$$

This fomulation defines the fractal dimension of the Sierpinski triangle as $D = log3/log2$. At each iteration level, we divide the map into four square units, in order to substract the bottom right part for each unit. When the initiator H_0 is a square and the generator H_1, the square without its bottom-right quadratic part, the Sierpinski triangle set is obtained as seen figure 2.

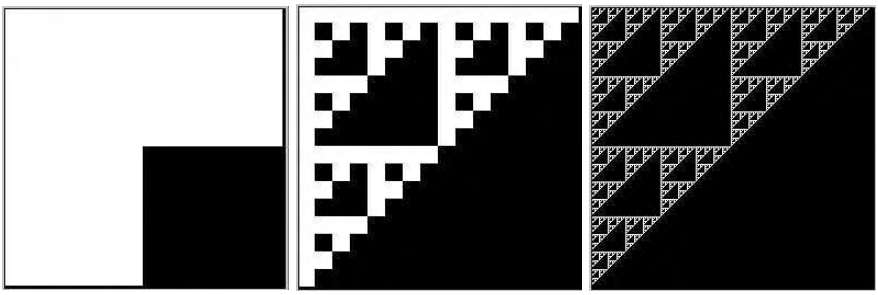


Fig. 2. Sierpinski triangle Generator, and 8 th & 16 th iteration levels.

The structure factor is constructed as a product of n scaled replicas of the periodic signals with scaling coefficients ν_i as seen equation 4. Similarity in

the construction of the multiplicative reflection structure itself and its Fourier Transform is a consequence of the reciprocity between structure and wave interaction, under the conditions of reciprocal dimensioning (wavelengths vs. reflection plane distances). It is known that such multiplicative cascades for real non-negative values of the generator matrix produce scaling symmetry for regular fractals [6]. As an obvious example of scaling symmetry, the Sierpinski triangle answers the well-known iteration law $g(r/n) = n^D g(r)$.

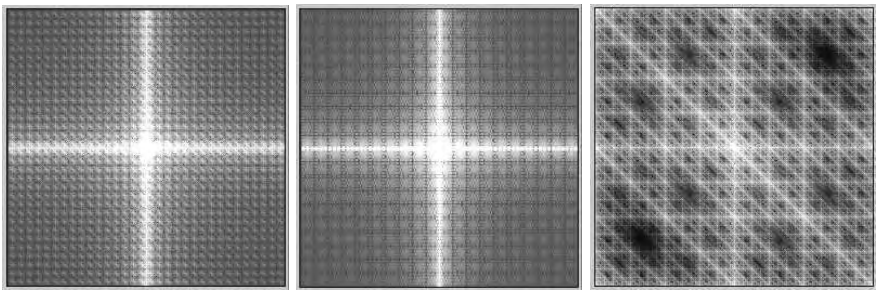


Fig. 3. Space Fourier Transform of the Sierpinski triangle Generator, 8 th & 16 th iteration).

As an indicator of the indentation frequency, the Fourier transform discriminates clearly the structure of the reflection surface, revealing the spatial occurrences of the roughness peaks. The Fourier transform of the structure shows quasiperiodic behaviour of its structure factor, at every iteration of the shape. In addition, we can note that, for high iteration levels, the fractal structure becomes invariant to its Fourier transform. The angular scattering distribution function is then defined through the structure factor computation of the surface, and indicates the fractal scattering behaviour for a particular direction of the incident wave.

3.2 Scattering functions

As scattering intensity is known as the Fourier transform of the multiplicative reflection structure of the Sierpinski triangle, the scaling property of the scattered field is constructed through the same multiplicative procedure as the original structure, which leads to the following expression of the scattered intensity field:

$$I(r) = n^D I(r/n) = \sum_{i=0}^{n-1} H_i(r) I_0(r - i\lambda) \tag{8}$$

We can discriminate an anomalous quasiperiodic behaviour of the scattering angular and its spatial factors, even when the characteristic Fraunhofer

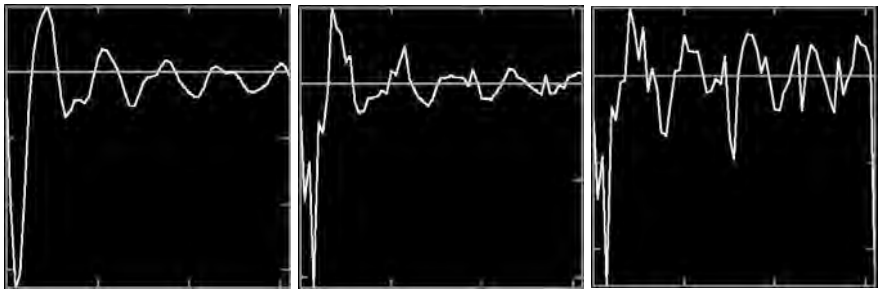


Fig. 4. Sierpinski triangles scattering angular distribution functions at $\pi/2rd$ for $\nu = \lambda$ (structure distance vs. power density function).

diffraction region pattern remains obvious. The angular density function is remaining quasi-periodic at any distance from the triangle, signing a quasi-perfect diffusive structure. In that case, the angular distribution of propagating energy won't depend on the distance from the plane structure. Moreover, the spatial scattering distribution function offers a tridimensional representation of the scattering distribution function, which presents the angular-distance 3-D map distribution of the surface's scatterers figure 5.

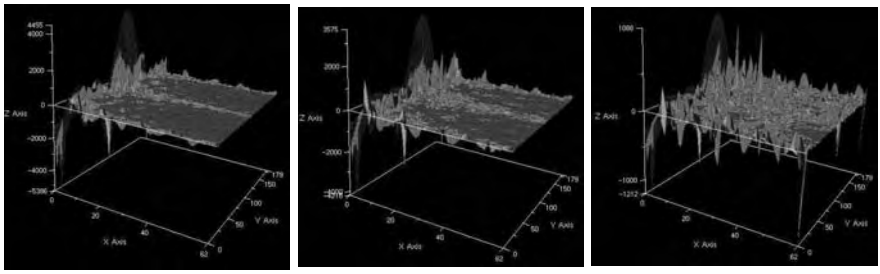


Fig. 5. Sierpinski triangles spatial scattering distribution functions (structure distance vs. diffraction angle vs. power density function).

Results displays a high spatial diffractance persistence for high-level iterated Sierpinski triangle. This behaviour traduces the capacity of the Sierpinski structure diffusion at high iteration levels. Spatial scattering distribution function data also provides scattering intensity through a cross-distribution plot, as presented figure 6 for the three iteration levels of the triangle.

As expected, the Sierpinki triangle far field scattering intensity cross-distribution shows self-similar diffraction patterns for several spatial frequency bands: after propagation of a coherent plane wave through this fractal grating, the computed acoustic field obtains self-similar properties. This diffraction pattern may be interpreted as the spatial filter response of the analysed structure. The location of mainlobe peak notifies in which direction we can

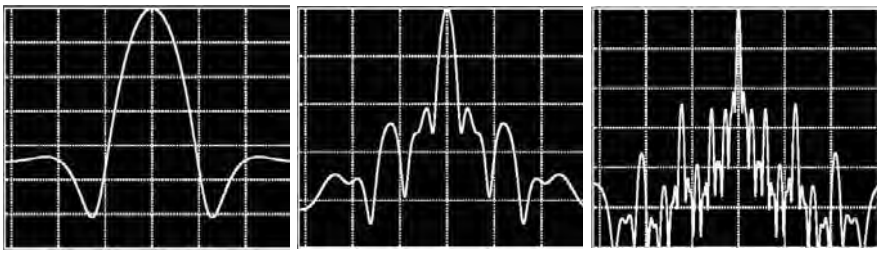


Fig. 6. Sierpinski triangles scattering intensity cross-distributions (diffraction angle vs. power density function).

get maximum response of the spatial structure from the acoustic solicitation. Mainlobe (highest peak) is then similar to the pass-band of a spatial filter, which diffracts acoustic energy in these particular directions [3] .

4 Application on a facade scattering characterization

4.1 The urban facade model

The spatial configuration is measured by a numerical 3-D model for a neo-classical facade of an urban street of Nantes (France), belonging to a 19th urban morphology type, with windows, doors, and freestone casting. One of the main characteristics of this type of architecture is the relative exuberance of its facade structure, which conforms to the neo-classical composition.

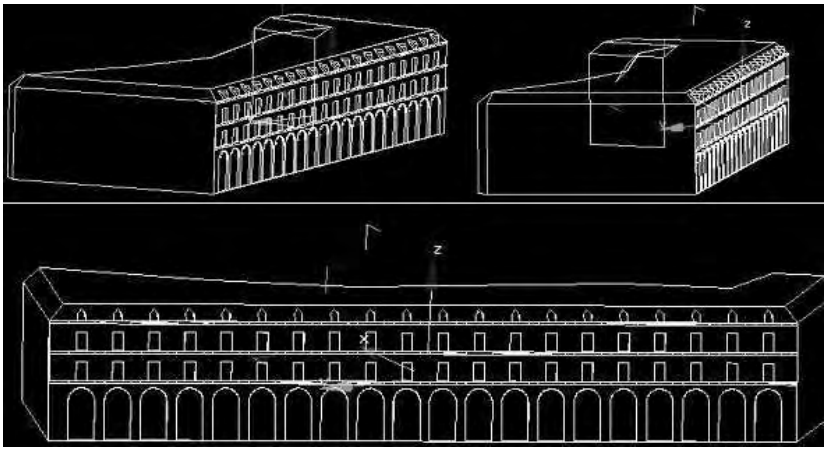


Fig. 7. Facade 3D model. (simplified)

4.2 Facade's vertex densitomentry

A geometrical analysis was carried out through the Minkowski mathematical operator, in order to compute the dilation factor of the structure. Results provides indicators of the facade's complexity, such as its angular distribution function and its spatial scattering function as shown figure 8.

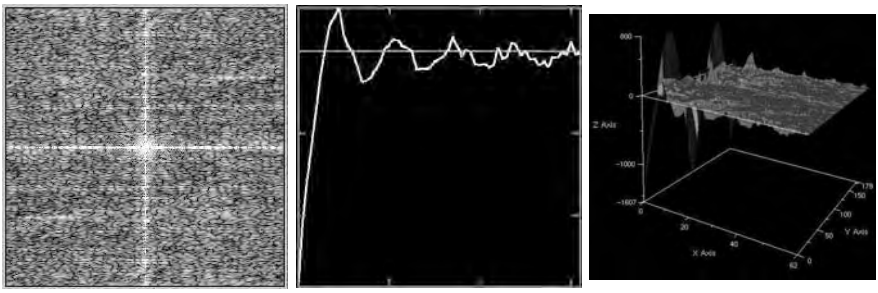


Fig. 8. Fourier transform (1), angular distribution function (2) and spatial scattering function (3) of the facade indented surface.

As an indicator of the indentation frequency, the Fourier transform discriminates clearly the structure of the facade, which reveals the spatial occurences of the roughness peaks. The angular distribution function of vertexes is defined through the structure factor computation of the surface, and represents the facade scattering behaviour for a particular direction of the incident beam. Moreover, the spatial scattering function allows a tridimensional representation of the angular distribution function, with displaying the distribution of the surface's scatterers along every incidence angle of the acoustic source. The facade scattering intensity cross-distribution is then computed for each incidence angle of the surface, for localisation lengths from 0.05 to 10 m, which corresponds to frequencies ranging from 25 Hz to 8 kHz. These densitometries correspond to the characteristic directions of scattering, through the calculation of the density distribution ρ for each incidence angle.

Global polar responses displayed 9 show a globally decreasing diffusivity with increasing localisation lengths. This reveals a bilobe distribution structure of the biggest scatterers, a pseudo-Gaussian for middle-sized ones and very characteristic peaks for high frequency roughness. The angular evaluation of the vertex distribution shows azimutal densitometries due to interreflexions of the corners and the freestone casting along three windows depth, in agreement with the lateral active diffraction zone composition [10].

4.3 Experimental validation: In situ measurements

In order to validate this geometrical scattering characterization model, we attempted to define a new method of measurement, applied to the neoclassical

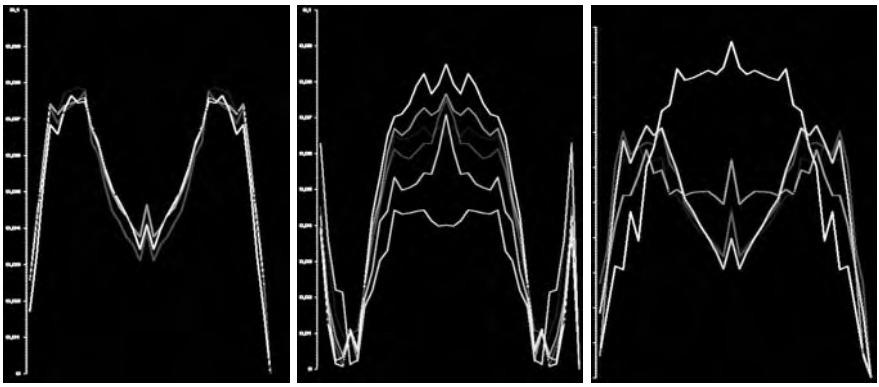


Fig. 9. urban facade scattering intensity cross-distributions (diffraction angle vs. power density function).

facade, which corresponds to the analyzed 3-D model presented figure 7. With exploiting a maximum-lenght sequences stimulus (MLS) signal, we obtained the facade impulse response, treated in order to pull the incident wave away from the rest of the signal by time windowing. The content of this window is analyzed in frequency domain, using the Fourier transform of the acoustic signal, with provides its Transfer Function. The reflection law is carried out by several positions of source and microphone. [11].

Fig. 10. In Situ experimental system.

It is noted that all measured reflection laws verify the specularity mode at low frequencies, but a diffuse reflection behaviour is observed for high frequencies. Figure 11 shows the impulse response provided with a normal positioning of the microphone successively in front and at the center of the neoclassical surface. After the direct contribution peak A, the stonework (B), the windows (C) and the guardrail (E) scattering contributions can be discriminated.

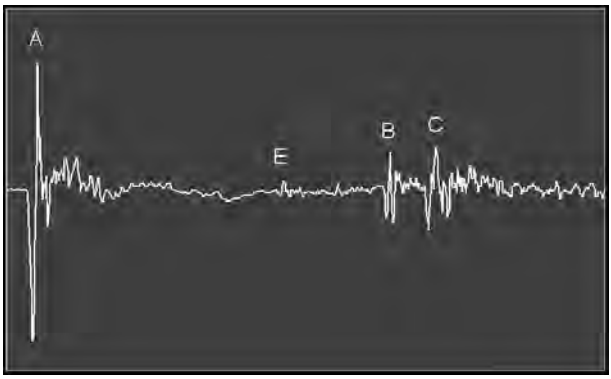


Fig. 11. Impulse response from the neoclassical facade.

5 Results comparison and discussion

Scattering intensity cross-distribution is then compared to experimental measurement results at each incidence angle. The measured reflection laws for each specified frequency band describes the early reflected sound energy level (the first 10 milliseconds backscattered signal) for each reception angle varying from 15 degree to 165 degree. Agreement between geometrical analysis figure 9 and metrological results figure 12 is clearly readable through the behavior of the whole spectrum, especially for high frequencies which diffusivity peaks emerges in the incidence direction and at characteristic angles from 60 degree to 75 degree (symetrically 105-170 degree), and for grazing angles from 15 degree to 35 degree (145-165 degree).

6 Conclusion

It has been shown that diffusion capacity of a complex plane can be directly computed from the Fourier transform of its spatial structure. Applied to a fractal plane structure, the resulting spatial intensity distribution reveals the self similar behaviour of the scattered acoustic field angular distribution. Applied to an indented surface of an urban neo-classical facade, measurement

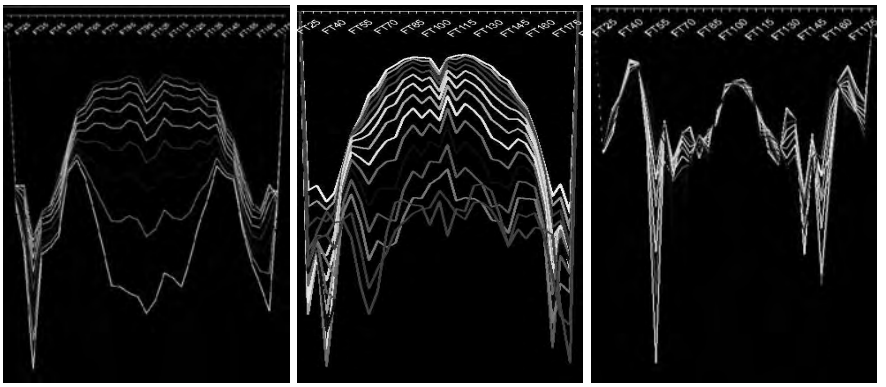


Fig. 12. Measured reflexion laws. Sound pressure level vs incidence angle.

results and intensity cross-distribution computation shows a good agreement for frequencies from 15 Hz to 6 kHz at non-grazing incidence angles. Moreover, this Fraunhofer diffraction region characterisation allows scattering prediction for many types of spatial configurations, under the condition of interferencial acoustic field.

For developping the architectural design tools for urban acoustics, this method allows a good evaluation of the acoustical reaction of an urban surface by computing the architectural geometry characterization. This confirms the diffusion process as a geometrical-dependant phenomenon, from micro- (optics) to macro-scale (acoustics). Furthermore, it enables a morphological computation of a spatial structure to predict its far-field diffusion behaviour, especially concerning the built structure influence on urban acoustics.

References

1. Hollander A. Dynamic structure factor in a random diffusion model. *Journal of Statistical Physics*, 76(5:6), 1994.
2. Allain C. and Cloitre M. Spatial spectrum of a general family of self-similar arrays. *Phys. Rev. A*, 36(12):5751–5757, 1987.
3. Hamilton M. F. and Blackstock D.T. *Nonlinear Acoustics*, chapter Sound Beams, pages 233–261. Academic Press, Boston, 1988.
4. Barnsley M. *Fractals Everywhere*. Academic Press, Boston, 1988.
5. Giona M. Analytic expression for the structure factor and for the moment-generating function of fractal sets and multifractal measures. *J. Phys. A*, (30):4293–4312, 1997.
6. Umberto M., Bettolo M., and Alberto P. Domain growth on self-similar structures. *Phys. Rev.*, 55(2), 1997.
7. Benoit Mandelbrot. *The Fractal Geometry of Nature*. Freeman, 1982.
8. Beckmann P. and Spizzichino A. *The Scattering of Electromagnetic Waves from Rough Surfaces*. Artec House, Norwood, 1987.

9. Woloszyn P. Is fractal estimation of a geometry worth for acoustics? In Miroslav M. Novak, editor, *Emergent Nature*, pages 423–425. World Scientific Publishing Singapore, 2002.
10. Woloszyn P. Geometrical scattering indicators for urban sound diffusion. *Ultragarsas*, 48(3):102–107, 2003.
11. Woloszyn P., Suner B., and Bachelier J. Angular characterisation of the urban frontages diffusivity factor. In *17th International Congress of Acoustics*, Rome, 2001.
12. Lord Rayleigh. *The Theory of Sound*. Dover, New-York, 1945.
13. Alieva T. and Calvo M-L. Paraxial diffraction on strucures generated by multiplicative iterative procedures. *J. Opt. A: Pure Appl. Opt.*, (5):324–328, 2003.
14. Vicsek T. *Fractal Growth Phenomena*. World Scientific, Singapore, 1989.

Turbulent $k - \epsilon$ model of flute-like musical instrument sound production

Rolf Bader

University of Hamburg, Institute of Musicology, Neue Rabenstr. 13, 20354 Hamburg, Germany R_Bader@t-online.de

Summary. The sound production of flute-like musical instruments like the transvers flute is governed by the coupling between the mouth or embouchure hole in which the flute player blows and the flute tube. Here the flute tubes eigenfrequencies forces the self-sustained oscillation of the generator region at the blowing hole into the tubes resonance frequencies. This paper supposes an explanation for this behaviour. Experiments show a very small amount of energy supplied by the players blowing actually getting into the tube of about 3.5%. So most of the air flow is blown into the room outside the flute. The modelling of the flutes presented here shows a turbulent description of the process as consistent with the experimental findings. The flute tube, which forces the flow in its direction leads to a large directional change of the flow, which results in a strong turbulent viscous damping of the system. So there is strong evidence, that in the nonlinear coupled flute system of blowing and tube the tube forces the blowing system in the tubes eigenfrequencies because the tubes air column is much less damped than the generator region at the soundhole.

1 Flute-like instrument sound production

Flute-like instruments, like the transvers flute, are blown instruments [Fletcher and Rossing 2000]. The player places his/her lips near the so called embouchure hole at a precise distance to that hole. This distance is important for the player, as just in a certain distance range, he is able to produce a sound at all. But it also influences the exact pitch being produced and the tone color of the musical sound [Benade and French 1965] [Coltman 1973]. The flute player then blows with a certain pressure to the lip of the embouchure hole at one of its sides. This lip or cut in transvers flutes is increased in thickness above the normal wall thickness of the flute of about $1mm$ to up to about $5mm$. It is much easier for the player to handle an instrument with an increased embouchure hole lip thickness in terms of production of sound and in its control.

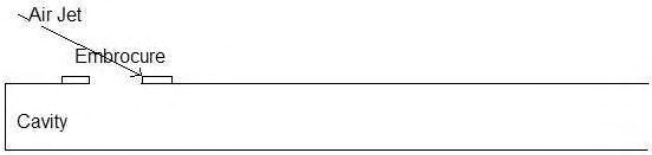


Fig. 1. Schematic view of a flute-like instrument. The air jet out of the players mouth hits the embouchure lip or cut. The embouchure wall height is enlarged compared with the tubes thickness. The tube is closed at the left and open on the right end (with a boundary condition of zero pressure). The cavity on the left end of the tube plays an important role in the tuning of the overblown pitches of the flute (which is not a topic of the present paper).

If the correct pressure is not applied with the flute, no sound is produced [Coltman 1968] [Coltman 1969]. This is because the impedance Z of the flute defined via the blowing pressure p needed to achieve a certain flow v like

$$Z = \frac{p}{v} \ . \tag{1}$$

So if high pressure is needed to get a certain flow, the tube 'resists', it has a high impedance. This impedance of the flute is complex, means there can occur phase shifts between pressure and flow.

Figure 2 shows the impedance behaviour of the flute-like instruments [Coltman 1968] with respect to blowing pressure. As the spiral gets larger, the pressure increases. It is measured with an artificial driving mechanism at the open end of the flute, applying different sinusodial frequencies to the tube. The applied pressure then corresponds to a measured flow in the tube. The air inside the tube is damped out via an acoustic resistance in the tube (i.e. small glass capillaries) so that the air column in the tube can not go into eigenfrequencies. The impedance has a real and an imaginary part. The real part corresponds to the resistive (positive) or generative (negative) behaviour of the tube. The imaginary part tell us about the phase shift between pressure and flow. So the only possible operating point of the flute is the largest negative point on the real axes. Here, the generation is highest. There is no phase shift between pressure and velocity. So the curve shows, that there is just one possible pressure region to generate the desired pitch, the player must blow with this precise pressure to produce the sound. Of course this region changes a bit for each pitch, so the flutist must change the blowing pressure while playing melodies.

But despite this generation process, in contrast with organ flute pipes, where the blowing around a lip produces a self-sustained oscillation [Dequand

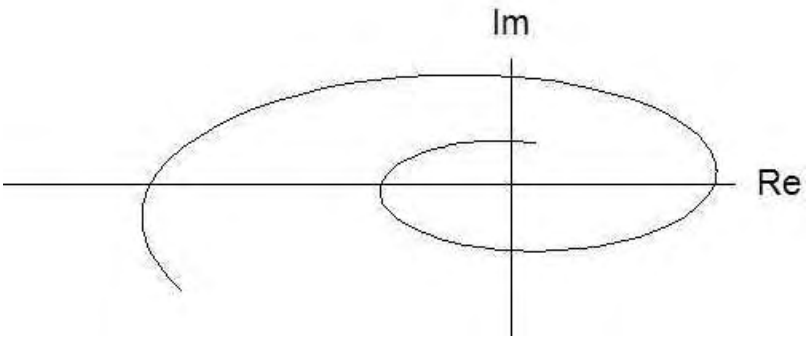


Fig. 2. Complex impedance of a flute-like instrument depending on the blowing pressure, increasing as the spiral gets larger. The positive range of the real axes corresponds to a resistive, the negative real axes to a generative behaviour of the flute. The positive range of the imaginary axes corresponds to an inductive, the negative to a capacitive phase difference between pressure and flow. So the main operating point of the flute is at the negative (or left) real side near the zero point of the imaginary axis without phase shift between pressure and velocity. This point corresponds to a fixed blowing pressure.

2000] [Miklós and Angster 2000] [Ségoufin et al. 2000], here the oscillation is determined by the length of the flutes tube, it depends on the open finger holes. If the sound production was the same as in the organ pipe, just one pitch could be played with the instrument. So even as there is a self-sustained oscillation at the hole embouchure of the flute, it is overridden by the travelling pressure pulse along the tube, which determines the pitch of the played sound (for a detailed descripion of the self-sustained oscillation of a flute see [Coltman 1968].

So the process of sound production has to be described as follows. The player blows into the instrument, in which a pressure pulse travels up to the first open finger hole. Here, because of a discontinuity of the boundary conditions at the finger hole, the pressure impulse is reflected and returns to the embouchure hole. Now this change in pressure in the hole forces the air jet to blow over the hole for a short amount of time. Then, with the returned pressure pulse gone, the jet blows into the tube again and produces a new impulse, which travels along the tube as the first one did. The fact, that there is just a small pressure blowing range for which this process can happen means, the two processes, the generation and the travelling impulse, have to fit.

The comparison between the transverse flute and the organ flue pipe sound production may be analog to the relation between reed instruments, reed blown instruments (saxophone, clarinet) on the one side and reed organ pipes on the other. Here, again we have i.e. the saxophone pitch being determined

by the length of the tube, while the reed organ pitch depends on the oscillation of the reed itself. The organ tube is just a resonator, while in the reed blown instruments the reed is just a generator, which eigenfrequencies do not determine the instruments pitch. The reason here is one of the principles of self-organisation, which in Synergetics is called the slavery-principle. If two oscillators are nonlinearly coupled - here the reed and the tube - then the pitch of the hole system is determined by the one, which is less damped. The other oscillator loses its eigenfrequencies completely and is now forced to vibrate with the frequency of the other. In the saxophone, the reed is very stiff and so much more damped than the air column in the saxophones tube [Aschoff 1936]. With the reed organ pipe, the reed is much more freely vibrating. One can compare this by an easy test. If the fixed reeds of the saxophone (or clarinet) and of the organ are displaced by hand and set free again, the saxophone reed will stop vibrating immediately, while the organ reed can still be heard after one or two seconds. So the organ reed is much less damped and so wins the game and forces the organ pipe air column to resonate in the reeds frequency.

Now in the case of flute instruments and organ flue pipes, there may be a similar concept underlying. As these systems are so complicated that an analytical solution is not always possible, we many have a look at the damping of the systems. Here we concentrate on the damping of the transverse flute and would expect it to be high in the region of sound generation, the embouchure hole.

Experimentally, it has been shown, that in the sound production of the transverse flute, where the air jet blows into the tube at the embouchure lip or cut, most of the flow is distributed in space outside the flute [Coltman 1968]. Just a small part actually finds its way in the tube, about 2.4% of the total blowing energy [1]. This is necessary to keep the process going. If too much flow would travel along the tube, the pressure impulse reflected at the finger hole had no change to reach again the embouchure region and therefore the produced pitch would purely be determined by the blowing pressure of the player, which is not the case. Now the tube of i.e. a transvers flute is about $30mm$ in diameter and the embouchure hole - normally elliptical - is about $12mm$ wide in its larger side.

So the question here is, if a normal Navier-Stokes model fits to describe the air jets splitting into the part going into the tube and the one going outside or if a turbulent model may lead to better results. If a large turbulent eddy viscosity would be found, there were a strong evidence for the similar situation as with the reed instruments, where the pitch is determined by the less damped part of the coupled oscillator system, here the air column and

[1] The radiated energy is even less. Only about 3.5% of the energy in the tube is radiated at frequencies around $440Hz$ and can then be heard. If we multiplicate this with the blowing efficiency, only about .0008% of the blowing energy produced by the player actually becomes radiated sound

a reason would be found, why the lip blowing system of the transverse flute does not lead to a self-sustained oscillation as with the flue organ pipes. So a numerical solution of a turbulent $k-\epsilon$ model is compared with a Navier-Stokes model.

2 The $k-\epsilon$ model

The $k-\epsilon$ model of turbulence is derived from the assumptions Kolmogorov made about turbulent eddy energy flow [Kolmogorov 1941]. If a laminar flow becomes turbulent, energy from the laminar flow starts to form large turbulent eddies. These eddies split into smaller ones, thus continue the energy flow from the larger to the smaller eddies. As the velocity of smaller eddies increase, there must exist a smallest possible eddy length, under which even smaller eddies will be dissipated immediately because of viscosity. So an energy cascade can be assumed from larger to smaller eddies. Kolmogorov proposed two variables for the turbulent energy k (mass weighted) and the turbulent energy dissipation ϵ. So in terms of units the mass weighted turbulent energy is

$$k = \left[\frac{length^2}{time^2}\right] \tag{2}$$

and the energy dissipation, as energy loss in time is

$$\epsilon = \left[\frac{length^2}{time^3}\right] \ . \tag{3}$$

From these assumptions, the energy of eddies of size r is

$$E(r) \sim (r\ \epsilon)^{2/3} \tag{4}$$

because of dimensional analysis. So the energy can be assumed to grow as a $2/3$ power law. Transferred into fourier space, this becomes a $-5/3$ power law which could be verified experimentally [Saddoughi and Veeravalli 1994].

Second, the $k-\epsilon$ model is a statistical model. It averages over time. We are not considered with the large variety of eddies in the exact fine structure of the turbulence, but instead assume the flow $\hat{u}$ to consist of a mean flow U and a fluctuating flow u as

$$\hat{u} = U + u \ . \tag{5}$$

So the normal Navier-Stokes equation NS is

$$\partial_t\ \hat{u}_i + \hat{u}_j\ \partial_j\ \hat{u}_i = -\frac{1}{\rho}\ \partial_i\ p + \nu\nabla^2\ \hat{u}_i \tag{6}$$

with the incompressibility condition

$$\partial_t \ \hat{u}_i = 0 \tag{7}$$

with $i = 1, 2, 3$, the pressure p, the fluid density ρ and the viscosity ν. As we just discuss the generator region, we are only concerned with flow. The acoustical behaviour of the system could later be added in two ways. Either the compressible NS is used or over the whole geometry an acoustical differential equation is used, taking into account the pressure distribution and being coupled to the NS. As the acoustical behaviour of the flute is well known [Fletcher and Rossing 2000] we concentrate on the flow only.

Now the NS becomes the Reynolds-Averaged-Navier-Stokes equation RANS

$$\partial_t \ U_i + U_j \ \partial_j \ U_i = -\frac{1}{\rho} \ \partial_i \ p + \nu \nabla^2 \ U_i - \partial_j \ \overline{u_j \ u_i} \tag{8}$$

again with the incompressibility condition

$$\partial_t \ U_i = 0 \ . \tag{9}$$

Note, that the RANS equation differs from the NS equation just in the last term on the right hand side, the Reynolds stress tensor

$$\begin{bmatrix} \overline{u_1 \ u_1} & \overline{u_1 \ u_2} & \overline{u_1 \ u_3} \\ \overline{u_2 \ u_1} & \overline{u_2 \ u_2} & \overline{u_2 \ u_3} \\ \overline{u_3 \ u_1} & \overline{u_3 \ u_2} & \overline{u_3 \ u_3} \end{bmatrix} \ . \tag{10}$$

The equation of the Reynolds stress tensor is

$$\partial_t \ \overline{u_i \ u_j} + U_k \ \partial_k \ \overline{u_i \ u_j} = \tag{11}$$

$$\begin{array}{ll} -\frac{1}{\rho}(\overline{u_j \ \partial_i \ p} + \overline{u_i \ \partial_j \ p}) & \textit{redistribution} \\ -2\nu \ \overline{\partial_k \ u_i \ \partial_k \ u_j} & \textit{dissipation} \\ -\partial_k \ \overline{u_k \ u_i \ u_j} & \textit{turbulent transport} \\ -\overline{u_j \ u_k} \ \partial_k \ U_i - \overline{u_i \ u_k} \ \partial_k \ U_j & \textit{production} \\ +\nu \ \nabla^2 \ \overline{u_i \ u_j} & . \end{array}$$

Here the dissipation term is of interest, when dealing with the impedance of flute-like instruments. It considers the derivatives of the fluctuating flows and the viscosity. The other terms describe the production of turbulent energy out of the laminar flow, the turbulent transport within the directions of the eddies and the redistribution of turbulence as a pressure derivative.

These equations implies a closure problem. Here, we have to average over the three flow directions. This average is different from the multiplication of each averaged flow, so

$$\overline{u_i} \ \ \overline{u_j} \neq \overline{u_i \ u_j} \ . \tag{12}$$

We have ten variables in four equations here, so the system is unclosed. To close it, we have to make some assumptions and take experimentally measured constants to the system.

The first assumption is, that the turbulent production equals the turbulent dissipation. Otherwise, turbulence would grow or die out. Secondly experimental results suggest, that the stress-intensity ratio is

$$\overline{u_i\ u_j}/k \approx 0.3\ . \tag{13}$$

In combination with the fist assumption of turbulent production being equal to turbulent dissipation and introducing a turbulent viscosity ν_T, we arrive at

$$\nu_T = C_\mu k^2/\epsilon \tag{14}$$

with $C_\mu = 0.09$.

The third assumption is, that the turbulent redistribution term is small, because of the small pressure gradient in the small eddy scales. Nevertheless, the redistribution term and the turbulent transport term are replaced by a gradient turbulent energy model including turbulent viscosity as

$$-\frac{1}{\rho}(\overline{u_j\ \partial_i\ p} + \overline{u_i\ \partial_j\ p}) - \partial_k\ \overline{u_k\ u_i\ u_j} \approx \partial_j(\nu_T\ \partial_j\ k)\ . \tag{15}$$

From this we arrive at an equation of the turbulent energy k

$$\partial_t k + U_j\ \partial_j\ k = -\overline{u_i\ u_j}\ \partial_i\ U_i - \epsilon + \partial_j((\nu + \frac{\nu_T}{\sigma_k})\partial_j\ k)\ . \tag{16}$$

The equation of ϵ is assumed as an analogy to the equation of turbulent energy k as

$$\partial_t\ \epsilon + U_j\ \partial_j\ \epsilon = \frac{C_{\epsilon 1}(-\overline{u_i\ u_j}\ \partial_i\ U_i) - C_{\epsilon 2}\epsilon}{T} + \partial_j\left((\nu + \frac{\nu_T}{\sigma_\epsilon})\partial_j\ \epsilon\right)\ . \tag{17}$$

Finally the equation of flow is

$$\partial_t\ U_i + U_j\ \partial_j\ U_i = -\frac{1}{\rho}\partial_i(p + 2/3\rho\ k) + \partial_j((\nu + \nu_T)(\partial_j\ U_i + \partial_i\ U_j))\ . \tag{18}$$

The experimental constants are

$$C_\mu = 0.09,\ \ C_{\epsilon 1} = 1.44,\ \ C_{\epsilon 2} = 1.92,\ \ \sigma_k = 1,\ \ \sigma_\epsilon = 1.3\ . \tag{19}$$

3 Method

The three equations of turbulent flow U, energy k and dissipation ϵ together with the condition of incompressibility, have been implemented in a Finite-Element model. As the aim of the study is to get a quantitative result of the importance of turbulence in flute-like impedance problems, only a stationary model is used, so the time derivatives in the above equations are not used. The model is compared to a Navier-Stokes model without turbulent assumptions for comparison on the same geometry

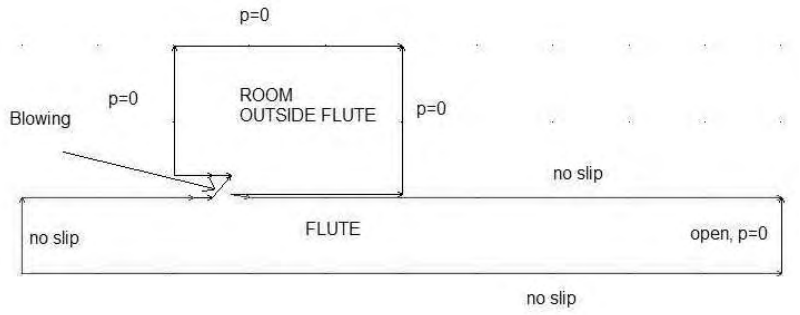


Fig. 3. The geometry of the used Finite-Element model. The geometry is split into the flute and the room outside the flute. As the room has to be finite, we assume the boundary conditions of the room of zero pressure. The same boundary condition holds for the right open end of the flute. The walls of the flute itself have a no-slip boundary condition (zero flow along the boundary). The blowing through the player happens at the left side of the flute hole.

Fig 3 shows the used geometry. The flute walls have boundary conditions of no-slip in the velocity field, meaning they do not only have zero flow perpendicular to the boundary, but also zero flow parallel to it. The left end of the flute is closed, the right end is open with a boundary condition of zero pressure. The blowing happens at the left corner of the flute hole and is directed to the right lip point of the hole. The room outside the flute has to be modelled of finite size. So its boundary conditions are taken to be of zero pressure.

4 Results

Two models, one of a normal Navier-Stokes model and one of a $k - \epsilon$ model are compared.

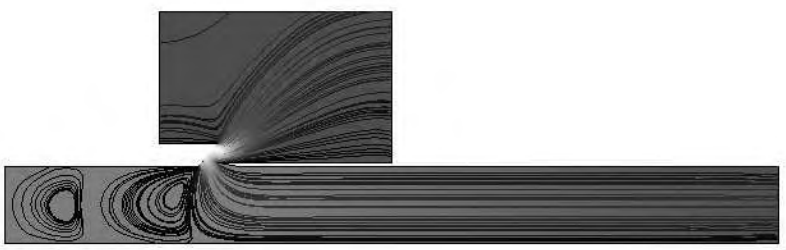


Fig. 4. Navier-Stokes model. The background represent the pressure ranging to up to 500 N/m^2 in the tube and up to $1400N/m^2$ at the inblowing hole (white). The streamlines represent the velocity field ranging from $0m/s$ (black) to $0.696m/s$ (white) at the inblown region. The velocity at the right tubes end is about $0.1m/s$, at the left tubes end about $.01m/s$.

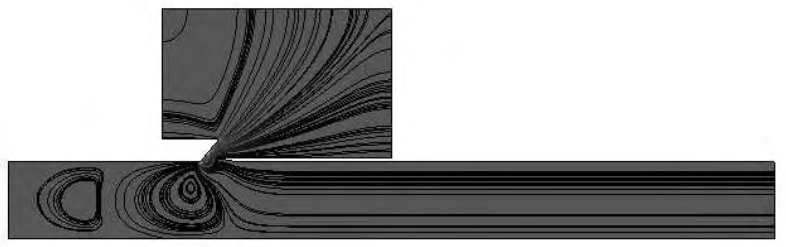


Fig. 5. Turbulence model. The background represent the pressure ranging from $-.5N/m^2$ just outside the tube (black) to about $1N/m^2$ in the tube (white). High pressure values of up to $200N/m^2$ are just reached right at the inblowing whole. The streamlines represent the velocity field ranging from $.001m/s$ at the left tubes side (black), $.005m/s$ at the right tubes side and $0.1m/s$ outside the tube up to $0.172m/s$ right at the inblowing point (white).

In Figure 4 and Figure 5 the results for the Navier-Stokes model and the turbulent model are shown respectively. The background show the pressure distribution, the streamlines represent the velocity field. The overall behaviour of both models is about the same. Both show the flow of the player separating into a flow into the flute and a flow into the room outside the flute. Additionally, both have two large eddies at the closed end of the flute.

But also the differences can clearly be seen. The major difference is with the ratio of the flow into and outside the flute. In the Navier-Stokes model, the amount of flowlines inside the flute is about the same, as the amount of flowlines going outside into the room. In the turbulent model, the flowlines going into the flute are much less, than those flowing into the room. In numbers,

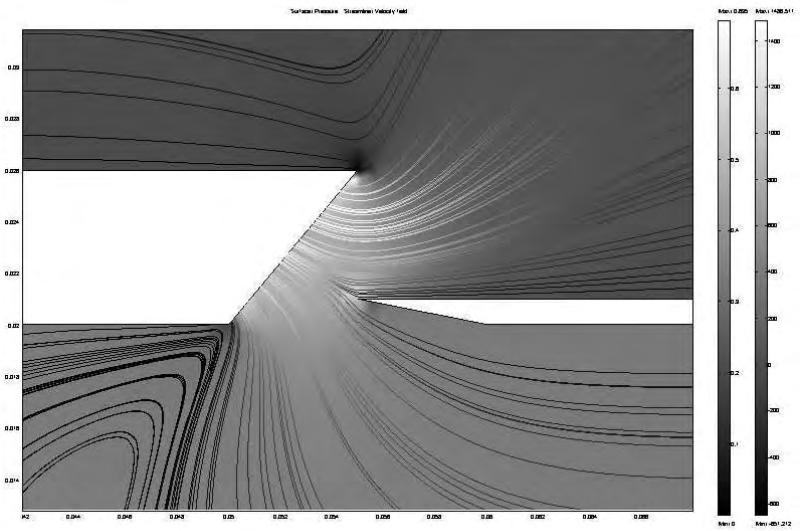


Fig. 6. Navier-Stokes model, blowing hole. The splitting of the inblowing flow of the flute player in two nearly equal parts, one inside the flute and one in the outside room are shown here.

in the NS model 43% of the flow gets into the flute. To compare these value with the turbulent model, we do also have to consider the amount of turbulent dissipated energy. Comparing the outflow of the right end of the flute between the two models, we get a ratio of *turbulent / not turbulent* ≈ 0.01269. If we take this ratio in comparison with the split in the NS model, we get as a result, that the amount of energy coming from the players blowing, which gets into the flute is about

Navier-Stokes model	$E_{intoflute}/E_{blowing} \approx \mathbf{43}\%$
Turbulent model	$E_{intoflute}/E_{blowing} \approx \mathbf{0.55}\%$

Remembering the experimental values of 2.4% of blowing energy getting into the tube, we must assume, that a turbulent modelling of flute-like instruments is appropriate.

Taking a closer look to the flute hole, the behaviour can be studied in more detail. Figure 6 and Figure 7 show the inblowing area. In the NS model, the splitting between the flows can clearly be seen. In the turbulent model,

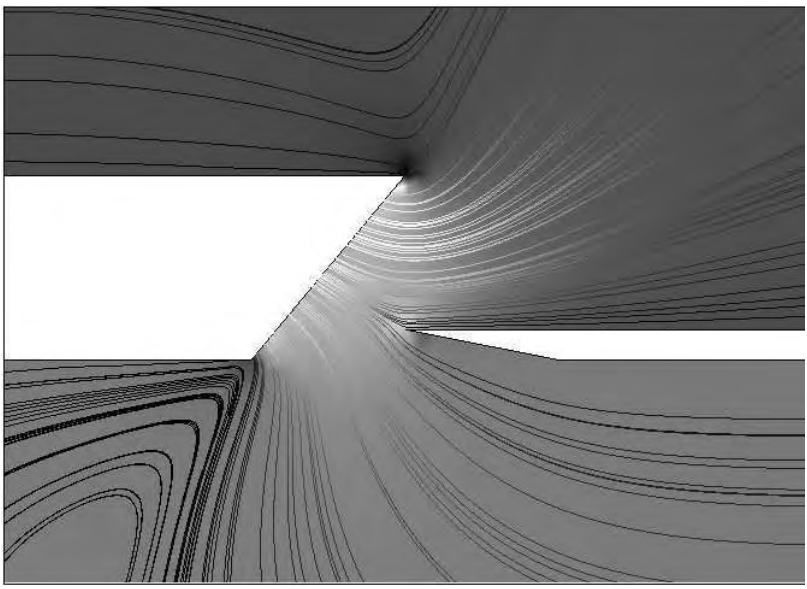


Fig. 7. Turbulent model, blowing hole. The splitting in two flows now favors the outflow in the outside room. Just the flowlines, which are at the very bottom of the inblowing area make it inside the flute.

Fig. 8. Turbulent model. The turbulent viscosity distribution shows a much larger dissipation inside the flute around the blowing region than in the outside room. So physically speaking, the inblowing flow takes the easier way out of the tube, which is much less damped than the way inside the flute. Here, the ratio between inflow and outflow energy is in the same region as experimental results, which does not hold for the Navier-Stokes modelling without turbulence. Range from 0.0685 kg / m s (black) to 0.0707 kg / m s (white).

only the velocity flow lines, with are at the very bottom of the blowing area make it into the flute. Figure 8 shows the reason for this behaviour. Here the turbulent viscosity shows a large region around the inblowing area which is much stronger in the flute than outside of it. So indeed it seems to be the turbulent dissipation, which forces the blowing flow out of the tube, where dissipation is much less.

5 Conclusions

The modelling of flute-like instruments with turbulence viscous damping seem appropriate. The model without turbulence leads to results, which do not correspond to experimental results concerning the flutes impedance [Coltman 1968], the amount of blowing energy of the flute player really making its way into the flute. The model used shows the reason in the distribution of turbulent viscous damping in such an inblowing situation. The turbulent viscosity in the flute is much larger than the viscosity outside of the flute. Taking into account the Reynolds stress tensor, we find as a reason for this damping a coupling of the different fluctuating flows. So the damping is caused by the flow changing direction. This change of direction in the case of the flute takes place much more in the flute, as it is a closed system, where the flow is forced to change direction into the flute tubes length. The flow going out into the outside room is not forced in such a way to flow into just one direction. So here, the Reynolds stress tensor is much smaller and the flow which is blown into the tube takes easier direction outside of the flute, as it is free to choose.

With these results in mind, it now seems appropriate to assume, that the reason for the tube of flute-like instruments forcing the self-sustained oscillation of the generator region at the blowing hole to vibrate with the eigenfrequencies of the tube seems to be the different values for damping of these two subsystems. The much more damped generator looses the game and the much less damped air column in the flutes tube takes over the systems overall behaviour. Just with a generator region of this behaviour we can have a flute, which can be played in different pitches. If the generator region would force the tube in the generators self-sustained eigenfrequencies, there would be just one pitch in which the flute could be played, the self-sustained oscillation of the sound hole. So if i.e. the sound hole is too large and so damping is decreased, the player just gets the very high pitches generator frequencies out of the flute which can be heard is just the generator region played without the tube fixed to it, a phenomenon every flute player, saxophonist or clarinetist know very well.

References

[Aschoff 1936] Aschhoff, V.:Experimentelle Untersuchungen an einer Klarinette. In: Akustische Zeitschrift 1, 77-93, 1936.

[Benade and French 1965] Benade, A.H., & French, J.W.: Analysis of the flute head joint. In: JASA 37, 679-91.

[Coltman 1969] Coltman, J.W.: Sound radiation from the mouth of an organ pipe. In: Journal of the Acoustical Society of America, 46, 477, 1969.

[Coltman 1973] Coltman, J.W.: Mouth resonance effects in the flute. In: Journal of the Acoustical Society of America 54, 417-420, 1973.

[Coltman 1976] Coltman, J.W.: Jet drive mechanisms in edge tones and organ pipes. In: Journal of the Acoustical Society of America, 60, 725-33, 1976.

[Coltman 1968] Coltman, J.W.: Acoustics of the flute. In: Physics Today 21, 11, 25-32, 1968.

[Dequand 2000] Dequand, Sylvie: Duct Aeroacoustics: from Technological Applications to the Flute. Eindhoven: Technische Universiteit Eindhoven, 2000.

[Durbin and Pettersson 2001] Durbin, P.A. & Pettersson, R.: Statistical Theory and Modeling for Turbulent Flows. John Wiley & Sons, 2001.

[Fletcher and Rossing 2000] Fletcher, N.H. & Rossing, Th.D.: The Physics of Musical Instruments. Springer 2000.

[Fletcher et al. 1982] Fletcher, N.H., Strong, W.J. & Silk, R.K.: Acoustical characterization of flute head joints. In: Journal of the Acoustical Society of America 71, 1255-60, 1982.

[Hughes 1987] Hughes, J.R.: The Finite Element Method. Linear Static and Dynamic Finite Element Analysis. Dover Publications, Mineola, 1987.

[Kolmogorov 1941] Kolmogovov, A.N.: The local structure of turbulence in incompressible viscous fluid for vary large Reynolds number. In: Dokl. Akad. Nauk SSSR 30, 301-5, 1941.

[Miklós and Angster 2000] Miklós, A. & Angerster, J.: Properties of the Sound of Flue Organ Pipes. In: Acustica 86, 4, 611-622, 2000.

[Saddoughi and Veeravalli 1994] Saddoughi, S.G. & Veeravalli, V.S.: Local isotropy in turbulent boundary layers at high Reynolds number. In: Journal of Fluid Mechanics, 268, 333-372, 1994.

[Ségoufin et al. 2000] Ségoufin, C., Fabre, B., Verge, M.P., Hirschberg, A. & Wijnands, A.P.J.: Experimantal Study of the Influence of the Mouth Geometry on Sound Production in a Recorder-like Instrument: Windway Length and Chamfers. In: Acustica, 86, 4, 649-61, 2000.

[Verge et al. 1994] Verge, M-P, Caussé, R., Fabre, B., Hirschberg, A., Wijnands, A.P.J. & van Steenbergen, A.: Jet oscillations and jet drive in recorder-like instruments. In: Acta Acoustica 2, 403-19, 1994.

[Wilcox 2004] Wilcox, D.C.: Turbulence MOdeling for CFD. Second Edition. DCW Industries, Inc. 2004.

3

CHEMICAL ENGINEERING

A simple discrete stochastic model for laser-induced jet-chemical etching

Alejandro Mora[1], Thomas Rabbow[2], Bernd Lehle[3], Peter J. Plath[2], and Maria Haase[1]

[1] Institut für Höchstleistungsrechnen (IHR), University of Stuttgart, 70569 Stuttgart, Germany
[2] Institut für Angewandte und Physikalische Chemie, Chemische Synergetik, University of Bremen, 28334 Bremen, Germany
[3] vFlow Engineering GmbH, 70499 Stuttgart, Germany
e-mail:ica2am@csv.ica.uni-stuttgart.de, mh@ica.uni-stuttgart.de

Summary. Recently developed processes based on laser-induced liquid jet-chemical etching provide efficient methods for high resolution microstructuring of metals. Like in other abrasive techniques (water-jet cutting, laser cutting, ion sputtering etc.) a spontaneous formation of ripples in the surface morphology has been observed depending upon the choice of system parameters. In this paper we present a discrete stochastic model describing the joint action of removal of material by chemical etching and thermally activated diffusion initiated by a moving laser leading to structure formation of a surface. Depending on scan speed and laser power different surface morphologies are observed ranging from rough surface structures to the formation of ripples. The continuum equation associated to the discrete model is shown to be a modified Kuramoto-Sivashinsky equation in a frame comoving with the laser beam. Fourier and wavelet techniques as well as large deviation spectra are used for a characterization of the surfaces.

1 Introduction

Laser-induced liquid jet-chemical etching processes for direct high-precision micromachining of metals without using masking techniques are of significant interest for various applications. One of the reasons is a high resolution and surface quality of the etched microstructures. In addition, the process temperature is sufficiently low, which helps for example in machining superelastic alloys to maintain the material properties, and since clean room facilities are not necessary the costs of production are moderate. The field of applications ranges from the fabrication of micro-tools for biomedical analysis and medicine like micro-grippers to metal master molds for the production of micro-optical components [1]. In a recently developed technique a moving laser beam which is guided through a coaxially expanding stream of liquid etchant locally heats

the passivated metal surface, inititiates a chemical etching process and leads to a selective removal of metal in the neighbourhood of the laser spot [2].

Depending upon various process parameters like the scan speed, the laser power, the specific etching acid and its concentration different surface morphologies are observed ranging from rough surfaces to the formation of often unwanted ripples. These ripple structures can either be triggered externally for example by an etchant pump or they can be generated intrinsically due to a spontaneous formation of a front instability which is known to occur also in other abrasive techniques like water-jet cutting [3].

In order to optimize and control the etching process in view of an efficient production of microstructures with high surface quality one has to gain insight into the dynamics of the interacting thermochemical and hydrodynamical processes and to set up adequate models. Since details of the dynamics of the involved processes are elusive we propose to mimic the main mechanisms like erosion, i.e. removal of material due to chemical reactions, and thermally activated diffusion processes initiated by a localized moving laser beam by means of a discrete stochastic model. In a first model we neglect the influence of transport and diffusion limitation induced by the reaction products. Our model is a modification of a stochastic model originally introduced for ion-sputtering by Cuerno et al. [4]. In this model periodic structures are formed at an early stage at the onset of instability. At later stages nonlinear effects lead to a crossover to a rough surface. For laser-induced etching we have to account for the moving localized energy source provided by the laser beam. Depending upon the scan speed of the laser beam we demonstrate, that for our discrete model either rough surfaces or patterns of 'frozen' ripples are formed.

A Langevin equation for the spatiotemporal evolution of the surface morphology corresponding to our discrete stochastic model is then derived from a master equation which is based on the discrete erosion and diffusion rules used in the model. The master equation can be approximated by a Fokker-Planck equation. In the case of small slopes a coarse-graining procedure allows then to derive a continuous Langevin equation which has the form of a modified noisy Kuramoto-Sivashinsky (KS) equation in a frame comoving with the laser beam.

The paper is organized as follows. In section 2, we briefly describe the principle scheme of the experimental setup of the laser-induced wet-etching together with experimentally observed surface morphologies. The discrete model is presented in section 3. The derivation of the associated master equation follows in section 4 together with the smoothing procedure used to find a continuum equation for the evolution of the height profile. In section 5, Fourier and wavelet techniques as well as large deviation spectra are used to characterize the surface profiles. Section 6 presents our conclusions and some perspectives for further investigations.

2 Experimental setup and some observed kerfs

The experimental setup used for the laser-induced etching consists of an etching cell, where the sample is mounted horizontally and submerged in the acid. Fresh etching acid is supplied by a jet-stream perpendicular to the workpiece. The light of a focussed laser beam running coaxially to the jet is absorbed at the surface of the sample that is passivated by the etchant at ambient temperature of the etchant. Due to heat conduction adjacent layers of the etchant will be heated up, which initiates the etching process by forming micro-cracks in the passive layer. At these sites where the passive layer has been removed the etchant subsequently can react with the metal leading to an ablation of the material. The whole basin is mounted onto computer-controlled xy-stages allowing a relative movement of the sample with respect to the laser beam. Details of the experimental setup are described elsewhere [2].

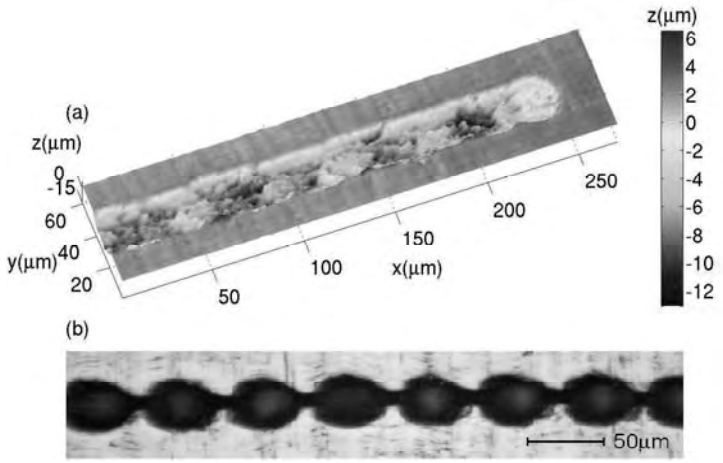


Fig. 1. Characteristic surface morphologies machined with laser-induced wet etching. a) UBM measurement of a rough kerf, b) Incident light-microscopy image of an etched kerf showing periodic keyhole structures [2].

The surface topography is measured by a 3D-laser-focus scanner (UBM) with resolutions of $1\mu m$ in horizontal x and y directions and $0.01\mu m$ in vertical height direction. In addition incident light-microscopy images show the surface morphology. In fig. 1 two typical etched kerfs are displayed. Depending upon process parameters like laser type and power, scan speed, concentration of etching acid, specific alloy of the metal foil different morphologies can be observed.

3 Description of the discrete model

We suppose that in wet-chemical etching like in other cases of nonequilibrium interface evolution [3–5] a competition between roughening and smoothing mechanisms is responsible for the structure formation of the surface. The laser beam heats the surface and initiates chemical reactions which lead to an ablation of material at the metal surface. Due to the heat transport between metal and liquid etchant higher temperatures are build-up within troughs than at peaks. Therefore, we argue that the erosion rate is curvature dependent. Experiments corroborate this assumption showing that material is preferentially removed in troughs. Similar observations are made in other abrasive techniques like ion-sputtering or water-jet cutting [3,4] and can be explained by a model proposed by Sigmund [6]. This phenomenon leads to a destabilization of the surface and can be described as a negative surface tension. The counteracting stabilizing mechanism is attributed to thermally activated surface diffusion where metal molecules move to those nearest neighbor sites where the binding energy is minimal.

Cuerno et al. proposed a discrete model for ion-sputtering which is based on the interplay of erosion and diffusion [4]. We adapt and modify this model for the laser-induced etching process. In the 1+1 dimensional case the material to be eroded is represented by a lattice composed of cells of horizontal width a and vertical width b. The interface is assumed to be periodic with heights described by integer values $h_i(t)$, $i = 1, \ldots, L$ where L is the system size.

Chemical etching and diffusion are activated by the localized laser beam the intensity of which is assumed to be a Gaussian. The active zone may hence be modelled as

$$G(x,t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-x_0)^2/\,2\sigma^2}. \tag{1}$$

where σ corresponds to the standard deviation and x_0 denotes the center of the laser beam which is moving with a constant velocity u. Within this active zone a site i of the lattice with coordinate $x_i = i\,a$ is chosen with probability $G_i = G(x_i)$ and subject either to erosion (with probability p) or to diffusion (with probability $1 - p$).

The probability for a particle at site i to be *removed by etching* is estimated as the product $p_{ei} = p_{ci}\,Y_i\,G_i$ where p_{ci} denotes the aforementioned curvature dependent probability for a particle to be removed. Y_i is the value of a nonlinear yield function at site i which is attributed to the dependency of the absorption of laser energy in the material on the slope of the surface. Fig. 2 shows typical curves for the the angle between laser beam and normal vector of the profile [8], see also [7]. In analogy to the case of parallel polarized light we approximate the absorptivity at site i as

$$Y_i = y_0 + y_1\varphi_i^2 + y_2\varphi_i^4 \tag{2}$$

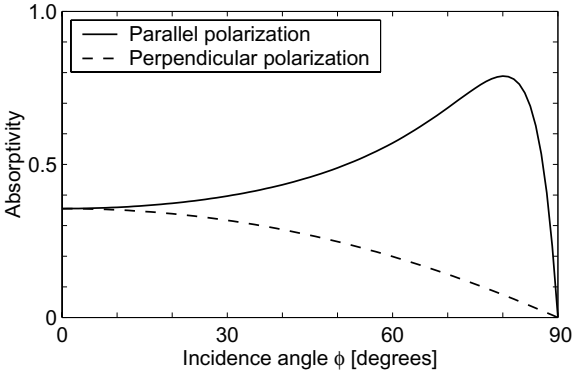


Fig. 2. Absorptivity of polarized light versus incidence angle for a flat iron surface. The solid (dashed) line shows the case when the plane of polarization is parallel (perpendicular) to the incidence plane [8].

where $\varphi_i = \arctan(\nabla h_i)$ is the slope of the surface and $\nabla h_i = (h_{i+1} - h_{i-1})/2a$. For the function $Y(\varphi)$ we have chosen the parameters $y_0 = 0.5$, $y_1 = 0.979$ and $y_2 = -0.479$.

In Cuerno's model p_{ci} is computed from a 3×3 box centered at i as $1/7$ times the number of occupied sites. In this model the curvature is estimated by the second order derivative $\nabla^2 h_i$ where $\nabla^2 h_i = (h_{i-1} - 2h_i + h_{i+1})/a^2$ denotes the discrete Laplacian. The erosion rule therefore restricts the curvature range to only 7 discrete values in an interval $[-2, 4]$. We propose to modify this box rule by replacing the 7 discrete probability values by a spectrum taken from an interval $[p_0, 1]$ with $0 < p_0 \ll 1$ which corresponds to a chosen interval of curvatures $[-\kappa_m, \kappa_m]$. The curvature at site i is evaluated as $\kappa_i = \nabla^2 h_i (1 + (\nabla h_i)^2)^{-3/2}$. Due to the discreteness of heights h_i the angles φ_i and curvatures κ_i usually vary considerably between neighboring sites. In order to avoid spurious fluctuations, φ_i and κ_i are replaced by local averages of these quantities taken at adjacent sites $i-1$, i, $i+1$.

For the *diffusion rule* a particle at site i is assumed to move to a randomly chosen nearest neighbor column with a hopping rate

$$w_i{}^{\pm} = \frac{1}{1 + e^{\beta \Delta \mathcal{H}_{i \to i \pm 1}}} \tag{3}$$

where

$$\mathcal{H} = J/b^2 \sum\nolimits_{i=1}^{L-1} (h_i - h_{i+1})^2 \tag{4}$$

denotes the energy and $\Delta \mathcal{H}_{i \to i \pm 1}$ is the energy difference between final and initial state of the move, J is a coupling constant and β denotes the inverse temperature [4].

As an example, we study the discrete model for the following set of system parameters: $a = 2$, $b = 1$, $p = 0.1$, $\beta J = 5$, $p_0 = 0.15$ and $\kappa_m = 0.0825\, b/a^2$. The

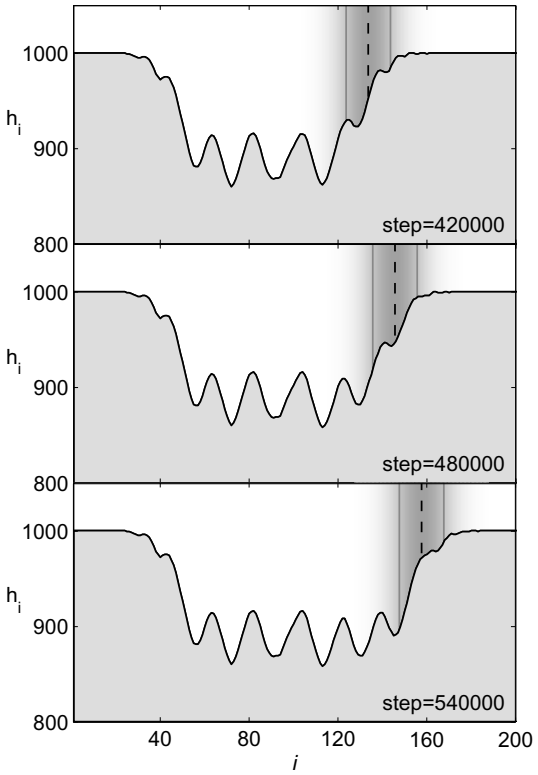


Fig. 3. Temporal evolution of the moving etching front: 3 stages demonstrate the formation of a single ripple structure. The center of the moving Gaussian beam is represented by the vertical dashed line.

standard deviation of the active zone is choosen as $\sigma = 10a$, the intensity as $\frac{1}{2}$ and the scan speed as $u = 0.002a$ per time step. Starting with an initially flat surface the evolution of the moving etching front for 3 subsequent time instants is shown in fig. 3. The characteristic time intervall where all structures develop is given by

$$\Delta T = \frac{d}{u} \tag{5}$$

where d denotes the diameter of the active spot. During ΔT the surface is exposed to the combined action of erosive and diffusive processes which create instabilities at the moving front. In fig. 3 the generation process of a single ripple structure in the front profile is demonstrated. Typically, a shallow cavity which is randomly created at the forefront of the beam becomes unstable. The local erosion rate increases as the laser passes through the valley due to its positive curvature while peaks are etched at smaller rates. As a result, ripples are formed continuously and keep their shape when the influence of the

laser ceases in analogy with the repassivation of the metal surface observed in experiments. A more detailed description of the discrete model can be found in [21].

4 Master equation and Langevin equation

In a series of papers the master equation approach has been used to derive a continuous Langevin equation for the evolution of the surface profile from the underlying discrete stochastic model [10–13]. This procedure consists of two steps. First, a master equation is set up and approximated by a Kramers-Moyal expansion [14]. If higher order terms can be neglected this expansion reduces to a Fokker-Planck equation which is equivalent to a system of Langevin equations for the temporal evolution of $h_i(t)$ at discrete sites i. In a second step, a coarse graining procedure is performed leading to a continuous Langevin description a prerequisite being that slopes are small along the profile.

4.1 Master equation

In what follows we briefly summarize the results for our discrete model. The gain and loss type of the master equation describes the evolution of the joint probability $P(\mathbf{H}, t)$ that the surface adopts the configuration $\mathbf{H}$ at time t

$$\frac{\partial P(\mathbf{H}, t)}{\partial t} = \sum_{\mathbf{H}'} W(\mathbf{H}, \mathbf{H}') P(\mathbf{H}', t) - \sum_{\mathbf{H}'} W(\mathbf{H}', \mathbf{H}) P(\mathbf{H}, t) \tag{6}$$

where $W(\mathbf{H}', \mathbf{H})$ is the transition rate per unit time to come from configuration $\mathbf{H}$ to $\mathbf{H}'$ and consists of two contributions corresponding to the erosion and diffusion rule.
The transition probability per unit time τ for the erosion can be written in the form

$$W_e(\mathbf{H}', \mathbf{H}) = \frac{p}{\tau} \sum_{i=1}^{L} p_{ci} Y_i G_i \, \delta(h_i', h_i - b) \prod_{j \neq i} \delta(h_j', h_j) \tag{7}$$

In the original model [4]

$$p_{ci} = \frac{1}{7}(5 + \frac{a^2}{b} \nabla^2 h_i + \frac{1}{b} \Theta_i). \tag{8}$$

describes the mapping of the curvature range to the interval of discrete probabilities based on the height steps h_i. Θ_i contains terms visible only in a box larger than 3×3:

$$\Theta_i = -\theta(h_{i-1}-h_i-2b)[h_{i-1}-h_i-b] - \theta(h_{i+1}-h_i-2b)[h_{i+1}-h_i-b] \\ +\theta(h_i-h_{i-1}-3b)[h_i-h_{i-1}-2b] + \theta(h_i-h_{i+1}-3b)[h_i-h_{i+1}-2b] \quad (9)$$

and $\theta(x)$ denotes the Heaviside function. Instead of controlling the curvature range by the box size in our model the interval of accessible curvatures $[-\kappa_m, \kappa_m]$ can be prescribed arbitrarily. The probability $p_{ci}(\kappa_i)$ is conveniently written in the form expressed as a smooth function

$$p_{ci}(\kappa_i) = \frac{1+p_0}{2} + \frac{1-p_0}{2}\tanh\frac{\kappa_i}{\kappa_m}. \quad (10)$$

The transition rate per unit time τ for the diffusion rule is a little more complex since for each diffusion step two neighboring sites of the lattice are involved and one has to consider all possible transitions

$$W_d(\mathbf{H}', \mathbf{H}) = \frac{1-p}{2\tau}\sum_{i=1}^{L}\big[\, w_i^+ G_i\,\delta(h_i', h_i-b)\,\delta(h_{i+1}', h_{i+1}+b) + \\ w_{i+1}^- G_{i+1}\,\delta(h_i', h_i+b)\,\delta(h_{i+1}', h_{i+1}-b)\,\big]\prod_{j\neq i,i+1}\delta(h_j', h_j) \quad (11)$$

where the energy difference has the form

$$\Delta\mathcal{H}_{i\to i\pm1} = 2J(3 - \frac{a^3}{b}\nabla^3 h_i). \quad (12)$$

and the hopping rates given in eq. (3) can be written as

$$w_i^+ = \big[1+q\,\exp(-\gamma\nabla^3 h_i)\big]^{-1} \quad \text{and} \quad w_{i+1}^- = \big[1+q\,\exp(\gamma\nabla^3 h_i)\big]^{-1} \\ \text{with} \qquad q = \exp(6J\beta) \quad \text{and} \quad \gamma = 2J\beta a^2/b. \quad (13)$$

The total rate is then given as $W(\mathbf{H}', \mathbf{H}) = W_e(\mathbf{H}', \mathbf{H}) + W_d(\mathbf{H}', \mathbf{H})$.

4.2 Fokker-Planck and discrete Langevin equation

In order to expand the master equation in a Kramers-Moyal expansion we determine the transition moments

$$K_i^{(1)} = \sum_{\mathbf{H}'}(h_i'-h_i)W(\mathbf{H}', \mathbf{H}) \\ K_{ij}^{(2)} = \sum_{\mathbf{H}'}(h_i'-h_i)(h_j'-h_j)W(\mathbf{H}', \mathbf{H}), \quad (14)$$

where the sums extend over all configurations $\mathbf{H}'$. The moments can be evaluated directly using eq. (13). Introducing the vector of weighted hopping rates $\mathbf{W}_i = \{G_{i-1}w_{i-1}^+ \quad G_i w_i^- \quad G_i w_i^+ \quad G_{i+1}w_{i+1}^-\}$ and the diagonal matrix $\mathbf{D} = \dfrac{(1-p)\,b}{2\tau}\lceil 1 \;\; b \;\; -b \;\; -b\rfloor$ the vector of first and second order transition

moments due to diffusion $\mathbf{K}_{d\,i} = \{K^{(1)}_{d\,i} \quad K^{(2)}_{d\,ii} \quad K^{(2)}_{d\,i-1,i} \quad K^{(2)}_{d\,i,i+1}\}$ can be determined from

$$\mathbf{K}_{d\,i} = \mathbf{D} \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \mathbf{W}_i. \tag{15}$$

In the small gradient approximation the transition moments can be simplified as follows

$$\begin{aligned} K^{(1)}_i &= -\frac{b}{\tau} G_i \left[p\, p_{ci}\, Y_i \; + \frac{(1-p)\, q\, \gamma}{(1+q)^2}\, a^2\, \nabla^4 h_i \right] \\ K^{(2)}_{ij} &= \; \frac{b^2}{\tau} G_i \left[p\, p_{ci}\, Y_i\, \delta_{ij} - \frac{1-p}{1+q}\, \nabla^2 \delta_{ij} \right] \end{aligned} \tag{16}$$

where $\nabla^2 \delta_{ij} = \delta_{i-1,j} - 2\delta_{ij} + \delta_{i+1,j}$. In the limit of decreasing lattice spacing $b \to 0$ the transition moments decrease like $K^{(n+1)}/K^{(n)} \sim \mathcal{O}(b)$ for increasing order n. Therefore, we follow the argumentation given in [11] which allows us to truncate the Kramers-Moyal expansion leading to a Fokker-Planck equation

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial h_i}(K^{(1)}_i P) + \frac{1}{2}\frac{\partial^2}{\partial h_i\, \partial h_j}(K^{(2)}_{i,j} P) \tag{17}$$

where the first term $K^{(1)}_i$ describes the deterministic part of the evolution and is denoted as *drift* and $K^{(2)}_{i,j}$ as *diffusion* term. Using Itô's definition the temporal evolution of h_i at site i for a stationary laser beam is described by the system of associated Langevin equations

$$\frac{dh_i}{dt} = K^{(1)}_i(h_i, t) + \eta_i(t) \tag{18}$$

where $\eta_i(t)$ is a Gaussian white noise with zero average and variance $\langle\, \eta_i(t)\eta_j(t')\,\rangle = K^{(2)}_{i,j}\, \delta(t-t')$. Coupling of heights between neighboring sites arises due to the diffusion term in the stochastic part of the equations. A numerical integration of the system can be performed without any assumptions about smallness of slopes etc. evaluating curvatures and hopping rates occuring in the expressions for the transitions moments directly from eqs. (10,13).

4.3 Continuous Langevin equation

In order to derive a continuum equation for the evolution of the surface front from the transition rules of the discrete stochastic model one has to perform regularization and coarse-graining procedures. A rigorous derivation is known to be a highly nontrivial task [11, 13], the main problem being the regularization of the step function in the discrete equations. In the following we restrict our consideration to an *ad hoc* derivation with the aim to determine

the most relevant leading terms in the evolution equation. We assume that for small lattice widths a and b the subsequent series expansions are allowed. The heights $h_i(t)$ at discrete sites i are replaced by a smooth function $h(x,t)$ with $h(ia,t) = h_i(t)$ and

$$h_{i\pm1}(t) = h_i(t) + \sum_{n=1}^{\infty} \frac{(\pm a)^n}{n!} \frac{\partial h(x,t)}{\partial x^n}\bigg|_{x=ia}. \tag{19}$$

For small gradients the yield function, eq. (2), can be written as

$$Y_i = y_0 + y_1(\nabla h_i)^2 + (y_2 - 2y_1/3)(\nabla h_i)^4 + \ldots \tag{20}$$

and the leading terms in the hopping rates of the diffusion rule, eq. (13), are given by

$$w_i^+ = \frac{1}{1+q}\left(1 - \frac{q\gamma}{1+q}\nabla^3 h_i + \ldots\right). \tag{21}$$

In the Cuerno model the Heaviside function in eq. (9) has to be replaced by a smooth function. Various regularizations have been proposed in the literature, which, however, can lead to different results [12]. For our modification we have already chosen a smooth function p_{ci} in eq. (10) thus circumventing the regularization step. For $\kappa_i/\kappa_m < \pi/2$ the curvature dependent transition probability can be written in the form

$$p_{ci} = \frac{1+p_0}{2} + \frac{1-p_0}{2}\nabla^2 h_i(1 - 3/2(\nabla h_i)^2) + \ldots. \tag{22}$$

Finally, eqs. (19)-(22) are inserted into the system of Langevin equations eq. (18). In a coordinate system comoving with the laser beam with constant velocity u we obtain the continuous Langevin equation

$$\frac{\partial h}{\partial t} + u\nabla h = G(x)\left[v_0 + \nu\nabla^2 h - D\nabla^4 h + \right. \\ \left. + \lambda(\nabla h)^2 + c_1\nabla^2 h(\nabla h)^2 + \ldots\right] + \eta(x,t). \tag{23}$$

The leading terms have the form of a modified noisy Kuramoto-Sivashinsky (KS) equation. Here, v_0 is the vertical mean velocity of the profile describing the ablation rate for the infinitely extended homogeneous case $G(x) \equiv 1$, ν is a negative surface tension coefficient resulting from erosion, D is a positive coefficient arising from thermally activated surface diffusion, λ is a negative coefficient describing the lateral shrinking of the profile due to erosion and c_1 is responsible for the observed coarsening of the profile in the case of $Y(\varphi) \equiv 1$ for increasing time [9]. An equation similar to eq. (23) has been proposed to model water-jet cutting processes [3].

A straightforward linear stability analysis for $G(x) \equiv 1$ explains the evolution of ripples in the moving front. A stationary solution of the homogeneous system is given by $h_0(x) = C_0 + C_1 x$ where the scan speed is related to v_0

by $u\,C_1 = v_0 + \lambda\,C_1^2$. Small perturbations of the form $Ce^{\omega t+ikx}$ lead to a dispersion relation

$$\omega(k) = -iku - \nu k^2 - Dk^4. \tag{24}$$

For $\nu < 0$ and $D > 0$ linear waves with wave numbers $0 < k < k_0 = \sqrt{|\nu|/D}$ are unstable, while modes with $k > k_0$ are stable. The mode with the fastest growing amplitude has a wave length $l_m = 2\pi\sqrt{2D/|\nu|}$. It is well known that the noisy KS equation develops ripples in an early stage followed by a crossover to a rough surface belonging to the universality class of the Kardar-Parisi-Zhang equation [11]. In the case of laser-induced etching the surface is exposed to erosion and diffusion only during a time interval $\Delta T = d/u$ (eq. 5). It is therefore perspicuous that the feed velocity of the confined laser beam can be chosen such that ripples have time to develop but are not degraded during ΔT. The competition between erosion and diffusion leading to a convective instability [3] is therefore believed to be the basic mechanism leading to the ripple formation in laser-induced etching.

5 Surface analysis

In fig. 4a three surface profiles generated by the discrete model for different scan speeds are shown. All other system parameters are identical to those used in the simulation described in section 3, fig. 3. Again as an initial condition we start with a flat surface for $t = 0$. The profile at the top of fig. 4a with the smallest width has been obtained for the highest velocity $u = 4 \times 10^{-3}a$/step. In this early stage of surface evolution ripples have not enough time to develop. In contrast, for $u = 4 \times 10^{-4}a$/step the instability leads to fast growing amplitudes and a selection of modes with a wave length of about $20a$ which is of the order of magnitude of the active spot diameter. For lower beam velocities a crossover to rough surfaces can be observed. In fig. 4b the corresponding wavelet power spectra are shown. The power spectra are obtained from the wavelet transform

$$Wh(l,x) = \frac{1}{l}\int h(\xi)\,\overline{\psi}\left(\frac{\xi - x}{l}\right)\mathrm{d}\xi, \tag{25}$$

using the Morlet wavelet $\psi(\xi) = \dfrac{\mathrm{d}^2}{\mathrm{d}\,\xi^2}\left(\mathrm{e}^{-\xi^2/2}\,\mathrm{e}^{i\omega_0\xi}\right)$ with $\omega_0 = 10$ [15]. In eq. (25) l denotes the scale and x the shift parameter and $\overline{\psi}(\xi)$ is the complex conjugate of $\psi(\xi)$. The wavelet power spectrum is defined as

$$P_h^W(l) = \int |Wh(l,x)|^2 dx. \tag{26}$$

The plotted curves are averages of 10 realizations each with 8192 points. For a scan speed of $u = 4 \times 10^{-4}a$/step the wavelet power spectrum shows a dominant wavelength which is in contrast to corresponding spectrum of the

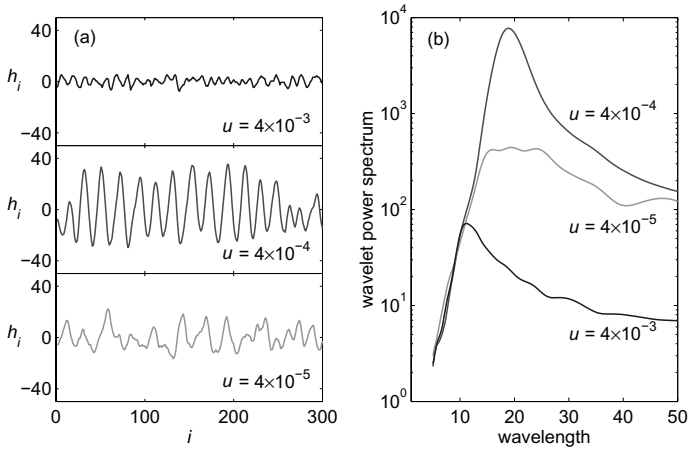


Fig. 4. Surface morphology for different velocities (given in units a/time step) of the laser beam: a) height profiles and b) corresponding wavelet power spectra.

rough surface pertaining to the lowest velocity $u = 4 \times 10^{-5} a$/step. For small scales the power spectra of the 3 profiles have almost the same slope indicating similar fractal properties.

Power spectra either obtained from Fourier or wavelet transform only allow to estimate a *global* Hölder exponent H [17, 20]. Local fluctuations in the degree of roughness call for location-dependent Hölder exponents $H(x)$. In turbulence, the standard way to extract the multiscaling properties of a function $h(x)$ is to study the scaling behaviour of the nth order structure functions $S_n(r) = <\delta h_r^n> \sim r^{\zeta_n}$ of the increments $\delta h_r = h(x+r/2) - h(x-r/2)$. In fig. 5a the top view of the etched kerf from fig. 1a is displayed. The structure functions $S_n(r)$ for $n = 2, 4, 6, 8$ are evaluated as averages of the height profiles at the bottom of the kerf and are shown in fig. 5b. There is a clear separation between scaling at small scales and structures on larger scales. Multifractal behaviour leads to a nonlinear scaling exponent ζ_n. The spectrum $f(H)$ of Hölder exponents is obtained by Legendre transforming the exponents ζ_n leading to $f(H) = \min_n(nH - \zeta_n + 1)$ which is, however, restricted to $n \geq 0$ and $0 < H < 1$.

For the evaluation of the spectrum of Hölder exponents on small scales from the experimental data we first applied the wavelet transform modulus maxima (WTMM) method using Mexican hat wavelets which allows a reliable estimation of the full spectrum of Hölder exponents if the singularities are non-oscillating and phase transitions do not occur [16, 17, 19, 20]. In this method, a partition function $Z(n, l)$ is calculated from the skeleton of maxima lines of the wavelet transform

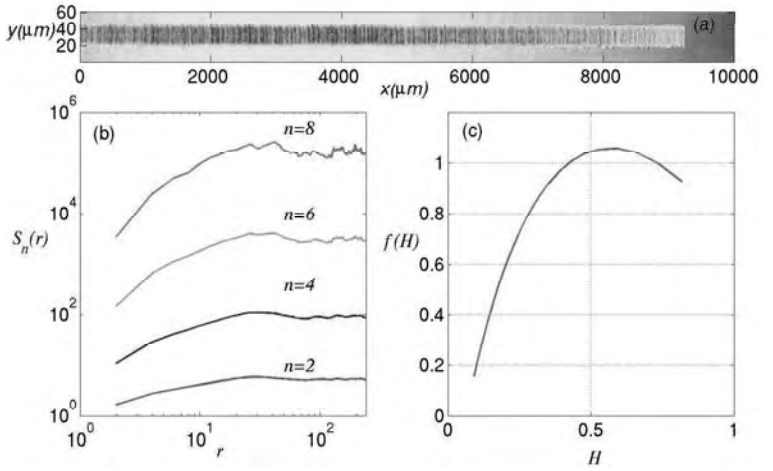


Fig. 5. Multifractal properties of surface profile: a) Top view of the etched kerf shown in fig. 1a, b) structure functions and c) the spectrum of Hölder exponents.

$$Z(n,l) = \sum_{x_i \in \max.\text{lines}} \left(\sup_{l' \leq l} |Wh(l', x_i)| \right)^n \tag{27}$$

From the power law behaviour of the partition function, $Z(n,l) \sim l^{\tau(n)}$ for $l \to 0^+$, the whole spectrum of Hölder exponents $f(H)$ is obtained by a direct method proposed by Chhabra and Jensen [22] thus avoiding Legendre transforming the exponent $\tau(n)$. For details the reader is referred to [17,19,20]. The result is shown in fig. 5c. Although the range of scales that can be analyzed is very small due to a limited resolution of the experimental measurements the surface at the bottom of the kerf clearly shows a multifractal structure.

In the present case, however, an estimation of the multifractal spectrum by means of the large deviation spectrum $f_g(H)$ is more adequate since different processes might be involved [18]. The large deviation spectrum reflects the scaling behaviour of coarse grain Hölder exponents on a sequence of interval partitions. The spectra shown in fig. 6 are calculated with the software tool FracLab for three resolutions [18]. Fig. 6a shows $f_g(H)$ for the range $4\mu m$ to $32\mu m$ corresponding to the scaling range of the structure functions displayed in fig. 5b. In fig. 6b the scaling range $15\mu m$ to $40\mu m$ corresponds to the range chosen for the WTMM method. With the WTMM method the surface roughness seems to be underestimated. For the discrete model the scaling range for the rough surface is even smaller preventing us to analyse the multifractal properties and to compare it with the measured data. Here, further investigations are necessary.

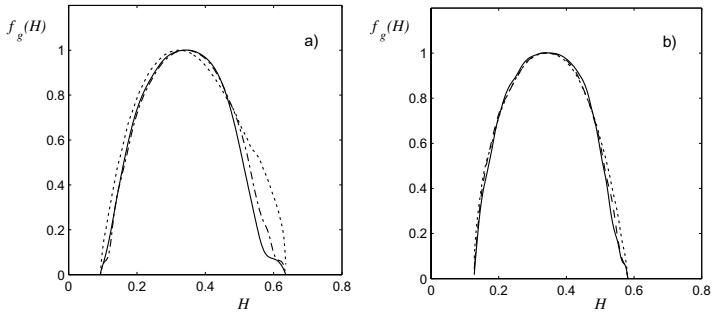


Fig. 6. Large deviation spectra $f_g(H)$ of the etched kerf shown in fig. 1a: a) scaling range $4\mu m$ to $32\mu m$ corresponding to the scaling range of the structure functions shown in fig. 5b, b) scaling range $15\mu m$ to $40\mu m$ corresponding to the range chosen for fig. 5a.

6 Conclusions

In summary, we have presented a simple discrete stochastic model for the description of laser-induced wet-chemical etching. Depending upon the feed velocity of the laser beam the surface profile is either rough or displays a ripple structure. The formation of ripples is explained by a linear instability of the moving front evolving for a range of laser velocities in the active zone of the beam and a subsequent 'freezing' of the structures. The continuum equation derived for the microscopic model takes the form of a modified noisy Kuramoto-Sivashinsky equation in a comoving coordinate frame. For a validation of the model the surface structure of the etched kerfs is analyzed. In its present form the discrete model does not take into account the effect of transport limitations due to a layer of reaction products above the surface. This effect will be included in an extended $2+1$ dimensional model which also has to be compared quantitatively to experimental observations. Our aim is to develop strategies for the etching process which avoid the formation of ripples.

We gratefully acknowledge the financial support of the Volkswagen-Stiftung (grant I/74151) and thank for the fruitful discussions with Rudolf Friedrich, Stefan J. Linz, Alexei Kouzmitchev, Andreas Stephen, Simeon Metev and Ferenc Kun.

References

1. S. Metev, A. Stephen, J. Schwarz, C. Wochnowski, Laser-induced chemical micro-treatment and synthesis of materials, RIKEN Review **50**, 47–52 (2003).
2. T.J. Rabbow, A. Mora, M. Haase, P.J. Plath, Selforganised structure formation in organised microstructuring by laser-jet etching, to be published in Int. J. Bif. Chaos.

3. R. Friedrich, G. Radons, T. Ditzinger, A. Henning, Ripple formation through an interface instability from moving growth and erosion sources, Phys. Rev. Lett. **85**, 4884–4887 (2000).
4. R. Cuerno, H.A. Makse, S. Tomassone, S.T. Harrington, H.E. Stanley, Stochastic model for surface erosion via ion sputtering: dynamical evolution from ripple morphology to rough morphology, Phys. Rev. Lett. **75**, 4464–4467 (1995).
5. A.-L. Barabási, H.E. Stanley: *Fractal Concepts an Surface Growth*, Cambridge Univ. Press (1995).
6. P. Sigmund, Theory of Sputtering. I. Sputtering Yield of Amorphous and Polycrystalline Targets, Phys. Rev. **184**, 383–416 (1969).
7. L. Bergmann, C. Clemens, *Lehrbuch der Experimentalphysik*, Bd.**3** Optik, Gruyter (2004).
8. Y. Lawrence Yao, *Laser Machining processes*, Section 2.9: Reflection and Absorption of Laser Beams. http://www.columbia.edu/cu/mechanical/mrl/ntm/level2/ch02/html/l2c02s09.html
9. A.A. Golovin, A.A. Nepomnyashchy, S.H. Davis, M.A. Zaks, Convective Cahn-Hillard models: from coarsening to roughening, Phys. Rev. Lett. **86**, 1550 (2001).
10. D. D. Vvedensky, A. Zangwill, C. N. Luse, M. R. Wilby, Stochastic equations of motion for epitaxial growth, Phys. Rev. E **48**, 852–862 (1993).
11. K.B. Lauritsen, R. Cuerno, H.A. Makse, Noisy Kuramoto-Sivashinsky equation for an erosion model, Phys. Rev. E **54**, 3577–3580 (1996).
12. M. Předota, M. Kotrla, Stochastic equations for simple discrete models of epitaxial growth, Phys. Rev. E **54**, 3933–3942 (1996).
13. D. D. Vvedensky, Edward-Wilkinson equation from lattice transition rules, Phys. Rev. E **67**, 025102 (2003).
14. H. Risken, *The Fokker-Planck equation*, Springer, Berlin (1996).
15. S. Mallat,*A wavelet tour of signal processing*, Academic Press, San Diego (1998).
16. S. Jaffard, Some open problems about multifractal functions, in: *Fractals in Engineering* (J. Lévy Véhel E. Lutton, C. Tricot eds.), Springer, London (1997).
17. J.F. Muzy, E. Bacry and A. Arnéodo, The multifractal formalism revisited with wavelets, Int. J. Bif. Chaos **4**, 245–302 (1994).
18. C. Canus, J. Lévy Véhel, C. Tricot, Continuous large deviation multifractal spectrum: definition and estimation, in: *Fractals and Beyond* (M. M. Novak ed.), World Scientific, Singapore, 117–128 (1998). FracLab software: URL: http//www-rocq.inria.fr/fractales.
19. M. Haase, B. Lehle, Tracing the skeleton of wavelet transform maxima lines for the characterization of fractal distributions, in: *Fractals and Beyond* (M. M. Novak ed.), World Scientific, Singapore, 241–250 (1998).
20. M. Haase, A. Mora, B. Lehle, Multifractal and stochastic analysis of electropolished surfaces, in: *Thinking in Patterns* (M. M. Novak ed.), World Scientific, Singapore, 69–78 (2004).
21. A. Mora, M. Haase, T. Rabbow, P.J. Plath, A discrete model for laser driven etching and microstructuring of metallic surfaces, http://arxiv.org/abs/cond-mat/0503093 and submitted to Phys. Rev. E.
22. A. Chhabra, R.V. Jensen, Direct determination of the $f(\alpha)$ singularity spectrum, Phys. Rev. Lett. **62** 1327–1330 (1989).

Invariant structures and multifractal measures in $2d$ mixing systems

Massimiliano Giona, Stefano Cerbelli, and Alessandra Adrover

Dipartimento di Ingegneria Chimica, Facoltà di Ingegneria, Università di Roma "La Sapienza", via Eudossiana 18, 00184 Roma, Italy
max@giona.ing.uniroma1.it

Summary. This article analyzes the relationship between geometric invariant structures in two-dimensional mixing systems and measure-theoretical properties associated with the spatial distribution of stable/unstable manifolds of periodic points. Specifically, a connection is established between the Bowen measure associated with the spatial distribution of periodic points and the w-measures characterizing the distribution of stable/unstable leaves throughout the mixing space. This result is made possible through the introduction of the concept of *symmetric product* of two measures.

1 Introduction

Fluid mixing is a peculiar field of investigation for which the chaotic motion of fluid particles is advisable in that it is associated with efficient stirring [1, 2]. Basic mixing mechanisms are primarily influenced by invariant geometric templates represented by the unstable invariant manifolds of the periodic points [3–5].

Extensive numerical investigation has shown that the spatial structure of the unstable manifolds possesses multifractal nature (multifractal properties can be defined by introducing the w-measures associated with the spatial length distributions of the unstable fibers) [6, 7]. In point of fact, the occurrence of multifractal properties associated with the measure-theoretical characterization of invariant fibers seems to be a specific property of physically realizable mixing systems (differently from simple paradigms of mixing and uniform hyperbolicity such as Anosov toral automorphisms) [8].

It is therefore interesting to analyze in greater detail the properties of the multifractal measures arising in the characterization of the structure of the fibers of the stable/unstable foliations, and to derive connections between these measures and other invariant structure characterizing mixing dynamics (specifically the spatial distribution of the periodic points).

This article attempts to derive this connection, by focusing on the relationship between the spatial structure of the periodic points of period up to n

(at large n), expressed by the Bowen measure [9,10], and the w-measures associated with the stable and unstable foliations. The analysis is developed by considering a prototypical two-dimensional model on the torus, which shares the main relevant phenomenological properties with physically realizable two-dimensional mixing systems.

The investigation of the relation between the Bowen measure and the w-measures associated with stable/unstable manifolds leads to the introduction of the concept of symmetric product of two measures, which is e.g. a useful tool for generating multifractal measures.

The article is organized as follows. Section 2 describes the model system considered and its properties. Section 3 defines the problem and reviews the measure-theoretical characterization of the stable/unstable fibers. Section 4 analyzes the measure theoretical properties of the invariant structures. Section 5 introduces the concept of symmetric product of two measures and applies it to obtain a relation between the w-measures and the Bowen measure.

2 Prototypical model for two-dimensional mixing

As a prototypical model for two-dimensional area-preserving systems possessing nonuniform chaotic behavior (roughly speakly, this means that their statistical features are different from those of uniformly hyperbolic systems), we consider a continuous transformation $\mathcal{H} : T^2 \to T^2$ of the two-dimensional torus defined by

$$\mathcal{H}(\mathbf{x}) = \mathcal{H}(x,y) = \begin{pmatrix} x + f(y) \\ y + x + f(y) \end{pmatrix} \quad \text{mod. } 1\,, \tag{1}$$

where $f(\xi) : [0,1] \to [0,1]$ is the periodicized tent map ($f(\xi) = 2\xi$ for $0 < \xi \leq 1/2$ and $f(\xi) = 2 - 2\xi$ for $1/2 < \xi \leq 1$), (x,y) are coordinates on the unit square interval $I^2 = [0,1) \times [0,1)$ equipped with periodic boundary conditions ("mod. 1"), which represents a global projection chart for the two-torus T^2. The transformation $\mathcal{H}$ is area-preserving and globally continuous. The map $\mathcal{H}$ expressed by Eq. (1) can be regarded as the stroboscopic map of a time-continuous flow resulting from the periodic alternation of two mutually orthogonal unidirectional steady flows on the two-torus T^2, namely $\mathbf{v}_1(\mathbf{x}) = (v_{1,x}, v_{1,y}) = (f(y), 0)$ acting for a time $\tau = 1$, followed by a linear shear, $\mathbf{v}_2 = (v_{2,x}, v_{2,y}) = (0, x)$ along the x-axis also acting for a unit time, as depicted in Fig. 1.

The torus meridians $y = 0$ and $y = 1/2$ are singularity lines for the differential $\mathcal{H}^*(\mathbf{x})$ of $\mathcal{H}$. In fact, $\mathcal{H}^*(\mathbf{x})$ is piecewise constant and equal to

$$\begin{aligned} H_1 &= \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix} \quad \text{for } 0 \leq x < 1,\ 0 < y < 1/2, \\ H_0 &= \begin{pmatrix} 1 & -2 \\ 1 & -1 \end{pmatrix} \quad \text{for } 0 \leq x < 1,\ 1/2 < y < 1. \end{aligned} \tag{2}$$

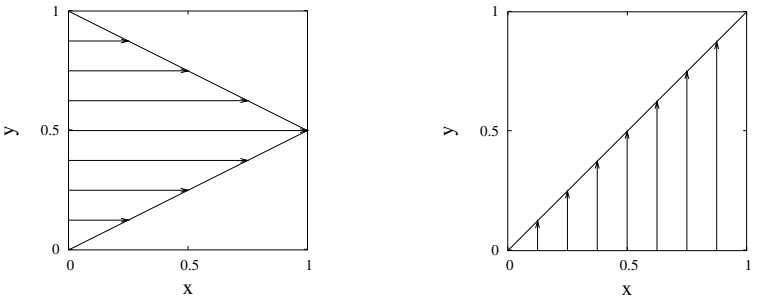


Fig. 1. The two steady flows that generate the time-periodic flow on the two torus by blinking alternately on and off for a unit time. The toral homeomorphism $\mathcal{H}$ is the time-2 map of the time-periodic flow so defined.

An important property that allows us to derive analytically several results is that the torus T^2 can be decomposed in three disjoint subsets A, B, C (see Fig. 2 (A)), undergoing a Markov-chain dynamics: $\mathcal{H}(A) \subset A \cup B$, $\mathcal{H}(B) = C$, and $\mathcal{H}(C) \subset A$. Specifically, stemming from this decomposition,

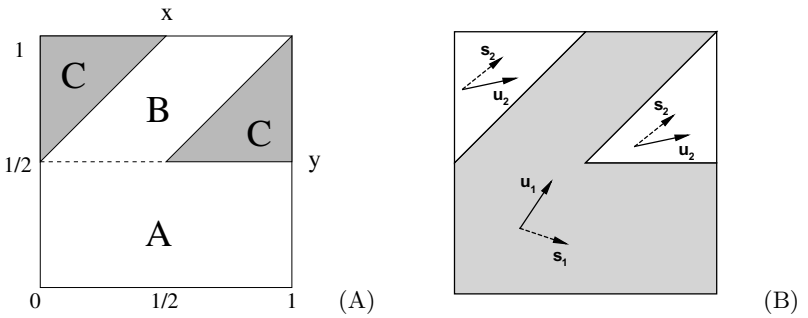


Fig. 2. (A) Decomposition of the two-torus into three disjoint subsets A, B, C undergoing Markov dynamics. (B) Structure of the invariant fields of dilating and contracting directions of the map $\mathcal{H}$.

it is possible to prove that the map $\mathcal{H}$ fulfils the basic properties of a "chaotic" transformation of T^2, namely mixing (topological and measure-theoretical), existence of a positive Lyapunov exponent, $\Lambda = (1/2)\log(2+\sqrt{3})$, positivity of the topological entropy h_{top} which proves strictly greater than Λ [11]. The strict inequality $h_{\text{top}} > \Lambda$ is the macroscopic outcome of complex correlations of stretching events associated with typical trajectories, which is a feature commonly observed in physically realizable mixing systems.

As it regards the focus of this paper, the map $\mathcal{H}$ is particularly appealing for theoretical investigation in that it is possible to prove that the tangent

bundle is the direct sum of two invariant sub-bundles $\mathcal{E}^u = \{E^u_{\mathbf{x}}\}_{\mathbf{x}\in T^2}$ and $\mathcal{E}^s = \{E^s_{\mathbf{x}}\}_{\mathbf{x}\in T^2}$, the stable and unstable sub-bundles respectively, and to obtain an explicit expression for these sub-bundles:

$$E^u_{\mathbf{x}} = \begin{cases} \text{span}\{\mathbf{u}_1\} & \text{if } \mathbf{x} \in A \bigcup B \\ \text{span}\{\mathbf{u}_2\} & \text{if } \mathbf{x} \in C \end{cases} \qquad E^s_{\mathbf{x}} = \begin{cases} \text{span}\{\mathbf{s}_1\} & \text{if } \mathbf{x} \in A \bigcup B \\ \text{span}\{\mathbf{s}_2\} & \text{if } \mathbf{x} \in C \end{cases} . \tag{3}$$

where $\mathbf{u}_h$, $\mathbf{s}_h$, $h = 1, 2$ are given by:

$$\mathbf{u}_1 = \begin{pmatrix} 2 \\ \lambda_u - 1 \end{pmatrix}, \qquad \mathbf{s}_1 = \begin{pmatrix} 2 \\ \lambda_s - 1 \end{pmatrix}. \tag{4}$$

where $\lambda_u = 2 + \sqrt{3}$ and $\lambda_s = 2 - \sqrt{3}$, and

$$\mathbf{u}_2 = H_0 \cdot \mathbf{u}_1, \qquad \mathbf{s}_2 = H_0 \cdot \mathbf{s}_1, \tag{5}$$

Vectors belonging to $E^s_{\mathbf{x}}$ and $E^u_{\mathbf{x}}$ shrink to vanishing norm by the iterative application of the forward and backward tangent dynamics, respectively.

The explicit expression for the stable and unstable sub-bundles is useful for unveiling the invariant geometric structure associated with $\mathcal{H}$, since it is straightforward to construct the elements (leaves or fibers) of the stable ($\mathcal{F}^s$) and unstable ($\mathcal{F}^u$) foliations, which are the Lipshitz curve tangent a.e. to $E^s_{\mathbf{x}}$ and $E^u_{\mathbf{x}}$, respectively. More precisely, elements of the foliation $\mathcal{F}^u$ are the invariant manifolds $\mathcal{W}^u_{\mathbf{x}^*}$ associated with any periodic point of $\mathcal{H}$ of period p. The manifold $\mathcal{W}^u_{\mathbf{x}^*}$ is invariant under $\mathcal{H}^p$. For points $\mathbf{x} \in \mathcal{W}^u_{\mathbf{x}^*}$ it follows that $\mathcal{H}^{-n\,p}(\mathbf{x}) \to \mathbf{x}^*$, for $n \to \infty$. Invariant manifolds of hyperbolic periodic points do not exhaust all the possible elements of $\mathcal{F}^u$, since the foliation $\mathcal{F}^u$ is formed also by all the other integral manifolds tangent to the unstable sub-bundle, the points of which do not converge neither forward ($n \to \infty$) or backward ($n \to -\infty$) to any periodic point. These integral manifolds can be referred to as the unstable wandering manifolds. The collection of all the unstable manifolds associated with periodic points of $\mathcal{H}$ and of all the unstable wandering manifold is $\mathcal{F}^u$, and $\mathcal{F}^u$ is invariant under $\mathcal{H}$. Its elements are referred to as unstable leaves (or fibers). Similar observations apply to the stable foliation $\mathcal{F}^s$ (with $n \mapsto -n$).

To give an example, Fig. 3 depicts a portion of a unstable and of a stable leaf (Panels (A) and (B) respectively).

Since the leaves of the unstable foliation constitute the basic geometric template around which fluid mixing is organized, it is of practical and conceptual importance to understand their geometric properties. This is developed in the next section.

3 Invariant foliations and w-measures

This section review briefly the concept of w-measure and states the main problem addressed in this article.

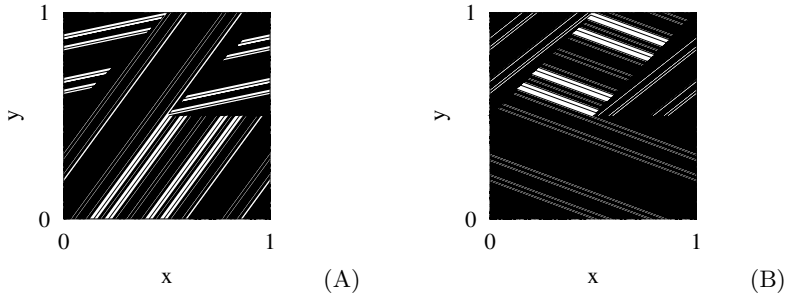


Fig. 3. (A) Unstable leaf and (B) stable leaf associated with the map $\mathcal{H}$.

Let $\{k_n\}_{n=1}^{\infty}$ be a monotonically increasing and diverging sequence of positive numbers, and let $\mathcal{W}^u_{\mathbf{x}^*}(k_n)$ be a finite portion of $\mathcal{W}^u_{\mathbf{x}^*}$ possessing length k_n. Let $L(\mathcal{W})$ be the length-function, which returns the overall length of a finite collection $\mathcal{W}$ of curve arcs.

The w-measure $\mu_{w_u}(D)$ of a measurable set D is the limit for $n \to \infty$

$$\mu_{w_u}(D) = \lim_{n\to\infty} \frac{L(\mathcal{W}^u_{\mathbf{x}^*}(k_n) \cap D)}{L(\mathcal{W}^u_{\mathbf{x}^*}(k_n))} \,. \tag{6}$$

Alternatively, one can introduce an equivalent "dynamic" definition of μ_{w_u} by considering the iterates of a local unstable manifold $\mathcal{W}^u_{\mathbf{x}^*,loc}$ (of finite length). In this case, the w-measure μ_{w_u} can be expressed as

$$\mu_{w_u}(D) = \lim_{n\to\infty} \frac{L(\mathcal{H}^n(\mathcal{W}^u_{\mathbf{x}^*,loc}) \cap D)}{L(\mathcal{H}^n(\mathcal{W}^u_{\mathbf{x}^*,loc}))} \,. \tag{7}$$

It follows from Eqs. (6)-(7) the geometrical meaning of μ_{w_u}: $\mu_{w_u}(D)$ is the length fraction of any sufficiently long arc of a generic leaf of $\mathcal{F}^u$ falling within the set D. Analogous definitions can be given for the w-measure μ_{w_s} associated with the stable foliation (of course n in Eq. (7) should be replaced by $-n$).

The w-measures are normalized (probability) measures (i.e. $\mu_{w_u}(T^2) = \mu_{w_s}(T^2) = 1$), and their estimate is independent of the choice of the leaf of the foliation [6, 7].

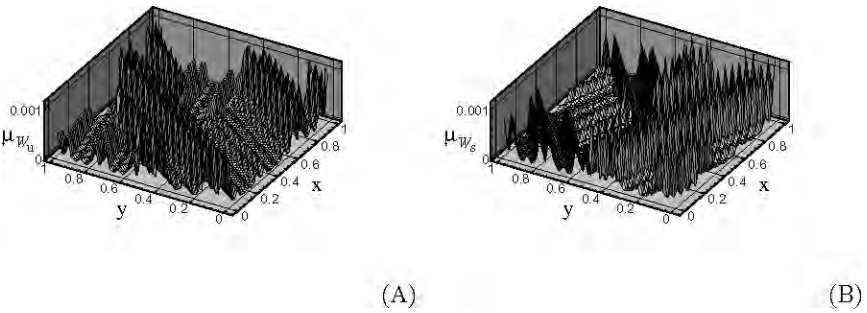
We are now able to state the basic issue of this paper. Essentially, this can be formulated as follows: *what is the relation between the w-measures and the structure of the periodic points?* Alternatively, this issue can be reformulated as follows: *how are the w-measures and the Bowen measure associated with the periodic points related?*

4 Multifractal properties of w- and Bowen measures

This Section analyzes the multifractal properties of the w- and of the Bowen measures associated with the prototypical map $\mathcal{H}$.

4.1 w-measures

Figures 4 (A)-(B) show the structure of the w-measures of the stable and unstable foliations of $\mathcal{H}$. In these figures, the measures $\mu_{w_u}(Q_{ij})$, $\mu_{w_s}(Q_{ij})$ are depicted, where Q_{ij} is a partition of T^2 into $N \times N$ equal squares ($N = 64$). One observes the highly nonuniform and supposedly singular structure of these measures.



(A) (B)

Fig. 4. w-measures associated with the transformation $\mathcal{H}$. (A) μ_{w_u} on a 64×64 grid. (B) μ_{w_s} on a 64×64 grid.

It is useful to connect the w-measures with a statistical characterization of the intersection of a leaf of the unstable foliation with a generic cross-section Γ. Given a generic cross section Γ (i.e. a smooth curve, parametrized with respect to the curvilinear abscissa ξ, $\xi \in [0, L_\Gamma]$) , and a sufficiently long piece of unstable leaf $\mathcal{W} = \mathcal{W}_{\mathbf{x}^*}(k_n)$, consider the intersections of $\mathcal{W}$, with Γ. The local distribution of these intersection points along Γ can be described by means of a normalized intersection measure μ^*, the support of which is the interval $[0, L_\Gamma]$. For simplicity, let $\mu^*(\xi) = \mu^*([0, \xi])$. By definition, $\mu^*(\xi)$ is the fraction of intersections of $\mathcal{W}$ and Γ falling within the interval $[0, \xi]$. Let us further introduce a "coarse" Probability Density Function (PDF) $\rho(\xi, \Delta)$ associated with μ^*, and with a characteristic size $\Delta = 2^{-n} L_\Gamma$, where the cross-section Γ is the circumference $\Gamma = \{ (x, y) \mid 0 \leq x < 1 \quad y = 1/4\}$.

Figure 5 shows the PDF $\rho(x, \Delta)$ of such intersections (in this figure, x is the curvilinear abscissa of the cross section) at two different discretization levels, $n = 10$ and $n = 14$. The data indicate clearly that the PDFs do not converge to any limit function (note that the scale of y-axis is different in the two panels). Conversely, the intersection measures $\mu^*(x)$ estimated by using a partition of the cross section into intervals of length 2^{-n} (with $n = 10$, 14) converge to a unique distribution (see Fig. 5 (C)-(D)). This is a qualitative indication that

the intersection measure, and the w-measures, are not Lebesgue-absolutely continuous.

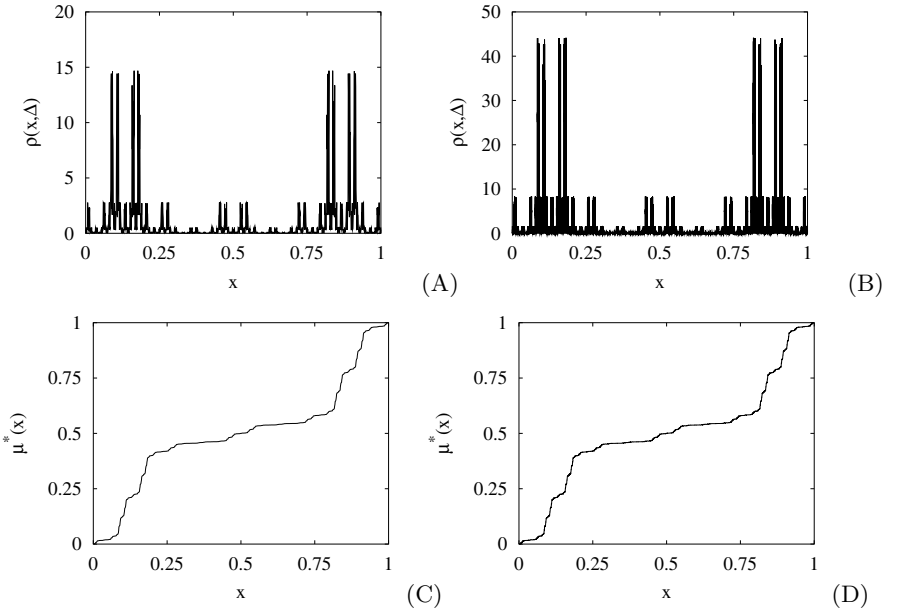


Fig. 5. Panels (A) and (B): intersection PDF for $\Delta = 2^{-10}$ (A), and for $\Delta = 2^{-14}$ (B). Panels (C) and (D): invariant measure $\mu^*(x)$ associated with the intersections of the unstable leaves with the curve $y = 1/4$ for $\Delta = 2^{-10}$ (C), and for $\Delta = 2^{-14}$ (D).

In the light of the analysis developed in [6], these qualitative properties suggest that the singularity of the w-measures and of the intersection measure can be described within the multifractal formalism, i.e. by considering the spectrum of generalized dimension $D(q)$ and the multifractal spectrum $f(\alpha)$ [12,13]. Specifically, the $f(\alpha)$-spectrum of a measure is the fractal dimension of the measure support characterized by a local Hölder (singularity) exponent α [12,13].

Figure 6 (A) shows the spectrum of generalized dimensions $D(q)$ as a function of the parameter q for the w-measures associated with the stable and unstable foliations. The scaling analysis for the multifractal properties has been performed by considering a partition of the w-measure up to 1024×1024 boxes, by analyzing invariant fibers possessing length order of 10^8, and number of intersections with the given cross-section order of 10^7.

As expected, the spectra of μ_{w_u} and μ_{w_s} coincide. This means that the singular structure associated with the spatial distribution of the stable leaves

is identical to that of the unstable manifolds, even though the two w-measures are different (see Fig. 4). This result appears to be a typical feature of area-preserving chaotic transformations, for which the stretching along the unstable directions is compensated by the shrinking along the stable sub-bundle.

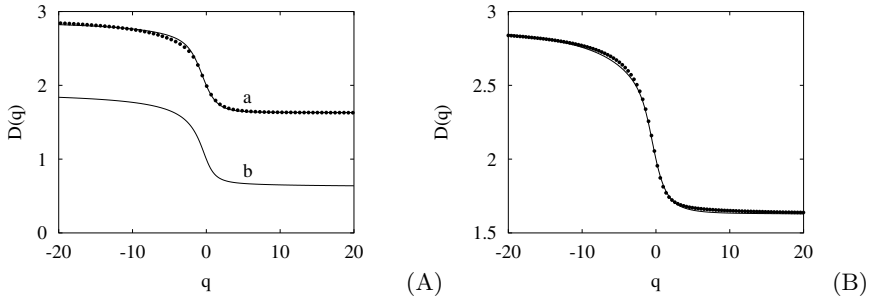


Fig. 6. Spectrum of generalized dimensions $D(q)$ vs q. (A) Line (a) and dots (•) w-measures associated with the unstable foliation (continuous line) and stable foliation (•). Line (b) intersection measure with the cross section at $y = 1/4$ associated with the unstable foliation. (B) Comparison of the spectrum of generalized dimensions associated with the stable foliation (continuous line) and $D(q)+1$ deriving from the intersection measure at $y = 1/4$ associated with the unstable foliation (dots •).

Figure 6-(A) also shows the spectrum of generalized dimensions associated with the intersection measure μ^* along the cross section $y = 1/2$ (line b). Let $D_w(q)$ be the generalized-dimension spectrum of the w-measures and $D_i(q)$ that of the intersection measure (with a generic cross section). Geometrical observations suggest the following relation:

$$D_w(q) = D_i(q) + 1 \,, \tag{8}$$

which is a consequence of the fact that unstable (stable) invariant fibers are locally smooth and rectifiable. The validity of Eq. (8) is confirmed by the data depicted in Fig. 6 (B).

This result is further supported by the analysis of the $f(\alpha)$ spectra depicted in Fig. 7 (A)-(B). The condition corresponding to Eq. (8) for the $f(\alpha)$ spectra is

$$f_w(\alpha) = f_i(\alpha + 1) + 1 \,, \tag{9}$$

where $f_w(\alpha)$ and $f_i(\alpha)$ are the $f(\alpha)$-spectra of the w-measure and of the intersection measure, respectively. The validity of Eq. (9) is supported by the data depicted in Fig. 7.

4.2 Bowen measure

An important motivation for introducing the map $\mathcal{H}$ is that it provides a simple archetype of nonuniform chaos for which many significant properties

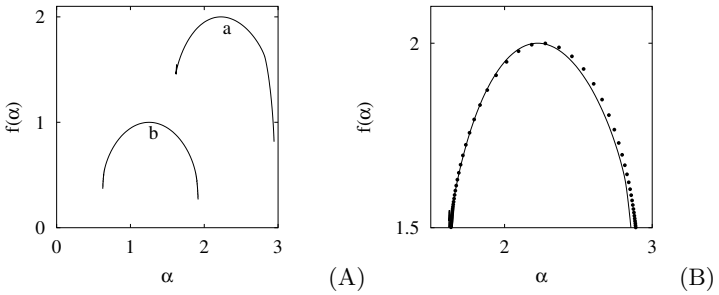


Fig. 7. Multifractal spectrum $f(\alpha)$ vs α. (A) Line (a) w-measure associated with the stable foliation, line (b) intersection measure at $y = 1/4$ associated with the unstable foliation. (B) Comparison of the $f(\alpha)$-spectrum of the w-measure associated with the stable foliation (solid line) and the curve $f(\alpha) + 1$ vs $\alpha + 1$ associated with the intersection measure depicted in (A).

are suitable to a direct analytic approach. Moreover, even when analytical results cannot be easily derived, the formal simplicity of $\mathcal{H}$ permits to approach geometrical and measure-theoretical properties with high (in some cases arbitrarily high) numerical accuracy. This is for example the case of the structure of the periodic points, and the connection with the geometry of the stable/unstable foliations. Starting from the articles by Procaccia and coworkers [14,15], it is clear that the dynamic properties of chaotic systems are organized around the structure of the periodic orbits, which are dense within the chaotic region. Periodic orbits form the skeleton around which chaotic dynamics is organized. For instance, shadowing properties, and the characterization of stretching dynamics can be related to the properties of the periodic points.

A considerable amount of work in the recent literature has been focused on the analysis of periodic points of chaotic maps. The mathematical literature analyzed the convergence properties of the Bowen measure [9,16]. To explain shortly how the Bowen measure is constructed, consider the set $Per(n)$, of the periodic points of period less than or equal to n. A sequence of measures, $\mu_{per}^{(n)}$ associated with this set at increasing n can be defined as:

$$\mu_{Per}^{(n)} = \frac{1}{N_{Per}(n)} \sum_{\mathbf{x}_i \in Per(n)} \delta(\mathbf{x} - \mathbf{x}_i) \; , \tag{10}$$

where $N_{Per}(n)$ is the cardinality of $Per(n)$, i.e. the total number of periodic points of period $\leq n$. This sequence of measures converges in the weak-$*$ limit to a measure μ_{Bow}, which characterizes the spatial distribution of periodic points [10,17]. The measure μ_{Bow} is referred to as the *Bowen measure*. Therefore, the Bowen measure is a normalized (probability) measure that yields the

fraction of period points of period less than or equal to n that fall within the set D in the limit where n diverges to infinity.

While the connection between the Bowen measure and the invariant ergodic maximal measure (which for area-preserving transformations is the uniform Lebesgue measure) has received great attention, no results have been produced as it regards the relationship of the Bowen measure and the invariant geometric properties, to the best of our knowledge. The remainder of this Section is devoted to the analysis of this connection, by taking the map $\mathcal{H}$ as an example.

Firstly, the property that the map $\mathcal{H}$ results from gluying two linear transformations associated with integer-valued matrices (H_0 and H_1, defined in Eq. (2)) simplifies significantly the estimate of the periodic points, due to the fact that each integer lattice Z_q

$$Z_q = \{\ (x,y)\ |\ x = i/q\,,\ y = j/q\quad i,j,q \in \mathbb{N}\,,\quad 0 \le i,j < q\ \} \tag{11}$$

is an invariant set for the map $\mathcal{H}$, which is formed exclusively by periodic orbits of $\mathcal{H}$. Within each Z_q, the action of $\mathcal{H}$ can be transformed into an exact integer dynamics. Consequently, periodic orbits can be estimated exactly by considering the restriction of $\mathcal{H}$ to the integer lattice Z_q. We have considered the restriction $\mathcal{H}|_{Z_q}$ to integer lattices Z_q, where q ranges up to $q_{\max} = 80.000$. Periodic points of low period are localized within integer lattices possessing relatively low values of q. For example, no new periodic points of prime period $n = 15$ arise for $q > 40.000$, and this lends confidence that the computation performed captures all of the periodic points up to $n = 15, 16$.

Figures 8 (A)-(B)-(C) show the spatial distribution of the periodic points of period n for different values of $n = 10, 12, 14$. The spatial distribution of periodic points is highly nonuniform and possesses self-affine symmetries. The resulting limit measure, namely the Bowen measure, is therefore highly nonuniform and concentrated mainly in those regions where both the stable and the unstable w-measure are preferentially concentrated. In fact, the latter observation is crucial in the light of a connection between the distribution of the periodic orbits and the invariant manifold structure.

We suggest that the Bowen measure could be related to the distribution of the intersections of the stable and unstable invariant manifolds. In order to support this idea, consider a sequence of finite length approximations of the stable and unstable manifolds, $\mathcal{W}^s_{\mathbf{x}^*}(k_m)$ and $\mathcal{W}^u_{\mathbf{x}^*}(k_m)$, respectively associated with a fixed point $\mathbf{x}^*$ of $\mathcal{H}$. Here, k_m indicates the length of the manifold portions, and the sequence k_m diverges to infinity as $m \to \infty$. Let $I_m = \{\mathbf{x}_i^{(m)}\}_{i=1}^{N_m}$ be the intersections of $\mathcal{W}^s_{\mathbf{x}^*}(k_m)$ with $\mathcal{W}^u_{\mathbf{x}^*}(k_m)$. Figure 8 (D) shows the spatial distribution of these intersection points. The analogy between this distribution and that of periodic points depicted in Fig. 8 (A)-(C) is striking. The measure of the intersections between stable and unstable leaves, to which we refer shortly as the *intersection measure* $\mu_i^{(u,s)}$, can thus be defined as

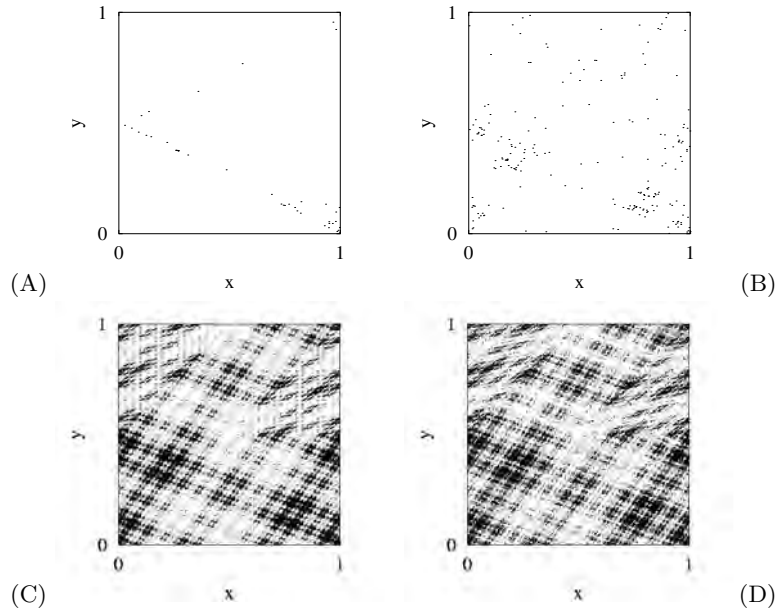


Fig. 8. Periodic points of period n of the map $\mathcal{H}$. (A) $n = 10$, (B) $n = 12$, (C) $n = 14$, (D) Intersection of the stable and unstable leaves.

$$\mu_i^{(u,s)} = \lim_{m\to\infty} \frac{1}{N_m} \sum_{\mathbf{x}_i \in I_m} \delta(\mathbf{x} - \mathbf{x}_i^{(m)}) \,, \tag{12}$$

where the convergence is understood in the weak-$*$ meaning [17]. The intersection measure $\mu_i^{(u,s)}$ is independent of the choice of the leaves within the stable and the unstable foliations, i.e. identical results are obtained by choosing any pair of leaves from the stable and unstable foliations, be them invariant manifolds of any periodic point or wandering leaves. The intersection measure permits to identify the connection between the structure of periodic points and the geometry of the invariant manifold *via* the relation:

$$\mu_{Bow} = \mu_i^{(u,s)} \,, \tag{13}$$

stating that the Bowen measure coincides with the intersection measure of the stable unstable foliations.

A quantitative indication of the validity of Eq. (13) stems from the analysis of the Fourier transforms of the measures μ_{Bow}, $\mu_i^{(u,s)}$. For any measure μ with support I^2, its Fourier transform is given by

$$C(h,k) = \int_{I^2} \exp[-i2\pi(hx + ky)]\, d\mu(\mathbf{x}) \,, \qquad h,k \in \mathbb{Z} \,. \tag{14}$$

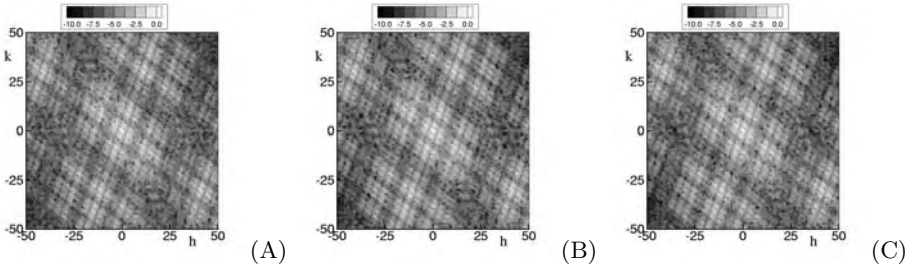


Fig. 9. Contour plots of the logarithm of the modulus of the Fourier transform of the measure vs h and k. (A) Bowen measure at $n = 15$; (B) Intersection measure of the stable and unstable manifolds. (C) Symmetric product measure Eq. (25).

Figures 9 (A)-(B) depict the contour plot of the logarithm of the modulus of the Fourier transforms associated with the Bowen measure (Fig. 9 (A)) and with the intersection measures of the stable and unstable leaves (Fig. 9 (B)). The agreement between these two transforms is excellent, and this is the best quantitative confirmation of the validity of Eq. (13).

5 Symmetric product measure

The final step is to connect directly the Bowen measure to the w-measures associated with the stable and unstable foliations. This goal can be achieved by introducing the concept of *symmetric product* of two measures.

Firstly, let us introduce the concept of symmetric product for measures defined on the real line. Without loss of generality, we can consider measures defined on the unit interval $I = [0, 1]$.

Let μ_1, μ_2 be two probability measures on I (a probability measure μ on I satisfies the condition $\mu(I) = 1$). Let $\{B_i\}_\varepsilon$ an ε-covering of I with intervals B_i, the maximum width of which is less than or equal to ε. If $B_i = (x_i^1, x_i^2)$ with $x_i^2 > x_i^1$ let $b_i = x_i^2 - x_i^1$.

The symmetric product measure of μ_1, μ_2, is a normalized measure $\mu^{(p)}$ defined as follows. Consider the quantity $\mu^{(p)}(A, \{B_i\}_\varepsilon)$ defined as:

$$\mu^{(p)}(A, \{B_i\}_\varepsilon) = \frac{\sum_i \mu_1(A \cap B_i)\, \mu_2(A \cap B_i)/b_i}{\sum_i \mu_1(B_i)\, \mu_2(B_i)/b_i} \tag{15}$$

Let $\mu^{(p)}(A, \varepsilon)$ be the infimum of $\mu^{(p)}(A, \{B_i\}_\varepsilon)$ over all the ε-coverings of I. The symmetric product measure $\mu^{(p)}$ of μ_1 and μ_2 is the limit of $\mu^{(p)}(A, \varepsilon)$ for $\varepsilon \to 0$:

$$\mu^{(p)}(A) = \lim_{\varepsilon \to 0} \mu^{(p)}(A, \varepsilon) \tag{16}$$

Henceforth, we will indicate the symmetric product of two measures with the symbol "$\odot$", i.e.

$$\mu^{(p)} = \mu_1 \odot \mu_2 \tag{17}$$

The construction of the symmetric product measure is similar to the classical procedure applied in the definition of the Hausdorff measure [12]. Indeed, since $\mu^{(p)}(A, \varepsilon)$ is a non decreasing function of ε, and since $\mu^{(p)}(A, \varepsilon) \leq 1$, it follows that the limit Eq. (16) exists.

The symmetric product measure satisfies the following elementary properties. (I) It is a normalized measure, i.e.

$$\mu^{(p)}(I) = 1 \ , \tag{18}$$

(II) It is symmetric with respect to the two measures μ_1, μ_2, i.e.

$$\mu_1 \odot \mu_2 = \mu_2 \odot \mu_1 \ . \tag{19}$$

(III) If μ_1, μ_2 are Lebesgue absolutely continuous, i.e. there exist the densities $p_1(x)$, $p_2(x)$, such that $d\mu_i(x) = p_i(x)\, dx$ $(i = 1, 2)$, then

$$\mu^{(p)}(A) = \int_A p_1(x)\, p_2(x)\, dx \Big/ \int_I p_1(x)\, p_2(x)\, dx \ , \tag{20}$$

which follows straightforwardly from Eqs. (15)-(16).

The definition of symmetric product measure introduced above can be straightforwardly extended to higher dimensional domains. Consider two measures μ_1, μ_2 possessing supports $S_1, S_2 \subseteq D \subseteq R^d$, where D is a measurable d-dimensional set $(d \geq 1)$, possessing non-zero Lebesgue measure. Let $\{B_i\}_\varepsilon$ be a ε-covering of D. This means that the diameters $|B_i|$ of the subsets B_i are smaller than or equal to ε, $(|B_i| \leq \varepsilon)$. Let $A \subseteq D$ be a generic measurable set. Instead of Eq. (15) one may consider the quantity

$$\mu^{(p)}(A, \{B_i\}_\varepsilon) = \frac{\sum_i \mu_1(A \cap B_i)\mu_2(A \cap B_i)/|B_i|^d}{\sum_i \mu_1(B_i)\mu_2(B_i)/|B_i|^d} \ . \tag{21}$$

For $d = 1$ one recovers Eq. (15) from Eq. (21).

The symmetric product measure $\mu^{(p)}(A) = \mu_1 \odot \mu_2(A)$ is defined as the limit for $\varepsilon \to 0$ of the infimum over all the possible ε-coverings of $\mu^{(p)}(A, \{B_i\}_\varepsilon)$:

$$\mu^{(p)}(A) = \mu_1 \odot \mu_2(A) = \lim_{\varepsilon \to 0} \inf_{\{B_i\}_\varepsilon} \mu^{(p)}(A, \{B_i\}_\varepsilon) \ . \tag{22}$$

There is a further generalization of SPM that is worth addressing. Consider a continuous function $q(\mathbf{x})$ defined on D, and attaining non negative values $(q(\mathbf{x}) \geq 0$, for $\mathbf{x} \in D)$. The q-weighted SPM (referred to as q-SPM) of μ_1 and μ_2 can be defined as follows. For any B_i of the ε-covering $\{B_i\}_\varepsilon$, let $\xi_i \in B_i$ a generic internal point. The q-weighted approximation $\mu^{(p)}_{q(\mathbf{x})}(A, \{B_i\}_\varepsilon)$ for the product measure of μ_1 and μ_2 on the covering $\{B_i\}_\varepsilon$ is given by:

$$\mu^{(p)}_{q(x)}(A,\{B_i\}_\varepsilon) = \frac{\sum_i q(\xi_i)\,\mu_1(A\cap B_i)\mu_2(A\cap B_i)/|B_i|^d}{\sum_i q(\xi_i)\,\mu_1(B_i)\mu_2(B_i)/|B_i|^d}\ , \tag{23}$$

and the q-weighted SPM $\mu^{(p)}_{q(\mathbf{x})}$ is given by:

$$\mu^{(p)}_{q(\mathbf{x})}(A) = (\mu_1 \odot \mu_2)_{q(\mathbf{x})}(A) = \lim_{\varepsilon\to 0}\inf_{\{B_i\}_\varepsilon} \mu^{(p)}_{q(\mathbf{x})}(A,\{B_i\}_\varepsilon)\ . \tag{24}$$

We are now able to state an important result connecting the Bowen measure to the w-measures. Geometrical arguments (not developed here for intrinsic space limitations) indicate that the intersection measure $\mu_i^{(u,s)}$ of the stable and unstable leaves can be expressed as the symmetric product measure of the two w-measures μ_{w_u}, μ_{w_s}, weighted with respect to the geometric factor $|\mathbf{e}_u(\mathbf{x})\times\mathbf{e}_s(\mathbf{x})|$, where $\mathbf{e}_u(\mathbf{x})$, $\mathbf{e}_s(\mathbf{x})$ are the unit tangent vectors spanning $E^u_{\mathbf{x}}$, $E^s_{\mathbf{x}}$ at point $\mathbf{x}$, respectively. This implies

$$\mu_{Bow} = \mu_i^{(u,s)} = (\mu_{w_u}\odot\mu_{w_s})_{|\mathbf{e}_u(\mathbf{x})\times\mathbf{e}_s(\mathbf{x})|}\ . \tag{25}$$

Eq. (25) is the main result relating the structure of the periodic points to the w-measures of the stable/unstable foliations. A quantitative confirmation of this result is depicted in Fig. 9 (C) which shows the Fourier transform of the symmetric product measure $(\mu_{w_u}\odot\mu_{w_s})_{|\mathbf{e}_u(\mathbf{x})\times\mathbf{e}_s(\mathbf{x})|}$.

Eq. (25) can be proved for simple dynamical systems, such as Anosov diffeomorphisms of T^2 that are smoothly conjugated with a linear hyperbolic toral automorphism. For generic nonlinear systems, no rigorous mathematical proof of Eq. (25) has been provided, so far. However, it is rather straightforward to give an elementary justification of Eq. (25). Consider two families of parallel fibers, such that each family is oriented according to a constant tangent vector (say $\mathbf{e}_u$, and $\mathbf{e}_s$). The number of intersections $N_{int}(D)$ between the fibers of the two families falling within a given domain D scales as $N_{int}(D)\sim L_u(D)\,L_s(D)\,|\mathbf{e}_u\times\mathbf{e}_s|$, where $L_u(D)$ and $L_s(D)$ are the overall lengths of the fiber segments of the two families falling within D. This result may be used locally, and applied to the case of the intersections of stable and unstable manifolds of dynamical systems, by considering that the leaves of $\mathcal{F}^u$ and $\mathcal{F}^s$ are locally rectifiable, and by observing that, in this transposition, $L_u(D)$ and $L_s(D)$ are quantified by the w-measures. The rigorous proof of Eq. (25), that we conjecture to hold for the class of nonuniformly hyperbolic systems (in the meaning envisaged by Pesin), is still to be prodiced, and is an important mathematical issue left over for further research.

6 Concluding remarks

Chaotic dynamics in two-dimensional mixing systems can be viewed as organized along the unstable invariant manifolds or around the periodic points. There is a strong connection between these two "qualitative" views, and this is

compactly expressed by Eq. (25). Eq. (25) is the relation connecting the spatial distribution of the periodic points and the invariant geometric properties associated with the stable and unstable leaves. A proof of Eq. (25) has been given only for particular cases, and a geometrical justification (supported by numerical evidence) has been provided for nonuniformly chaotic model systems. This is a satisfactory results for the application of Eq. (25) to physical problems (e.g. associated with fluid mixing), albeit the rigorous proof of Eq. (25) for nonuniformly hyperbolic system is a mathematical challenge left to future research.

A singular (i.e. not Lebesgue absolutely continuous) Bowen measure arises as a consequence of the multifractal spatial distribution of the leaves of $\mathcal{F}^u$ and $\mathcal{F}^s$. The non uniformity and singularity of the Bowen measure is a property characterizing nonuniformly chaotic systems (such that $h_{top} > \Lambda$).

Furthermore, the definition of symmetric product of two measure is interesting in itself. It finds application in the analysis of measure-theoretical properties of regular (Lipshitz) fibers, although it may be expect that it could be fruitfully applied for other purposes related to multifractal characterization of singular measures.

References

1. Aref H (1984) J Fluid Mech 143:1–21
2. Ottino J M (1989) The kinematics of mixing: stretching, chaos and transport, Cambridge Univ. Press, Cambridge
3. Beigie D, Leonard A, Wiggins S (1994) Chaos Solitons & Fractals 4:749–868
4. Rom-Kedar V, Leonard A, Wiggins S (1990) J Fluid Mech 214:347–394
5. Giona M, Adrover A, Muzzio F J, Cerbelli S, Alvarez, M M (1999) Physica D 132:298–324
6. Giona M, Adrover A (1998) Phys Rev Lett 81:3864–3867.
7. Adrover A, Giona M (1999) Phys Rev E 60:357–362
8. Giona M, Cerbelli S, Muzzio F J, Adrover A (1998) Physica A 253 451–465
9. Bowen R (1971) Trans Amer Math Soc 154:377–397
10. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems, Cambridge Univ. Press, Cambridge
11. Cerbelli S, Giona M (2004) A continuous archetype of nonuniform chaos (unpublished)
12. Halsey T C, Jensen M H, Kadanoff L P, Procaccia I, Shraiman B I, (1986) Phys Rev A 33:540–550
13. Falconer K (1990) Fractal geometry, John Wiley & Sons, New York
14. Auerbach D, Cvitanovic P, Eckmann J-P, Gunaratne G, Procaccia I (1987) Phys Rev Lett 58:2387–2389
15. Gunaratne G, Procaccia I (1987) Phys Rev Lett 59.1377–1380
16. Bowen R (1972) American J Math 94:1–30
17. Parthasarathy K R (1967) Probability measures on metric spaces, Academic Press, New York

4

FINANCE

Long range dependence in financial markets

Rama Cont

Centre de Mathématiques appliquées, Ecole Polytechnique, France.
www.cmap.poytechnique.fr/~rama/

Summary. The notions of self-similarity, scaling, fractional processes and long range dependence have been repeatedly used to describe properties of financial time series: stock prices, foreign exchange rates, market indices and commodity prices. We discuss the relevance of these properties in the context of financial modelling, their relation with the basic principles of financial theory and possible economic explanations for their presence in financial time series.

1 Introduction

The study of statistical properties of financial time series has revealed a wealth of interesting stylized facts which seem to be common to a wide variety of markets, instruments and periods [15, 21, 30, 59]:

- **Excess volatility:** many empirical studies point out to the fact that it is difficult to justify the observed level of variability in asset returns by variations in "fundamental" economic variables. In particular, the occurrence of large (negative or positive) returns is not always explainable by the arrival of new information on the market [18].
- **Heavy tails:** the (unconditional) distribution of returns displays a heavy tail with positive excess kurtosis.
- **Absence of autocorrelations in returns:** (linear) autocorrelations of asset returns are often insignificant, except for very small intraday time scales ($\simeq$ 20 minutes) where microstructure effects come into play.
- **Volatility clustering:** as noted by Mandelbrot [48], "large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes." A quantitative manifestation of this fact is that, while returns themselves are uncorrelated, absolute returns $|r_t|$ or their squares display a positive, significant and slowly decaying autocorrelation function: $\mathrm{corr}(|r_t|, |r_{t+\tau}|) > 0$ for τ ranging from a few minutes to a several weeks.

- **Volume/volatility correlation:** trading volume is positively correlated with market volatility. Moreover, trading volume and volatility show the same type of "long memory" behavior [43].

The dependence properties of asset returns and the phenomenon of volatility clustering have especially intrigued many researchers and oriented in a major way the development of stochastic models in finance –GARCH models and stochastic volatility models are intended primarily to model this phenomenon. Also, it has inspired much debate as to whether there is long-range dependence in volatility.

Since the 1990s we have witnessed a surge of interest in this topic with the availability of new sources of financial data. A large number of empirical studies on asset prices have investigated long range dependence properties of asset returns. The concepts of self-similarity, scaling, fractional processes and long range dependence have been repeatedly used to describe properties of financial time series such as stock prices, foreign exchange rates, market indices and commodity prices.

While there is a vast literature on long range dependence in asset prices, most authors tackle the questions either from a purely theoretical perspective or from a purely empirical one, rarely both. We will attempt to discuss the relevance of these notions in the context of financial modelling both at a conceptual level, in relation with the basic principles of financial theory, and at an empirical level, by comparing them to properties of market data. Finally, we will briefly discuss some possible economic explanations for the presence of such properties in financial time series.

2 Dependence properties of financial time series

Denote by S_t the price of a financial asset — a stock, an exchange rate or a market index — and $X_t = \ln S_t$ its logarithm. Given a *time scale* Δ, the log return at scale Δ is defined as:

$$r_t = X_{t+\Delta} - X_t = \ln(\frac{S_{t+\Delta}}{S_t}). \tag{1}$$

Δ may vary between a minute (or even seconds) for tick data to several days. Observations are sampled at discrete times $t_n = n\Delta$. Time lags will be denoted by the Greek letter τ; typically, τ will be a multiple of Δ in estimations. For example, if Δ =1 day, $\text{corr}[r_{t+\tau}, r_t]$ denotes the correlation between the daily return at period t and the daily return τ periods later.

2.1 Empirical behavior of autocorrelation functions

A typical display of daily log-returns is shown in figure 1: the volatility clustering feature is seen graphically from the presence of sustained periods of high

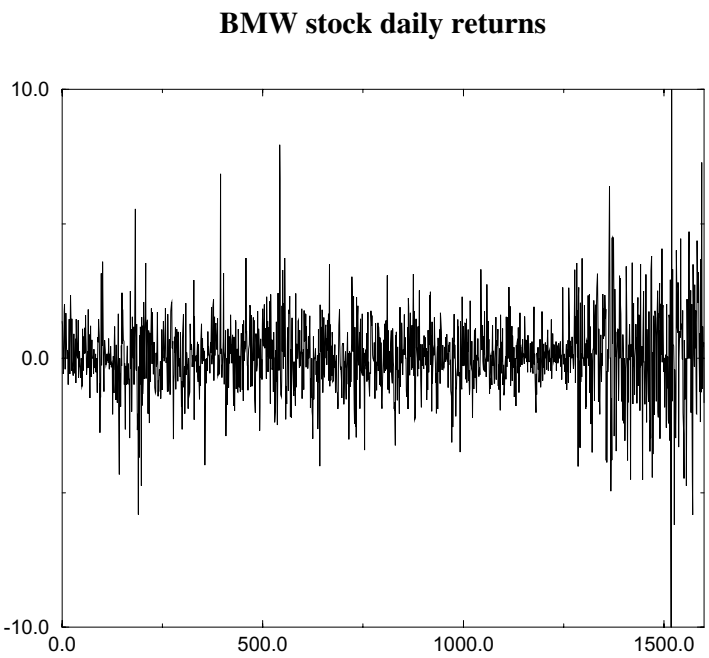


Fig. 1. Large changes cluster together: BMW daily log-returns. $\Delta = 1$ day.

or low volatility. As noted above, the autocorrelation of returns is typically insignificant at lags between a few minutes and a month. An example is shown in figure 2 (left). This "spectral whiteness" of returns can be attributed to the activity of arbitrageurs who exploit linear correlations in returns via trend following strategies [49]. By contrast, the autocorrelation function of absolute returns remains positive over lags of several weeks and decays slowly to zero: figure 2 (right) shows this decay for SLM stock (NYSE). This observation is remarkably stable across asset classes and time periods and is regarded as a typical manifestation of volatility clustering [11, 16, 21, 30]. Similar behavior is observed for the autocorrelation of squared returns [11] and more generally for $|r_t|^\alpha$ [16, 21, 22] but it seems to be most significant for $\alpha = 1$ i.e. absolute returns [21].

GARCH models [11, 24] were among the first models to take into account the volatility clustering phenomenon. In a GARCH(1,1) model the (squared) volatility depends on last periods volatility:

$$r_t = \sigma_t \mathrm{var}\epsilon_t \quad \sigma_t^2 = a_0 + a\sigma_{t-1}^2 + b\mathrm{var}\epsilon_t^2 \qquad 0 < a + b < 1 \tag{2}$$

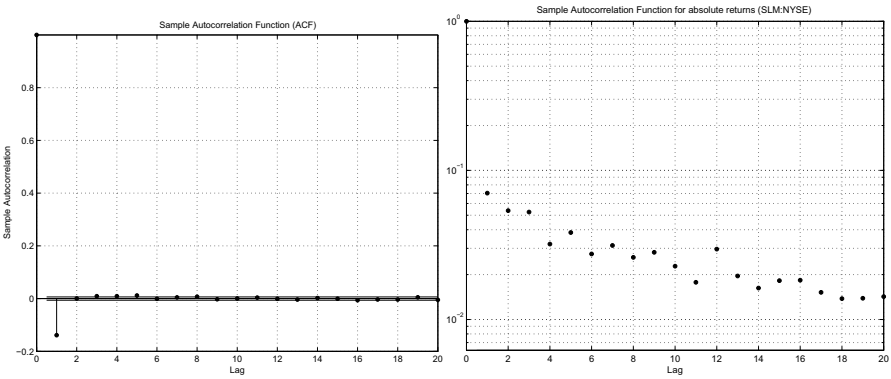


Fig. 2. SLM stock, NYSE, $\Delta = 5$ minutes. Left: autocorrelation function of log-returns. Right: autocorrelation of absolute log-returns.

leading to positive autocorrelation in the volatility process σ_t, with a rate of decay governed by $a + b$: the closer $a + b$ is to 1, the slower the decay of the autocorrelation of σ_t. The constraint $a + b < 1$ allows for the existence of a stationary solution, while the upper limit $a + b = 1$ corresponds to the case of an integrated process. Estimations of GARCH(1,1) on stock and index returns usually yield $a + b$ very close to 1 [11]. For this reason the volatility clustering phenomenon is sometimes called a "GARCH effect"; one should keep in mind however that volatility clustering is a "non-parametric" property and is not intrinsically linked to a GARCH specification.

While GARCH models give rise to exponential decay in autocorrelations of absolute or squared returns, the empirical autocorrelations are similar to a power law [16, 30]:

$$C_{|r|}(\tau) = \text{corr}(|r_t|, |r_{t+\tau}|) \simeq \frac{c}{\tau^\beta}$$

with an exponent $\beta \leq 0.5$ [5, 16], which suggests the presence of "long-range" dependence in amplitudes of returns, discussed below.

2.2 Long range dependence

Let us recall briefly the commonly used definitions of long range dependence, based on the autocorrelation function of a process:

Definition 1 (Long range dependence). *A stationary process Y_t (with finite variance) is said to have* long range dependence *if its autocorrelation function $C(\tau) = \text{corr}(Y_t, Y_{t+\tau})$ decays as a power of the lag τ:*

$$C(\tau) = \text{corr}(Y_t, Y_{t+\tau}) \underset{\tau \to \infty}{\sim} \frac{L(\tau)}{\tau^{1-2d}} \qquad 0 < d < \frac{1}{2} \tag{3}$$

where L is slowly varying at infinity, i.e. verifies $\forall a > 0, \frac{L(at)}{L(t)} \to 1$ *as* $t \to \infty$.

By contrast, one speaks of "short range dependence" if the autocorrelation function decreases at a geometric rate:

$$\exists K > 0, c \in]0,1[, \ |C(\tau)| \leq Kc^{\tau} \tag{4}$$

Obviously, (3) and (4) are not the only possibilities for the behavior of the autocorrelation function at large lags: there are many other possible decays rates, intermediate between a power decay and a geometric decay. However, it is noteworthy that in all stochastic models used in the financial modeling literature, the behavior of returns and their absolute values fall within one of the two categories.

Although there had been considerable development of statistical methods for processes with long-range dependence in the physical sciences, especially hydrology and agronomy, it was Granger [29] in 1966 who alerted the econometrics community to the ubiquity of time series with preponderance of spectral power near the origin, referring to this property as determining "the typical spectral shape of an economic variable".

2.3 Long range dependence and self-similarity

The long range dependence property (3) hinges upon the behavior of the autocorrelation function at *large* lags, a quantity which may be difficult to estimate empirically [9]. For this reason, models with long-range dependence are often formulated in terms of self-similar processes, which allow to extrapolate across time scales and deduce long time behavior from short time behavior, which is more readily observed. A stochastic process $(X_t)_{t\geq 0}$ is said to be self-similar if there exists $H > 0$ such that for any scaling factor $c > 0$, the processes $(X_{ct})_{t\geq 0}$ and $(c^H X_t)_{t\geq 0}$ have the same law:

$$(X_{ct})_{t\geq 0} \stackrel{d}{=} (c^H X_t)_{t\geq 0}. \tag{5}$$

H is called the self-similarity exponent of the process X. Note that a self-similar process cannot be stationary, so the above definition of long-range dependence cannot hold for a self-similar process, but eventually for its increments (if they are stationary).

In 1968 Mandelbrot and Van Ness [53] provided the connection between self-similar processes and long-range dependence in stationary time series via fractional Gaussian noise, and produced its spectral density $f(\lambda) \sim c_H|\lambda|^{1-2H}(\frac{1}{2} < H < 1)$ with an integrable pole at the origin, leading to the notion of "$1/f$-noise". Fractional Brownian motion is a typical example of self-similar process whose increments exhibit long range dependence: a fractional Brownian motion with self-similarity exponent $H \in]0,1[$ is a real centered Gaussian process with stationary increments $(B_t^H)_{t\geq 0}$ with covariance function:

$$\text{cov}(B_t^H, B_s^H) = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}). \tag{6}$$

For $H = 1/2$ we recover Brownian motion. For $H \neq 1/2$, the covariance of the increments decays very slowly, as a power of the lag; for $H > 1/2$ this leads to long-range dependence in the increments [53, 65].

But self-similarity does not imply long-range dependence in any way: α-stable Lévy processes provide examples of self-similar processes with *independent* increments. Nor is self-similarity implied by long range dependence: Cheridito [13] gives several examples of Gaussian processes with the same long range dependence features as fractional Brownian noise but with no self-similarity (thus very different "short range" properties and sample path behavior).

Comparing fractional Brownian motions and α-stable Lévy processes shows that self-similarity can have very different origins: it can arise from high variability, in situations where increments are independent and heavy-tailed (stable Lévy processes) or it can arise from *strong dependence* between increments even in absence of high variability, as illustrated by the example of fractional Brownian motion. These two mechanisms for self-similarity have been called the "Noah effect" and the "Joseph effect" by Mandelbrot [50]. By mixing these effects, one can construct self-similar processes where both long range dependence and heavy tails are present: fractional stable processes [3,65] offer such examples.

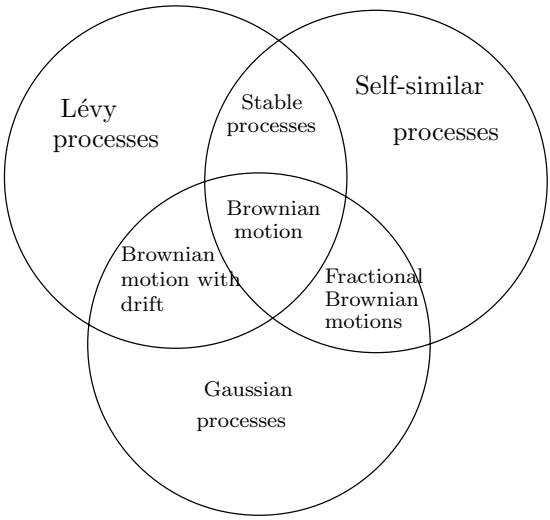


Fig. 3. Self-similar processes and their relation to Lévy processes and Gaussian processes.

2.4 Are stock prices self-similar?

As noted above, the example of fractional Brownian motion is thus misleading in this regard, since it conveys the idea that these two properties are associated. When testing for long range dependence in a model based on fractional Brownian motion, we thus test the joint hypothesis of self-similarity *and* long-range dependence and strict self-similarity is not observed to hold in asset returns [15, 16]. One should therefore distinguish general tests for self-similarity from tests of particular parametric models (such as α-stable Lévy processes or fractional Brownian motions).

A consequence of selfsimilarity is that for any $c, t > 0$, X_{ct} and $c^H X_t$ have the same distribution. Choosing $c = 1/t$ yields

$$\forall t > 0, \quad X_t \stackrel{d}{=} t^H X_1, \tag{7}$$

so the distribution of X_t, for any t, is completely determined by the distribution of X_1:

$$F_t(x) = \mathbb{P}(t^H X_1 \leq x) = F_1(\frac{x}{t^H}). \tag{8}$$

In particular if the tail of F_1 decays as a power of x, then the tail of F_t decays in the same way:

$$\mathbb{P}(X_1 \geq x) \underset{x\to\infty}{\sim} \frac{C}{x^\alpha} \Rightarrow \ [\forall t > 0, \ \mathbb{P}(X_1 \geq x) \underset{x\to\infty}{\sim} C\frac{t^{\alpha H}}{x^\alpha} = \frac{C(t)}{x^\alpha}]. \tag{9}$$

If F_t has a density ρ_t we obtain, by differentiating (8), the following relation for the densities:

$$\rho_t(x) = \frac{1}{t^H}\rho_1\left(\frac{x}{t^H}\right). \tag{10}$$

Substituting $x = 0$ in (10) yields the following scaling relation:

$$\forall t > 0, \ \rho_t(0) = \frac{\rho_1(0)}{t^H}. \tag{11}$$

Let us now consider the moments of X_t. From (7) it is obvious that $E[|X_t|^k] < \infty$ if and only if $E[|X_1|^k] < \infty$ in which case

$$E[X_t] = t^H E[X_1], \quad \text{var}(X_t) = t^{2H}\text{var}(X_1), \tag{12}$$

$$E[|X_t|^k] = t^{kH} E[|X_1|^k]. \tag{13}$$

Assume that the log-price $X_t = \ln S_t$ is a process with *stationary increments*. Since $X_{t+\Delta} - X_t$ has the same law as X_Δ, the density and moments of X_Δ can be estimated from a sample of increments.

The relation (11) has been used by several authors to test for self-similarity and estimate H from the behavior of the density of returns at zero: first one

estimates $\rho_t(0)$ using the empirical histogram or a kernel estimator and then obtains an estimate of H as the regression coefficient of $\ln \rho_t(0)$ on $\ln t$:

$$\ln \hat{\rho}_t(0) = H \ln \frac{t}{\Delta} + \ln \hat{\rho}_\Delta(0) + \epsilon. \tag{14}$$

Applying this method to S&P 500 returns, Mantegna and Stanley [54] obtained $H \simeq 0.55$ and concluded towards evidence for an α-stable model with $\alpha = 1/H \simeq 1.75$. However, the scaling relation (11) holds for any self-similar process with exponent H and does not imply in any way that the process is a (stable) Lévy process. For example, (11) also holds for a fractional Brownian motion with exponent H — a Gaussian process with correlated increments having long range dependence! Scaling behavior of $\rho_t(0)$ is simply a necessary but not a sufficient condition for self-similarity: even if (11) is verified, one cannot conclude that the data generating process is self-similar and even less that it is an α-stable process.

Another method which has often been used in the empirical literature to test self-similarity is the "curve collapsing" method: one compares the aggregation properties of empirical densities with (10). Using asset prices sampled at interval Δ, one computes returns at various time horizons $n\Delta, n = 1 \ldots M$ and estimates the marginal density of these returns (via a histogram or a smooth kernel estimator). The scaling relation (10) then implies that the densities $\hat{\rho}_{n\Delta}(x)$ and $\frac{1}{n^H}\hat{\rho}_\Delta(\frac{x}{n^H})$ should coincide, a hypothesis which can be tested graphically and also more formally using a Kolmogorov–Smirnov test.

Although self-similarity is not limited to α-stable processes, rejecting self-similarity also leads to reject the α-stable Lévy process as a model for log-prices. If the log-price follows an α-stable Lévy process, daily, weekly and monthly returns should also be α-stable (with the same α). Empirical estimates [1, 10] show a value of α which increases with the time horizon. Finally, various estimates of tail indices for most stocks and exchange rates [1, 33, 35, 44–46] are often found to be larger than 2, which rules out infinite variance and stable distributions.

2.5 Dependence in stock returns

The volatility clustering feature indicates that asset returns are not independent across time; on the other hand the absence of linear autocorrelation shows that their dependence is nonlinear. Whether this dependence is "short range" or "long range" has been the object of many empirical studies.

The idea that stock returns could exhibit long range dependence was first suggested by Mandelbrot [49] and subsequently observed in many empirical studies using R/S analysis [52]. Such tests have been criticized by Lo [42] who pointed out that, after accounting for short range dependence, they might yield a different result and proposed a modified test statistic. Lo's statistic highly depends on the way "short range" dependence is accounted for and

shows a bias towards rejecting long range dependence [67]. The final empirical conclusions are therefore less clear [68].

However, the absence of long range dependence in returns may be compatible with its presence in absolute returns or "volatility". As noted by Heyde [32], one should distinguish long range dependence in signs of increments, when $\text{sign}(r_t)$ verifies (3), from long range dependence in amplitudes, when $|r_t|$ verifies (3). Asset returns do not seem to possess long range dependence in signs [32]. Many authors have thus suggested models, such as Fractionally Integrated GARCH models [6], in which returns have no autocorrelation but their amplitudes have long range dependence [5, 23].

It has been argued [7, 39] that the decay of $C_{|r|}(\tau)$ can also be reproduced by a superposition of several exponentials, indicating that the dependence is characterized by multiple time scales. In fact, an operational definition of long range dependence is that the time scale of dependence in a sample of length T is found to be of the order of T: dependence extends over the whole sample.[1] Interestingly, the largest time scale in [39] is found to be of the order of...the sample size, a prediction which would be compatible with long-range dependence!

Many of these studies test for long range dependence in returns, volatility,.. by examining sample autocorrelations, Hurst exponents etc. but if time series of asset returns indeed possess the two features of heavy tails *and* long range dependence, then many of the standard estimation procedures for these quantities may fail to work [9, 62]. For example, sample autocorrelation functions may fail to be consistent estimators of the true autocorrelation of returns in the price generating process: Resnick and van der Berg [63] give examples of such processes where sample autocorrelations converge to *random* values as sample size grows! Also, in cases where the sample ACF is consistent, its estimation error can have a heavy-tailed asymptotic distribution, leading to large errors. The situation is even worse for autocorrelations of squared returns [56]. Thus, one must be cautious in identifying behavior of *sample* autocorrelation with the autocorrelations of the return process.

Slow decay of sample autocorrelation functions may possibly arise from other mechanism than long-range dependence. For example, Mikosch & Starica [57] note that nonstationarity of the returns may also generate spurious effects which can be mistaken for long-range dependence in the volatility. However, we will not go to the extreme of suggesting, as in [57], that the slow decay of sample autocorrelations of absolute returns is a pure artefact due to non-stationarity. "Non-stationarity" does not suggest a modeling approach and it seems highly unlikely that unstructured non-stationarity would lead to such a robust, stylized behavior for the sample autocorrelations of absolute returns, stable across asset classes and time periods. The robustness of these empirical facts call for an explanation, which "non-stationarity" does not provide. Of course, these mechanisms are not mutually exclusive: a recent

[1] On this point, see also [51].

study by Granger and Hyng [27] illustrates the interplay of these two effects by combining an underlying long memory process with occasional structural breaks.

3 Fractional processes and arbitrage constraints

A fallacy often encountered in the literature is that "long range dependence in returns is incompatible with absence of arbitrage", therefore ruled out by financial theory. This idea is so widespread that it is worthwhile discussing it here.

The problem stems from the fact that fractional Brownian motions and several related fractional processes do not belong to the class of *semimartingales*. We will review this notion briefly and discuss its implications for fractional models in finance.

3.1 Stochastic integrals and trading gains

Let us consider a financial asset whose price is modeled by a stochastic process S_t defined on a probability space $(\Omega, \mathcal{F}_t, \mathbb{P})$. If an investor trades at times $T_0 = 0 < T_1 < \cdots < T_n < T_{n+1} = T$, detaining a quantity ϕ_i of the asset during the period $]T_i, T_{i+1}]$ then the capital gain resulting from fluctuations in the market price is given by

$$\sum_{i=0}^{n} \phi_i (S_{T_{i+1}} - S_{T_i}). \tag{15}$$

This nonanticipative quantity, which represents the capital gain of the investor following the strategy ϕ, is called the stochastic integral of the process

$$\phi = \sum_{i=0}^{n} \phi_i 1_{]T_i, T_{i+1}]} \tag{16}$$

with respect to S and denoted by $\int_0^T \phi_t dS_t$. Here the trading times T_i can be nonanticipative random times –buys or sells can be triggered by recent price behavior– and ϕ_i are nonanticipative bounded random variables. ϕ is then called a *simple predictable process*: such processes are the mathematical representations of realistic trading strategies, which consist in buying and selling a finite number of times in $[0, T]$. Denote the set of simple predictable processes by $\mathbb{S}([0, T])$.

In the setting of Ito integration theory, stochastic integration is developed with respect to a class of stochastic processes known as *semimartingales*: these processes can be defined via their decomposition as a bounded variation process (signal) plus a local martingale (noise) [20] or, alternatively,

as processes S for which the stochastic integral defined by (15) is continuous[2] [55, 61] in the following sense: for any sequence of simple predictable processes $(\phi^n) \in \mathbb{S}([0,T])$ if

$$\sup_{(t,\omega)\in[0,T]\times\Omega} |\phi_t^n(\omega) - \phi_t(\omega)| \underset{n\to\infty}{\to} 0 \text{ then } \int_0^T \phi^n dS \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} \int_0^T \phi dS. \quad (17)$$

This stability property of stochastic integrals then allows to extend the set of integrands to processes ϕ which can be expressed as limits of *nonanticipative Riemann sums* –as in (15)– of simple predictable processes. This allows to consider more general strategies in a set $\mathcal{A}$ larger than $\mathbb{S}([0,T])$ and allow for "continuous trading". An important point is that the Ito integral is both nonanticipative and can be interpreted, as in (15), in terms of the gains from trading. This important remark, first pointed out in [31], isolates Ito integration theory as the appropriate one for use in financial modelling.

Fractional Brownian motion with $H \neq 1/2$ is *not* a semimartingale. This result can be shown in several ways, see [64]. The implication is that, in a model where the stock price is described by a function of a FBM, one cannot extend the construction of the gains process (15) to strategies beyond $\mathbb{S}([0,T])$ in a continuous way. Alternative constructions of the stochastic integral do exist: various extensions have been proposed which allow to construct stochastic integrals with respect to fractional Brownian motion and several authors have attempted to construct financial models using them. But it should be noted from the onset that such approaches are doomed to produce results whose financial interpretation is dubious: the *only* integral which can be interpreted in terms of the capital gain of a trading strategy is (15), and any other one will either anticipate on the future and/or not coincide with the gains of simple strategies.

However, it should be kept in mind that the set $\mathbb{S}([0,T])$ already contains any reasonable trading strategy. And, for computing the gain $\int_0^T \phi.dS$ of such strategies there is no need for S to be a semimartingale. Thus fractional processes with any sample path structure can be used as long as we limit ourselves to constructive problems (as opposed to existence theorems such as martingale representations) based on simple strategies.

In fact, $\mathbb{S}([0,T])$ already contains too many unrealistic strategies, which require to trade very frequently since the number of trades n can be arbitrarily large. One can define restricted sets where such infinitely frequent trading is excluded (see below).

3.2 Martingales, semi-martingales and arbitrage

An arbitrage strategy is defined as a strategy $\phi \in \mathcal{A}$ which realizes a possibly non-zero gain by starting from a zero initial capital:

[2] The fact that these two definitions of semimartingales coincide is a deep result, due to Dellacherie-Mokobodski-Meyer, see [61].

$$\phi_0 = 0 \qquad \mathbb{P}(\int_0^T \phi_t dS_t) > 0 \tag{18}$$

Note that the definition of arbitrage depends on the set of possible strategies $\mathcal{A}$ and on the definition of the stochastic integral.

Arbitrage pricing theory is based on a fundamental result of Harrison and Pliska [31], who show a model of price evolution is arbitrage-free if and only if the (discounted) price S_t of any asset can be represented as the conditional expectation of its final value S_T with respect to some probability measure $\mathbb{Q} \sim \mathbb{P}$:

$$\exists \mathbb{Q} \sim \mathbb{P}, \qquad S_t = E^{\mathbb{Q}}[S_T|\mathcal{F}_t] \tag{19}$$

In particular, S_t is a martingale under $\mathbb{Q}$. A precise statement of this results involves the specification of the set of admissible strategies $\mathcal{A}$, which is taken to be much larger than $\mathbb{S}([0,T])$ [19], usually containing all predictable processes ϕ such that $\int_0^t \phi dS$ is bounded.

Under the (real-world) model $\mathbb{P}$, S_t is not a martingale necessarily, but it is a still a semi-martingale: this property is preserved under equivalent changes of measure [61]. Since fractional Brownian motion is not a semimartingale, a model in which the (log)-price are described by a fractional Brownian motion is not arbitrage-free, in the sense that there exists a strategy $\phi \in \mathcal{A}$ verifying (18).

But this result and the fact that fractional Brownian motions fail to be semimartingales crucially depend on the *local* behavior of its sample paths, not on its long range dependence property. Cheridito [12] and Rogers [64] give several examples of Gaussian processes with the same long range dependence features as fractional Brownian motion, but which are semimartingales and lead to arbitrage-free models. A starting point for such constructions is the moving average represenation for fractional Brownian motion:

$$B_t^H = k \int_{-\infty}^{\infty} \{((t-s)^+)^{H-1/2} - (-s)^{H-1/2}\} dW_s,$$

where W_t is a Brownian motion, $H \in (0,1)$ is the self-similarity parameter and k is a suitable normalizing constant. Rogers [64] proposes a model which has the same long range dependence properties as Brownian motion but *is* a semimartingale, in the following way:

$$X_t = \int_{-\infty}^{t} \phi(t-s) dW_s + \int_{-\infty}^{0} \phi(-s) dW_s.$$

where $\phi \in C^2(\mathbf{R}), \phi(0) = 1,\ \phi'(0) = 1$ and $\lim_{t\to\infty} \phi"(t) t^{5/2-H} \in (0,\infty)$. An example of a kernel verifying this property is

$$\phi(t) = (\epsilon + t^2)^{(2H-1)/4}$$

Also, a closer look shows that *even* fractional Brownian motion and fractional processes are not ruled out by arbitrage considerations. The fact that FBM is not a semimartingale implied the existence of $\phi \in \mathcal{A}$ verifying (18): Rogers [64] offers examples of such strategies. However, these strategies can only be performed if it is possible to buy and sell within arbitrarily small time intervals. Arbitrage can be ruled out from fractional Brownian models by introducing a minimal amount of time h ¿ 0 that must lie between two consecutive transactions i.e. considering strategies in

$$\mathbb{S}^h([0,T]) := \{\sum_{i=0}^{n} \phi_i 1_{]T_i,T_{i+1}]} \in \mathbb{S}([0,T]), \quad \inf_i (T_{i+1} - T_i) > h\} \tag{20}$$

As shown by Cheridito [12], no arbitrage can be constructed using strategies in $\bigcap_{h>0} \mathbb{S}^h([0,T])$, i.e. no matter how frequently one trades.

Thus, the semi-martingale property is more a question of theoretical convenience, allowing not to worry constantly about the class of admissible strategies in theoretical developments, rather than a constraint on the models to be used. Finally, we have noted that long-range dependence has no relation with the semi-martingale property –which is a property of the fine structure of sample paths– and only when it is coupled to self-similarity (in the case of fractional processes) does it interfere with the semimartingale property.

But the main conclusion of this discussion is that the question of the adequacy of stochastic processes with long range dependence, and in particular models based on fractional Brownian motion, for modeling asset prices is mainly an *empirical* one: theoretical restrictions imposed by arbitrage are quite weak and cannot be used as arguments to exclude a family of stochastic processes as possible models. On the empirical side, however, there is a lot of evidence pointing to positive dependence over large time horizons in *absolute* returns [5, 15, 16, 21, 43, 58] but not in the returns themselves, showing that it is more interesting to use fractional processes as models of *volatility* rather than for modeling prices directly [6, 14, 58].

4 Economic mechanisms for long range dependence

While fractional processes may mimick volatility clustering in financial time series, they do not provide any economic explanation for it. The fact that these observations are common to a wide variety of markets and time periods [15] suggest that common mechanisms may be at work in these markets. Many attempts have been made to trace back the phenomenon of long range dependence in volatility to economic mechanisms present in the markets generating this volatility.

Independently of the econometric debate on the "true nature" of the return generating process, one can take into account such empirical observations without pinpointing a specific stochastic model by testing for similar behavior

of sample autocorrelations in such economic models and using sample autocorrelations for indirect inference [26] of the parameters of such models.

4.1 Heterogeneity in time horizons of economic agents

Heterogeneity in agent's time scale has been considered as a possible origin for various stylized facts [30]. Long term investors naturally focus on long-term behavior of prices, whereas traders aim to exploit short-term fluctuations. Granger [28] suggested that long memory in economic time series can be due to the aggregation of a cross section of time series with different persistence levels. This argument was proposed by Andersen & Bollerslev [2] as a possible explanation for volatility clustering in terms of aggregation of different information flows.

The effects of the diversity in time horizons on price dynamics have also been studied by Lebaron [38] in an artificial stock market, showing that the presence of heterogeneity in horizons may lead to an increase in return variability, as well as volatility-volume relationships similar to those of actual markets.

4.2 Evolutionary models

Several studies have considered modeling financial markets by analogy with ecological systems where various trading strategies co-exist and evolve via a "natural selection" mechanism, according to their relative profitability [4,38]. The idea of these models, the prototype of which is the Santa Fe artificial stock market [4, 40], is that a financial market can be viewed as a population of agents, identified by their (set of) decision rules. A decision rule is defined as a mapping from an agents information set (price history, trading volume, other economic indicators) to the set of actions (buy, sell, no trade). The evolution of agents decision rule is often modeled using a genetic algorithm. The specification and simulation of such evolutionary models can be quite involved and specialized simulation platforms have been developed to allow the user to specify variants of agents strategies and evolution rules. Other evolutionary models represent the evolution by a deterministic dynamical system which, through the complex price dynamics it generate, are able to mimick some "statistical" properties of the returns process, including volatility clustering [34]. However due to the complexity of the models they are not amenable to a direct comparison with financial data.

4.3 Switching between trading strategies

Another mechanism leading to long range dependence is switching of agents trading behavior between two or more strategies. The economic literature contains examples where switching of economic agents between two behavioral

patterns leads to large aggregate fluctuations [36]: in the context of financial markets, these behavioral patterns can be seen as trading rules and the resulting aggregate fluctuations as large movements in the market price i.e. heavy tails in returns. Recently, models based on this idea have also been shown to generate volatility clustering [37, 47].

Lux and Marchesi [47] study an agent-based model in which heavy tails of asset returns and volatility clustering arise from behavioral switching of market participants between fundamentalist and chartist behavior. Fundamentalists expect that the price follows the fundamental value in the long run. Noise traders try to identify price trends, which results in a tendency to herding. Agents are allowed to switch between these two behaviors according to the performance of the various strategies. Noise traders evaluate their performance according to realized gains, whereas for the fundamentalists, performance is measured according to the difference between the price and the fundamental value, which represents the anticipated gain of a "convergence trade". This decision-making process is driven by an exogenous fundamental value, which follows a Gaussian random walk. Price changes are brought about by a market maker reacting to imbalances between demand and supply. Most of the time, a stable and efficient market results. However, its usual tranquil performance is interspersed by sudden transient phases of destabilization. An outbreak of volatility occurs if the fraction of agents using chartist techniques surpasses a certain threshold value, but such phases are quickly brought to an end by stabilizing tendencies. This behavioral switching is believed be the cause of volatility clustering and heavy tails in the Lux-Marchesi model [47].

Kirman and Teyssière [37] have proposed a variant of [36] in which the proportion $\alpha(t)$ of fundamentalists in the market follows a Markov chain, of the type used in epidemiological models, describing herding of opinions. Simulation of this model exhibit autocorrelation patterns in absolute returns with a behavior similar to those observed in returns.

Ghoulmie, Cont and Nadal [17] propose a model where agents compare a common information (signal) to an individual threshold, whose value is heterogeneous across agents. These thresholds are dynamically updated based on recent price volatility. It is shown in [17] that, without any chartist/fundamentalist competition nor any direct interaction between agents, this model is capable of generating volatility clustering while maintaining absence of linear correlations in returns. This model points to a link between investor inertia and volatility clustering and provide an economic explanation for the switching mechanism proposed in the econometrics literature as an origin of volatility clustering.

4.4 Investor inertia

As argued by Liu [41], though the presence of a Markovian regime switching mechanism in volatility can lead to volatility clustering, is not sufficient to generate long-range dependence in absolute returns. More important than the

switching is the fact the time spent in each regime –the duration of regimes– should have a heavy-tailed distribution [60, 66]. By contrast with Markov switching, which leads to short range correlations, this mechanism has been called "renewal switching".

Bayraktar et al. [8] study a model where an order flow with random, heavy-tailed, durations between trades leads to long range dependence in returns. When the durations τ_n of the inactivity periods have a distribution of the form $\mathbb{P}(\tau_n \geq t) = t^{-\alpha}L(t)$, conditions are given under which, in the limit of a large number of agents randomly submitting orders, the price process in this models converges to a a process with Hurst exponent $H = (3-\alpha)/2 > 1/2$. In this model the randomness (and the heavy tailed nature) of the durations between trades are both exogenous ingredients, chosen in a way that generates long range dependence in the returns. However, as noted above, empirical observations point to clustering and persistence in *volatility* rather than in returns so such a result does not seem to be consistent with the stylized facts.

By contrast, as noted above, regime switching in *volatility* with heavy-tailed durations could lead to volatility clustering. Although in the agent-based models discussed above, it may not be easy to speak of well-defined "regimes" of activity, but Giardina and Bouchaud [25] argue that this is indeed the mechanism which generates volatility clustering in the Lux-Marchesi [47] and other models discussed above. In these models, agents switch between strategies based on their relative performance; Giardina and Bouchaud argue that this (cumulative) relative performance index actually behaves in time like a random walk, so the switching times can be interpreted as times when the random walk crosses zero: the interval between successive zero-crossings is then known to be heavy-tailed, with a tail exponent 3/2.

Ghoulmie, Cont and Nadal [17] show that investor inertia can also result from a threshold behavior of agents: an agent will not trade in the market unless the discrepancy between his anticipation of the value of the financial asset and the current market price reaches a certain threshold, which may be heterogeneous across agents. Figure 4 displays the evolution of the portfolio $\pi_i(t)$ of a typical agent in this model: short periods of activity (trading) are separated by long periods of inertia, where the portfolio remains constant. Such "renewal switching" between periods of high and low activity, with long durations of periods, can lead to long range dependence in volatilities [66].

5 Conclusion

Volatility clustering, manifested through slowly decaying autocorrelations for absolute returns, is a characteristic property of most time series of financial asset returns. Whether this "slow" decay corresponds to long range dependence is a difficult question subject to an ongoing statistical debate. But is definitely an *empirical* question: first principles of financial theory –such as absence of arbitrage– cannot be invoked to give any response to it.

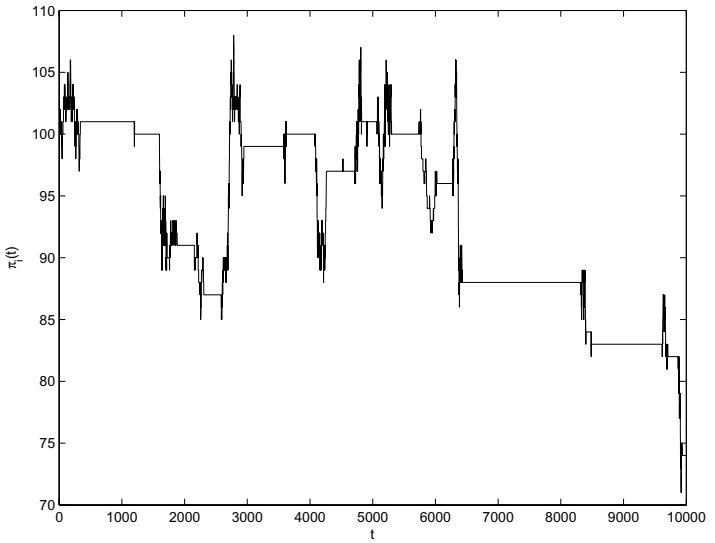


Fig. 4. Investor inertia: the evolution of the portfolio of a typical agent shows long periods of inactivity punctuated by bursts of activity.

As noted by many econometricians [57, 68], statistical analysis alone is not likely to provide a definite answer for the presence or absence of long-range dependence phenomenon in stock returns or volatility, unless economic mechanisms are proposed to understand the origin of such phenomena.

Agent-based models, which seek to explain volatility clustering in terms of behavior of market participants, have been proposed in order to explain long range dependence in volatility. A common feature of these models seems to be the "switching" of the market between periods of high and low activity, with long durations of periods. As we have noted, such "renewal switching" can lead to long range dependence in volatilities if the market switches between regimes of high and low volatility. The link between such economic models and the realm of stochastic models in finance is intriguing and remains an active topic of research at the time of writing.

References

1. V. Akgiray and G. Booth, *The stable law model of stock returns*, J. Business Econom. Statis., 6 (1988), pp. 51–57.
2. T. Andersen and T. Bollerslev, *Heterogeneous information arrivals and returns volatility dynamics*, Journal of finance, 52 (1997), pp. 975–1005.
3. V. V. Anh, C. C. Heyde, and N. N. Leonenko, *Dynamic models of long-memory processes driven by Lévy noise*, J. Appl. Probab., 39 (2002), pp. 730–747.

4. W. Arthur, J. Holland, B. LeBaron, J. Palmer, and P. Tayler, *Asset pricing under endogeneous expectations in an artificial stock market*, in The economy as an Evolving Complex System, W. Arthur, S. Durlauf, and D. Lane, eds., vol. II, Reading MA, 1997, Perseus Books, pp. 15–44.
5. R. T. Baillie, *Long memory processes and fractional integration in econometrics*, J. Econometrics, 73 (1996), pp. 5–59.
6. R. T. Baillie, T. Bollerslev, and H. O. Mikkelsen, *Fractionally integrated generalized autoregressive conditional heteroskedasticity*, J. Econometrics, 74 (1996), pp. 3–30.
7. O. E. Barndorff-Nielsen and N. Shephard, *Modelling by Lévy processes for financial econometrics*, in Lévy processes—Theory and Applications, O. Barndorff-Nielsen, T. Mikosch, and S. Resnick, eds., Birkhäuser, Boston, 2001, pp. 283–318.
8. E. Bayraktar, U. Horst, and K. R. Sircar, *A limit theorem for financial markets with inert investors*, working paper, Princeton University, 2003.
9. J. Beran, *Statistics for long-memory processes*, vol. 61 of Monographs on Statistics and Applied Probability, Chapman and Hall, New York, 1994.
10. R. Blattberg and N. Gonedes, *A comparison of the stable Paretian and Student t distributions as statistial models for prices*, Journal of Business, 47 (1974), pp. 244–280.
11. T. Bollerslev, R. Chou, and K. Kroner, *ARCH modeling in finance*, J. Econometrics, 52 (1992), pp. 5–59.
12. P. Cheridito, *Arbitrage in fractional Brownian motion models*, Finance Stoch., 7 (2003), pp. 533–553.
13. ———, *Gaussian moving averages, semimartingales and option pricing*, Stochastic Process. Appl., 109 (2004), pp. 47–68.
14. F. Comte and E. Renault, *Long memory continuous time models*, J. Econometrics, 73 (1996), pp. 101–149.
15. R. Cont, *Empirical properties of asset returns: Stylized facts and statistical issues*, Quant. Finance, 1 (2001), pp. 1–14.
16. R. Cont, J.-P. Bouchaud, and M. Potters, *Scaling in financial data: Stable laws and beyond*, in Scale Invariance and Beyond, B. Dubrulle, F. Graner, and D. Sornette, eds., Springer, Berlin, 1997.
17. R. Cont, F. Ghoulmié, and J.-P. Nadal, *Heterogeneity and feedback in an agent-based market model*, Journal of Physics: Condensed Matter, 17 (2005), pp. S1259–S1268.
18. D. Cutler, J. Poterba, and L. Summers, *What moves stock prices?*, Journal of Portfolio Management, (1989), pp. 4–12.
19. F. Delbaen and W. Schachermayer, *The fundamental theorem of asset pricing for unbounded stochastic processes*, Math. Ann., 312 (1998), pp. 215–250.
20. C. Dellacherie and P. Meyer, *Probabilités et Potentiel. Chapitres I à IV*, Hermann, Paris, 1975.
21. Z. Ding, C. Granger, and R. Engle, *A long memory property of stock market returns and a new model*, Journal of empirical finance, 1 (1983), pp. 83–106.
22. Z. Ding and C. W. J. Granger, *Modeling volatility persistence of speculative returns: a new approach*, J. Econometrics, 73 (1996), pp. 185–215.
23. P. Doukhan, G. Oppenheim, and M. S. Taqqu, eds., *Theory and applications of long-range dependence*, Birkhäuser Boston Inc., Boston, MA, 2003.
24. R. Engle, *ARCH models*, Oxford University Press, Oxford, 1995.

25. I. Giardina and J.-P. Bouchaud, *Bubbles, crashes and intermittency in agent based market models*, Eur. Phys. J. B Condens. Matter Phys., 31 (2003), pp. 421–437.
26. C. Gourieroux, A. Monfort, and E. Renault, *Indirect inference*, Journal of Applied Econometrics, (1993).
27. C. Granger and N. Hyng, *Occasional structural breaks and long memory with an application to the SP500 absolute stock returns*, Journal of Empirical Finance, 11 (2004), pp. 399–421.
28. C. W. J. Granger, *Long memory relationships and the aggregation of dynamic models*, J. Econometrics, 14 (1980), pp. 227–238.
29. C. W. J. Granger, *Essays in econometrics: collected papers of Clive W. J. Granger. Vol. II*, vol. 33 of Econometric Society Monographs, Cambridge University Press, Cambridge, 2001. Causality, integration and cointegration, and long memory, Edited by Eric Ghysels, Norman R. Swanson and Mark W. Watson.
30. D. Guillaume, M. Dacorogna, R. Davé, U. Müller, R. Olsen, and O. Pictet, *From the birds eye view to the microscope: A survey of new stylized facts of the intraday foreign exchange markets*, Finance Stoch., 1 (1997), pp. 95–131.
31. J. M. Harrison and S. R. Pliska, *Martingales and stochastic integrals in the theory of continuous trading*, Stochastic Process. Appl., 11 (1981), pp. 215–260.
32. C. C. Heyde, *On modes of long-range dependence*, J. Appl. Probab., 39 (2002), pp. 882–888.
33. M. Hols and C. De Vries, *The limiting distribution of extremal exchange rate returns*, Journal of Applied Econometrics, 6 (1991), pp. 287–302.
34. C. H. Hommes, A. Gaunersdorfer, and F. O. Wagener, *Bifurcation routes to volatility clustering under evolutionary learning*, working paper, CenDEF, 2003.
35. D. Jansen and C. De Vries, *On the frequency of large stock returns*, Rev. Econ. Stat., 73 (1991), pp. 18–24.
36. A. Kirman, *Ants, rationality, and recruitment*, Quarterly Journal of Economics, 108 (1993), pp. 137–156.
37. A. Kirman and G. Teyssiere, *Microeconomic models for long-memory in the volatility of financial time series*, Studies in nonlinear dynamics and econometrics, 5 (2002), pp. 281–302.
38. B. LeBaron, *Evolution and time horizons in an agent-based stock market*, Macroeconomic Dynamics, 5 (2001), pp. 225–254.
39. ———, *Stochastic volatility as a simple generator of apparent financial power laws and long memory*, Quant. Finance, 1 (2001), pp. 621–631.
40. B. LeBaron, B. Arthur, and R. Palmer, *Time series properties of an artificial stock market*, Journal of Economic Dynamics and Control, 23 (1999), pp. 1487–1516.
41. M. Liu, *Modeling long memory in stock market volatility*, Journal of Econometrics, 99 (2000), pp. 139–171.
42. A. Lo, *Long-term memory in stock market prices*, Econometrica, 59 (1991), pp. 1279–1313.
43. I. N. Lobato and Velasco, *Long memory in stock market trading volume*, J. Bus. Econom. Statist., 18 (2000), pp. 410–427.
44. F. Longin, *The asymptotic distribution of extreme stock market returns*, Journal of Business, 69 (1996), pp. 383–408.

45. M. Loretan and P. Phillips, *Testing the covariance stationarity of heavy-tailed time series*, Journal of empirical finance, 1 (1994), pp. 211–248.
46. T. Lux, *On moment condition failure in German stock returns*, Empirical economics, 25 (2000), pp. 641–652.
47. T. Lux and M. Marchesi, *Volatility clustering in financial markets: a microsimulation of interacting agents*, Int. J. Theor. Appl. Finance, 3 (2000), pp. 675–702.
48. B. B. Mandelbrot, *The variation of certain speculative prices*, Journal of Business, XXXVI (1963), pp. 392–417.
49. ———, *When can price be arbitraged efficiently? A limit to the validity of the random walk and martingale models*, Rev. Econom. Statist., 53 (1971), pp. 225–236.
50. ———, *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk.*, Springer, New York, 1997.
51. ———, *Stochastic volatility, power laws and long memory*, Quant. Finance, 1 (2001), pp. 558–559.
52. B. B. Mandelbrot and M. S. Taqqu, *Robust R/S analysis of long-run serial correlation*, in Proceedings of the 42nd session of the International Statistical Institute, Vol. 2 (Manila, 1979), vol. 48, 1979, pp. 69–99.
53. B. B. Mandelbrot and J. Van Ness, *Fractional Brownian motion, fractional noises and applications*, SIAM Review, 10 (1968), pp. 422–437.
54. R. Mantegna and H. Stanley, *Scaling behavior of an economic index*, Nature, 376 (1995), pp. 46–49.
55. M. Métivier and J. Pellaumail, *Stochastic integration*, Academic Press, New York, 1980. Probability and Mathematical Statistics.
56. T. Mikosch and C. Stărică, Limit theory for the sample autocorrelations and extremes of a GARCH $(1, 1)$ process, *Ann. Statist.*, 28 (2000), pp. 1427–1451.
57. T. Mikosch and C. Stărică, *Long-range dependence effects and ARCH modeling*, in Theory and applications of long-range dependence, Birkhäuser Boston, Boston, MA, 2003, pp. 439–459.
58. J. Muzy, J. Delour, and E. Bacry, *Modeling fluctuations of financial time series: From cascade processes to stochastic volatility models*, Eur. J. Phys. B, 17 (2000), pp. 537–548.
59. A. Pagan, *The econometrics of financial markets*, Journal of Empirical Finance, 3 (1996), pp. 15–102.
60. M. Pourahmadi, *Stationarity of the solution of $X_t = A_t X_{t-1} + \epsilon_t$ and analysis of non-Gaussian dependent random variables*, J. Time Ser. Anal., 9 (1988), pp. 225–239.
61. P. Protter, *Stochastic integration and differential equations*, Springer, Berlin, 1990.
62. S. Resnick, G. Samorodnitsky, and F. Xue, *How misleading can sample ACFs of stable MAs be? (Very!)*, Ann. Appl. Probab., 9 (1999), pp. 797–817.
63. S. Resnick and E. van den Berg, *Sample correlation behavior for the heavy tailed general bilinear process*, Comm. Statist. Stochastic Models, 16 (2000), pp. 233–258.
64. L. C. G. Rogers, *Arbitrage with fractional Brownian motion*, Math. Finance, 7 (1997), pp. 95–105.
65. G. Samorodnitsky and M. Taqqu, *Stable Non-Gaussian Random Processes*, Chapman & Hall, New York, 1994.

66. M. S. Taqqu and J. B. Levy, *Using renewal processes to generate long-range dependence and high variability*, in Dependence in probability and statistics (Oberwolfach, 1985), vol. 11 of Progr. Probab. Statist., Birkhäuser Boston, Boston, MA, 1986, pp. 73–89.
67. V. Teverovsky, M. S. Taqqu, and W. Willinger, *A critical look at Lo's modified R/S statistic*, J. Statist. Plann. Inference, 80 (1999), pp. 211–227.
68. W. Willinger, M. Taqqu, and V. Teverovsky, *Long range dependence and stock returns*, Finance and Stochastics, 3 (1999), pp. 1–13.

Financial Modelling by Multiscale Fractional Brownian Motion

Pierre Bertrand

Laboratoire de Mathématiques - UMR CNRS 6620, Université Blaise Pascal (Clermont-Ferrand II), 24 Avenue des Landais, 63117 Aubière Cedex, France.
Pierre.Bertrand@math.univ-bpclermont.fr

Summary. The multiscale fractional Brownian motion provides an example of process with long memory which is arbitrage free. After having recalled the definition of the long memory and the notion of arbitrage free price process, we derive the price of European options. This one is the Black-Scholes price using the "high-frequency" volatility parameter. This shows that the presence or absence of long memory is not a relevant question for pricing European options.

1 Introduction

The probabilistic properties of the multiscale fractional Brownian motion have been studied in Bardet and Bertrand (2003a). It provides a model of price process with long memory which is arbitrage free. Indeed, different empirical observations suggest the existence of the long memory property for the prices of financial assets, see for instance Willinger *et al.* (1999) and the references therein. In this work, we derive the option pricing formula for the multiscale fractional Brownian motion: the price is the Black-Scholes price using the "high-frequency" volatility parameter.

Before going further, we would like to discuss the apparent paradox: a model with long memory leads to the same price as a Black-Scholes case. First, the long memory component of the process disappears in the price of European option as it is the case for the drift term. Indeed, by using Girsanov Theorem, the discounted price process follows a stochastic differential equation depending only on the "high-frequency" volatility parameter. The long memory component is taken into account in the new probability measure and thus disappears from calculations, exactly like the drift term in the Black-Scholes model. Next, one obtains a result of robustness of the Black-Scholes formula with respect to the parameter of long memory or more generally an absence of sensitivity to the low frequency behavior of the model. This close the recurrent polemic on the presence of long memory for the financial assets,

since this presence or absence has no effect on the price of options. From a heuristic point of view, this can be explained by the fact that one considers a continuous time model. This authorizes to follow the market on very small intervals of time, therefore the price depends only on the high frequency behavior.

Anyway, this robustness must be moderated by the taking into account of the statistical aspect. We refer to Bardet and Bertrand (2003b) to a detailed statistical study of the multiscale fractional Brownian motion. One remarks that the long range dependency can bias the empirical estimator of the high frequency volatility, when this one is based on the quadratic variations. Thus the implied volatility appears as different from the estimated one. However, let us note that this bias on the estimated high frequency volatility would be avoid by using a method based on the wavelet analysis.

A last comment, our aim is to put the emphasize on the absence of influence of the long memory on the pricing formula and the need to use estimate of the high frequency volatility based on wavelet analysis. We do not pretend to furnish a realistic model for financial data. On the contrary, we restrict ourself to the case of processes with stationarity increments, since in this case the definition of long range dependence is more simple. Therefore, the volatility parameters remain constant and could not depend on time as in the literature on stochastic volatility. Anyway, it could be possible to define the long memory for generalized fractional motion with time dependent parameters, see for example Ayache *et al.* (2000), but this would lead to heavy calculations.

Our plan is the following. First we recall the definition of the long memory property for Gaussian processes with stationary increments. Next, we recall that the possible model for the financial processes should be arbitrage free and the consequences of this principle. As an example, we explain that the absence of arbitrage opportunity does not allow the price of share to be modelled by fractional Brownian motion. Then, we recall the definition of the multi-scale fractional Brownian motion and its properties. Finally, we derive the price of European option for this model.

2 Gaussian Processes with Long Memory. The Example of Fractional Brownian Motion.

In this section, we recall the definitions of long memory processes and of the fractional Brownian motion.

Gaussian Processes with long memory.

Let $X(t)$ be a centered Gaussian process with stationary increments, i.e.

$$\forall \delta > 0, \quad (X(t+\delta) - X(t))_{t \in \mathbb{R}} \stackrel{(d)}{=} (X(\delta) - X(0))_{t \in \mathbb{R}}$$

The increments of the process X defined by $Y(n) = X(n+1) - X(n)$, are a centered Gaussian process with correlation $r(n) = \mathrm{cov}\,(Y(n), Y(0)) = I\!E\,(Y(n)\,Y(0))$.

- When $\sum_{n>0} |r(n)| = +\infty$ X is called a long memory process
- When $\sum_{n>0} |r(n)| < +\infty$ X is called a short memory process.

Definition of the fractional Brownian motion.

A Fractional Brownian Motion (FBM) is a centered Gaussian process with stationary increments $(B_H(t),\, t \in I\!R)$ such that $B_H(0) = 0$ and $\forall (s,t) \in I\!R^2$

$$I\!E\,|B_H(t) - B_H(s)|^2 = \sigma^2 \times |t-s|^{2H}. \tag{1}$$

Let us stress that the FBM depends on two parameters : 1) the scale parameter σ, 2) the Hurst index H. When $H = 1/2$ and $\sigma = 1$, we get the standard Brownian motion. By using the relationship $r(n) = [\nu(n+1) - 2\nu(n) + \nu(n-1)]\,/2$ where $\nu(n) = I\!E\,|B_H(n) - B_H(0)|^2$, we deduce the behavior of the correlation. So,

$$r(n) = \sigma^2 \times \frac{H(2H-1)}{n^{2-2H}} + \mathcal{O}\left(n^{-(3-2H)}\right).$$

Moreover, when $H = 1/2$ we have $r(n) = 0$ for all integer n and the increments are independent. Eventually, we get the following landscape

- When $H = 1/2$, B_H is Brownian motion and its increments are independent. This corresponds to the case $r(n) = 0$, for all $n > 0$.
- When $H > 1/2$, then $\sum_{k=-\infty}^{+\infty} |r(k)| = +\infty$ and B_H is a long memory process
- When $H < 1/2$, then $\sum_{k=-\infty}^{+\infty} |r(k)| < +\infty$ and B_H is a short memory process.

3 The Consequences of the Principe of Absence of Arbitrage Opportunity.

In this section, we first recall the principle of Absence of Arbitrage opportunity. Then we precise three consequences of this assumption. The first one is related to the existence of the price of the European options. The second one is the fundamental theorem of asset pricing, which induces that the stock price should be a semi-martingale. Eventually, we recall that the FBM is not a semi-martingale except when $H = 1/2$ which corresponds to the case of a standard Brownian motion.

The Principle of Absence of Arbitrage Opportunity for Continuous Stock Price

A *"good"* financial modelling must prohibit the existence of arbitrage opportunity. To be precise, one considers a market with two assets: a risky one, say the price of a share, and a risk-free one, say a bond. The risky asset is modelled by a stochastic process $(S_t)_{t\in[0,T]}$ defined on a probability filtered space $(\Omega, \mathcal{F}, I\!P)$, where the underlying filtration $\mathcal{F}_t$ is the standard augmentation of the filtration generated by the stock prices $\mathcal{F}_t^S := \{S_u,\, u \le t\}$. The risk-free security is a deterministic positive process $(Z_t)_{t\in[0,T]}$, for instance $Z_t = e^{rt}$, where $r > 0$ is the instantaneous interest rate. A portfolio or a trading strategy is defined as a pair $\Phi = \left(\Phi^1,\, \Phi^2\right)$ of progressively measurable stochastic processes on the underlying probability space $(\Omega, \mathcal{F}, I\!P)$. At each time, the associate wealth is given by

$$Y_t(\Phi) = \Phi^1(t)\, S_t \,+\, \Phi^2(t)\, Z_t.$$

We consider portfolio without income of money between times 0 and T. These trading strategies are called self-financed portfolios. In the general framework, the price process S is not *a priori* forced to be a semi-martingale. This induces technical difficulties to define the stochastic integral. Following Delbaen and Schachermayer (1994), we restrict the class of admissible trading strategies to pair of piecewise constant progressively measurable stochastic processes. Thus we assume there exists a finite family of $\mathcal{F}^S$ stopping times $(\tau_i)_{i=1,\dots,N}$ with $0 = \tau_1 < \tau_2 < \cdots < \tau_N = T$ and $\Phi = \left(\Phi^1,\, \Phi^2\right)$ is defined by

$$\Phi^1 = f_0\, \mathbf{1}_{\{0\}} \,+\, \sum_{j=1}^{N-1} f_j\, \mathbf{1}_{(\tau_j,\tau_{j+1}]} \quad \text{and} \quad \Phi^2 = g_0\, \mathbf{1}_{\{0\}} \,+\, \sum_{j=1}^{N-1} g_j\, \mathbf{1}_{(\tau_j,\tau_{j+1}]}$$

where the random variables f_j and g_j are $\mathcal{F}_{\tau_j}^S$ adapted. In this framework, a portfolio is self financing if and only if

$$(f_j - f_{j-1})\; S_{\tau_j} + (g_j - g_{j-1})\; Z_{\tau_j} = 0 \qquad \text{for all } j = 1, \dots, N-1.$$

With these notations we are in position to define the notion of arbitrage and after of Absence Of Arbitrage opportunity (A.O.A.). Anyway, there are different notions of no-arbitrage depending on the set of admissible strategy and on the choice of topology on the space of random variables. Following Delbaen and Schachermayer (1994), the spot market $(S,\, Z)$ is said arbitrage free if its satisfies the NFLVR condition *(No Free Lunch at Vanishing Risk)* which is stated below.

Definition 3.1 *A price process S satisfies the condition NFLVR if there does not exist a sequence of admissible self financing strategies $(\Phi_n)_{n\in I\!N}$, a sequence of positive real number δ_n converging to 0 and a random variable X with values in $R^+ \cup \{+\infty\}$ such that*

1) $Y_0(\Phi_n) = 0$ *a.s. and* $Y_t(\Phi_n) \geq -M$ *for all* $t \in [0,T]$ *where* $M > 0$,
2) $\lim_{n\to\infty} Y_T(\Phi_n) = X$ *a.s.* ,
3) for all $n \in \mathbb{N}$, $Y_T(\Phi_n) \geq -\delta_n$,
4) $\mathbb{P}(X > 0) > 0$.

The price of European claim

One of the main object of mathematical finance is the calculation of the price of options and to find an hedging strategy. In this work, we consider only European options. An European option gives the right to its holder to exercise this one only at the expiry time and provides a payoff $h(S_T)$ where h is a given function. Remark that the payoff function h could be positive or negative. Let us stress that if the principle of absence of arbitrage opportunity is not satisfied, then one cannot define the price of the options. Therefore, a model for stock price which does not insure that the market is arbitrage free is not a coherent model.

Fundamental Theorem of Asset Pricing

When S is a continuous process, the fundamental theorem of asset pricing, due to Delbaen and Schachermayer (1994), asserts that the condition NFLVR is equivalent to the existence of a martingale measure. By definition, a martingale measure for the price process is a measure $\mathbb{Q}$ equivalent to $\mathbb{P}$ and such that the discounted stock price process S^*, which is defined by the formulas $S_t^* = S_t\, Z_t^{-1}$ with $S_0^* = S_0$, follows a $\mathbb{Q}$-martingale; that is the equality $S_t^* = \mathbb{E}_{\mathbb{Q}}\,(S_s^* | \mathcal{F}_t)$ for all pair (s,t) with $0 \leq t < s \leq T$. Moreover, the Girsanov Theorem induces as a corollary that the price process should be a semi-martingale under the probability $\mathbb{P}$.

Why the stock price could not be modelled by a Fractional Brownian Motion

The last point, raises the following question: when the FBM is a semi-martingale ? The result is well-known from many years : if $H \neq 1/2$, then the FBM is not a semi-martingale. Therefore, the fractional Brownian motion is not a coherent model for the stock price except in the obvious case where it is a Brownian motion. Without entering the questions of precedences, we can refer for instance to Rogers (1997) or Cheridito (2003). Our matter rather consists in examining the demonstration. To put in a nutshell, the proof is based on the fact that the quadratic variation is 0 when $H > 1/2$ and ∞ when $H < 1/2$ while the quadratic variation is finite and does not vanish for a semi-martingale. Let us stress that it is a property related to the high frequency behavior of the process while, on the contrary, the long memory is a property related to the low frequency behavior.

Before detailing this point, we would like to reconsider the various properties of the FBM. Since B_H is a Gaussian process, the relationship (1) induces that the Hurst index H corresponds to the three following properties:

1. **The correlation of the increments:**
$$r(n) = \frac{H(2H-1)}{n^{2-2H}} + \mathcal{O}\left(n^{-(3-2H)}\right)$$
2. **The local regularity:**
Let
$$\alpha_X(t,\omega) = \sup\left\{\alpha, limsup_{h\to 0}\frac{|X(t+h,\omega)-X(t,\omega)|}{|h|^\alpha} = 0\right\}$$
then $\mathbb{P}\left(\alpha_X(t,\omega) = H,\ \forall t \in \mathbb{R}\right) = 1$.
3. **The Self-Similarity:**
$$\left(B_H(\lambda t)\right)_{t\in\mathbb{R}} \overset{(d)}{=} \left(\lambda^H B_H(t)\right)_{t\in\mathbb{R}}.$$

Let us insist on the fact that from one hand, the regularity of the sample paths and the finiteness of the quadratic variation could be interpreted as high frequency behaviors, while from the other hand the long memory of the increments correspond to a low frequency behavior. Yet, as mentioned by Mandelbrot and Van Ness (1968), for the fractional Brownian motion *" the concept of self-similarity is a form of invariance with respect to changes of time scale "*. For this reason, the behaviors at high and low frequencies are bound and controlled by the same parameter H. However, in financial applications, this link could appear as artificial. We are thus led to seek processes which are at the same time semi-martingale and with a long memory. Arises then the question of the price of the options for such a model of the stock prices. This is the aim of the next section.

4 Multiscale Fractional Brownian Motion

In this section, we first recall the definition of the multiscale fractional Brownian motion. Then, we give the high frequency behavior of these processes, namely the path regularity and a condition necessary and sufficient so that this process is a semi-martingale. Next, we specify its properties of long memory. Afterwards, we derive the price of European option for such a model of stock prices. Eventually, we give a numerical simulation of a path.

4.1 Definition of the Multiscale Fractional Brownian Motion

Fractional Brownian motions (F.B.M.) are Gaussian processes with stationary increments with a covariance structure given by (1). The F.B.M. has different representations which could alternately be used as definition : the

moving average one due to Mandelbrot and Van Ness (1968) and the harmonizable one introduced by Kolmogorov (1940). We propose Samorodnitsky and Taqqu (1994), chapter 7 as a reference on the matter. Various generalizations of F.B.M. have been proposed these last years to fill the gap between the mathematical modelling and real data. They are based on the different representation of the F.B.M. or on wavelet random series, see for example Ayache and Lévy Véhel (2000). Anyway, the harmonizable representation is the best adapted to our purpose and is defined by

$$B_H(t) = \int_{\mathbb{R}} \frac{\left(e^{it\xi} - 1\right)}{|\xi|^{H+1/2}} \widehat{W}(d\xi) \tag{2}$$

where $W(dx)$ is a Brownian measure and $\widehat{W}(d\xi)$ its Fourier transform, namely for any function $f \in L^2(\mathbb{R})$ one has almost surely, $\int_{\mathbb{R}} f(x)W(dx) = \int_{\mathbb{R}} \widehat{f}(\xi)\,\widehat{W}(d\xi)$, with the convention that $\widehat{f}(\xi) = \int_{\mathbb{R}} e^{-i\xi\,x}\, f(x)\, dx$ when $f \in L^1(\mathbb{R}) \bigcap L^2(\mathbb{R})$.

Let us stress that the spectral density $|\xi|^{-(H-1/2)}$ follows a power law with the same index at every frequency ξ and that this corresponds precisely to the self similarity property. Thus, a natural generalization consists to replace in (2) the Hurst index H by a piecewise function depending on the frequency $\xi \mapsto H(\xi)$. In Bardet and Bertrand (2003), this process is called the multiscale fractional Brownian motion and it is defined as follows:

Definition 4.1 *For $J \in \mathbb{N}$, a (M_J)-multiscale fractional Brownian motion (simplify by (M_J)-F.B.M.) $X = (X(t), t \in \mathbb{R}_+)$ is a process such that*

$$X(t) = \int_{\mathbb{R}} \frac{(e^{it\xi} - 1\ )}{\rho(\xi)}\, \widehat{W}(d\xi) \quad \textit{for } t \in \mathbb{R}_+ \tag{3}$$

with

- *$W(d\xi)$ is a standard Brownian measure and $\widehat{W}(d\xi)$ its Fourier transform in the distribution meaning.*
- *for $j = 0, \ldots, J$, there exist $(\omega_j, \sigma_j, H_j) \in (\mathbb{R}_+, \mathbb{R}_+^*,]0,1[)$ such that $\rho(\xi) = \sigma_j^{-1}|\xi|^{H_j+1/2}$ for $|\xi| \in [\omega_j, \omega_{j+1}[$ with $\omega_0 = 0 < \omega_1 < \cdots < \omega_J < \omega_{J+1} = +\infty$ by convention.*

Let us notice that $J \in \mathbb{N}$ is the number of frequency changes. Therefore, when $J = 0$, a (M_0)-FBM corresponds to a FBM with Hurst index $H = H_0$.

4.2 High Frequency Behavior of the Multiscale Fractional Brownian Motion

The following properties of the multiscale fractional Brownian motion have been proved in Bardet and Bertrand (2003):

Property 4.1
1. X is a Gaussian centered process with stationary increments.
2. We have the decomposition

$$X(t) = \sigma_J \, B_{H_J}(t) + \sigma_J \, R(t), \tag{4}$$

where B_{H_J} is a FBM with Hurst index H_J and R a continuous process with finite variation. Moreover, we get

$$B_{H_J}(t) = \int_{I\!R} \frac{\left(e^{it\xi} - 1\right)}{|\xi|^{H_J+1/2}} \widehat{W}(d\xi)$$

and

$$R(t) = -\int_{|\xi|\le\omega_J} \frac{\left(e^{it\xi} - 1\right)}{|\xi|^{H_J+1/2}} \widehat{W}(d\xi) \; + \; \sum_{j=0}^{J-1} \frac{\sigma_j}{\sigma_J} \int_{\omega_j<|\xi|\le\omega_{j+1}} \frac{\left(e^{it\xi} - 1\right)}{|\xi|^{H_j+1/2}} \widehat{W}(d\xi) \tag{5}$$

Since R is a continuous process with finite variation, Formula (4) gives us the high frequency behavior of the process. This is stated in the following immediate corollary:

Corollary 4.1 *Let X be a (M_J)-FBM, then*

- $I\!P\left(\alpha_X(t,\omega) = H_J, \; \forall t \in I\!R\right) = 1.$
- *X is a semi-martingale if and only if $H_J = 1/2$.*
- *X is locally self-similar with index H_J*

$$\lim_{\epsilon\to 0} \left(\frac{X(t+\epsilon u) - X(t)}{|\epsilon|^{H_J}} \right)_{u\in I\!R} \overset{(d)}{=} \left(B_{H_J}(u)\right)_{u\in I\!R}$$

where B_{H_J} is a FBM with Hurst index H_J and the convergence is in distribution on the space of continuous functions with the topology of uniform convergence on the compacts.

In short, the high frequency behavior of the multiscale fractional Brownian motion is controlled by the high frequency Hurst index H_J.

4.3 Long Memory of the Multiscale Fractional Brownian Motion

One could expect that the long term dependence of its increments is controlled by H_0. This is essentially true but a little bit more complicated. This point is explained in the following property, which has been proved in Bardet and Bertrand (2003 a) :

Property 4.2 *Let X be a (M_J)-F.B.M. with $K \ge 1$, Y be its increments defined by $Y = (Y(t), t \in I\!R_+)$ and $r(n) = cov\left(Y(n), Y(0)\right)$ be the correlogram of the increments of X. Then*

$$\text{if } H_0 > \frac{1}{2}, \;\; r(n) = \left(\frac{\pi\sigma_0^2(2H_0-1)}{\Gamma(2H_0)\sin(\pi H_0)} \right) \frac{1}{n^{2-2H_0}} + \mathcal{O}\left(\frac{1}{n}\right), \tag{6}$$

$$\text{if } H_0 \le \frac{1}{2}, \;\; r(n) = \frac{P(n)}{n} + \mathcal{O}\left(\frac{1}{n^{2-2H_0}}\right), \tag{7}$$

with $P(n)$ a trigonometric polynomial depending only on $(H_i, \omega_i, \sigma_i)_{0\leq i\leq K}$ such that

$$P(n) = 8\sum_{j=1}^{J} \sin^2\left(\frac{\omega_j}{2}\right) \times \left(\frac{\sigma_{j-1}^2}{\omega_{j-1}^{2H_j-1}} - \frac{\sigma_j^2}{\omega_j^{2H_j}}\right) \times \sin(n\omega_j). \tag{8}$$

From this property we derive the asymptotic behavior of the dependence of the increments of X. Then, the long-range property of the increments of the process Y depends on H_0, but not only on this value. More precisely, if $H_0 > 1/2$, the long-range behavior of Y is the same that the behavior of a fractional Gaussian noise with parameter H_0. Thus, Y is a long-range dependent process. If $H_0 \leq 1/2$, the process Y is a long-range dependent process except when $(H_j, \omega_j, \sigma_j)_{0\leq j\leq J}$ verifies a relation such that the trigonometric polynomial $P(n)$ of (8) is vanished. In this case, the frequency changes induce long memory.

4.4 Pricing an European Option when the Stock Price is modelled by a Multi-Scale Fractional Brownian Motion

In this subsection, we consider a log normal model for the prices of share S_t and assume that the interest rate of the risk free security B_t is constant. Therefore we have the following model

$$Z_t = e^{rt}, \tag{9}$$

$$S_t = S_0 \times \exp(X_t) \tag{10}$$

where the Gaussian process X is a (M_J)-F.B.M. As usual, we define the discounted stock price process $\widetilde{S}$ by $\widetilde{S}_t = S_t/Z_t$. We have the following proposition :

Proposition 4.1

1. *The condition NFLVR implies that S is a semi-martingale.*
2. *The process S given by (10) is a semi-martingale if and only if $H_J = 1/2$. Moreover, in this case*

$$dS_t = S_t \times \sigma_J \times \{dW_t^* + (\mu(t) + \frac{\sigma_J}{2})\, dt\} \tag{11}$$

where

$$W_t^* = \int_{I\!R} \left(e^{it\xi} - 1\right) \times |\xi|^{-1}\, \widehat{W}(d\xi) \tag{12}$$

and

$$\mu(t) = -\int_{|\xi|\leq\omega_J} \frac{i\,\xi\, e^{it\xi}}{|\xi|}\, \widehat{W}(d\xi) + \sum_{j=0}^{J-1} \frac{\sigma_j}{\sigma_J} \int_{\omega_j<|\xi|\leq\omega_{j+1}} \frac{i\,\xi\, e^{it\xi}}{|\xi|^{(H_j+1/2)}}\, \widehat{W}(d\xi). \tag{13}$$

3. *There exists one probability measure $\mathbb{Q}$ on $(\Omega, \mathcal{F}_T)$ equivalent to $\mathbb{P}$ for which the discounted stock price process $\widetilde{S}$ is a martingale. Furthermore, the process $\widetilde{W}$ defined by*

$$\widetilde{W}_t = W_t^* + \int_0^t \left(\mu(u) + \frac{\sigma_J}{2} - \frac{r}{\sigma_J}\right) du, \quad \text{for } t \in [0,T] \tag{14}$$

is a standard Brownian motion with respect to $(\Omega, \mathcal{F}, Q)$ and

$$d\widetilde{S}_t = \sigma_J \, \widetilde{S}_t \, d\widetilde{W}_t. \tag{15}$$

Proof. The first point follows from the fundamental theorem of asset pricing for continuous time processes, see Delbaen and Schachermayer (1994). Item 2 now is considered. From the one hand, when $H_J \neq 1/2$, the process X is not a semi-martingale. More precisely, when $H_J < 1/2$, the process X has almost surely infinite quadratic variation like S, thus S could not be a semi-martingale. When $H_J > 1/2$, the process X has almost surely a quadratic variation equal to 0. This implies that the quadratic variation of S is also zero, and so (if S were a semi-martingale) S must be a finite-variation process. But, the finite variation of X is almost surely infinite like the one of S. This contradicts the almost-sure finiteness of the quadratic variation of S, assuming S is a semi-martingale. Either way, if $H_J \neq 1/2$, S is not a semi-martingale. From the other hand, when $H_J = 1/2$, by using Property 4.1 item 2, one has decomposition (4) with

$$R(t) = -\int_{|\xi| \leq \omega_J} \frac{\left(e^{it\xi} - 1\right)}{|\xi|} \widehat{W}(d\xi) + \sum_{j=0}^{J-1} \frac{\sigma_j}{\sigma_J} \int_{\omega_j < |\xi| \leq \omega_{j+1}} \frac{\left(e^{it\xi} - 1\right)}{|\xi|^{H_j + 1/2}} \widehat{W}(d\xi)$$

One wishes to show that $R(t) = \displaystyle\int_0^t \mu(t)\, dt$ almost surely. Let us remark that $\mu(t) = i \int_{\mathbb{R}} f(t,\xi)\, \widehat{W}(d\xi)$ where f is a bounded function. Indeed, the only difficulty could occur into $\xi = 0$, but $|f(t,0)| = 1$ for all $t \in \mathbb{R}$. One can then apply Fubini's Theorem for stochastic integral which is stated in Revuz and Yor, (5.17) p.176. Since $R(0) = 0$, we deduce that $R(t) = \displaystyle\int_0^t \mu(t)\, dt$ almost surely. Thus, X is a It process, more precisely, we have

$$dX_t = \sigma_J \, dW_t^* + \sigma_J \, \mu(t)\, dt.$$

Then, by using It's Formula, we deduce Formula (11). This implies that S is a semi-martingale.

Finally, we will prove the third point. An easy calculation shows that

$$d\widetilde{S}_t = \widetilde{S}_t \times \left[\sigma_J \, dW_t^* + \left(\sigma_J \, \mu(t) + \frac{\sigma_J^2}{2} - r\right) dt\right].$$

One would like to find a probability $\mathbb{Q}$ equivalent to $\mathbb{P}$ under which the process $\widetilde{W}$ defined by (14) is a standard Brownian motion. This follows from Girsanov's Theorem, see for example Th.5.1, p.191 in Karatzas and Shreve (1991). The Novikov condition is a sufficient condition to apply the Girsanov theorem, see corollary 5.14, p.199 in Karatzas and Shreve (1991). In this case, the Novikov condition is

$$\mathbb{E}\exp\left(\frac{1}{2}\int_{t_{n-1}}^{t_n}\left(\mu(u)+\frac{\sigma_J}{2}-\frac{r}{\sigma_J}\right)^2 du\right)<\infty \quad \text{for all } n=1,\ldots,N_0$$

where (t_n) is a family of real numbers with $0=t_0<t_1<\cdots<t_{N_0}$. But, by using $(a+b)^2\leq 2a^2+2b^2$ for all real numbers a and b, we deduce $(\mu(u)+\sigma_J/2-r/\sigma_J)^2\leq 2\mu(u)^2+2\,(\sigma_J/2-r/\sigma_J)^2$ and after

$$\begin{aligned}\mathbb{E}\exp\left(\frac{1}{2}\int_{t_{n-1}}^{t_n}\left(\mu(u)+\frac{\sigma_J}{2}-\frac{r}{\sigma_J}\right)^2 du\right) &\leq \exp\left((t_n-t_{n-1})\left(\frac{\sigma_J}{2}-\frac{r}{\sigma_J}\right)^2\right)\\ &\quad\times\mathbb{E}\exp\left(\int_{t_{n-1}}^{t_n}\mu(u)^2 du\right)\\ &<\infty,\end{aligned}$$

where the last bound result from the finiteness of the real numbers t_n-t_{n-1}, r, σ_J and the lemma 4.1 below. To finish with, one get $d\widetilde{S}_t=\sigma_J\,\widetilde{S}_t\,d\widetilde{W}_t$. This induces that $\widetilde{S}$ is a $(\Omega,\mathcal{F},Q)$ martingale. ■

The existence of one martingale measure $\mathbb{Q}$ for the discounted stock price is a consequence of the Novikov Condition which results from the following lemma. The proof of this lemma is technical and it is given in appendix.

Lemma 4.1 *Let μ be defined by (13), then there exists a family of real numbers $(t_n)_{n=0,\ldots,N_0}$ with $0=t_0<t_1<\cdots<t_{N_0}$, such that*
$\mathbb{E}\exp\left(\int_{t_{n-1}}^{t_n}\mu(u)^2 du\right)<\infty$ *for all* $n=1,\ldots,N_0$.

Let us emphasize that the Novikov condition stated in lemma 4.1 is the key tool for the pricing and hedging of any European option. Indeed, after the change of probability, one obtain the same relationship (15) as in the Black and Scholes model, that is when $dS_t=S_t\,(\sigma_0\,dW_t+\mu_0\,dt)$ where W is a standard Brownian motion and σ_0, μ_0 are two real numbers. Therefore, one can derive the whole theory of pricing and hedging European option as in the Black and Scholes model, we refer for instanced to the chapter 5 in Musiela and Rutkowski, pp.109-134. The calculations are straightforward but long and tedious. The only important point consists in noting that all the results are valid by replacing volatility σ_0 by high frequency volatility σ_J. Also, we will give only the statement of the result without any proof.

Theorem 4.1 *Let X be a (M_J)-F.B.M. with $H_J = 1/2$ and $S_t = S_0 \exp X(t)$. Then there exists one and only one martingale measure $\mathbb{Q}$ on $(\Omega, \mathcal{F}_T)$ for which the discounted stock price process is a martingale and this is the probability measure given in Proposition 4.1. For any European option there exists one and only one self-financing portfolio Φ which replicates the option, that is $Y_T(\Phi) = h(S_T)$ a.s. Its wealth at time $t \in [0,T]$ is the arbitrage free price of this option and is given by $Y_t(\Phi) = F(t, S_t)$ where*

$$F(t,x) = e^{-r(T-t)} \int_{I\!R} h\left(x \exp^{(r-\sigma_J^2)(T-t)+\sigma_J u\sqrt{T-t}}\right) \frac{e^{-u^2/2}}{\sqrt{2\pi}}\, du,$$

Moreover we have

$$\Phi^1(t) = \frac{\partial F}{\partial x} F(t, S_t) \quad \textit{and} \quad \Phi^2(t) = e^{-rt}\left(F(t,S_t) - \Phi^1(t)\, S_t\right).$$

Let us consider an example in a particular case. Assume that the final pay-off is $h(S) = (S - K_1)^+$, what is equivalent to say that the option is a call option with maturity T and a strike price K_1. Then the price at time $t \in [0,T]$ is $C(t, S_t)$ where

$$C(t,x) = x\, N(d_1) \;-\; K e^{-r(T-t)} N(d_2), \qquad N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du$$

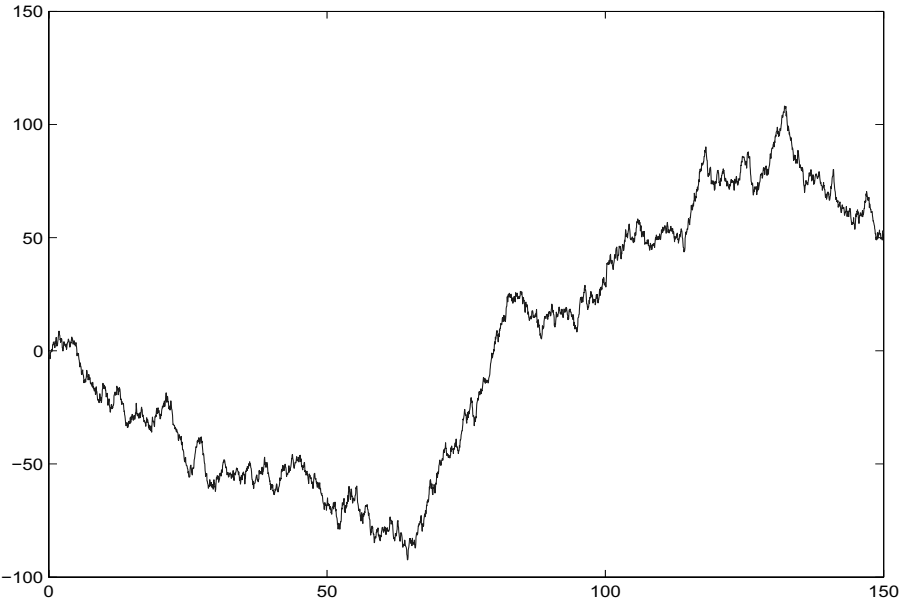
$$d_1 = \frac{\log(x/K_1) + (r + \sigma_J^2/2)(T-t)}{\sigma_J\sqrt{T-t}}, \quad \text{and} \quad d_2 = \frac{\log(x/K_1) + (r - \sigma_J^2/2)(T-t)}{\sigma_J\sqrt{T-t}}.$$

The replicating portfolio is $\Phi = (\Phi^1, \Phi^2)$ with $\Phi^1(t) = N(d_1)$ and $\Phi^2(t) = -K_1 e^{rT} N(d_2)$.

Remark 4.1 *In conclusion, we have build semi-martingales with long range dependance of order $n^{-(2-2H)}$ for all $H \geq 1/2$. In this framework, the price of any European option does not depend on long memory, but only of the volatility at high frequency σ_J. The statistical processing is out of the scope of this paper. We refer to Bardet and Bertand (2003b) for a detailed statistical study of the (M_J) F.B.M. In two words, let us mention that the long range dependency can bias the empirical estimator $\widehat{\sigma}$ of the high frequency volatility σ_J, when this one is based on the quadratic variations. This makes it possible to explain the difficulties encountered by the financial experts to estimate volatility when the time mesh is not small enough. To avoid this problem, in Bardet and Bertand (2003b), we have developed a method based on the wavelet analysis.*

4.5 An example of a path of a (M_1) Multiscale Fractional Brownian Motion

We give a numerical simulation of a discretized path $(X(0.05), X(0.10), \cdots, X(150))$ of a (M_1)-F.B.M., with $\sigma_0 = \sigma_1 = 5$, $H_0 = 0.6$, $H_1 = 0.5$ and $\omega_1 = 0.5$.



Following the previous discussion, the corresponding process provides an example of semi martingale with long range correlation.

A Appendix : Proof of the Novikov Condition

In this subsection, we prove the lemma 4.1.

Let $S = 2\omega_J + \sum_{j=0}^{J-1} \left(\frac{\sigma_j}{\sigma_J}\right)^2 \int_{\omega_j}^{\omega_{j+1}} |\xi|^{1-2H_j}\, d\xi$,

there exists a family $(t_n)_{n=0,\ldots,N_0}$ with $0 < (t_n - t_{n-1}) < 1/(4S)$ for all $n = 1, \ldots, N_0$. To simplify, one denotes $\tau = \max\{(t_n - t_{n-1}),\ \text{for } n = 1, \ldots, N_0\}$. Since μ is a stationary process, its suffices to prove that

$$I\!E \exp\left(\int_0^\tau \mu(u)^2\, du\right) < \infty \quad \text{when} \quad \tau < 1/(4S). \tag{16}$$

To obtain the bound (16), we will use the two following arguments :

1) The integral $\int_0^\tau \mu(u)^2\, du$ could be approximate by its Riemann sum $\sum_{k=1}^{N_1} \left(\frac{\tau}{N_1}\right) \mu\left(\frac{k\tau}{N_1}\right)^2$.

2) Some properties of Gaussian vectors and Gaussian laws. Indeed, the vector $\mathbf{Z} = \left(\mu\left(\frac{k\tau}{N_1}\right)\right)_{k=1,\ldots,N_1}$ is a centered Gaussian vector. Therefore, there exits

a $N_1 \times N_1$ orthonormal matrix U such that $\mathbf{V} = U\mathbf{Z}$ is a centered Gaussian vector with independent components. Moreover, by using that $I\!E\mu\left(\frac{k\tau}{N_1}\right)^2 = S$ for all k, we deduce

$$\sum_{k=1}^{N_1} I\!E V_k^2 = \sum_{k=1}^{N_1} I\!E Z_k^2 \;=\; N_1\, S. \tag{17}$$

We will also use that when X is a centered Gaussian random variable with variance $\varsigma < 1/2$, then $I\!E\exp(X^2) = 1/\sqrt{1-2\varsigma}$.

With these two tools, we will turn us now to the proof of the bound (16). By using the independency of the component of $\mathbf{V}$, one get that for every integer N_1

$$\begin{aligned} I\!E\exp\left(\sum_{k=1}^{N_1}\left(\frac{\tau}{N_1}\right)\mu\left(\frac{k\tau}{N_1}\right)^2\right) &= I\!E\prod_{k=1}^{N_1}\exp\left(\left(\frac{\tau}{N_1}\right)\mathbf{V}_k^2\right) \\ &= \prod_{k=1}^{N_1} I\!E\exp\left(\left(\frac{\tau}{N_1}\right)\mathbf{V}_k^2\right) \\ &= \prod_{k=1}^{N_1}\left(1-2\left(\frac{\tau}{N_1}\right) I\!E\mathbf{V}_k^2\right)^{-1/2} \end{aligned}$$

where the last equality is valid if and only if $I\!E\left(\left(\frac{\tau}{N_1}\right)\mathbf{V}_k^2\right) < 1/2$ for all $k = 1, \ldots, N_1$. But, by using successively (17) and $\tau < 1/(4S)$, for all $k = 1, \ldots, N_1$, one have

$$I\!E\left(\frac{\tau}{N_1}\mathbf{V}_k^2\right) \;\le\; \left(\frac{\tau}{N_1}\right)\sum_{k=1}^{N_1} I\!E\mathbf{V}_k^2 \;=\; \left(\frac{\tau}{N_1}\right)\times N\,S \;=\; \tau\,S \;<\; 1/4.$$

One can prove that $\displaystyle\prod_{k=1}^{N_1}(1-x_k)^{-1/2} \le \left(1-\sum_{k=1}^{N_1} x_k\right)^{-1/2}$ for all family of real numbers $(x_1, \ldots, x_{N_1}) \in (0,1)^{N_1}$. Combining this remark with (17), one obtains

$$I\!E\exp\left(\sum_{k=1}^{N_1}\left(\frac{\tau}{N_1}\right)\mu\left(\frac{k\tau}{N_1}\right)^2\right) \;=\; \left(1-2\frac{\tau}{N_1}\sum_{k=1}^{N_1} I\!E\mathbf{V}_k^2\right)^{-1/2} \;\le\; (1-2\tau\,S)^{-1/2}\,. \tag{18}$$

It remains to deduce the bound with the integral from the bounds for the Riemann sums. For every integer N, we denote by μ_N the process defined by

$$\mu_N(t) = \sum_{k=1}^{N}\mu\left(\frac{k\tau}{N}\right)\mathbf{1}_{[(k-1)\tau/N,\, k\tau/N)}(t).$$

We have

$$\lim_{N\to\infty} I\!E \int_0^T |\mu(t) - \mu_N(t)|^2 \, dt = 0. \tag{19}$$

Indeed, one can remark that $\mu(t) = \displaystyle\int_{-\omega_J}^{\omega_J} e^{it\xi} \, \varphi(\xi) \, \widehat{W}(d\xi)$, where φ is a continuous map bounded by M_1.

One first deduces that $\mu_N(t) = \displaystyle\int_{-\omega_J}^{\omega_J} \varphi(\xi) \left(\sum_{k=1}^{N} e^{k\tau/N} \, \mathbf{1}_{[(k-1)\tau/N,\, k\tau/N)}(t) \right) \widehat{W}(d\xi)$ and after that

$$\begin{aligned}
I\!E \int_0^T |\mu(t) - \mu_N(t)|^2 \, dt &= \int_0^T \int_{-\omega_J}^{\omega_J} |\varphi(\xi)|^2 \left| e^{it\xi} - \sum_{k=1}^{N} e^{k\tau/N} \, \mathbf{1}_{[(k-1)\tau/N,\, k\tau/N)}(t) \right|^2 d\xi \, dt \\
&\leq M_1^2 \sum_{k=1}^{N} \int_{-\omega_J}^{\omega_J} \int_{(k-1)\tau/N}^{k\tau/N} \left| e^{it\xi} - e^{k\tau/N} \right|^2 d\xi \, dt \\
&= 4M_1^2 \times N \int_0^{\omega_J} \int_0^{\tau/N} \left| e^{itu} - 1 \right|^2 d\xi \, du \\
&= 4M_1^2 \times N \int_0^{\omega_J} \left(\frac{1}{N} - \frac{1}{\xi} \sin(\xi/N) \right) d\xi \\
&\leq Ctte \times 1/N^2.
\end{aligned}$$

Now, from (19), we deduce the existence of a subsequence $(N_p)_{p\in I\!N}$ for which one have almost surely $\lim_{p\to\infty} \displaystyle\int_0^T \mu_N(t)^2 \, dt = \int_0^T \mu(t)^2 \, dt$. Hence, it follows from Fatou's lemma and the bound (18) that

$$\begin{aligned}
I\!E \exp\left(\int_0^T \mu(t)^2 \, dt \right) &\leq \liminf_{p\to\infty} I\!E \exp\left(\int_0^T \mu_{N_p}(t)^2 \, dt \right) \\
&= \liminf_{p\to\infty} I\!E \exp\left(\sum_{k=1}^{N_1 p} \left(\frac{\tau}{N_p} \right) \mu\left(\frac{k\tau}{N_p} \right) \right) \\
&\leq (1 - 2\tau \, S)^{-1/2}.
\end{aligned}$$

This induces (16) and finishes the proof of the lemma 4.1. ■

References

1. Ayache, A. and Lévy Véhel, J., 2000. "The Generalized Multifractional Brownian Motion." *S.I.S.P.*, Vol. 3, Issue 1/2, p. 7-18.
2. Ayache, A.; Cohen, S. and Lévy Véhel, J., 2000. "The covariance structure of multifractional Brownian motion." ICASSP (2000).

3. Bardet J.M. and Bertrand, P., 2003 a. Definition, properties and wavelet analysis of multiscale fractional Brownian motion, Preprint of L.S.P., Université Toulouse III.
4. Bardet J.M. and Bertrand, P., 2003 b. Identification of the multiscale fractional Brownian motion with biomechanical applications. Preprint of L.S.P., Université Toulouse III.
5. Cheridito, P., 2003. Gaussian moving averages, semimartingales and option pricing *Stochastic Processes and their Applications*, 109(1), p. 47-68.
6. Delbaen, F. and Schachermayer, W., 1994. "A general version of the fundamental theorem of asset pricing", *Math. Annals*, (1994) p. 463-520.
7. Karatzas, I., and Shreve, S. E., 1991, *Brownian Motion and Stochastic Calculus*, 2nd edition, Springer Verlag, New-York.
8. Kolmogorov, "Wienersche Spiralen und einige andere interessante Kurven in Hilbertchen Raume", *Doklady*, **26**, p.115-118, 1940.
9. Mandelbrot, B. and Van Ness J., 1968. Fractional Brownian motion, fractional noises and applications. *SIAM review* 10, p.422-437.
10. Musiela, M. and Rutkowski, M., 1998, *Martingale Methods in Financial Modelling*, 2nd ed. Springer Verlag, New-York.
11. Revuz, A. and Yor, M., 1999, *Continuous Martingales and Brownian Motion*, 3rd ed. Springer Verlag, New-York.
12. Rogers, L.C.G., 1997. Arbitrage with fractional Brownian motion. *Mathematical Finance* 7, p. 95-105.
13. Samorodnitsky, G. and Taqqu M. S., 1994. *Stable non-Gaussian Random Processes*, Chapman & Hall.
14. Willinger, W., Taqqu,M.S. and Teverovsky, V., 1999. Stock market price and long-range dependence, *Finance and Stochastics*, 1, p. 1-14

5

INTERNET TRAFFIC

Limiting Fractal Random Processes in Heavy-Tailed Systems

Ingemar Kaj

Dept. of Mathematics, Uppsala University, Box 480, SE 751 06 Uppsala, Sweden
ikaj@math.uu.se

Summary. We give an overview of limit results for a selection of stochastic models that involve heavy-tailed distributions, exhibit long-range dependence and are naturally parametrized by a tail-index. Under aggregation of independent subsystems and simultaneous time or space rescaling, the asymptotic behavior of such systems varies considerably. It is the relative speed of aggregation degree and rescaling that determines the nature of the limit process, ranging from fractional Brownian motion to stable Lévy processes. The limits obtained for a generalized example based on a spatial Poisson grains model include a fractional Brownian field. We are particularly interested in the intermediate scaling regime, bridging Gaussian and stable asymptotics. One feature shared by all limit processes is identified as an aggregate-similarity property.

1 Introduction

The purpose of this paper is to give an overview of a class of convergence results for scaled random processes. The building blocks of the systems we study are real-valued standard random processes in continuous time, such as renewal, Lévy, or M/G/∞ models. Some higher-dimensional examples include the use of spatial Poisson point processes. The common feature is that all models involve heavy-tailed distributions, exhibit long-range dependence, and are naturally parametrized by the corresponding tail-index. To avoid an excessive number of variations of similar models we focus on those that have stationary increments. We are interested in the stochastic fluctuations that build up when a large number of independent subsystems are super-positioned and simultaneously scaled in time (or, if applicable, in the space variable). In other words, our results concern convergence in distribution when performing double limits simultaneously. To compensate for an increase in the degree of aggregation we rescale time, and it is the relative speed of the two which determines the asymptotic result. The proper scalings involve the tail-index and it turns out that that there exist essentially three scaling regimes, each with a different asymptotic behavior.

It has been shown in various earlier studies that properly centered and normalized processes with long-range dependence have Gaussian as well as stable asymptotic regimes, depending on the particular choice in performing the double limit scaling. As a rule, the limit is a fractional Brownian motion if the degree of aggregation grows fast enough relative to the speed of time rescaling and the limit is a stable Lévy process if the degree of aggregation is slow compared to the change of time scale. In the former case the aggregation of independent subsystems is the dominating effect and yields a Gaussian limit in line with ordinary central limit theory. The effect of scaling time is merely in shaping the resulting covariance function. In the latter case the time scale is so large that the heavy tails no longer are able to carry long memories. The increments become asymptotically independent and generate summation schemes in the domain of attraction of stable laws. The motivation for many of these studies were attempts to understand the characteristics of flows and fluctuations in packet network computer traffic. The double limits correspond to increasing the capacity as well as the degreee of aggregation of packet streams on high-speed links and the two limit regimes correspond to fast or slow growth of aggregation relative to capacity (time), [15–17, 19]

In the next section we begin with a brief presentation of six different models for which there are known results on the asymptotic behavior under double limit scaling, and where the intermediate regime linking Gaussian and stable behavior has been investigated. This includes super-positioned renewal counting processes [3], the sum of inverse Lévy subordinators [8], infinite source Poisson models [10], a self-similar rate model [9], a spatial interference model for wireless communication [4], and a spatial model for positively correlated mass configurations [7]. In Sect. 3 we first discuss the three scaling regimes of fast, slow and intermediate growth, and give some heuristics for the corresponding limit processes. Then we state and compare the scaling limit results for the six different models. Section 4 contains some proofs that can not be found elsewhere.

2 A Selection of Models with Long-Range Dependence

Models 1 to 5 in the following list are one-dimensional examples and model 6 is a higher-dimensional spatial Poisson germs and grains model which is constructed to have long-range dependence over spatial distances.

2.1 Renewal Process with Heavy-Tailed Interarrival Times

Consider a renewal counting process $N = \{N(t), t \geq 0\}$ associated with the sequence of independent interrenewal times $U_1, U_2, \ldots$. The variables $(U_k)_{k\geq 2}$ are identically distributed and U_1 has the corresponding equilibrium distribution, so that N has stationary increments. The interrenewal times are heavy-tailed in the sense of possessing a regularly varying tail,

$$P(U > t) \sim t^{-\gamma} L(t), \quad 1 < \gamma < 2, \tag{1}$$

where $L(t)$ is a slowly varying function and the tail index γ is such that the expected value $\mu = E(U_k)$, $k \geq 2$, is finite but the variance is infinite. Letting $(N^{(i)})_{i\geq 1}$ be independent copies of N, we form for each aggregation level $m \geq 1$ the centered superposition process

$$Y_1^{(m)}(t) = \frac{1}{b_m} \sum_{i=1}^{m} (N^{(i)}(a_m t) - \frac{a_m t}{\mu}), \quad t \geq 0,$$

which is scaled and normalized using sequences $a_m, b_m \to \infty$, $m \to \infty$.

2.2 Inverse Lévy Process

Let $\{\widetilde{X}_t, t \geq 0\}$, $\widetilde{X}_0 = 0$, denote a Lévy subordinator with right-continuous paths and Laplace transform given by $-\ln E(e^{-u\widetilde{X}_t}) = t\Phi(u)$, $u \geq 0$, where $\Phi(u) = \int_0^\infty (1 - e^{-ux})\, \nu(\mathrm{d}x)$ is the Laplace exponent. Here, the Lévy measure $\nu(\mathrm{d}x)$ has regularly varying tail such that

$$\int_x^\infty \nu(\mathrm{d}y) \sim x^{-\gamma} L(x), \quad 1 < \gamma < 2, \quad \mu = \int_0^\infty x\nu(\mathrm{d}x) < \infty.$$

Consider the delayed subordinator $X_t = X_0 + \widetilde{X}_t$ where $P(X_0 \leq x) = \frac{1}{\mu} \int_0^x \int_y^\infty \nu(\mathrm{d}s)\, \mathrm{d}y$ and define the first passage time process

$$T_x = \inf\{t \geq 0 : X_t > x\}, \quad x \geq 0.$$

Then $ET_x = x/\mu$ and $\{T_x, x \geq 0\}$ has stationary increments, [5,8]. As in the previous example we consider the rescaled aggregation process

$$Y_2^{(m)}(t) = \frac{1}{b_m} \sum_{i=1}^{m} (T_{a_m x}^{(i)} - \frac{a_m x}{\mu}), \quad x \geq 0.$$

2.3 Infinite Source Poisson Models

Let $(N_t^\lambda)_{t\geq 0}$ be a Poisson process with intensity λ and denote by $(S_i)_{i\geq 1}$ the epochs of the successive Poisson events. Let $(U_i)_{i\geq 1}$ be independent, non-negative random variables, independent of the Poisson process, with distribution function $G(t)$ such that the tail satisfies $P(U > t) \sim \gamma^{-1} t^{-\gamma} L(t)$. Here, for convenience the heavy tail asymptotics is slightly modified compared to (1). The expected value EU is finite. Independently of N and $(U_i)_{i\geq 1}$, let $(V_i)_{i\geq 1}$ denote an i.i.d. sequence of random variables with the corresponding equilibrium distribution $G_{\mathrm{eq}}(t) = \frac{1}{EU} \int_0^t P(U > s)\, \mathrm{d}s$ and let M be Poisson with expected value $\lambda\, EU$. Then

$$\Lambda_\lambda(t) = \sum_{i=1}^{N_t^\lambda} 1_{\{S_i+U_i>t\}} + \sum_{i=1}^{M} 1_{\{V_i>t\}}, \quad t \geq 0,$$

is a stationary version of the system size in the standard M/G/∞ queuing model with service time distribution G. The infinite source Poisson model is defined as the corresponding accumulated flow process

$$W_{\mathrm{eq},\lambda}(t) = \int_0^t \Lambda_\lambda(s)\,\mathrm{d}s = \sum_{i=1}^{N_t^\lambda}(t-S_i)\wedge U_i + \sum_{i=1}^{M} t\wedge V_i. \tag{2}$$

The double limit scaling problem is to find the asymptotic behavior of

$$Y_3^{(\lambda)}(t) = \frac{1}{b_\lambda}\int_0^{a_\lambda t} (\Lambda_\lambda(s) - \lambda\, EU)\,\mathrm{d}s, \quad t \geq 0,$$

where a_λ and b_λ are scaling parameters such that $a_\lambda, b_\lambda \to \infty$ as $\lambda \to \infty$.

An alternative approach to this class of models is based on Poisson point measures, [1, 10, 12]. Let $N(\mathrm{d}s, \mathrm{d}u)$ be a Poisson point measure on $\mathbf{R}\times\mathbf{R}_+$ with intensity measure $\lambda \mathrm{d}s\, G(\mathrm{d}u)$. Then

$$\Lambda_\lambda(t) \stackrel{d}{=} \int_{-\infty}^{\infty}\int_0^{\infty} 1_{\{s<t<s+u\}}\, N(\mathrm{d}s, \mathrm{d}u),$$

and hence, letting

$$K_t(s,u) = \int_0^t 1_{\{s<y<s+u\}}\,\mathrm{d}y, \quad (s,u)\in\mathbf{R}\times\mathbf{R}_+,\ t\geq 0, \tag{3}$$

we have

$$Y_3^{(\lambda)}(t) \stackrel{d}{=} \frac{1}{b_\lambda}\int_{-\infty}^{\infty}\int_0^{\infty} K_{a_\lambda t}(s,u)\,\widetilde{N}(\mathrm{d}s,\mathrm{d}u),$$

where $\widetilde{N}(\mathrm{d}s,\mathrm{d}u) = N(\mathrm{d}s,\mathrm{d}u) - \lambda \mathrm{d}s\, G(\mathrm{d}u)$ is the compensated Poisson measure associated with N.

2.4 Self-Similar Rate Model

One of the possible variations of the model in Sect. 2.3 involves a further random process $\xi(t)$ with increasing sample paths and stationary increments, which is self-similar with self-similarity index H_0 such that $1/\gamma < H_0 < 1/(2-\gamma)$, [9, 14]. A copy of ξ is supposed to run independently during each Poisson session $[s, s+u]$. This model has been suggested to capture some aspects of the TCP protocol for transmission control of web traffic. The sessions $[s, s+u]$ correspond to download requests at an Internet server occuring at time s and with duration u for the flow of IP-packets generated by each request. The additional process ξ accounts for random transfer rate of the flows on the short

time scales of single packet return transmission times, thought to originate from the feedback control mechanism of TCP. The fluctuation process for this model has the representation

$$Y_4^{(\lambda)}(t) \stackrel{d}{=} \frac{1}{b_\lambda} \int_{-\infty}^{\infty} \int_0^{\infty} \int_{\mathcal{D}} \xi(K_{a_\lambda t}(s,u))\, \widetilde{N}(\mathrm{d}s, \mathrm{d}u, \mathrm{d}\xi).$$

In this case $\widetilde{N}(\mathrm{d}s, \mathrm{d}u, \mathrm{d}\xi)$ is a compensated Poisson point process on $\mathbf{R} \times \mathbf{R}_+ \times \mathcal{D}$, where $\mathcal{D}$ is the trajectory space of càdlàg paths, with intensity measure given by $\lambda \mathrm{d}s\, G(\mathrm{d}u)\mu_H(\mathrm{d}\xi)$ where $\mu_H(\mathrm{d}\xi)$ is the distribution of ξ on $\mathcal{D}$.

2.5 Spatial Interference Model

To model a spatial communication system we have suggested in [4] the following scenario. Imagine signal transmitters placed at random locations in a set $\mathbf{S} \subset \mathbf{R}^d$ that are actively transmitting calls during a time interval of random length. The initial time s, the location x and the call holding time u of each transmission is given by a point (s, x, u) of a Poisson point measure $N(\mathrm{d}s, \mathrm{d}x, \mathrm{d}u)$ on $\mathbf{R} \times \mathbf{S} \times \mathbf{R}_+$ with intensity $\lambda \mathrm{d}s\mathrm{d}x\, G(\mathrm{d}u)$. We are interested in the total spatial interference in the system, measured by the received power at a given fixed observation point in $\mathbf{S}$, which we call the origin, resulting from the superposition of all signal sources scattered in space and time. Two types of fading reduce the received signal power, lognormal fading and Rayleigh fading. Both mechanisms have their origin in the intrinsic nature of wave propagation. In our case lognormal fading is considered to be a multiplicative wave shadowing phenomena, which makes the signal power lognormally distributed and fixed throughout the duration of the transmission session. The expected power is a decreasing function of the spatial distance between source and receiver. Rayleigh fading produces additional short time random variations due to multipath interactions, conditional on the lognormal session average. In [4] we argue that the Lévy gamma process is a natural model for Rayleigh fading. The arguments are based on interesting links between Gaussian complex waves, Bessel diffusions with fractal dimension, and Lévy subordinators.

Let $g(x) = g_0/(1 + |x|)^\beta$, $\beta > d$, be the attenuation function, which describes the reduction in average signal power received at the origin when transmitted in location x. Let V with distribution $F_x(dv)$ have the lognormal distribution with expected value $g(x)$ and second moment $EV^2 = g(x)^2 \exp \sigma_L^2$, where $\sigma_L^2 > 0$ is a parameter for lognormal fading variation. Given $V = v$, let $Q_0^v(d\gamma)$ be the distribution in $\mathcal{D}$ of the Lévy subordinator $\{\Gamma_v(t), t \geq 0\}$ with Lévy measure $\nu(dy) = y^{-1}e^{-y/v}\, dy$. Then, we take $\sigma_R^2\, \Gamma_v(t)$ to represent the Rayleigh variation with variance parameter σ_R^2 over a time interval of length t. As a result of these assumptions we can express the time evolution of the accumulated interference in terms of a Poisson measure $N(\mathrm{d}s\mathrm{d}x, \mathrm{d}u, \mathrm{d}v, \mathrm{d}\gamma)$ with intensity measure $\lambda \mathrm{d}s\mathrm{d}x\, G(\mathrm{d}u)F_x(\mathrm{d}v)Q_0^v(\mathrm{d}\gamma)$. In close analogy to the

models in Sects. 2.3, 2.4, we obtain the centered and scaled fluctuations of accumulated interference in the form

$$Y_5^{(\lambda)}(t) = \frac{1}{b_\lambda} \int_{\mathbf{R}\times\mathbf{R}^d} \int_0^\infty \int_0^\infty \int_{\mathcal{D}} \sigma_R^2 \, \gamma(K_{a_\lambda t}(s,u)) \, \widetilde{N}(\mathrm{d}s\mathrm{d}x, \mathrm{d}u, \mathrm{d}v, \mathrm{d}\gamma), \quad t \geq 0,$$

where we write again $\widetilde{N}$ for the compensated Poisson measure.

2.6 Fractional Gaussian Field

In search of a natural generalization of (positively correlated) fractional Brownian motion to a spatial setting with strong depenendence over long spatial distances, it is proposed in [7] to study a heavy-tailed Poisson grains model.

We start from random configurations of mass in $\mathbf{R}^d$ obtained by placing balls $B(x,r)$ of random radius r centered at the locations x of a Poisson point measure with Lebesgue intensity measure $\mathrm{d}x$. Each ball has independent radius R drawn from a heavy-tailed distribution $F(\mathrm{d}r)$, such that $P(R > r) \sim \gamma^{-1} r^{-\gamma} L(r)$ with index γ satisfying $d < \gamma < 2d$. Let $N(\mathrm{d}x, \mathrm{d}r)$ be a Poisson point measure with intensity $\mathrm{d}x\, F(\mathrm{d}u)$. The superposition of all grains (balls, in this simplest case) contains a mixture of large grains that will keep the mass configurations rigid, and small grains that will cause non-trivial noise effects. One technique for studying such random mass configurations is to use finite Borel measures μ on $\mathbf{R}^d$ as an indexing tool. By assigning to each measure μ the aggregated μ-measure of all balls in a Poisson configuration, we obtain the map

$$\mu \mapsto W(\mu) = \int_{\mathbf{R}^d} \int_0^\infty \mu(B(x,r)) \, N(\mathrm{d}x, \mathrm{d}r),$$

which is viewed as a random field indexed by the collection of measures μ. The stochastic integral is well defined a a limit in probability of elementary integrals, [11] Lemma 12.13. Moreover, $EW(\mu) = \mu(\mathbf{R}^d) E(R^d) |B_d|$, where $B_d = B(0,1)$ is the unit ball in R_d.

Let $M_f^{\pm}$ denote the set of finite signed Borel measures on $\mathbf{R}^d$ and $|\mu|$ the total variation of μ. The extended mapping $\mu \to W(\mu)$, $\mu \in M_f^{\pm}$, is linear. Now we restrict to signed measures with finite α-energy, to obtain an index set for which $W(\mu)$ has suitable properties. The α-energy of $\mu \in M_f^{\pm}$ is the functional

$$I_\alpha(\mu) = \int_{\mathbf{R}^d} \int_{\mathbf{R}^d} \frac{\mu(\mathrm{d}y)\mu(\mathrm{d}y')}{|y - y'|^\alpha}.$$

For $0 \leq \alpha \leq d$ let

$$\mathcal{M}_\alpha = \left\{ \mu \in M_f^{\pm} : I_\alpha(|\mu|) < \infty \right\}$$

denote the family of finite signed Borel measures with finite α-energy (so that $\mathcal{M}_0 = M_f^{\pm}$). The relevant scaling for this model is to replace the Poisson

intensity by $\lambda \mathrm{d}x$ and the radius R by ρR, $\rho > 0$. Write $N_{\lambda,\rho}$ for the corresponding Poisson measure and $F_\rho(\mathrm{d}r)$ for the distribution of ρR. The centered and scaled quantity of interest is

$$J_{\lambda,\rho}(\mu) = \int_{\mathbf{R}^d} \int_0^\infty \mu(B(x,r)) \left(N_{\lambda,\rho}(\mathrm{d}x, \mathrm{d}r) - \lambda \mathrm{d}x F_\rho(\mathrm{d}r)\right), \quad \mu \in \mathcal{M}_d,$$

with $\rho = \rho_\lambda \to 0$ as $\lambda \to \infty$.

3 Asymptotic Scaling Regimes and Results

All of the one-dimensional models in Sects. 2.1–2.5 are parametrized by a regularly varying tail index γ, $1 < \gamma < 2$. For the d-dimensional model in Sect. 2.6 we have $d < \gamma < 2d$. To obtain asymptotic results for the one-dimensional models we must select a time scale a that tends to infinity at a certain speed relative to m (or λ), and then choose a proper normalization sequence b. For the spatial model 2.6 the goal is to find the asymptotic behavior in the limit of "many-small-balls". It is thus required to rescale space by letting the radius ρ tend to 0 at a proper speed, relative to which the Poisson intensity λ increases the number of balls. Next we discuss three natural limit regimes that arise in all of these examples and give some heuristics for the underlying dependence structure.

3.1 Fast, Intermediate and Slow Growth Rates

In some of the earlier works first pointing out the dichotomy of Gaussian and stable limit regimes, the limit operations were performed separately and the result would depend on the order in which the limits were taken, [18]. Much recent work have focused on performing the two limits jointly. Typically, for the models in one dimension, it is the asymptotic behavior of the ratio $mL(a)/a^{\gamma-1}$ (or $\lambda L(a)/a^{\gamma-1}$) which determines the normalization b and the limit process. To understand this property we discuss the renewal model introduced in Sect. 2.1. Out of m counting processes observed at a fixed time t, let $\#(m,a)$ be the number of processes with residual renewal time exceeding a, that is $\#(m,a) = \sum_{i=1}^m 1_{\{V_i > a\}}$ where V_i has the equilibrium distribution associated with the interrenewal time U. Hence, for large a the expected number of large gaps between renewal times observed at time t, measured using residual waiting times, is given by

$$E\#(m,a) = mP(V > a) \sim mL(a)/\mu a^{\gamma-1}.$$

Suppose $m\,P(V > a)$ is large asymptotically with a and m. Then we expect to have at any given cut point t a large number of overlapping empty gaps between renewal jumps among the m counting processes. This is a sign of strong positive dependence in the trajectories of the summation process. On

the contrary, if $mP(V > a)$ vanishes asymptotically there is a lack of overlapping interrenewal intervals and hence a tendency for the increments of the summation process to become independent. The further possibility is to have a balance between m and a such that $mP(V > a)$ remains constant in the asymptotic limit. Similar arguments apply for the other models. In Sect. 2.3, the analog equibilrium quantity $\#(\lambda, a) = \sum_{i=1}^{M} 1_{\{V_i > a\}}$ counts the number of ongoing long sessions at a fixed time. For large λ and a, the expected number $E\#(\lambda, a)$ is asymptotically $\lambda L(a)/a^{\gamma-1}$. For the model in Sect. 2.6, the expected number of balls containing the origin and with volume exceeding that of the unit ball is proportional to $\lambda\rho^{\gamma}$. This reasoning makes it natural to distinguish three cases of fast, intermediate and slow growth and introduce a parameter $c > 0$ to quantify the relative speed of rescaling, summarized as follows:

Models	$2.1 - 2.2$	$2.3 - 2.5$	2.6
fast growth:	$\frac{mL(a)}{a^{\gamma-1}} \to \infty$	$\frac{\lambda L(a)}{a^{\gamma-1}} \to \infty$	$\lambda\rho^{\gamma}L(1/\rho) \to \infty$
intermediate growth:	$\frac{mL(a)}{a^{\gamma-1}} \to \mu\, c^{\gamma-1}$	$\frac{\lambda L(a)}{a^{\gamma-1}} \to c^{\gamma-1}$	$\lambda\rho^{\gamma}L(1/\rho) \to c^{\gamma}$
slow growth:	$\frac{mL(a)}{a^{\gamma-1}} \to 0$	$\frac{\lambda L(a)}{a^{\gamma-1}} \to 0.$	$\lambda\rho^{\gamma}L(1/\rho) \to 0$

We will see below how the scaling parameter c enters in the limit results, and discuss its interpretation further in Sect. 3.6.

3.2 Fast and Slow Growth Limits, One-Dimensional Case

All models in Sects. 2.1-2.5 have the same generic behavior. Assuming fast growth scaling the limit is fractional Brownian motion with Hurst index a function of the heavy tail parameter γ. Under slow growth rescaling the corresponding limit is γ-stable Lévy. It should be pointed out that if we consider more general models, such as renewal-reward processes with heavy-tailed rewards, then the simple picture of Gaussian and stable regimes described here is no longer true. For example, fast growth scaling could yield in the limit a stable process called the Telecom process [10, 13].

Fast Growth

Under fast growth conditions the proper normalizing sequences b are given by $b_m^2 = ma^{3-\gamma}L(a)$ and $b_\lambda^2 = \lambda a^{3-\gamma}L(a)$. In particular, $b/a \to \infty$. As m and a or λ tend to infinity, the rescaled processes $Y_k^{(\cdot)}$, $k = 1, \ldots, 5$, converge in finite-dimensional distributions to multiples of fractional Brownian motion $B_H(t)$. For $k = 1, 2, 3, 5$ the Hurst index is $H = (3-\gamma)/2 \in (1/2, 1)$, [3, 4, 8, 15]. The limit process of $Y_4^{(\lambda)}$ in Sect. 2.4 is fractional Brownian motion with Hurst index $H = (1-\gamma/2)H_0 + 1/2 \in (1/2, 1)$, [9].

We illustrate these results by indicating the proof for model 2.1 given in [3] based on convergence of cumulants, taking for simplicity $\mu = 1$. There are constants α_{kj} such that

$$D_k^{(m)}(t) = m\frac{d^k}{d\theta^k} \ln E \exp \theta Y_1^{(m)}(t)\Big|_{\theta=0} = m \sum_{j=0}^{k} \alpha_{kj} E(N_{at} - at)^j / b^j.$$

By analyzing the moments for heavy-tailed renewal processes and using the fast growth asymptotics it follows that $D_2^{(m)}(t) \to \sigma_\gamma^2\, t^{3-\gamma}$, where $\sigma_\gamma^2 = 2/(\gamma - 1)(2-\gamma)(3-\gamma)$, and $D_k^{(m)} \to 0$ if $k > 2$. Since $D_1^{(m)} = 0$ this shows that the cumulants of $Y_1^{(m)}$ converge to those of $\sigma_\gamma B_H$ with $H = (3-\gamma)/2$. Moreover, the covariance function of the process $Y_1^{(m)}$ converges to that of $\sigma_\gamma B_H$. Indeed, since the increments are stationary,

$$\begin{aligned} E(Y_1^{(m)}(t) Y_1^{(m)}(s)) &= \frac{1}{2}(D_2^{(m)}(t) + D_2^{(m)}(s) - D_2^{(m)}(t-s)) \\ &\to \frac{\sigma_\gamma^2}{2}(t^{3-\gamma} + s^{3-\gamma} - (t-s)^{3-\gamma}). \end{aligned}$$

To relate these results for fast growth to the situation under slow and intermediate growth we mention also a representation of fractional Brownian motion which uses the function K_t in (3). Let $M_2(\mathrm{d}s, \mathrm{d}u)$ be a Gaussian measure on $\mathbf{R} \times \mathbf{R}_+$ with control measure $\mathrm{d}s\, u^{-\gamma-1}\mathrm{d}u$. Then

$$\sigma_\gamma B_H(t) = \int_{-\infty}^{\infty} \int_0^{\infty} K_t(s,u)\, M_2(\mathrm{d}s, \mathrm{d}u).$$

Slow Growth

The normalization sequence in this case is obtained from the relation $\lambda a P(U > b) = 1$, which we can assume is invertible in b. If we ignore the slowly varying function this is simply $b = (\lambda a)^{1/\gamma} \to \infty$. Moreover, $b/a \to 0$.

Each of the models $Y_k^{(\cdot)}$, $k = 1, 3, 4, 5$ converges in the sense of finite dimensional distributions to a multiple of a γ-stable Lévy process [2–4, 9, 15]. We think the same is true for $Y_2^{(m)}$. To illustrate this type of convergence we discuss the case $Y_3^{(\lambda)}$ and follow the approach in [10]. Using the scaling properties of the function $K_t(s,u)$ introduced in (3), it is seen that

$$\frac{1}{b}\int_{-\infty}^{\infty}\int_0^{\infty} K_{at}(s,u)\, \widetilde{N}(\mathrm{d}s, \mathrm{d}u) = \int_{-\infty}^{\infty}\int_0^{\infty} K_{(a/b)t}((a/b)s, u)\, \widetilde{N}(a\,\mathrm{d}s, b\,\mathrm{d}u).$$

Here, the compensator scales as $\lambda a \mathrm{d}s\, F(b\,\mathrm{d}u) \sim \mathrm{d}s\, u^{-1-\gamma}\mathrm{d}u$. Moreover, if we write $z = a/b$ then $z \to \infty$ and

$$K_{zt}(zs, u) = \int_0^u 1_{\{0<y+zs<zt\}}\, \mathrm{d}y \to \int_0^u 1_{\{0<s<t\}}\, \mathrm{d}y = u 1_{\{0<s<t\}}.$$

Thus, we expect the limit process to be given by

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} u1_{\{0<s<t\}}\,\big(N(\mathrm{d}s,\mathrm{d}u)-\mathrm{d}s\,u^{-1-\gamma}\,\mathrm{d}u\big)\stackrel{d}{=}\sigma_\gamma\int_0^t M_\gamma(\mathrm{d}s)\stackrel{d}{=}\Lambda_\gamma(t),$$

where σ_γ is an appropriate constant, $M_\gamma(ds)$ is a γ-stable random measure with control measure ds and $\Lambda_\gamma(t)$ is a Lévy-stable process with index γ. To formalize the argument one can verify that the corresponding rescaled characteristic function converges to that of the γ-stable limit.

3.3 Intermediate Growth Limit Process, One-Dimensional Case

In the intermediate scaling regime the proper normalization is simply to take $b = a$. A limit result for this case was first obtained for the renewal model $Y_1^{(m)}$ in [3].

Theorem 1. [3] *Consider the case of intermediate growth rate with $c = 1$ as $m \to \infty$, that is $mL(a)/a^{\gamma-1} \to \mu$. Then we have weak convergence of the process $\{Y_1^{(m)}(t), t \geq 0\}$ to $\{-\mu^{-1}Y_\gamma(t), t \geq 0\}$, where Y_γ is a zero mean, non-Gaussian and non-stable process with stationary increments. The limit process has continuous trajectories and is not self-similar. It has finite moments of all orders and positive skewness. The finite-dimensional distributions of the increments $\Delta Y_\gamma(t_i) = Y_\gamma(t_i) - Y_\gamma(t_{i-1})$ of Y_γ are characterized by the cumulant generating function*

$$\begin{aligned}\ln E\exp\Big\{\sum_{i=1}^n\theta_i\Delta Y_\gamma(t_i)\Big\} &= \frac{1}{\gamma-1}\sum_{i=1}^n\theta_i^2\int_0^{\Delta t_i}\int_0^v e^{\theta_i u}u^{-(\gamma-1)}\,\mathrm{d}u\mathrm{d}v\\ &+\frac{1}{\gamma-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\theta_i\theta_j\exp\Big\{\sum_{k=i+1}^{j-1}\theta_k\Delta t_k\Big\}\\ &\times\int_0^{\Delta t_i}\mathrm{d}v\int_0^{\Delta t_j}e^{\theta_j u}e^{\theta_i v}(t_{j-1}-t_i+u+v)^{-(\gamma-1)}\,\mathrm{d}u, \qquad (4)\end{aligned}$$

where $0 = t_0 < t_1 < \cdots < t_n$ and $\Delta t_j = t_j - t_{j-1}$.

It can be checked that if we take a general growth parameter $c > 0$ and assume $mL(a)/a^{\gamma-1} \to \mu c^{\gamma-1}$, then instead of $Y_\gamma(t)$ we obtain $cY_\gamma(t/c)$ in the limit.

A completely analogous result holds for the process $Y_2^{(m)}$ based on inverse Lévy subordinators with regularly varying Lévy measure of index $\gamma \in (1, 2)$.

Theorem 2. [8] *Suppose that the Lévy measure $\nu(\mathrm{d}x)$ has strictly positive lower index, that is*

$$\sigma = \sup\Big\{\alpha > 0 : \lim_{u\to\infty} u^{-\alpha}\Phi(u)\Big\} > 0.$$

Assuming intermediate growth $mL(a)/a^{\gamma-1} \to \mu c^{\gamma-1}$ *as* m *and* $a = a_m$ *tends to infinity, we have the weak convergence of continuous processes*

$$\Big\{\frac{1}{a}\sum_{i=1}^{m}(T_{ax}^{i} - \frac{1}{\mu}ax),\ x \geq 0\Big\} \Rightarrow \{-\mu^{-1}c\,Y_{\beta}(x/c),\ x \geq 0\}.$$

In spite of their similarity the proofs of convergence of the marginal distributions in Thms. 1 and 2 are rather different. The proofs of convergence of the multidimensional distributions are more closely related. In both cases they are based on recursive systems of integral equations satisfied by families of their moment generating functions.

The following limit theorem for the infinite source Poisson process $Y_3^{(\lambda)}$ valid for the intermediate scaling regime is derived in [10].

Theorem 3. [10] *Fix* $1 < \gamma < 2$ *and* $c > 0$*, and consider the intermediate growth condition* $\lambda L(a)/a^{\gamma-1} \to c^{\gamma-1}$ *as* $\lambda \to \infty$*. Then we have the weak convergence of continuous random processes*

$$\Big\{Y_3^{(\lambda)}(t), t \geq 0\Big\} \Rightarrow \Big\{\int_{-\infty}^{\infty}\int_{0}^{\infty} cK_{t/c}(s,u)\,(N(\mathrm{d}s,\mathrm{d}u) - \mathrm{d}s\,u^{-\gamma-1}\mathrm{d}u), t \geq 0\Big\},$$

where the limit process is well defined as a stochastic integral with respect to the compensated Poisson measure on $\mathbf{R}\times\mathbf{R}_+$ *with intensity measure* $\mathrm{d}s\,u^{-\gamma-1}\mathrm{d}u$.

Comparison of these quoted results led us to believe that the limits in Thms. 1,2 and in Thm. 3, respectively, are actually two representations of the same process in the sense of coinciding finite-dimensional distributions. That this is indeed the case has been confirmed in [6] (a partly outdated preprint which is replaced by the present work) and in [2] by persistent rewriting of the expression (4). Note also the opposite signs appearing in Thms. 1,2 on one hand and in Thm. 3 on the other. Long interarrival times in the renewal model yield long periods without positive jumps and hence no contributions to the counting processes. Same for the inverse Lévy processes. On the contrary, long sessions in the infinite source Poisson model correspond to a continuous flow of positive mass throughout the session.

Theorem 4. *The limit process* Y_γ *in Thms. 1,2 has the representation*

$$Y_\gamma(t) \stackrel{d}{=} \int_{-\infty}^{\infty}\int_{0}^{\infty} K_t(s,u)\,(N(\mathrm{d}s,\mathrm{d}u) - \mathrm{d}s\,\gamma u^{-\gamma-1}\mathrm{d}u),$$

where $K_t(s,u) = \int_0^t 1_{\{s<y<s+u\}}\,\mathrm{d}y$.

The extra factor γ in the Poisson intensity has been inserted to account for the choice of parameters in (1) used in Thms. 1 and 2. To comply with the notation in [10] we used the tail asymptotics $\gamma^{-1}t^{-\gamma}$ in Thm 3. We will discuss the proof of Thm. 4 in Sect. 4.

3.4 Further Variations in One Dimension

We mention briefly some additional results for intermediate growth rescaling by discussing the models $Y_4^{(\lambda)}$ and $Y_5^{(\lambda)}$. It is interesting to see that the short time variations in the self-similar rate model, represented by the self-similar process $\xi(t)$, survive in the scaling. The resulting limit process is a Poisson stochastic integral with a stochastic integrand:

$$Y_4^{(\lambda)}(t) \to \int_{-\infty}^{\infty}\int_0^{\infty}\int_{\mathcal{D}} K_t(s,u,\xi)\,(N(\mathrm{d}s,\mathrm{d}u,\mathrm{d}\xi) - \mathrm{d}s\,u^{-\gamma-1}\mathrm{d}u\,\mu_H(\mathrm{d}\xi)),$$

where the integrand is defined by

$$K_t(s,u,\xi) = (\xi(t+s)\wedge u - \xi(-s)\wedge v).$$

The interference model $Y_5^{(\lambda)}$ in Sect. 2.5 turns out to have the same asymptotic behavior as the integral

$$\widetilde{Y}_5^{(\lambda)}(t) = \frac{1}{b}\int_{\mathbf{R}\times\mathbf{R}^d}\int_0^{\infty}\int_0^{\infty} vK_{at}(s,u)\,(N(\mathrm{d}s\mathrm{d}x,\mathrm{d}u,\mathrm{d}v) - \mathrm{d}s\mathrm{d}x\,G(\mathrm{d}u)F_x(\mathrm{d}v)),$$

which can be analyzed in parallel to other infinite source Poisson models. Details for these models will appear elsewhere [4,9].

3.5 Limit Behavior of the Spatial Model

Turning to the Poisson grains model in Sect. 2.6 we describe briefly the results for convergence of finite-dimensional distributions obtained in [7]. For the case of intermediate growth, $\lambda\rho^\gamma L(1/\rho)\to c^\gamma$, there is no need to normalize further and we have for any $\mu\in\mathcal{M}_d$ the convergence $J_{\lambda,\rho}(\mu)\to Y_\gamma(\mu)$, where

$$Y_\gamma(\mu) = \int_{\mathbf{R}^d}\int_0^{\infty}\mu_c(B(x,r))\,(N_\gamma(\mathrm{d}x,\mathrm{d}r) - \mathrm{d}x\,r^{-\gamma-1}\mathrm{d}r),$$

and μ_c is defined by $\mu_c(A)=\mu(cA)$ for any Borel set A in $\mathbf{R}^d$, $cA=\{ca : a\in A\}$ (so that $\mu_1=\mu$). Such integrals are well defined on condition that

$$\int_{\mathbf{R}^d}\int_0^{\infty} |\mu|(B(x,r))\wedge|\mu|(B(x,r))^2\,\mathrm{d}x\,r^{-\gamma-1}\mathrm{d}r<\infty$$

([11]), which is the case for any $d<\gamma<2d$ when $\mu\in\mathcal{M}_d$. It is a pleasant property of these limiting integrals that the variance is given by the energy functional,

$$\mathrm{Var}(Y_\gamma(\mu)) = C_{\gamma,d}\,I_{\gamma-d}(\mu), \tag{5}$$

where $C_{\gamma,d}$ is a constant.

The limiting random field in the case of fast growth asymptotics can be constructed using a Gaussian measure $M_{2,\gamma}(\mathrm{d}x,\mathrm{d}r)$ on $\mathbf{R}^d\times\mathbf{R}_+$ with control

measure $\mathrm{d}x\, r^{-\gamma-1}\mathrm{d}r$. Indeed, if $\lambda\rho^\gamma L(1/\rho) \to \infty$ and $b = b_{\lambda,\rho}$ is the normalizing sequence such that $b^2 = \lambda\rho^\gamma L(1/\rho)$, then the limit of $b^{-1}J_{\lambda,\rho}(\mu)$ is a Gaussian random field $\mu \to B_H(\mu)$, such that

$$B_H(\mu) = \int_{\mathbf{R}^d}\int_0^\infty \mu(B(x,r))\, M_{2,\gamma}(\mathrm{d}x, \mathrm{d}r).$$

We have for any $c > 0$ the self-similarity property

$$B_H(\mu_c) \stackrel{d}{=} c^{dH} B_H(\mu), \quad H = \frac{3d-\gamma}{2d} \in (1/2, 1),$$

which makes it natural to think of $B_H(\mu)$ as a fractional Brownian field. The variance functional is the same as in (5) and the characteristic functional is given by

$$\ln E \exp\Big\{\mathrm{i}\int_{\mathbf{R}^d}\int_0^\infty \mu(B(x,r))\, M_{2,\gamma}(\mathrm{d}x, \mathrm{d}r)\Big\} = -\frac{1}{2} C_{\gamma,d}\, I_{\gamma-d}(\mu).$$

Finally, in the slow growth regime, if $\lambda\rho^\gamma L(1/\rho) \to 0$ and $b = b_{\lambda,\rho}$ is such that $\lambda\, P(\rho R > b^{1/d}) \to 1$, then

$$b^{-1}J_{\lambda,\rho}(\mu) \to |B_d| \int_{\mathbf{R}^d}\int_0^\infty D\mu(x)\, r^d\, (N_\gamma(\mathrm{d}x, \mathrm{d}r) - \mathrm{d}x r^{-\gamma-1}\, \mathrm{d}r),$$

where $D\mu$ is the Radon-Nikodym derivative of the absolutely continuous part of μ. The limit is a stable noise with stable index $\gamma/d \in (1,2)$.

3.6 A Similarity Property of Aggregates

Unlike limit processes under fast and slow growth scaling, intermediate growth limit processes are not self-similar. They are, however, in a certain sense similar to their aggregates. To capture this property we propose tentatively the following notion.

Definition 1. *Let $X = \{X(t),\, t \geq 0\}$, $X(0) = 0$, be a stochastic process in continuous time with $EX(t) = 0$ and stationary increments, and let $(X^{(i)})_{i\geq 1}$ denote i.i.d. copies of X. The process X is said to be aggregate-similar with rigidity-index ρ, if for each $m \geq 1$, we have the distributional identity*

$$\Big\{\sum_{i=1}^m X^{(i)}(t),\, t \geq 0\Big\} \stackrel{\text{f.d.d.}}{=} \{m^\rho\, X(t/m^\rho),\, t \geq 0\}. \tag{6}$$

The definition can be modified to cover also spatial models such as $J_{\lambda,\rho}(\mu)$.

Some standard examples of stationary increment processes are easy to check. Brownian motion is aggregate-similar with index $\rho = 1$. Fractional

Brownian motion B_H is aggregate-similar with index $\rho = 1/2(1-H)$. The α-stable Lévy process is aggregate-similar with index $\rho = 1/(\alpha - 1)$. Regarding the intermediate limit process Y_γ one can check that for any $c > 0$,

$$\ln E \exp\Big\{ \sum_{j=1}^{n} \mathrm{i}\, \theta_j c Y_\gamma(t_j/c) \Big\} = c^{1/\rho} \ln E \exp\Big\{ \mathrm{i} \sum_{j=1}^{n} \theta_j Y_\gamma(t_j) \Big\},$$

for arbitrary $n \geq 1$, time points $0 \leq t_1 \leq \cdots \leq t_n$ and real numbers θ_j, $1 \leq j \leq n$. Take c such that $c^{1/\rho} = m$ to see that Y_γ is aggregate-similar with rigidity index $\rho = 1/(\gamma - 1)$.

This property provides an interpretation of the intermediate growth scaling parameter c. If we choose c so that the number $c^{\gamma-1}$ is an integer m, then the limit process $cY_\gamma(\cdot/c)$ has the same distribution as the sum of m independent copies of Y_γ. Heuristically, for each number $E\#(m,a)$ of long overlapping interrenewal times, or each number $E\#(\lambda,a)$ of long ongoing M/G/∞ sessions, etc, we obtain in the limit one further independent copy of Y_γ.

4 Proof of Thm. 4

Consider the rescaled model

$$Y_3^{(\lambda)}(t) = \frac{1}{a_\lambda} \int_0^{a_\lambda t} (\Lambda_\lambda(s) - \lambda\, EU)\, \mathrm{d}s, \quad t \geq 0,$$

subject to intermediate growth conditions. Our goal is to show that its limit has distribution given by (4). We assume therefore that the distribution G of the activity periods U has a regularly varying tail given by (1). It is enough to consider $c = 1$. Thus, suppose $\lambda L(a)/a^{\gamma-1} \to 1$ as λ and a tend to infinity. It follows that V with distribution G_{eq} is such that

$$\lambda EU P(V > au) \to u^{-\gamma-1}/(\gamma - 1). \tag{7}$$

Standard calculations using the representations in (2) yield

$$\ln E \exp \theta W_{\mathrm{eq},\lambda}(t) = \lambda EU \int_0^t \theta e^{\theta u} P(V > u)\, \mathrm{d}u + \lambda \int_0^t \int_0^s \theta e^{\theta u} P(U > u)\, \mathrm{d}u\mathrm{d}s.$$

Subtract the expected value, simplify and perform an integration by parts to obtain

$$\ln E \exp \theta (W_{\mathrm{eq},\lambda}(t) - \lambda EU t) = \lambda EU \int_0^t \int_0^s \theta^2 e^{\theta u} P(V > u)\, \mathrm{d}u\mathrm{d}s.$$

Hence, in view of (7),

$$\ln E \exp \theta Y_3^{(\lambda)}(t) \to \frac{1}{\gamma - 1} \int_0^t \int_0^s \theta^2 e^{\theta u} u^{-(\gamma-1)}\, \mathrm{d}u\mathrm{d}s,$$

which is (4) for $n = 1$.

We will show now that the finite-dimensional distributions converge. In addition to the stationary process $W_{\mathrm{eq},\lambda}(t)$ introduced in (2) we need a separate notation $W_\lambda(t) = \sum_{j=1}^{N_t^\lambda}(t - S_j) \wedge U_j$ for the non-stationary flow. Some reflection shows that

$$W_\lambda(t) = \sum_{j=1}^{N_s^\lambda}(t - S_j) \wedge U_j + \widetilde{W}_\lambda(t - s), \quad s < t,$$

where the two terms on the right side are independent and $\widetilde{W}_\lambda$ is an independent copy of W_λ. Thus

$$\begin{aligned}\ln E \exp\Big\{ \sum_{i=1}^{n} \theta_i \Delta W_\lambda(t_i)\Big\} &= \lambda t_1 (g(\theta_{1,n}, t_{1,n}) - 1) \\ &+ \ln E \exp\Big\{ \sum_{i=2}^{n} \theta_i \big(W_\lambda(t_i - t_1) - W_\lambda(t_{i-1} - t_1)\big)\Big\}, \end{aligned} \tag{8}$$

where

$$g(\theta_{1,n}, t_{1,n}) = E \exp\Big\{ \sum_{i=1}^{n} \theta_i ((t_i - S) \wedge U - (t_{i-1} - S)_+ \wedge U)\Big\}$$

and S is uniformly distributed on $[0, t_1]$, and independent of U. Evaluate the expectations over S and U to get

$$\begin{aligned} g(\theta_{1,n}, t_{1,n}) &= 1 + \frac{1}{t_1} \int_0^{t_1} \Big\{ \int_0^{t_1 - s} \theta_1 e^{\theta_1 u} P(U > u)\, \mathrm{d}u \\ &+ \sum_{j=2}^{n} \exp\Big\{\theta_1 (t_1 - s) + \sum_{i=2}^{j-1} \theta_i \Delta t_i\Big\} \int_{t_{j-1} - s}^{t_j - s} \theta_j e^{\theta_j (u - t_{j-1} + s)} P(U > u)\, \mathrm{d}u\Big\} \mathrm{d}s. \end{aligned}$$

Next we observe

$$\begin{aligned} &\ln E \exp\Big\{ \sum_{i=1}^{n} \theta_i \Delta W_{\mathrm{eq},\lambda}(t_i)\Big\} - \ln E \exp\Big\{ \sum_{i=1}^{n} \theta_i \Delta W_\lambda(t_i)\Big\} \\ &= \lambda E U E\Big(\exp\Big\{ \sum_{i=1}^{n} \theta_i (t_i \wedge V - t_{i-1} \wedge V)\Big\} - 1\Big) \\ &= \lambda E U \sum_{j=1}^{n} \exp\Big\{ \sum_{i=1}^{j-1} \theta_i \Delta t_i\Big\} \int_{t_{j-1}}^{t_j} \theta_j e^{\theta_j (v - t_{j-1})} P(V > v)\, \mathrm{d}v. \end{aligned}$$

In combination with (8), we obtain

$$
\begin{aligned}
&\ln E\exp\Big\{\sum_{i=1}^{n}\theta_i \Delta W_{\mathrm{eq},\lambda}(t_i)\Big\}\\
&=\lambda EU\sum_{j=1}^{n}\exp\Big\{\sum_{i=1}^{j-1}\theta_i\Delta t_i\Big\}\int_{t_{j-1}}^{t_j}\theta_j e^{\theta_j(v-t_{j-1})}P(V>v)\,\mathrm{d}v\\
&+\lambda\int_0^{t_1}\Big\{\int_0^{t_1-s}\theta_1 e^{\theta_1 u}P(U>u)\,\mathrm{d}u\\
&+\sum_{j=2}^{n}\exp\Big\{\theta_1(t_1-s)+\sum_{i=2}^{j-1}\theta_i\Delta t_i\Big\}\int_{t_{j-1}-s}^{t_j-s}\theta_j e^{\theta_j(u-t_{j-1}+s)}P(U>u)\,\mathrm{d}u\Big\}\mathrm{d}s\\
&+\ln E\exp\Big\{\sum_{i=2}^{n}\theta_i\big(W_\lambda(t_i-t_1)-W_\lambda(t_{i-1}-t_1)\big)\Big\},
\end{aligned}
$$

But if we rewrite the last term using relation (9) with shifted time points and $\theta_1=0$, i.e.

$$
\begin{aligned}
&\ln E\exp\Big\{\sum_{i=2}^{n}\theta_i(W_{\mathrm{eq},\lambda}(t_i-t_1)-W_{\mathrm{eq},\lambda}(t_{i-1}-t_1))\Big\}\\
&\quad=\ln E\exp\Big\{\sum_{i=2}^{n}\theta_i\big(W_\lambda(t_i-t_1)-W_\lambda(t_{i-1}-t_1)\big)\Big\}\\
&\qquad+\lambda EU\sum_{j=2}^{n}\exp\Big\{\sum_{i=2}^{j-1}\theta_i\Delta t_i\Big\}\int_{t_{j-1}-t_1}^{t_j-t_1}\theta_j e^{\theta_j(v-t_{j-1}+t_1)}P(V>v)\,\mathrm{d}v,
\end{aligned}
$$

the result is a recursive equation for the cumulant generating functions of $W_{\mathrm{eq},\lambda}(t)$, namely

$$
\begin{aligned}
&\ln E\exp\Big\{\sum_{i=1}^{n}\theta_i \Delta W_{\mathrm{eq},\lambda}(t_i)\Big\} \qquad (9)\\
&\quad=\ln E\exp\Big\{\sum_{i=2}^{n}\theta_i\big(W_{\mathrm{eq},\lambda}(t_i-t_1)-W_{\mathrm{eq},\lambda}(t_{i-1}-t_1)\big)\Big\}\\
&\qquad+\lambda EU\int_0^{t_1}\theta_1 e^{\theta_1 v}P(V>v)\,\mathrm{d}v+\lambda\int_0^{t_1}\int_0^{t_1-s}\theta_1 e^{\theta_1 u}P(U>u)\,\mathrm{d}u\mathrm{d}s\\
&\qquad+\sum_{j=2}^{n}\theta_j\exp\Big\{\sum_{i=2}^{j-1}\theta_i\Delta t_i\Big\}\Big(I_1^j+I_2^j+I_3^j\Big),
\end{aligned}
$$

where

$$I_1^j = \lambda EU\, e^{\theta_1 t_1} \int_{t_{j-1}}^{t_j} e^{\theta_j (v - t_{j-1})} P(V > v)\, \mathrm{d}v$$

$$I_2^j = -\lambda EU \int_{t_{j-1}-t_1}^{t_j - t_1} e^{\theta_j (v - t_{j-1} + t_1)} P(V > v)\, \mathrm{d}v$$

$$I_3^j = \lambda \int_0^{t_1} e^{\theta_1 (t_1 - s)} \int_{t_{j-1}-s}^{t_j - s} e^{\theta_j (u - t_{j-1} + s)} P(U > u)\, \mathrm{d}u\, \mathrm{d}s.$$

are functions of t_1, t_{j-1} and t_j. Integration by parts show that

$$I_1^j + I_2^j + I_3^j = \lambda EU\theta_1 \int_0^{t_1} \int_0^{\Delta t_j} e^{\theta_1 s} e^{\theta_j u} P(V > u + t_{j-1} - t_1 + s)\, \mathrm{d}u\mathrm{d}s,$$

After centering and rescaling of (9) we arrive at

$$\begin{aligned}
\ln E \exp\Big\{ \sum_{i=1}^{n} \theta_i \Delta Y_3^{(\lambda)}(t_i) \Big\} &= \ln E \exp\Big\{ \sum_{i=2}^{n} \theta_i \Delta Y_3^{(\lambda)}(t_i - t_1) \Big\} \\
&+ \lambda EU \int_0^{t_1} \int_0^{s} \theta_1^2 e^{\theta_1 u} P(V > au)\, \mathrm{d}u\mathrm{d}s \\
&+ \lambda EU \sum_{j=2}^{n} \theta_1 \theta_j \exp\Big\{ \sum_{i=2}^{j-1} \theta_i \Delta t_i \Big\} \\
&\qquad \times \int_0^{t_1} \int_0^{t_j - t_{j-1}} e^{\theta_1 s} e^{\theta_j u} P(V > a(u + t_{j-1} - t_1 + s))\, \mathrm{d}u\mathrm{d}s.
\end{aligned}$$

In this representation, it follows from (7) that the integral terms on the right hand side have well-defined limits as $\lambda, a \to \infty$. Since the one-dimensional cumulant generating function converges, it follows by induction on n that in the asymptotic limit the cumulant generating functions of all finite-dimensional distributions of $Y_3^{(\lambda)}$ exist. We may therefore associate with the limit distribution a random process Y_γ and conclude that the finite-dimensional distributions of its increments must satisfy

$$\begin{aligned}
\ln E \exp\Big\{ \sum_{i=1}^{n} \theta_i \Delta Y_\gamma(t_i) \Big\} &= \ln E \exp\Big\{ \sum_{i=2}^{n} \theta_i \Delta Y_\gamma(t_i - t_1) \Big\} \\
&+ \frac{1}{\gamma - 1} \int_0^{t_1} \int_0^{s} \theta_1^2 e^{\theta_1 v} v^{-(\gamma-1)}\, \mathrm{d}v\mathrm{d}s \\
&+ \frac{1}{\gamma - 1} \sum_{j=2}^{n} \theta_1 \theta_j \exp\Big\{ \sum_{i=2}^{j-1} \theta_i \Delta t_i \Big\} \\
&\qquad \times \int_0^{t_1} \int_0^{\Delta t_j} e^{\theta_1 s} e^{\theta_j u} (u + t_{j-1} - t_1 + s)^{-(\gamma-1)}\, \mathrm{d}u\mathrm{d}s.
\end{aligned}$$

It is not difficult to see that the cumulant generating function (4) in Thm. 1 satisfies the above recursion. We can therefore identify the limit in distribution

of $Y_3^{(\lambda)}$ with the process Y_γ obtained in Thm. 1. On the other hand, the limit of $Y_3^{(\lambda)}$ has an integral representation as given in Thm. 3. Making the proper modification of the Poisson compensator we end up with the represention in Thm. 4.

References

1. Çağlar M (2004) A long-range dependent workload model for packet data traffic. Mathematics of Operations Research 29:1, 92-105
2. Gaigalas R (2004) A non-Gaussian limit process with long-range dependence. Uppsala Dissertations in Mathematics 33, Uppsala University
3. Gaigalas R, Kaj I (2003) Convergence of scaled renewal processes and a packet arrival model. Bernoulli 9, 671-703
4. Gunnarsson N, Kaj I (2005) Dynamics of the interference process in a spatial communication system. In preparation, Uppsala University
5. van Harn K, Steutel FW (2001) Stationarity of delayed subordinators. Stochastic Models 17(3), 369-374
6. Kaj I (2003) The long-memory infinite source Poisson model: scaling limit and representation. Preprint Uppsala University (outdated)
7. Kaj I, Leskelä L, Norros I, Schmidt V (2005) Scaling limits for random fields with long range dependence. Institut Mittag-Leffler Preprint Series 2004:f24
8. Kaj I, Martin-Löf A (2004) Scaling limit results for the sum of many inverse Lévy subordinators. U.U.D.M. 2004:13, Department of Mathematics, Uppsala University
9. Kaj I, Maulik K (2005) Manuscript in preparation, Uppsala University
10. Kaj I, Taqqu MS (2004) Convergence to fractional Brownian motion and to the Telecom process: the integral representation approach. U.U.D.M 2004:16, Uppsala University
11. Kallenberg O (2002) Foundations of Modern Probability. 2nd ed Springer-Verlag, New York
12. Kurtz TG (1996) Limit theorems for workload input. In: Kelly F.P, Zachary S, Ziedins I (eds) Stochastic Networks, theory and applications. Clarendon Press Oxford U.K.
13. Levy JB, Taqqu MS (2000) Renewal reward processes with heavy-tailed inter-renewal times and heavy-tailed rewards. Bernoulli 6:1, 23-44
14. Maulik K, Resnick S, Rootzn H (2003) A network traffic model with random transmission rate. Adv. Appl. Probab. 39 671-699
15. Mikosch Th, Resnick S, Rootzén H, Stegeman A (2002) Is network traffic approximated by stable Lévy motion or fractional Brownian motion? Ann. Appl. Probab. 12, 23-68
16. Pipiras V, Taqqu MS, Levy JB (2004) Slow, fast and arbitrary growth conditions for renewal reward processes when the renewals and the rewards are heavy-tailed. Bernoulli 10, 121-163
17. Taqqu MS (2002) The modeling of Ethernet data and of signals that are heavy-tailed with infinite variance. Scand. J. Stat. 29, 273-295
18. Taqqu MS, Willinger W, Sherman R (1997) Proof of a fundamental result in self-similar traffic modeling. Comput. Commun. Review 27:2, 5-23

19. Willinger W, Paxson V, Riedi RH, Taqqu MS (2003) Long-range dependence and data network traffic. In: Doukhan P Oppenheim G, Taqqu MS (eds) Theory and Applications of Long-Range Dependence, Birkhuser, Basel

A non-parametric test for self-similarity and stationarity in network traffic

Owen Dafydd Jones[1] and Yuan Shen[2]

[1] Department of Mathematics and Statistics, University of Melbourne
o.d.jones@ms.unimelb.edu.au
[2] Department of Statistics, University of Warwick
shen@stats.warwick.ac.uk

Summary. We develop a non-parametric statistical test for self-similarity based on the crossing tree and use simulation experiments to test its performance. It is applied to a number of packet traces both to determine the range of scales over which they appear self-similar and to detect temporal changes in the mean packet arrival rate and/or scaling behaviour.

1 Introduction

During the last decade packet traces collected from both Local Area Networks (LAN) and Wide Area Networks (WAN) have been extensively analysed in the framework of self-similar processes [14]. Not only is there ample empirical evidence supporting the existence of self-similarity in network traffic data [8, 9], but also mathematical models have been established which suggest physical explanations for observed self-similarity [12, 13]. Of course in practice self-similarity can only hold over a finite range of scales, whence the need for a statistically founded test for determining the scaling range. A further difficulty that arises when modelling packet traces is distinguishing long-range correlation from trends in the mean packet arrival rate.

In what follows we develop a non-parametric statistical test for self-similarity and use it both to determine scales over which we have a constant scaling regime and to determine time intervals wherein the intensity process appears to be stationary. The method is then applied to some publicly available LAN and WAN traces. These traces all exhibited self-similarity over a range of around 5 seconds to 5 minutes. The change in scaling below 5 seconds is most likely due to the effect of the network protocols used to transmit packets. Above 5 minutes we may be seeing the effect of finite network capacity, which puts an upper limit on the traffic intensity. Some of the traces we looked at also showed changes in the mean packet arrival rate and/or scaling

behaviour over time. The length of stationary periods varied from 5 minutes to 120 minutes (the longest trace we considered).

1.1 The Crossing Tree

Our approach uses the crossing tree, previously introduced in [7]. Let X be a self-similar process with Hurst index H, so that $X(t) \stackrel{\mathrm{d}}{=} a^{-H}X(at)$ for all $a > 0$. We will also assume that $X(0) = 0$ and that X has stationary and ergodic increments. We fix a base scale δ then let T_k^n be the start time of the k-th crossing of size $\delta 2^n$. That is for $n \geq 0$, $T_0^n = 0$ and $T_{k+1}^n = \inf\{t > T_k^n : X(t) \in \delta 2^n \mathbb{Z}, X(t) \neq X(T_k^n)\}$. There is a natural tree structure to the crossings, as each crossing of size $\delta 2^n$ can be decomposed into a sequence of crossings of size $\delta 2^{n-1}$. The nodes of the crossing tree are crossings and the offspring of any given crossing are the corresponding set of subcrossings at the level below. An example of a crossing tree is given in Figure 1. Note that we do not include the first apparent crossing at each level, from T_0^n to T_1^n. This is because for X non-Markov the path from T_0^n to T_1^n is not a true crossing. Thus the k-th crossing at level n is that from $X(T_k^n)$ to $X(T_{k+1}^n)$.

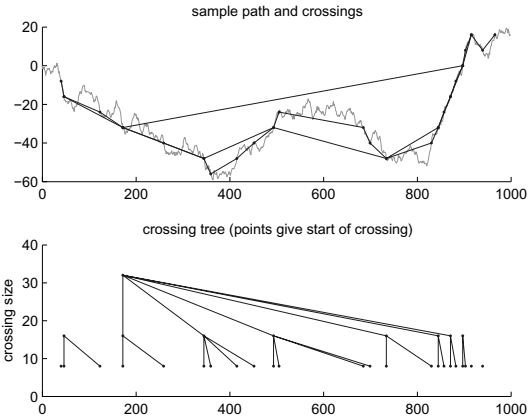


Fig. 1. Formation of the crossing tree from a sample path. The upper figure shows the sample path of a stochastic process and superimposed upon it three approximations made up of crossings of size 8, 16 and 32 respectively. The lower figure illustrates the tree structure of the crossings. The x-axis gives the starting point of a crossing and the y-axis its size. Crossings are linked if one is a subcrossing of the other.

Let Z_k^n be the number of subcrossings of size $\delta 2^{n-1}$ that make up the k-th crossing of size $\delta 2^n$ and let $W_k^n = T_{k+1}^n - T_k^n$ be the duration of the k-th size

$\delta 2^n$ crossing. If we let $N(n)$ be the total number of crossings of size $\delta 2^n$ then the Z_k^n must satisfy $N(n) \geq \sum_{k=1}^{N(n+1)} Z_k^{n+1}$.

It can be shown that if X is self-similar and cadlag then the sequences $Z^n = \{Z_k^n\}_{k=1}^{\infty}$ are identically distributed and the sequences $2^{-n/H}W^n = \{2^{-n/H}W_k^n\}_{k=1}^{\infty}$ are identically distributed [5]. We make the following additional assumption.

A1 The sequences Z^n and W^n are stationary and ergodic with finite means.

It can be shown that the Z^n and W^n are stationary and ergodic if X is continuous with stationary and ergodic increments [5]. We also have

Theorem 1. *Given Assumption A1*

$$H = \log 2 / \log \mu \text{ where } \mu = \mathbb{E} Z_k^n.$$

A proof (of a slightly more general result) is given in the Appendix.

We form scale-specific estimators for H. Let $\hat{\mu}_n = \sum_{k=1}^{N(n)} Z_k^n / N(n)$, then our estimator for the Hurst index at scale $\delta 2^n$ (alternatively at level n) is $\hat{H}_n = \log 2 / \log \hat{\mu}_n$. Clearly $\hat{\mu}_n$ is an unbiased and, given Assumption A1, consistent estimator of μ. See [7] for a discussion of how to estimate the variance of $\hat{\mu}_n$, which can be used to correct for the transform bias in $\hat{H}_n$ and to give a confidence interval. We call $\hat{H}_n$ the EBP-estimator, for Embedded Branching Process.

More generally we can form estimators for μ and thus H by averaging Z_k^n over a range of scales and/or restricted to a given time period.

1.2 The Integrated Packet Arrival Process

A network packet trace consists of a sequence of packet sizes (in bytes) and time-stamps (in seconds) at which the packets arrive. Let r be the network's maximum transmission rate (in bytes per second), then a packet of size x which arrives at time t is received over the time interval $[t, t + x/r]$. Let A be the 0-1 process given by $A(t) = 1$ if a packet is being received at time t and $A(t) = 0$ otherwise. Let ν be the average byte arrival rate then we define the integrated packet arrival process to be the continuous piecewise-linear process given by $\tilde{X}(t) = \int_0^t (A(s)r - \nu)\, ds$. When packets are arriving the process increases at rate $r - \nu$ and otherwise decreases at rate ν. If we take increments of this process, for example $\tilde{Y}(k) = \tilde{X}(kh) - \tilde{X}((k-1)h)$ for some h (typically around 10 milliseconds), then we obtain the usual traffic intensity process (scaled to have mean 0). In Figure 2 we plot $\tilde{X}$ and $\tilde{Y}$ for the first second of the Bellcore LAN packet trace pAug89 [6].

2 A Test for Self-Similarity

Let X be a continuous process and for each n let Z^n be the sequence of family sizes $\{Z_1^n, Z_2^n, \ldots\}$. Suppose that for each n the sequence $Z^n = \{Z_1^n, Z_2^n, \ldots\}$

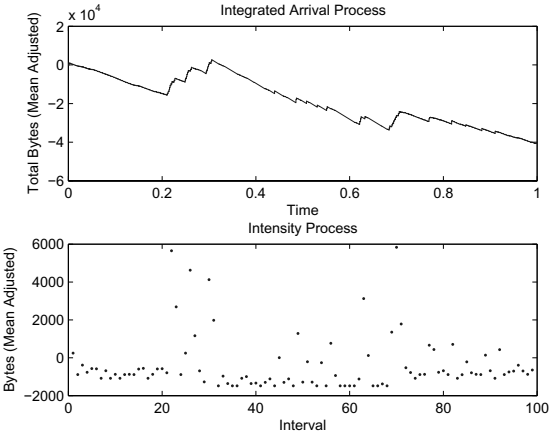


Fig. 2. Plot of the integrated packet arrival and intensity process for the first second of the Bellcore trace pAug89 [6]. The intensity process uses intervals of size 0.01 seconds.

is stationary and ergodic. We consider first the problem of determining those scales over which X is self-similar, that is we want to find levels $p < q$ such that

$$Z^p \stackrel{\mathrm{d}}{=} Z^{p+1} \stackrel{\mathrm{d}}{=} \cdots \stackrel{\mathrm{d}}{=} Z^q.$$

Here $\stackrel{\mathrm{d}}{=}$ indicates equality for all finite dimensional distributions. We will say that such a process is self-similar over scales $\delta 2^p$ to $\delta 2^q$ or equivalently self-similar over levels p to q.

Let $p^n(x) = \mathbb{P}(Z_k^n = x)$ and put $\mathbf{p}^n = \{p^n(2), p^n(4), \ldots\}$. Also let $\hat{\mathbf{p}}^n$ be the empirical distribution of Z_k^n obtained from Z^n. As Z^n is stationary and ergodic, $\hat{p}^n(x)$ is a consistent estimator of $p^n(x)$. To test for self-similarity we could use any test for the equality of distributions applied to $\hat{\mathbf{p}}^p, \ldots, \hat{\mathbf{p}}^q$. We chose to use a contingency table (see for example [4]). Take h bins $\{2\}, \{4\}, \ldots, \{2h-2\}, \{2h, 2h+2, \ldots\}$, let $\hat{p}^n[k]$ be the frequency Z^n falls in bin k, and let $\hat{p}[k]$ be the frequency the combined sequences $Z^p \cup \cdots \cup Z^q$ fall into bin k. Then the test statistic used is

$$T^{p,q} = \sum_{n=p}^{q} \sum_{k=1}^{h} N(n) \frac{(\hat{p}^n[k] - \hat{p}[k])^2}{\hat{p}[k]}. \tag{1}$$

We reject the hypothesis of self-similarity over the levels p to q if $T^{p,q}$ is too large.

An approximate critical region for $T^{p,q}$ can be obtained by assuming that the Z_k^n are all independent, in which case $T^{p,q}$ is asymptotically chi-squared distributed, with $(q-p-1)(h-1)$ degrees of freedom. (As a rule of thumb, this asymptotic approximation is reasonable provided $N(n)\hat{p}[k] \geq 5$ for all n and

k.) Theoretically we know that the Z_k^n are not independent: if nothing else they are constrained by the structure of the tree to satisfy $N(n) \geq \sum_{k=1}^{N(n+1)} Z_k^{n+1}$. However in practice we found that this approximation worked reasonably well.

For a fixed n we can test for dependence within the sequence Z^n directly. We used both the runs test and the gaps test (see for example [11]) and in none of the examples we considered did we find significant evidence of dependence. None-the-less we do believe there is dependence present, but that it is too small to be statistically significant in these samples. To show that dependence will be present in general we used repeated simulations of fractional Brownian motion (FBM) to form accurate estimates of the autocorrelation of Z^n. This was shown to be significantly different from 0, though still very small. For example for FBM with $H = 0.8$ we estimated $\mathrm{Corr}\,(Z_k^n, Z_{k+1}^n) = 0.04$ and for FBM with $H = 0.6$ we estimated $\mathrm{Corr}\,(Z_k^n, Z_{k+1}^n) = 0.01$. For $H = 0.5$ (classical Brownian motion) it can be shown that the Z_k^n are independent.

We can assess the effect of the dependence between the sequences $Z^p, \ldots, Z^q$ using bootstrapping. As above let $\hat{\mathbf{p}}$ be the subcrossing size distribution obtained from the combined sequences. Given $\hat{\mathbf{p}}$ and $N(p)$ we first generate an i.i.d. sequence $\bar{Z}^p = \{\bar{Z}_1^p, \ldots, \bar{Z}_{N(p)}^p\}$ from the distribution $\hat{\mathbf{p}}$. We then generate an i.i.d. sequence $\bar{Z}^{p+1} = \{\bar{Z}_1^{p+1}, \ldots, \bar{Z}_{\bar{N}(p+1)}^{p+1}\}$ from $\hat{\mathbf{p}}$, where $\bar{N}(p+1)$ is the largest k such that $\sum_{j=1}^{k} \bar{Z}_j^{p+1} \leq N(p)$. We continue in this fashion: for each $n < q$, given $\bar{N}(n)$ we generate a sequence $\bar{Z}^{n+1} = \{\bar{Z}_1^{n+1}, \ldots, \bar{Z}_{\bar{N}(n+1)}^{n+1}\}$ from $\hat{\mathbf{p}}$, where $\bar{N}(n+1)$ is the largest k such that $\sum_{j=1}^{k} \bar{Z}_j^{n+1} \leq \bar{N}(n)$. The sequences $\bar{Z}^p, \ldots, \bar{Z}^q$ come from a crossing tree for which both $\mathbf{p}^n = \hat{\mathbf{p}}$ for all n and, within each level, the subcrossing family sizes are independent of each other. In fact the crossing tree is the tree of a Galton-Watson branching process. Let $\bar{T}^{p,q}$ be the test statistic obtained from the bootstrap sample. Repeated sampling allows us to estimate percentage points of the distribution of $\bar{T}^{p,q}$; comparing these with the approximations given by the $\chi^2_{(q-p-1)(h-1)}$ distribution we see that for reasonable sample sizes the dependence introduced by the tree structure is negligible.

Note that the bootstrapped tree as described does not take account of the effect of deleting the false crossings from T_0^n to T_1^n. It is straightforward to do this, just a bit fiddly, and in any case for reasonable sample sizes the effect of doing so is negligible.

In summary we reject the hypothesis of self-similarity across scales $\delta 2^p$ to $\delta 2^q$, at the $100\alpha\%$ level, if the observed value of $T^{p,q}$ is larger than the $\chi^2_{(q-p-1)(h-1)}(\alpha)$ percentage point.

2.1 Testing for Stationarity

Suppose that X is self-similar over scales $\delta 2^m$ to $\delta 2^p$, as in the previous section. We will also suppose that $\mu := \mathbb{E}Z_k^n$, $m \leq n \leq p$, and $f(t) := \mathbb{E}X(t)$ are finite. Let $H = \log 2/\log \mu$ and $g(t) = f'(t)$. We will consider the problem of determining changes in H and/or g over time.

We consider X as a model for the integrated packet trace $\tilde{X}$. If the mean arrival rate ν changes by ϵ say, the effect is to subtract a trend ϵt from $\tilde{X}$. For the model X this corresponds to a change of $-\epsilon$ in g.

Let $\mathbf{p}$ be the common distribution of Z_k^n for $m \leq n \leq p$. Clearly an increase in H results in a decrease in the mean μ of $\mathbf{p}$. An increase in $|g|$ also results in a decrease in μ. Accordingly to detect changes in H or g over time we look for changes in the distribution $\mathbf{p}$ of Z_k^n as k varies.

We describe our stationarity test in the context of a fixed sample, however it can easily be adapted to dynamic sampling.

Our approach is to partition the observation period into intervals, form a crossing tree for each interval, then compare the empirical subcrossing family distribution of each interval using a contingency table. Intervals could be chosen of fixed length, but we prefer an adaptive approach. Fix a level q then take as our time intervals the level q crossings, that is the k-th interval is $[T_k^q, T_{k+1}^q)$. In the dynamic setting, if there is a sudden change in g then the length of the intervals will reduce allowing a more rapid detection of the change.

Let $\hat{\mathbf{p}}_i^{m,p}$ be the empirical distribution of Z_k^n obtained from levels m to p of the crossing tree within the i-th interval, and let $\hat{\mathbf{p}}^{m,p}$ be the empirical distribution obtained by combining all the intervals, that is, obtained from $Z^m \cup \cdots \cup Z^p$. Let $N(m,p;i)$ be the total number of crossings on levels m through to p in interval i. In practice we found that best results were obtained when $N(m,p;i)$ was at least 150 for all i. Adjacent intervals can be amalgamated if necessary to achieve this.

Take h bins $\{2\}, \{4\}, \ldots, \{2h-2\}, \{2h, 2h+2, \ldots\}$ and let N be the number of intervals, then the test statistic used is

$$T_B^{m,p} = \sum_{i=1}^{N} \sum_{k=1}^{h} N(m,p;i) \frac{(\hat{p}_i^{m,p}[k] - \hat{p}^{m,p}[k])^2}{\hat{p}^{m,p}[k]}. \tag{2}$$

We reject the hypothesis that $\mathbf{p}$ is constant if $T_B^{m,p}$ is too large. We again use a chi-squared approximation to the distribution of $T_B^{m,p}$, so that we reject the hypothesis of stationarity, at the $100\alpha\%$ level, if the observed value of $T_B^{m,p}$ is larger than the $\chi^2_{(N-1)(h-1)}(\alpha)$ percentage point.

If there is significant non-stationarity then we can use sequential testing to estimate the points at which $\mathbf{p}$ changes. There are a variety of ways of doing this; we used the following algorithm, which easily adapts to the dynamic sampling situation.

1. Start with $i_L(1) = 1$ and $i_R(1) = 2$ and $k = 1$.
2. Test intervals $i_L(k)$ to $i_R(k)$, if the test is not significant increase $i_R(k)$ by 1. If the test is significant then put $i_L(k+1) = i_R(k)$, $i_R(k+1) = i_R(k)+1$ and increase k by 1.
3. Repeat the previous step until $i_R(k) = N + 1$.

The algorithm arranges the intervals into groups with constant **p**. As we perform multiple tests, we need to adjust the level of significance used, to avoid false rejections. Using a Bonferroni adjustment (see for example [3] Ch. 9), given that we perform $N-1$ tests we reduce the significance level from α to $\alpha/(N-1)$.

Note that if we perform sequential testing using the integrated packet arrival trace $\tilde{X}$, we should at each step recalculate the mean arrival rate ν for the intervals $i_L(k)$ to $i_R(k)$, which necessitates recalculating the crossing tree for that period.

2.2 Simulation Experiments

We performed three sets of simulation experiments to test the ability of our methodology to detect changes in H and g. It proved to be moderately good at detecting changes in H but very good at detecting changes in g. However the false positive rate — detecting changes when none are present — was higher than desired.

For all of the following experiments we used fractional Brownian motion, simulated using the Wood-Chan algorithm [15]. The simulated traces are observations at fixed times $t = 0, 1, \ldots, 2^{21}$, with the variance of each increment $X(k+1) - X(k)$ set to be 1. We set our base scale δ to 1 and the smallest crossings we consider are those at level 5 (with subcrossings at level 4). This is necessary because observing the process at fixed times means we do not observe all the crossings at small scales. Time intervals were determined by the crossings at level $q = 10$. Observing crossings at levels $m = 5$ to $p = 10$ we get, for $H = 0.8$, on average approximately 136 observations of Z_k^n per interval.

The significance level α was set to 0.01 for all experiments. The choice of α gives a trade-off between the sensitivity of the method to changes in H and g, and the tendency to produce false positives.

Firstly to test the rate of false positives we applied our methodology to 10 simulated FBM traces of length 2^{21} with $H = 0.8$. In two 2 out of the 10 cases the test was significant, incorrectly indicating non-stationarity of the crossings. This indicates that our test is more sensitive than it should be. This is consistent with the subcrossing family sizes Z_k^n having positive correlation not accounted for by our chi-squared approximation of $T_B^{m,p}$.

Secondly, to test the ability of the methodology to detect a change in H we considered four types of composite process. Each consisted of a sequence of 0.8-FBM of length 2^{20} followed by a sequence of 0.6-FBM of length 2^{20}, $2^{20}/3$, $2^{20}/5$ or $2^{20}/10$ respectively, followed by a sequence of 0.8-FBM of length 0, $2^{20}2/3$, $2^{20}4/5$ or $2^{20}9/10$ respectively. 10 copies of each composite process were simulated and our methodology applied. The results are illustrated in Figure 3. Here for each group of intervals identified as stationary we calculated the H-estimate $\log 2/\log \hat{\mu}$, where $\hat{\mu}$ is the average subcrossing family size for that episode.

We see that the change in H is reasonably well detected in all but the last case, in which only 3 of the 10 samples produced a good result. In this case the time period with lower H was on average covered by less than 20 intervals.

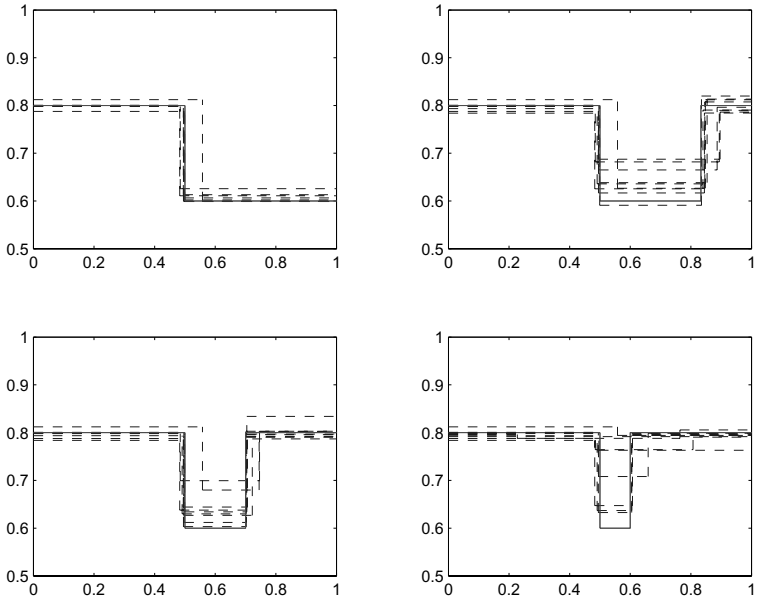


Fig. 3. Each panel gives time indexed estimates of the Hurst index H for 10 simulated time series with changing H. In each case the solid line gives the known H-value of the simulations and the dashed lines are the estimates. Note that the time axis on each panel is normalised to $[0, 1]$

Finally we tested the ability of the methodology to detect a change in g. To do this we again simulated FBM with $H = 0.8$, obtaining a trace of length 2^{21}, then added a small trend to an interior section. The length of the section with non-constant mean was $2^{21}/3$, $2^{21}/5$, $2^{21}/10$ or $2^{21}/50$. The original trace was scaled so that $\mathrm{Var}\,(X(t+1) - X(t)) = 1$, then the added trend was given a slope in the range 0.25 to 1.

In Figure 4 we illustrate four cases with $g = 1$ in an interior section of varying length. In each case we simulated 10 traces and estimated the points at which g changes. These are plotted together with the true change points in the figure; we see that they give a good match in all four cases. Overall we found that we could reliably detect the change in trend over periods as small as $2^{20}/50$, corresponding to approximately 4 level q intervals. For a period of this length, the test starts to find it difficult to detect the change in g when it is of magnitude 0.25.

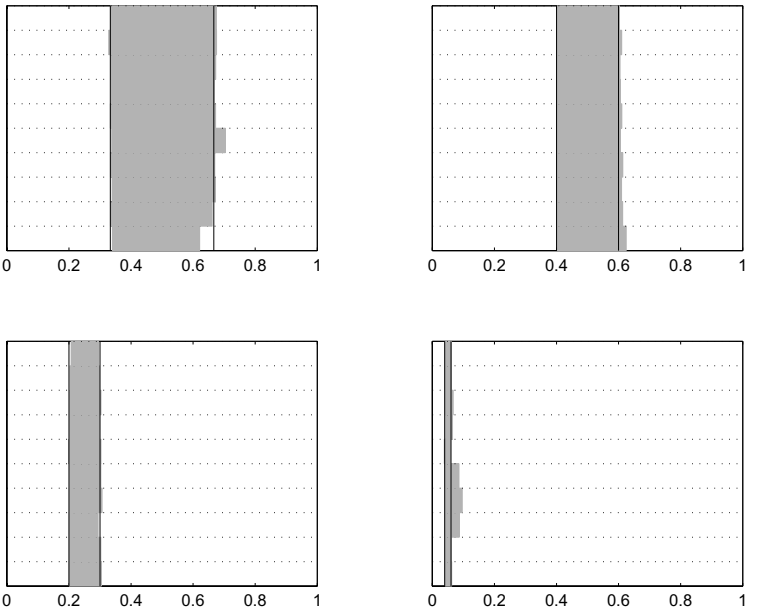


Fig. 4. Each panel gives estimates of the point at which the slope g changes for 10 simulated traces. The true change points are given by solid vertical lines. Note that the time axis on each panel is normalised to $[0, 1]$

3 Application to Packet Traces

We consider here two Bellcore LAN packet traces, pAug89 and pOct89, and three Lawrence TCP packet traces, lbl-pkt-3, lbl-pkt-4 and lbl-pkt-5. All five traces are publicly available from the Internet Traffic Archive [6]. The total duration and the mean byte arrival rate ν of each trace is given in Table 1.

Table 1. The duration T, mean byte arrival rate ν, upper and lower scaling levels and crossing times and estimated global H with 95% error bound, for 5 packet traces.

data set	T (min)	ν (bytes/μs)	$n_{\min}$	$\mathbb{E}W_k^{n_{\min}}$ (s)	$n_{\max}$	$\mathbb{E}W_k^{n_{\max}}$ (min)	H
pAug89	52.4	0.14	5	2.4	11	6.4	0.82 ± 0.01
pOct89	29.3	0.36	7	9.1	11	3.0	0.93 ± 0.03
lbl-pkt-3	120	0.03	6	13.0	10	4.9	0.89 ± 0.02
lbl-pkt-4	60	0.04	5	4.1	11	7.7	0.88 ± 0.03
lbl-pkt-5	60	0.03	6	15.6	10	6.8	0.85 ± 0.05

For each trace we formed the integrated packet arrival process, then for the crossing tree we fixed the base scale δ to be the root mean square of the height of the linear segments of the process.

Initially, we assume stationarity and examine each trace for a range of self-similar scaling. We begin by calculating for each level n the scale-specific H-estimate $\hat{H}_n = \log 2/\log \hat{\mu}_n$, these are plotted in Figure 5, with error bars calculated using the method of [7]. It is clearly seen that, for all 5 traces, $\hat{H}_n$ remains almost constant over scales from $2^5\delta$ to $2^{11}\delta$. We determined the precise self-similar region using the self-similarity test statistic (1) with significance level $\alpha = 0.05$ and $h = 3$. The upper and lower level of the self-similar region are given by $n_{\max}$ and $n_{\min}$ in Table 1, together with the average crossing times at these levels. We also give a global estimate of H given by $\log 2/\log \hat{\mu}_{m,n}$ where $\hat{\mu}_{m,n}$ is the average subcrossing family size calculated using $Z^m \cup \cdots \cup Z^n$ for $m = n_{\min}$ and $n = n_{\max}$.

We see that the Hurst index of all 5 packet traces is large, indicating strong self-similarity in these packet traces. In each case there is a clear dip in H at scales between $2^3\delta$ and $2^4\delta$, possibly caused by the action of routers on packet traffic. The rise in H at scales $2^1\delta$ and $2^2\delta$ can be attributed to long linear sections appearing in the observed integrated packet arrival trace.

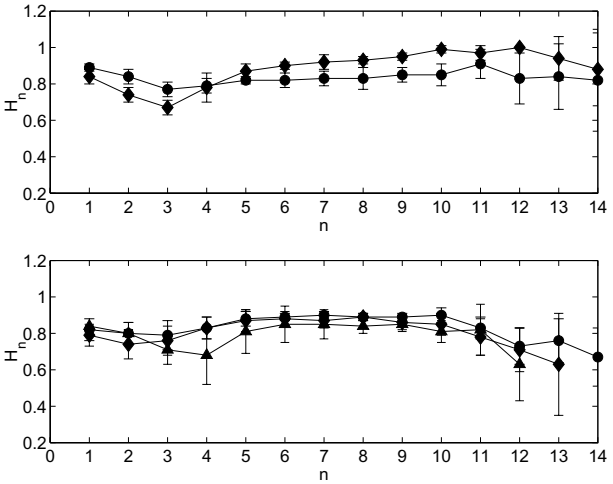


Fig. 5. Scale-specific Hurst index estimates $\hat{H}_n$ for different levels n for five network packet traces: pAug89 (circle) and pOct89 (diamond) in the upper panel and lbl-pkt-3 (circle), lbl-pkt-4 (diamond) and lbl-pkt-5 (triangle) in the lower panel. The error bars display estimated 95% confidence intervals for H.

Next we investigated possible changes in H and ν over time. Time intervals were determined using level q crossings, for q varying from 10 to 14. We used

the range of levels $m = 6$ to $p = 10$ to calculate the $\hat{\mathbf{p}}_i^{m,p}$, chosen to coincide with the self-similar region.

In Table 2 we give, for various values of q, the stationary episodes of each packet trace and the average length of the intervals used. The stationary episodes were determined using the test statistic (2) with an overall significance level of $\alpha = 0.01$ and $h = 3$. For each episode we give an estimate of the Hurst index based on that episode alone, with error bars calculated using the method of [7]. We note that there is in general a good correspondence between the episodes calculated using intervals of different resolution. In general we prefer to use smaller intervals as this gives a finer resolution. For reference the intensity process for each trace is plotted in Figure 6.

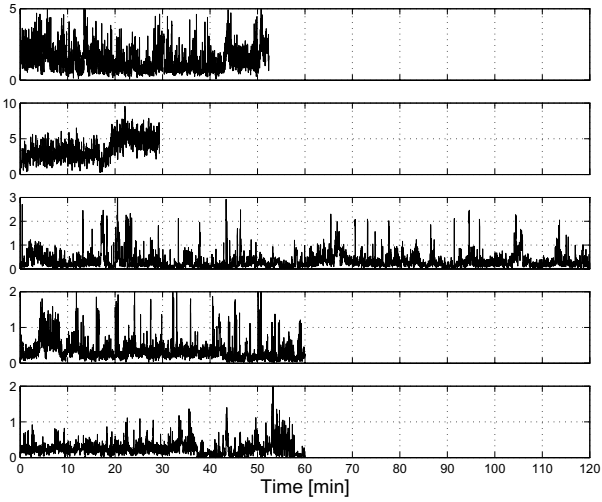


Fig. 6. The intensity process for the traces pAug89, pOct89, lbl-pkt-3, lbl-pkt-4 and lbl-pkt-5.

The TCP packet traces produced no surprises. Trace lbl-pkt-3 is shown to be time homogenous, while for the other two TCP traces changes in H and ν were detected, consistently for different choices of the interval resolution.

The packet trace pAug89 is shown to have constant H and ν, as has been reported by other authors [2]. The other LAN packet trace pOct89 is decomposed into two regions with different H-values. However, the precise episodes identified and the H estimates obtained for each episode depend on the size of intervals used. Let $T_1 = 14.5$, $T_2 = 20.0$ and $T = 29.3$ minutes. For $q = 10, 11, 12$ we get (roughly) episodes $[0, T_2)$ and $[T_2, T)$ and for $q = 13, 14$ we get episodes $[0, T_1)$ and $[T_1, T)$. Here there are really three episodes, $[0, T_1)$, $[T_1, T_2)$ and $[T_2, T)$, but our choice of α has resulted in a test not sensitive enough to pick up the short episode $[T_1, T_2)$. The arrival rate ν is steady in

Table 2. Time homogeneous episodes and their Hurst indices for each of 5 packet traces. Homogeneity is checked at different crossing scales, and therefore at different time resolutions, indicated by the mean crossing time.

data set	interval level q	mean crossing time [min]	end points of episodes [min]	H
pAug89	14	19.1	(0.0, 52.4]	0.82 ± 0.02
	13	9.5	(0.0 52.4]	0.82 ± 0.02
	12	7.3	(0.0 49.9]	0.82 ± 0.02
	11	6.4	(0.0, 50.8]	0.82 ± 0.02
	10	5.1	(0.0, 51.0]	0.82 ± 0.02
pOct89	14	15.3	(0.0, 14.5]	0.77 ± 0.05
			(14.5, 29.3]	0.87 ± 0.03
	13	9.2	(0.0, 14.5]	0.77 ± 0.05
			(14.5, 29.3]	0.87 ± 0.03
	12	3.5	(0.0, 20.9]	0.81 ± 0.03
			(20.9, 28.3]	0.74 ± 0.03
	11	3.0	(0.0, 20.9]	0.81 ± 0.03
			(20.9, 29.3]	0.74 ± 0.03
	10	2.9	(0.0, 20.0]	0.81 ± 0.03
			(20.0, 28.7]	0.74 ± 0.03
lbl-pkt-3	14	50.8	(0.0, 120.0]	0.90 ± 0.02
	13	14.3	(0.0, 118.1]	0.90 ± 0.02
	12	8.2	(0.0, 118.1]	0.90 ± 0.02
	11	5.8	(0.0, 120.0]	0.90 ± 0.02
	10	4.9	(0.0, 120.0]	0.90 ± 0.02
lbl-pkt-4	13	17.8	(0.0, 49.9]	0.89 ± 0.02
			(49.9, 60.0]	0.84 ± 0.03
	12	16.5	(0.0, 56.0]	0.89 ± 0.03
			(56.0, 60.0]	0.79 ± 0.14
	11	7.7	(0.0, 54.3]	0.90 ± 0.02
			(54.3, 60.0]	0.81 ± 0.10
	10	6.0	(0.0, 54.3]	0.90 ± 0.02
			(54.3, 60.0]	0.81 ± 0.10
lbl-pkt-5	12	12.0	(0.0, 35.7]	0.80 ± 0.03
			(35.7, 60.0]	0.89 ± 0.03
	11	5.9	(0.0, 33.9]	0.80 ± 0.04
			(33.9, 57.4]	0.87 ± 0.03
	10	6.8	(0.0, 33.9]	0.80 ± 0.04
			(33.9, 57.4]	0.87 ± 0.03

the first and last episode, but rapidly increasing during the middle episode. If we split the trace into these three episodes and then calculate the H estimate for each episode we get estimates of 0.77, 0.89 and 0.74. The EBP estimator of H is sensitive to trends, so if you amalgamate the middle episode with either the first or last episode, the effect is to introduce a trend and thus increase the H estimate for the amalgamated region. We can interpret this as a trend mimicking long-range correlation and leading to an over-estimation of the true Hurst index.

We also note that the H values in the episodes $[0, T_1)$ and $[T_2, T)$ are comparable even though the mean arrival rate is markedly higher in the episode

$[T_2, T)$. That is, a change of network condition such as an increase in the mean arrival rate does not necessarily imply a change of H value.

3.1 Comparison with Other Estimators

The H estimates obtained above using the Embedded Branching Process (EBP) estimator were compared to those obtained using two other commonly used methods: Detrended Fluctuation Analysis (DFA) [10] and Wavelet Analysis (WAV) [1]. For each trace except pOct89 we applied each estimator to the different episodes identified using level 10 intervals. For pOct89 we used the episodes $[0, T_1)$, $[T_1, T_2)$ and $[T_2, T)$. The results are given in Table 3.

In general there is not a great deal of agreement between any of the estimators, in a number of cases with differences of up to 0.1, and there is no readily discernable pattern to the differences.

Unlike the WAV estimator, both the EBP and DFA estimates are sensitive to local linear trends. Possible evidence for this is given by comparing the global H estimates with the restricted H estimates in the example pOct89, in which case the EBP and DFA estimates for the whole trace appear to be inflated. However we would also expect to see this in traces lbl-pkt-4 and lbl-pkt-5 and we do not. A quickly changing trend probably also explains why the EBP and DFA estimates are both notably higher than the WAV estimate in the second episode of the pOct89 trace. However this can not explain why both are notably lower than the WAV estimate in the first and third episodes of that trace.

A possible cause for some of the observed discrepancies is our use of a piecewise constant model for the mean arrival rate.

Table 3. The EBP, DFA, and WAV Hurst estimates for the stationary episodes of the 5 packet traces.

data	method	whole trace	first episode	second episode	third episode
pAug89	EBP	0.82	–	–	–
	DFA	0.83	–	–	–
	WAV	0.81	–	–	–
pOct89	EBP	0.93	0.77	0.89	0.74
	DFA	0.94	0.78	0.98	0.76
	WAV	0.87	0.85	0.83	0.84
lbl-pkt-3	EBP	0.89	–	–	–
	DFA	0.88	–	–	–
	WAV	0.96	–	–	–
lbl-pkt-4	EBP	0.88	0.90	0.81	–
	DFA	0.84	0.84	0.71	–
	WAV	0.93	0.94	0.84	–
lbl-pkt-5	EBP	0.85	0.80	0.87	–
	DFA	0.85	0.77	0.87	–
	WAV	0.84	0.83	0.83	–

4 Conclusions

In the examples considered the subcrossing family sizes Z_k^n had no discernable dependencies, so the test statistics (1) and (2) were well approximated by chi-squared distributions, making them very easy to implement. The resulting tests are objective and quantitative and can be readily automated. Their sensitivity is easily controlled by the choice of significance level α; we found values in the range 0.01 to 0.05 worked well.

In the examples considered we were able to identify the range of scales over which they were self-similar and identify the time-scales at which changes in the Hurst index H and the mean arrival rate ν can be seen. These scales were of the same order for all of the examples.

While the crossing-tree based test was sensitive to changes in the mean arrival rate ν, the crossing-tree based EBP-estimator is reliant on this having been done successfully, unlike the wavelet based estimator.

5 Appendix

The following theorem is a generalisation of Theorem 1, in that does not assume a priori that Z_k^n has a finite mean.

Theorem 2. *Suppose that* $\{W_k^1\}$, $\{W_k^0\}$ *and* $\{Z_k^0\}$ *are stationary and ergodic and that* $\mathbb{E}W_k^1 = a < \infty$ *and* $\mathbb{E}W_k^0 = b < \infty$, *then* $\mu := \mathbb{E}Z_1^0 < \infty$.

Moreover, if $W_k^1 \stackrel{\mathrm{d}}{=} 2^{1/H} W_k^0$ *then* $H = \log 2 / \log \mu$.

Proof Consider the first n crossings of size 2. Let $n_j = |\{k : 1 \le k \le n, Z_k^1 = j\}|$. The crossings of size 1 can be grouped into families based on which size 2 crossing they lie in. Let $W^0_{\sigma(j,k)}$ be the k-th size 1 crossing that comes from a size j family, then, for $N = Z_1^1 + \cdots + Z_n^1$,

$$\frac{\sum_{k=1}^n W_k^1}{n} = \frac{\sum_{k=1}^N W_k^0}{n} = \sum_j \sum_{k=1}^{jn_j} \frac{W^0_{\sigma(j,k)}}{jn_j} \frac{jn_j}{n}. \tag{3}$$

Since $\{W_k^0\}$ and $\{W_k^1\}$ are ergodic, we have $\sum_{k=1}^n W_k^1/n \to a$ and $\sum_{k=1}^{jn_j} W^0_{\sigma(j,k)}/jn_j \to b$ a.s. as $n \to \infty$. Also, since $\{Z_k^1\}$ is ergodic, $n_j/n \to p_j := \mathbb{P}(Z_1^1 = j)$ a.s. as $n \to \infty$. Now if $\mu = \infty$ then $\sum jp_j = \infty$, which implies that the RHS of (3) must diverge, a contradiction. Thus $\mu < \infty$ as required.

If $W_k^1 \stackrel{\mathrm{d}}{=} 2^{1/H} W_k^0$ then $a = 2^{1/H} b$. We expect the RHS of (3) to converge to $b\mu$, which would give $\mu = 2^{1/H}$ since the LHS converges to a. However because $\sum_{k=1}^{jn_j} W^0_{\sigma(j,k)}/jn_j \to b$ at a different rate for each j, we can not conclude this directly.

For any $\epsilon > 0$ let n_ϵ be such that for $n \ge n_\epsilon$, $n(\mu - \epsilon) \le \sum_{k=1}^n Z_k^1 \le n(\mu + \epsilon)$ with probability $\ge 1 - \epsilon$. Then, with probability $1 - \epsilon$,

$$\sum_{k=1}^{\lfloor n(\mu-\epsilon)\rfloor} W_k^0 \le T_n^1 \le \sum_{k=1}^{\lceil n(\mu+\epsilon)\rceil} W_k^0.$$

Dividing by n and sending $n \to \infty$ gives $(\mu-\epsilon)b \le 2^{1/H}b \le (\mu+\epsilon)b$ with probability $1-\epsilon$. Dividing by b and sending $\epsilon \to 0$ gives $\mu = 2^{1/H}$ as required. □

We also give a more checkable condition for $\mathbb{E}W_k^n < \infty$. The heavier the tails of the marginal distributions of X the smaller W_k^n will be, so intuitively requiring W_k^n to have finite mean is quite mild.

Lemma 1. *Let* $Y = \sup_{0\le s\le 1}|X(s)|$, *then* $\mathbb{E}(W_1^0)^\alpha < \infty$ *iff* $\mathbb{E}Y^{-\alpha/H} < \infty$, *for all* $\alpha > 0$.

Proof Note that with probability 1, $T_1^0 < W_1^0 < T_2^0$. Now consider

$$\begin{aligned}(T_1^0)^\alpha > t &\iff \sup_{0\le s\le t^{1/\alpha}} |X(s)| < 1\\ &\iff \sup_{0\le s\le t^{1/\alpha}} t^{H/\alpha}|X(s/t^{1/\alpha})| < 1\\ &\iff Y < t^{-H/\alpha}.\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}(T_1^0)^\alpha &= \int_0^\infty \mathbb{P}((T_1^0)^\alpha > t)\,dt\\ &= \int_0^\infty \mathbb{P}(Y^{-\alpha/H} > t)\,dt \;=\; \mathbb{E}Y^{-\alpha/H},\end{aligned}$$

So $\mathbb{E}Y^{-\alpha/H} = \infty \Rightarrow \mathbb{E}(W_1^0)^\alpha = \infty$. Similarly we have

$$\mathbb{P}(T_2^0 > t^{1/\alpha}) \le \mathbb{P}(\sup_{0\le s\le t^{1/\alpha}} |X(s)| < 2) = \mathbb{P}(Y < 2t^{-H/\alpha}).$$

Thus $\mathbb{E}(W_1^0)^\alpha \le 2^{\alpha/H}\mathbb{E}Y^{-\alpha/H}$. So $\mathbb{E}Y^{-\alpha/H} < \infty \Rightarrow \mathbb{E}(W_1^0)^\alpha < \infty$. □

References

1. P. Abry, P. Gonçalvès and P. Flandrin Wavelets, spectrum estimation and $1/f$ processes, in "Wavelets and Statistics, Lecture Notes in Statistics" (A. Antoniadis and G. Oppenheim, eds.), Springer Verlag, New York (1995), 15–30.
2. P. Abry and D. Veitch, Wavelet analysis of long-range-dependent traffic, IEEE Transactions on Information Theory **44** (1998), 2–15.
3. D.G. Altman, Practical Statistics for Medical Research. Chapman and Hall, London (1991).
4. J.E. Freund, Mathematical Statistics (Fifth Edition), Prentice-Hall, New Jersey (1992).

5. B.M. Hambly, O.D. Jones and Y. Shen, The crossing tree of a continuous self-similar process, In preparation.
6. Association for Computing Machinery, Special Interest Group on Data Communications, The Internet Traffic Archive, http://ita.ee.lbl.gov/, Accessed 1 Dec 2004.
7. O.D. Jones and Y. Shen, Estimating the Hurst index of a self-similar process via the crossing tree, Signal Processing Letters **11** (2004), 416–419.
8. W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, On the self-similar nature of ethernet traffic (extended version), IEEE/ACM Transactions on Networking **2** (1994), 1–15.
9. V. Paxson and S. Floyd, Wide area traffic: the failure of Poisson modeling, EEE/ACM Transactions on Networking **3** (1995), 226–244.
10. C.K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley and A.L. Goldberger, Mosaic organization of DNA nucleotides, Physical Review E **49** (1994), 1685–1689.
11. B.D. Ripley, Stochastic Simulation, Wiley (1987).
12. W. Willinger, V. Paxson and M.S. Taqqu, Self-similarity and heavy tails: structural modeling of network traffic, in "A Practical Guide to Heavy Tails: Statistical Techniques and Applications" (R. Adler, R. Feldman and M.S. Taqqu, eds.), Birkhauser, Boston (1998), 27–53.
13. W. Willinger, M.S. Taqqu, R. Sherman and D.V. Wilson, Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level, IEEE/ACM Transactions on Networking **5** (1997), 71–86.
14. W. Willinger and M.S. Taqqu and A. Erramilli, A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks, in "Stochastic Networks"(F. P. Kelly, S. Zachary and I. Ziedins, eds.), Oxford University Press, Oxford (1996), 339–366.
15. A.T.A. Wood and G. Chan, Simulation of stationary Gaussian processes in $[0,1]^d$, Journal of Computational and Graphical Statistics **3** (1994), 409–432.

6

IMAGE PROCESSING

Continuous evolution of functions and measures toward fixed points of contraction mappings

Jerry L. Bona[1] and Edward R. Vrscay[2]

[1] Department of Mathematics, Statistics and Computer Science, The University of Illinois at Chicago, Chicago, Illinois, USA 60607-7045
bona@math.uic.edu

[2] Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
ervrscay@uwaterloo.ca

Summary. Let T be a contraction mapping on an appropriate Banach space $B(X)$. Then the evolution equation $y_t = Ty - y$ can be used to produce a continuous evolution $y(x,t)$ from an arbitrary initial condition $y_0 \in B(X)$ to the fixed point $\bar{y} \in B(X)$ of T. This simple observation is applied in the context of iterated function systems (IFS). In particular, we consider (1) the Markov operator M (on a space of probability measures) associated with an N-map IFS with probabilities (IFSP) and (2) the fractal transform T (on functions in $L^1(X)$, for example) associated with an N-map IFS with greyscale maps (IFSM), which is generally used to perform fractal image coding. In all cases, the evolution equation takes the form of a nonlocal differential equation.

Such an evolution equation technique can also be applied to complex analytic mappings which are not strictly contractive but which possess invariant attractor sets. A few simple cases are discussed, including Newton's method in the complex plane.

1 Introduction

In this paper we introduce a class of evolution equations associated with contraction mappings on a Banach space of functions $B(X)$. The purpose is to produce a *continuous evolution* toward the fixed point $\bar{x}$ of a contraction map T, as opposed to the usual discrete sequence of iterates $x_n = T^n x_0$ that converge to $\bar{x}$.

The original motivation to devise such evolution equations arose from a desire to perform continuous, yet fractal-like, (i.e., spatially-contracted and greyscale distorted) "touch-up" operations on images. A fractal-based evolution method could be used to produce arbitarily small local alterations to

an image $u(x)$ in a neighbourhood of a point $x_0 \in X$ that depend upon the behaviour of u at other regions of X.

In addition to applying the idea to contractive fractal transform operators, we show briefly that such methods can also be applied to to other mappings, e.g., complex analytic mappings, that possess invariant attractor sets. While a few simple cases are examined here, a more thorough investigation will be reported in a future paper.

The structure of this paper is as follows. In Section 2, we introduce the class of evolution equations which take the form of differential equations involving functions or measures on an appropriate complete metric space (X, d). Section 3 reviews the basics of N-map Iterated Function Systems with probabilities (IFSP) and associated invariant measures. In Section 4 is constructed an evolution equation scheme so that time-dependent probability measures will evolve continuously (in time) to the invariant measure $\bar{\mu}$ of an IFSP. In Section 5, we consider IFS-type fractal transform operators on functions and illustrate the evolution scheme on a fractally-coded image. In Section 6, some simple applications of the evolution method to complex analytic dynamics are briefly considered, namely (i) iteration of quadratic maps and (ii) Newton's method in the complex plane.

2 Evolution equations associated with contractive mappings on Banach spaces

In all that follows, X will denote a closed and bounded subset of $\mathbf{R}^n$, $n = 1, 2, \cdots$, with d the Euclidean metric on X. Let $B(X)$ denote a Banach space of functions defined on X. Suppose that $T : B(X) \to B(X)$ is a contraction mapping, which is to say

$$\| Tu - Tv \| \ \leq \ c_T \ \| u - v \|, \quad \text{for all } u, v \in B(X), \tag{1}$$

where $c_T \in [0, 1)$. From Banach's fixed point theorem, there exists a unique $\bar{y} \in B(X)$ such that $\bar{y} = T\bar{y}$. Moreover, if, for any $y_0 \in B(X)$, we define the iteration sequence $y_{n+1} = Ty_n$, $n = 0, 1, \cdots$, then $\| y_n - \bar{y} \| \to 0$ as $n \to \infty$.

The goal is to produce a continuous evolution from the initial condition y_0 to the fixed point $\bar{y}$. In other words, from the initial condition y_0, we aim to find a function $y(t)$ that converges to $\bar{y}$ as $t \to \infty$, i.e., $\| y(t) - \bar{y} \| \to 0$ as $t \to \infty$. Consider the evolution equation

$$\frac{\partial y}{\partial t} = Ty - y, \tag{2}$$

where $y = y(x, t)$, $x \in X$, $t \geq 0$. Clearly the fixed point function $\bar{y}$ is a solution of this equation – an *equilibrium solution* in the parlance of dynamical systems theory. One immediately wants to know whether this equilibrium solution

is globally asymptotically stable, i.e, whether all solutions $y(t)$ to Eq. (2) converge to $\bar{y}$.

Example: Let $Ty = \frac{1}{2}y + \frac{1}{2}$ for $y \in L^1([0,1])$, say. Then $\bar{y}(x) \equiv 1$ is the unique fixed point of T. In this simple case, where T does not involve any operations on x, the evolution of y is rather simple. For a given $y(x,0)$, the unique solution to Eq. (2) is

$$y(x,t) = [y(x,0) - 1]e^{-t/2} + 1, \tag{3}$$

and we see that $y(x,t) \to \bar{y}(x)$ for all $x \in [0,1]$.

Theorem 1. *Let $B(X)$ be a real Banach space and $T : B(X) \to B(X)$ a contraction map on $B(X)$ with fixed point function $\bar{y}$. Then for any initial value $y(x,0) = y_0(x) \in B(X)$, the solution $y(t)$ converges exponentially rapidly to $\bar{y}$ as $t \to \infty$.*

Remark 1. By a solution of Eq. (2), we mean a C^1-curve $y : [0,\infty) \to B(X)$ such that (2) is satisfied for each $t > 0$.

Proof. First, rewrite Eq. (2) as

$$\frac{\partial y}{\partial t} + y = Ty. \tag{4}$$

Duhamel's formula leads in the usual way to the equation

$$y(t) = y_0 e^{-t} + e^{-t} \int_0^t e^s (Ty)(s) ds. \tag{5}$$

In the special case $y_0 = \bar{y}$, we have $y(t) = \bar{y} = T\bar{y}$ and

$$\bar{y} = \bar{y} e^{-t} + e^{-t} \int_0^t e^s (T\bar{y})(s) ds. \tag{6}$$

Subtracting (6) from (5), one obtains

$$y(t) - \bar{y} \;=\; \big(y_0 - \bar{y}\big)e^{-t} + e^{-t} \int_0^t e^s \big(Ty(s) - T\bar{y}\big) ds. \tag{7}$$

By Minkowski's integral inequality, it follows that

$$\| \, y(t) - \bar{y} \, \| \;\leq\; \| \, y_0 - \bar{y} \, \| \, e^{-t} + e^{-t} \int_0^t e^s \, \| \, Ty(s) - T\bar{y} \, \| \, ds, \tag{8}$$

where the norm is that of $B(X)$. Contractivity of T gives

$$\| \, y(t) - \bar{y} \, \| \;\leq\; \| \, y_0 - \bar{y} \, \| \, e^{-t} + c_T e^{-t} \int_0^t e^s \, \| \, y(s) - \bar{y} \, \| \, ds, \tag{9}$$

Gronwall's lemma then implies that for $t > 0$,

$$\| \, y(t) - \bar{y} \, \| \;\leq\; \| \, y_0 - \bar{y} \, \| \, e^{(c_T - 1)t}. \tag{10}$$

Since $0 \leq c_T < 1$, it follows that $\| \, y(t) - \bar{y} \, \| \to 0$ as $t \to \infty$ and exponentially rapidly, in fact, and the theorem is proved.

2.1 Practicalities and the discretization of the evolution equation

Attention is now turned to a few practicalities of solving the evolution equation in (2). In the computations reported below, we have employed a very simple forward Euler scheme to compute $y(t)$. If $h > 0$ denotes the time step and y_0 the initial condition, then letting $y_n = y(nh)$, for $n = 1, 2, \cdots$, we have

$$y_{n+1} = y_n + (Ty_n - y_n)h, \quad n = 0, 1, 2, \cdots . \tag{11}$$

Note that for $h = 1$, the above scheme reduces to the usual iteration procedure $y_{n+1} = Ty_n$. In fact, the above recursion relation may be rewritten as

$$\begin{aligned} y_{n+1} &= Uy_n \\ &= hTy_n + (1-h)y_n, \quad n = 0, 1, 2, \cdots , \end{aligned} \tag{12}$$

which is a linear interpolation between y_n and Ty_n. If T is a contraction with Lipschitz factor $c_T \in [0, 1)$, then U has Lipschitz factor $c_U = 1 - h(1 - c_T)$. If $0 < h \leq 1$, then U is a contraction and the fixed point of U is $\bar{y}$, the fixed point of T. In Section 3, we shall apply this method to contractive fractal transforms (Iterated Function Systems with greyscale maps) on $L^1(X)$.

Note that the repeated application of the Euler operator U can be written as

$$U^n = \sum_{k=1}^{n} \binom{n}{k} (1-h)^{n-k}\, h^k\, T^k. \tag{13}$$

This can have some interesting consequences on the evolution of $y(t)$ to the fixed point $\bar{y}$ of T. For example, suppose that $h = 1/N$ where $N > 1$. Then the Euler scheme over a fixed time interval, say between $t = 0$ and $t = 1$, will involve higher iterations of the operator T than the simple discrete iteration process $y_1 = Ty_0$ corresponding to $h = 1$. In the same way, a higher-order scheme for the integration of the evolution equation (2) will also yield a more subtle approximation than does straightforward iteration.

Suppose that the starting function y_0 is "flat," for example, a constant function on X. Then in the case that T is a fractal transform operator which produces spatial contractions of a function (see Section 5), the resulting function $U^N y_0$ could have much more spatial detail than the iterate $u_1 = Tu_0$, although the higher-frequency components will probably have rather low amplitude. This clearly indicates that the C^1-curve $y(t)$ does not have to pass through, i.e., interpolate, the discrete sequence $y_n = T^n y_0$, $n \geq 0$. The prospect for enhanced spatial detail may become even more pronounced when higher order integration schemes, e.g., Runge Kutta methods, are used to approximate the differential equation. The points that arise from such considerations are currently being investigated.

In the next section, we introduce Iterated Function Systems in their original setting, namely, applied to probability measures on a compact set X.

3 Iterated Function Systems and invariant measures

The idea of defining an operator through the parallel action of a set of contraction maps can be traced back to a number of papers, for example, [16]. However, the use of such systems of maps to construct fractal sets and supporting measures was described independently by Hutchinson [11] and Barnsley and Demko [2]. The latter paper introduced the appelation "Iterated Function Systems".

In what follows, (X, d) denotes a compact metric "base space," typically $[0,1]^n$. Let $\mathbf{w} = \{w_1, \cdots, w_N\}$ be a set of contraction maps $w_i : X \to X$, to be referred to as an N-map IFS. Let $c_i \in [0,1)$ denote the contraction factors of the w_i and define $c = \max_{1 \le i \le N} c_i$. Note that $c \in [0,1)$.

Now let $\mathcal{H}(X)$ denote the set of nonempty compact subsets of X and h the Hausdorff metric. Then $(\mathcal{H}, h)$ is a complete metric space [7]. Associated with the IFS maps w_i is a set-valued mapping $\hat{\mathbf{w}} : \mathcal{H}(X) \to \mathcal{H}(X)$ the action of which is defined to be

$$\hat{\mathbf{w}}(S) = \bigcup_{i=1}^{N} w_i(S), \qquad S \in \mathcal{H}(X), \tag{14}$$

where $w_i(S) := \{w_i(x), x \in S\}$ is the image of S under w_i, $i = 1, 2, \cdots, N$.

It is a standard result that $\hat{\mathbf{w}}$ is a contraction mapping on $(\mathcal{H}(X), h)$ [11]; in fact,

$$h(\mathbf{w}(A), \mathbf{w}(B)) \le c\, h(A, B), \quad A, B \in \mathcal{H}(X). \tag{15}$$

Consequently, there exists a unique set $A \in \mathcal{H}(X)$, such that $\mathbf{w}(A) = A$, the so-called *attractor* of the IFS $\mathbf{w}$. Moreover, for any $S_0 \in \mathcal{H}(X)$, the sequence of sets $S_n \in \mathcal{H}(X)$ defined by $S_{n+1} = \hat{\mathbf{w}}(S_n)$ converges in Hausdorff metric to A.

Examples:

1. $X = [0,1]$, $N = 2$: $w_1(x) = \frac{1}{3}x$, $w_2(x) = \frac{1}{3}x + \frac{2}{3}$. Then the attractor A is the ternary Cantor set C on $[0,1]$.
2. $X = [0,1]$, $N = 2$: $w_1(x) = sx$, $w_2(x) = sx + (1-s)$ for $0 < s < 1$. Then $A = [0,1]$. If $s = 0$ then $A = \{0,1\}$.

Let $\mathcal{M}(X)$ denote the set of Borel probability measures on X and d_M the Monge-Kantorovich metric on this set:

$$d_M(\mu, \nu) = \sup_{f \in Lip_1(X, \mathbf{R})} \left[\int_X f(x) d\mu - \int_X f(x) d\nu \right], \tag{16}$$

where

$$Lip_1(X, \mathbf{R}) = \{f : X \to \mathbf{R} \mid |f(x_1) - f(x_2)| \le d(x_1, x_2),\ \forall x_1, x_2 \in X\}. \tag{17}$$

For $1 \le i \le N$, let $0 < p_i < 1$ be a partition of unity associated with the IFS maps w_i, so that $\sum_{i=1}^{N} p_i = 1$. Associated with the IFS with probabilities

(IFSP) $(\mathbf{w}, \mathbf{p})$ is the so-called *Markov operator*, $M : \mathcal{M}(X) \to \mathcal{M}(X)$, the action of which is

$$\nu(S) = (M\mu)(S) = \sum_{i=1}^{N} p_i \mu(w_i^{-1}(S)), \quad \forall S \in \mathcal{H}(X). \tag{18}$$

(Here, $w_i^{-1}(S) = \{y \in X \mid w_i(y) \in S\}$.) Then M is a contraction mapping on $(\mathcal{M}(X), d_M)$ [1, 7, 11]. Consequently, there exists a unique measure $\bar{\mu} \in \mathcal{M}(X)$, the so-called *invariant measure* of the IFSP $(\mathbf{w}, \mathbf{p})$, such that $\bar{\mu} = M\mu$. Moreover, for any $\mu_0 \in \mathcal{M}(X)$, the sequence of measures $\mu_n \in \mathcal{M}(X)$ defined by $\mu_{n+1} = M\mu_n$ converges in d_M-metric to $\bar{\mu}$.

Examples:

1. $X = [0, 1]$, $N = 2$: $w_1(x) = \frac{1}{3}x$, $w_2(x) = \frac{1}{3}x + \frac{2}{3}$, as in Example 1 above, with $p_1 = p_2 = \frac{1}{2}$. Then $\bar{\mu}$ is the Cantor-Lebesgue measure supported on the Cantor set $C \subset [0, 1]$.
2. $X = [0, 1]$, $N = 2$: $w_1(x) = \frac{1}{2}x$, $w_2(x) = \frac{1}{2}x + \frac{1}{2}$. The attractor of this IFS is $[0, 1]$. When $p_1 = p_2 = \frac{1}{2}$, the IFSP invariant measure is Lebesgue measure on $[0, 1]$.
3. $X = [0, 1]$, $N = 2$: $w_1(x) = \frac{1}{2}x$, $w_2(x) = \frac{1}{2}x + \frac{1}{2}$, with $p_1 = 0.4$, $p_2 = 0.6$. The IFSP invariant measure is pictured in Figure 1 (histogram approximation).

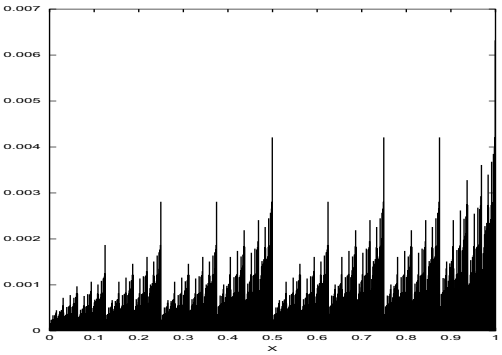


Fig. 1. Histogram approximation to the invariant measure for the IFSP in Example 3 above. (1000 bins on [0,1] were used.)

4 Continuous evolution of probability measures to fixed points of IFS Markov operators

As just mentioned, the iteration procedure,

$$\mu_{n+1} = M\mu_n, \quad n = 0, 1, 2, \cdots, \tag{19}$$

with initial condition $\mu_0 \in \mathcal{M}(X)$, produces a sequence of measures μ_i that converge in d_M-metric to the invariant measure $\bar{\mu} = M\bar{\mu}$ of the IFSP $(\mathbf{w}, \mathbf{p})$. We aim now to produce a continuous version of this evolution from μ_0 to $\bar{\mu}$. Consider measures $\mu \in \mathcal{M}(X)$ to be time-dependent, i.e., $\mu : X \times \mathbf{R} \to \mathbf{R}$. In what follows, we shall use the notations $\mu(S, t)$ and $\mu(t)$ interchangeably: the former specifically denotes the $\mu(t)$-measure of a Borel set $S \in X$.

The evolution of $\mu(S, t)$ is defined by the equation

$$\frac{\partial \mu}{\partial t} = M\mu - \mu. \tag{20}$$

This evolution equation for $\mu(S, t)$ may be interpreted as follows. Let $\mu_0 \in \mathcal{M}(X)$ be the initial value of this equation and let S be a Borel set in X. Then (20) at S is the ordinary differential equation

$$\begin{aligned} \frac{d\mu(S,t)}{dt} &= (M\mu)(S,t) - \mu(S,t) \\ &= \sum_{i=1}^{N} p_i \mu(w_i^{-1}(S), t) - \mu(S,t). \end{aligned} \tag{21}$$

Note that if $\mu_0 = \bar{\mu}$, the IFSP invariant measure, then $\mu(S, t) = \bar{\mu}(S)$ for all $t \geq 0$. It remains to show that if $\mu(0) \neq \bar{\mu}$, then $\mu(t) \to \bar{\mu}$ as $t \to \infty$ (in d_M-metric).

Eq. (21) may be integrated to yield

$$\mu(S,t) = \mu_0(S)e^{-t} + \sum_{i=1}^{N} p_i e^{-t} \int_0^t e^s \mu(w_i^{-1}(S), s) ds. \tag{22}$$

It is clear that $\mu(S, t)$ defined in (22) is a Borel measure on X. Evaluating at $S = X$ shows that $\mu(t)(X) = \mu(0)(X) = 1$. So $\mu(t) \in \mathcal{M}(X)$. Also note that if S has no preimages, i.e. $S \cap w_i(X) = \emptyset$ for $1 \leq i \leq N$, then $(M\mu)(S) = 0$ so that $\mu(S, t) = \mu_0(S)e^{-t} \to 0$ as $t \to \infty$.

Theorem 2. *Let $\bar{\mu}$ be the invariant measure of the IFSP $(\mathbf{w}, \mathbf{p})$ with associated Markov operator M, so that $M\bar{\mu} = \bar{\mu}$. Then all solutions $\mu(t)$ of Eq. (20) approach $\bar{\mu}$ exponentially as $t \to \infty$ in the d_M-metric.*

Proof. In the special case that $\mu(0) = \bar{\mu}$, it follows that the RHS of Eq. (20) is zero. This implies that $\mu(t) = \bar{\mu}$ for all $t \geq 0$. More generally, we proceed as in Section 1. Integrating Eq. (20), we obtain

$$\mu(t) = \mu(0)e^{-t} + e^{-t} \int_0^t e^s (M\mu)(s) ds. \tag{23}$$

When $\mu(0) = \bar{\mu}$, we have $\mu(t) = \bar{\mu}$ for all t and clearly,

$$\bar{\mu} = \bar{\mu}e^{-t} + e^{-t}\int_0^t e^s (M\bar{\mu})(s)ds. \tag{24}$$

Subtracting (24) from (23), multiplying the result by a function $f \in Lip_1(X, \mathbf{R})$ and integrating both sides of the resulting equation over X, there obtains

$$\begin{aligned}\left|\int_X f d\mu(t) - \int_X f d\bar{\mu}\right| \;&\leq\; e^{-t}\left|\int_X f d\mu(0) - \int_X f d\bar{\mu}\right| \\ &\quad + e^{-t}\int_0^t e^s \left|\int_X f d(M\mu(s)) - \int_X f d(M\bar{\mu})\right| ds. \end{aligned}\tag{25}$$

It follows immediately that

$$d_M(\mu(t), \bar{\mu}) \;\leq\; d_M(\mu(0), \bar{\mu})e^{-t} + e^{-t}\int_0^t e^s d_M(M\mu(s), M\bar{\mu})ds. \tag{26}$$

Contractivity of M on $\mathcal{M}(X)$ implies that

$$d_M(\mu(t), \bar{\mu}) \;\leq\; d_M(\mu(0), \bar{\mu})e^{-t} + ce^{-t}\int_0^t e^s d_M(\mu(s), \mu)ds. \tag{27}$$

As in Section 2, it follows that

$$d_M(\mu(t), \bar{\mu}) \;\leq\; d_M(\mu(0), \bar{\mu})e^{(c-1)t}, \qquad t > 0. \tag{28}$$

Since $c \in [0, 1)$, then necessarily $d_M(\mu(t), \bar{\mu}) \to 0$ as $t \to \infty$, once again exponentially rapidly, thus proving the theorem.

5 IFS operators on function spaces

IFS operators on function spaces $B(X)$ may be defined in a manner rather similar to that of measures [9, 10]. For the moment, we consider general function spaces $\mathcal{F}(X)$ supported on X. The essential components of a *fractal transform operator* are as follows.

1. A set of N one-to-one contraction maps $w_i : X \to X$; given a function $u \in \mathcal{F}(X)$, each map w_i produces a spatially-contracted copy of u, $u_i(x) = u(w_i^{-1}(x))$ supported on the subset $X_i = w_i(X)$. **Note:** In the case of functions, we demand that $\mathbf{w}(X) = \cup_{i=1}^N w_i(X) = X$ so that each point $x \in X$ has at least one preimage $w_k^{-1}(x)$.
2. The $u_i(x)$ are now modified by means of *greyscale maps* $\phi_i : \mathbf{R} \to \mathbf{R}$ that satisfy suitable conditions. Usually, they are assumed to be Lipschitz so that for each ϕ_i there exists a $K_i \geq 0$ such that

$$|\phi_i(t_1) - \phi_i(t_2)| \leq K_i|t_1 - t_2|, \quad \text{for all } t_1, t_2 \in \mathbf{R}. \tag{29}$$

The result is a set of spatially-contracted and greyscale-modified copies of u, namely, $g_i(x) = (\phi_i \circ u \circ w_i^{-1})(x)$. (The set of contraction maps w_i with associated greyscale maps ϕ_i is also referred to as an "Iterated Function System with greyscale maps" or an IFSM ($\mathbf{w}$, Φ) [9].)

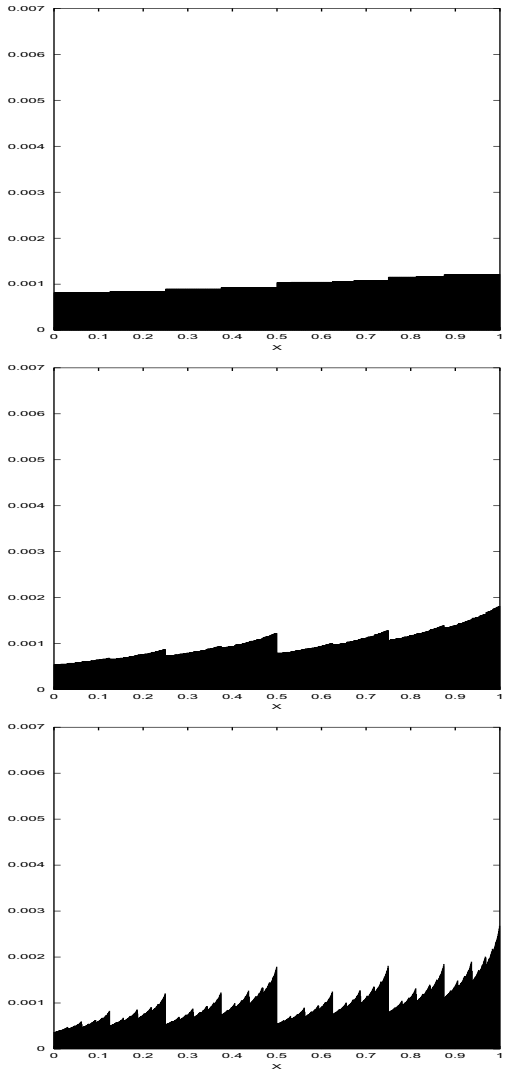


Fig. 2. Evolution of measures $\mu(t)$ toward the IFSP invariant measure of Figure 1, according to Eq. (20). Initial value is $\mu(0) = m$, Lebesgue measure on [0,1]. (a) $t = 1.0$, (b) $t = 3.0$, (c) $t = 5.0$. As in Figure 1, these are histogram approximations using 1000 bins on [0,1].

3. These *fractal components* $g_i(x)$ are then combined – with an operation that is suitable to the space in which we are working – to produce a new function $v \in \mathcal{F}(X)$. The natural operation to combine fractal components in L^p spaces is the summation operation.

The net result of all of these operations is summarized as $v = Tu$, where $T : \mathcal{F}(X) \to \mathcal{F}(X)$ is called the *fractal transform* operator, *viz.*

$$v(x) = (Tu)(x) = \sum_{i=1}^{N} \phi_i(u(w_i^{-1}))(x). \tag{30}$$

For a given $p \geq 1$ and with suitable conditions on the IFS map contraction factors c_i and the greyscale map Lipschitz constants K_i [9], the fractal transform T is contractive on $L^p(X)$ with fixed point function $\bar{u} \in L^p(X)$. The fixed point equation,

$$\bar{u}(x) = (T\bar{u})(x) = \sum_{i=1}^{N} \phi_i(\bar{u}(w_i^{-1}))(x), \tag{31}$$

indicates that $\bar{u}$ is "self-similar," i.e., that it can be written as a sum of spatially-contracted and greyscale-modified copies of itself. Moreover, for any $u_0 \in L^p(X)$, the sequence of iterates $u_{n+1} = Tu_n$, $n = 0, 1, 2, \cdots$, converges in L^p-metric to $\bar{u}$.

The continuous evolution equation analogous to Eq. (20) is then

$$\frac{\partial u(x,t)}{\partial t} = \sum_{i=1}^{N} \phi(u(w_i^{-1}(x))) - u(x), \quad x \in X. \tag{32}$$

In most applications, the greyscale maps are assumed to have an affine form, $\phi_i(t) = \alpha_i t + \beta_i$, so that the above equation becomes

$$\frac{\partial u(x,t)}{\partial t} = \sum_{i=1}^{N} \left[\alpha_i u(w_i^{-1}(x)) + \beta I_{X_i}(x)\right] - u(x), \quad x \in X, \tag{33}$$

where $I_A(s)$ denotes the characteristic function of a set $A \in X$, i.e., $I_A(x) = 1$ if $x \in A$ and zero otherwise.

Since the fractal transform T does not contain any differential operators, Eqs. (32) and (33) are ordinary differential equations in $u(x,t)$, involving only time derivatives. Nevertheless, because of the terms $w_i^{-1}(x)$, these DE's are *nonlocal* in that the time evolution of $u(x,t)$ is determined by values of u generally *not* at x. This can lead to rather complicated evolution.

This evolution scheme and its implementation are quite similar to that discussed for measures in the previous section. As such, we do not present any examples and proceed to consider block fractal transforms in the next section.

5.1 Fractal transform operators and block image coding

It is overambitious to expect that a general image, viewed as a function or measure, would be well approximated by a union of shrunken and distorted copies of itself. Following the idea of Jacquin [12], however, it has been found that an image u may be well approximated by a union of shrunken and distorted copies of *subsets* of itself. This has been the basis of block-based fractal image coding [3, 8, 13] which is reviewed very briefly below.

For simplicity, the support X of an image will be considered as either $[0,1]^2$ (continuous support) or an $n \times n$ pixel array (discrete support). Consider a partition of X into subblocks R_i with $X = \cup_i R_i$. The R_i are assumed to be "nonoverlapping," either intersecting only at common boundaries (continuous case) or not at all (pixel case). Associated with each "range block" R_i is a larger "domain" block $D_i \subset X$ so that $R_i = w_i(D_i)$ where w_i is a 1-1 contraction map. The image function $u|_{R_i}$ supported on each R_i is also found to be well approximated by a spatially-shrunken and greyscale-modified copy of $u|_{D_i}$:

$$u|_{R_i} \approx \phi_i \circ u|_{D_i} = \phi_i \circ u|_{D_i} \circ w_i^{-1}, \quad 1 \le i \le N. \tag{34}$$

Once again, the $\phi_i : \mathbf{R} \to \mathbf{R}$ are greyscale maps that are typically affine maps in practice.

Because the range blocks are nonoverlapping, we may write the above relation as

$$u(x) \approx (Tu)(x) = \sum_i \phi_i(u(w_i(x))), \quad x \in X. \tag{35}$$

The block fractal transform operator T has the same form as the operator in Eq. (30). It is a well-known result [1] that if the "collage distance" $\| u - Tu \|$ is small, then u is well approximated by the fixed point $\bar{u}$ of T. More precisely,

$$\| u - \bar{u} \| \ \le \ \frac{1}{1-c} \| u - Tu \|, \tag{36}$$

where c is the contraction factor of the fractal transform operator T. This allows the inverse problem of fractal image coding to be reformulated into a more tractable problem. Instead of searching for a T whose fixed point $\bar{u}$ is close to u, we search for a T that maps u close to itself.

In Figure 3 is presented the fixed point approximation $\bar{u}$ to the standard 512×512 Lena image (8 bits/pixel) using a partition of 8×8 nonoverlapping pixel blocks ($64^2 = 4096$ in total). The "domain pool" for each range block was the set of $32^2 = 1024$ 16×16 non-overlapping pixel blocks. (This is clearly not optimal.) This image was obtained by starting with the seed image $u_0(x) = 255$ (plain white image) and iterating $u_{n+1} = Tu_n$ to $n = 15$. Iterates u_1, u_2 and u_3 are also shown in this figure.

In Figure 5.1 we show the time evolution of images $u(x,t)$ as determined by $u_t = Tu - u$ where T is the fractal transform operator described above and used in Figure 3. In this case, the time evolution proceeds at a slower pace

Fig. 3. Starting at upper left and moving clockwise: The iterates u_1, u_2 and u_3 along with the fixed point $\bar{u}$ of the fractal transform operator T designed to approximate the standard 512×512 (8 bpp) "Lena" image. The "seed" image was $u_0(x) = 255$ (plain white). The fractal transform T was obtained by "collage coding" using 4096 8×8 nonoverlapping pixel range blocks. The domain pool consisted of the set of 1024 nonoverlapping 16×16 pixel blocks.

than in Figure 3, starting at $u_0 = 255$ (plain white image) and proceeding in time steps of 0.2 using a step size of $h = 0.1$ in the Euler method of Eq. (11).

6 Applications to complex analytic dynamics

Here we are concerned with complex iteration dynamics $z_{n+1} = R(z_n)$, where $z \in \mathbf{C}$ and $R : \mathbf{C} \to \mathbf{C}$ is a rational function of degree greater than or equal to two. The closure of the set of all repulsive k-cycles of R, $k \geq 1$, is the so-called *Julia set* J_R [4, 5].

For simplicity, the examples here will be taken from the well studied one-parameter family of quadratic maps $R(z) = z^2 + c$. When $c = 0$, the Julia set of $R(z) = z^2$ is the unit circle. The points $\bar{z} = 0$ and $\bar{z} = \infty$ are attractive fixed points of $R(z)$. The point $\bar{z} = 1$ is repulsive. The set J_R acts as a

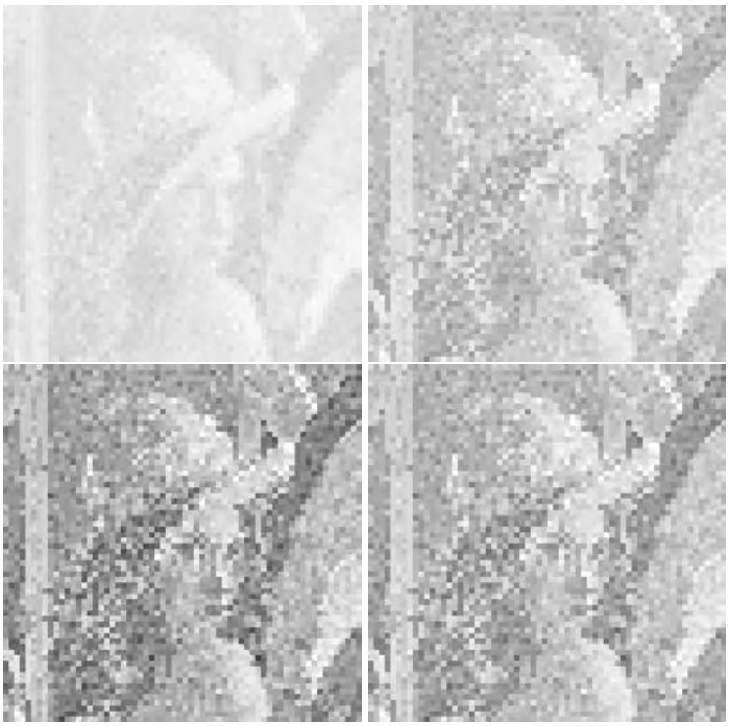


Fig. 4. Starting at upper left and moving clockwise: The images $u(x,t)$ at times 0.2, 0.4, 0.6, 0.8 produced from $u(x,0) = 255$ (plain white) under evolution by $u_t = Tu - u$ where T is the fractal transform whose discrete iteration was shown in Figure 3. Euler method, step-size $h = 0.1$.

boundary between the basins of attraction of $\bar{z} = 0$ and $\bar{z} = \infty$. As c decreases from 0, the Julia set becomes "crinkly" and the finite attractive fixed point moves leftward along the negative real axis. At $c = -3/4$, this fixed point becomes neutral. Further decrease of c produces an attractive two-cycle that becomes neutral at $c = -5/4$. This is the start of the famous period-doubling bifurcation associated with this iteration process. Much more can be written about the iteration dynamics, but we keep the discussion short here.

Following Eq. (2), the evolution equation associated with the quadratic map $R(z) = z^2 + c$ will be

$$\frac{dz}{dt} = z^2 + c - z, \tag{37}$$

where $z(t) \in \mathbf{C}$. If we let $z = x + iy$ and $c = c_1 + ic_2$, then Eq. (37) yields the system

$$\begin{aligned} \frac{dx}{dt} &= x^2 - y^2 - x + c_1, \\ \frac{dy}{dt} &= 2xy - y + c_2, \end{aligned} \tag{38}$$

of ordinary differential equations for $x(t)$ and $y(t)$. The results of the previous section do not apply globally since $R(z)$ is contractive only in neighbourhoods of its attractive fixed points. (In the quadratic case, $R(z)$ can have at most one finite attractive fixed point. Moreover, such finite attractive fixed points exist only for c values on a bounded set in $\mathbf{C}$ – the large principal cardioid region of the Mandelbrot set of $z^2 + c$.) Nevertheless, it is interesting to explore the dynamics of the system of ODEs in (38).

First consider the case $c = 0$. Simple linearization shows that the point $\bar{z} = 0$ is a locally asymptotically stable equilibrium solution and that $\bar{z} = 1$ is a locally unstable equilibrium solution of (38). Numerical experiments and analysis suggest that, in contrast to the discrete case (where the basin of attraction of $\bar{z} = 0$ is the open disc $|z| < 1$), the basin of attraction of $z = 0$ is the cut plane $\mathbf{C} \setminus [1, \infty)$. All points $(a, 0)$ with $a > 1$ travel on the positive real axis toward ∞.

For $c < 0$, linearization shows that the fixed point $\bar{z}_1 = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4c} < 0$ continues to be the asymptotically stable equilibrium point of (38) even when it is no longer attractive for the discrete iteration process, i.e., for $c \leq -\frac{3}{4}$. Numerical experiments show that its basin of attraction continues to be the cut complex plane. However, when c has a nonzero imaginary component, the cut disappears and all points, except the other fixed point $\bar{z}_2$, travel toward $\bar{z}_1$.

6.1 Newton's method in the complex plane

Recall that the Newton function associated with the complex valued function $f(z)$ is

$$N(z) = z - \frac{f(z)}{f'(z)}. \tag{39}$$

If $\bar{z}$ is a zero of $f(z)$, then it is also a fixed point of $N(z)$. In the discussion that follows, we assume that the zeros of $f(z)$ are simple. Then $N'(\bar{z}) = 0$, so that $\bar{z}$ is locally *superattractive*. For an initial seed $z_0 \in \mathbf{C}$ suitably close to $\bar{z}$, the iterates $z_{n+1} = N(z_n)$ converge quadratically to $\bar{z}$.

Consider the following evolution equation in the time-dependent complex variable $z(t)$ that corresponds to Eq. (2), namely,

$$\begin{aligned} \frac{dz}{dt} &= N(z) - z \\ &= -\frac{f(z)}{f'(z)}. \end{aligned} \tag{40}$$

Technically speaking, N is not necessarily a contraction mapping so that the results of Sections 2 to 5 do not apply globally. However, Eq. (40) can be integrated to give

$$f(z(t)) = f(z(0))e^{-t}. \tag{41}$$

Thus $f(z(t)) \to 0$ as $t \to \infty$. Assuming for simplicity that $f(z)$ is analytic, it follows that $z(t)$ tends to a zero of f. (There will, of course, be complications with the critical points of f.)

It is instructive to consider a couple of examples.

1. $f(z) = z^2 - 1$ with roots $\bar{z}_1 = 1$ and $\bar{z}_2 = -1$. In the classical Newton iteration method, the imaginary axis $I = N(I)$ serves as the boundary for the basins of attraction of the two roots, as originally shown by Cayley [6]. If we let $z = x + iy$, then Eq. (40) yields the system

$$\begin{aligned} \frac{dx}{dt} &= -\frac{x}{2} + \frac{1}{2}\frac{x}{x^2+y^2}, \\ \frac{dy}{dt} &= -\frac{y}{2} - \frac{1}{2}\frac{y}{x^2+y^2}, \end{aligned} \tag{42}$$

   of ODEs in $x(t)$ and $y(t)$. As expected, the y-axis is invariant (it also contains the critical point $z = 0$) and acts as a boundary for the basins of attraction of the two roots. If $x(0) = 0$, then $x(t) = 0$ for $t > 0$. If $x(0) > 0$, then $(x(t), y(t)) \to (1, 0)$ as $t \to \infty$. If $x(0) < 0$, then $(x(t), y(t)) \to (-1, 0)$ as $t \to \infty$.
2. $f(z) = z^3 - 1$ with roots $\bar{z}_1 = 1$ and $\bar{z}_{2,3} = -\frac{1}{2} \pm i\frac{1}{2}\sqrt{3}$. Here, the boundary between the basins of attraction for the Newton iteration method is a very complicated object – the Julia set of the Newton function. From the Julia-Fatou theory of iteration of rational functions, any open neighbourhood of a point z on the Julia set must contain subsets of *each* of the basins of attraction of the roots $\bar{z}_i$. The basins of attraction of the three roots are shown in Figure 6.1 (left).
   Once again letting $z = x + iy$, Eq. (40) yields the following system of ODEs in $x(t)$ and $y(t)$:

$$\begin{aligned} \frac{dx}{dt} &= -\frac{x}{3} + \frac{1}{3}\frac{x^2-y^2}{(x^2+y^2)^2}, \\ \frac{dy}{dt} &= -\frac{y}{3} - \frac{2}{3}\frac{xy}{(x^2+y^2)^2}. \end{aligned} \tag{43}$$

   The Julia set boundaries are eliminated in this scheme. The basin boundaries lie on the rays $\theta = \pi/3$, $\theta = \pi$ and $\theta = -\pi/3$ as shown in Figure 6.1 (right). (The critical point $z = 0$ is also included in this set of boundary points.)

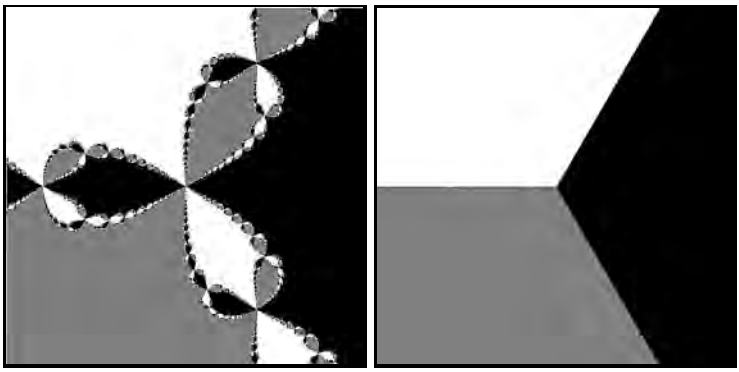


Fig. 5. Left: Basins of attraction of Newton's complex iteration method for the three cubic roots of unity. Right: Basins of attraction using the system of ODEs in Eq. (43). In both plots, the region of the complex plane being plotted is $-1 \leq Re(z) \leq 1$, $-1 \leq Im(z) \leq 1$.

7 Concluding Remarks

In this paper we have introduced a method to produce a continuous evolution of a quantity y to the fixed point $\bar{y}$ of an appropriate contraction mapping T. Such a continuous evolution replaces the usual discrete sequence of iterates $y_n = T^n y_0$ that converge to $\bar{y}$. The curve $y(t)$, however, does not generally interpolate the y_n. The evolution equation in Eq. (2) is not unique in that there are many possible ways of continuously "steering" the evolution of $y(t)$ to the fixed point $\bar{y}$ of the contractive operator T.

The method of Eq. (2) has been applied to two fundamental sets of contraction mappings associated with Iterated Function Systems: (1) the Markov operator M on probability measures associated to an IFS with probabilities (IFSP) and (2) the fractal transform operator T associated to IFS with greyscale maps (IFSM). We have also presented some preliminary results of applying the method to complex analytic dynamics: (1) the iteration of quadratic functions and (2) Newton's method in the complex plane.

As mentioned earlier, the original motivation to devise such a continuous evolution procedure arose from a desire to perform nonlocal, fractal-like (i.e., spatially-contracted and greyscale distorted) operations of a more continuous, i.e., "touch-up" nature. In other words, one could make arbitarily small alterations to an image function u in the neighbourhood of a point $x_0 \in X$ that would depend on values of $u(x)$ for x away from x_0.

In view of the vast research on partial differential equation methods in imaging, we envision combined fractal-based/PDE methods that could be used to produce small nonlocal alterations in the presence of diffusion-like processes. For example, one may wish to modify the evolution equation (20) by adding a small diffusion term, e.g.,

$$\frac{\partial y}{\partial t} - \epsilon \triangle y = Ty - y, \tag{44}$$

where ϵ could be either positive (diffusion) or negative (backward diffusion). Of course, such simple diffusion schemes are not generally employed in imaging. Instead, one considers "anisotropic diffusion" methods [14, 15].

Finally, the application of continuous evolution methods to complex analytic dynamics reported here is admittedly very preliminary and much remains to be explored.

Acknowledgements

We gratefully acknowledge support of this research in part by the National Science Foundation of the USA (JLB) and the Natural Sciences and Engineering Research Council of Canada (ERV).

References

1. M.F. Barnsley, *Fractals Everywhere*, Academic Press, New York (1988).
2. M.F. Barnsley and S. Demko, Iterated function systems and the global construction of fractals, *Proc. Roy. Soc. London* **A399**, 243-275 (1985).
3. M.F. Barnsley and L.P. Hurd, *Fractal Image Compression*, A.K. Peters, Wellesley, Massachusetts (1993).
4. P. Blanchard, Complex analytic dynamics on the Riemann sphere, *Bull. Amer. Math. Soc.* **11**, 85-141 (1984).
5. H. Brolin, Invariant sets under iteration of rational functions, *Ark. Math.* **6**, 103-144 (1966).
6. A. Cayley, Application of the Newton-Fourier method to an imaginary root of an equation. *Quart. J. Pure Appl. Math.* **16**, 179-185 (1879).
7. K. Falconer, *The Geometry of Fractal Sets*, Cambridge University Press, Cambridge (1985).
8. Y. Fisher, *Fractal Image Compression*, Springer Verlag, New York (1995).
9. B. Forte and E.R. Vrscay, Theory of generalized fractal transforms, in *Fractal Image Encoding and Analysis*, edited by Y. Fisher, NATO ASI Series F 159, Springer Verlag, New York (1998).
10. B. Forte and E.R. Vrscay, Inverse problem methods for generalized fractal transforms, in *Fractal Image Encoding and Analysis*, *ibid.*.
11. J. Hutchinson, Fractals and self-similarity, *Indiana Univ. J. Math.* **30**, 713-747 (1981).
12. A. Jacquin, *Image coding based on a fractal theory of iterated contractive image transformations*, *IEEE Trans. Image Proc.* **1** 18-30 (1992).
13. N. Lu, *Fractal Imaging*, Academic Press, New York (1997).
14. P. Perona and J. Malik, Scale-space and edge detection using anisotropic diffusion, *IEEE Trans. PAMI* **12**, 629-639 (1990).
15. G. Sapiro, *Geometric partial differential equations and image analysis*, Cambridge University Press, New York (2001).
16. R.F. Williams, Composition of contractions, *Bol. Soc. Brasil. Mat.* **2**, 55-59 (1971).

Various Mathematical Approaches to Extract Information from Textures of Increasing Complexities

Fahima Nekka[1] and Jun Li[2]

[1] Faculté de Pharmacie and Centre de Recherches Mathématiques, Université de Montréal C.P. 6128, Succ. Centre-ville, Montréal (Québec), Canada H3C 3J7
fahima.nekka@umontreal.ca

[2] Faculté de Pharmacie and Centre de Recherches Mathématiques, Université de Montréal C.P. 6128, Succ. Centre-ville, Montréal (Québec), Canada H3C 3J7
li@crm.umontreal.ca

Summary. Dealing with several structures of different complexities, we adopt various strategies to extract information. The general idea is to find appropriate tools to analyze the variation of the corresponding autocorrelation functions. First, for homogeneous media under different conditions, we recover, in a statistical way, a relationship between porosity and the autocorrelation function. Then, for low-complexity textures, we exploit this relationship to extract complementary parameters from the autocorrelation function beyond porosity using spectral analysis. For fractal-like structures, we process them according to their porosity. For fat fractals, usually used as synthetic models of porous media, we combine the regularization dimension, a method proposed to estimate the curve variation, with the autocorrelation function. This leads to a more robust classification. For fractals of negligible porosity, such as fractals of non-integer dimension, we discuss how the method HMSF we developed serves as an original means to estimate the Hausdorff dimension and how it can be exploited to give complementary characteristic parameters.

1 Introduction

New information and measurement technologies give access to signals of exceedingly complex nature, making it demanding for more sophisticated data analysis to efficiently extract information beyond the scope of the traditional methods. For example, the design of synthetic polymers has been revolutionized by the new achievements in high-resolution, broad-mass-range spectrometry [1]. Wave propagation and scattering through porous media and highly ramified materials lead to (spatial) signals which can be considered as defined on fractal systems [2,3]. As a matter of fact, the obtained mass spectra, which carry out microstructure information of great impact on the macroscopic physical properties, can be very complex. A central concern when processing such

complex data is to use tools that extract the maximum information with the least degeneracy. The autocorrelating process, expressed through the autocorrelation function is a classical mathematical method widely used in engineering and applied sciences [4]. It is a powerful process that accumulates and reorganizes intrinsic similarities hidden in a structure, allowing thus for a (pre-)processing of the signal as well as its analysis. Fractal methods already proved to be efficient to quantify complex information based on existing similarities. The major use of fractal analysis is to measure this complexity through fractal dimensions and other related indexes. However, inadequacy of traditional methods when processing complex systems and the known limitations of popular fractal methods led us to investigate, in parallel, these two different ways of information processing, and to combine them in order to create more powerful and less degenerate methods [6–9]. This resulted in an extension of classical methods, making them applicable within a nowadays context knowledge and in a benefit of fractal analysis from these classical and very popular methods [7].

This paper is organized as follows. In Section 2, we propose several statistical models of homogeneous media from which we recover the relationship between porosity and the autocorrelation function. In Section 3, we propose complementary parameters to porosity by exploiting various complexities of the autocorrelation functions of the signals. For low-complexity textures, we propose a measure of departure from homogeneity. For fractal-like textures exhibiting power law properties, usually used as synthetic porous media, we show how it is possible to classify them by combining the autocorrelation function with the regularization dimension [11]. We also consider the case of non-trivial structures having zero porosity (thin fractals). For these structures, we use what we previously called Hausdorff measure spectrum function (HMSF), which generalizes the classical autocorrelation function and proves to be more sensitive and suitable to describe this kind of scattered objects [6–8]. In section 4, we comment our proposed strategies and suggest some possibilities to pursue additional work in this direction.

2 A Statistical Relationship Between Porosity and the Autocorrelation Function

In signal processing, the autocorrelation function is often regarded as another aspect of a signal which facilitates its analysis. To characterize a porous structure, porosity is considered as one principal index reflecting its spatial occupancy. In this section, based on several statistical models of porous images, we discuss the relationship between their porosity and their corresponding autocorrelation function.

Assume that a signal $S(x)$ is of finite energy, we have the simplest form of the autocorrelation function:

$$r_S(t) = \int_{-\infty}^{\infty} S(x)S(x+t)dx. \tag{2.1}$$

This self crossing process sums all point-point similarities (by product) of the signal at distances t. Given the signal of an image set, the autocorrelation function is dedicated to measure the similarity in the set's geometric structure [4]. Though various forms of autocorrelation functions are widely used, this basic idea is always the same.

Study of porous media mainly involves porosity, which refers to the occupancy of the set, measured on samples having regular geometrical shapes. To facilitate porosity estimation, we generally select for the sampling, the familiar shapes such as cubes, cylinders, spheres, etc.

For a porous media of a large size, the concept of deterministic porosity is questionable. We can invoke two possible reasons: the boundary as well as the volume of the compartment where lives the porous media, generally having a complex geometrical form, are difficult to be determined or the measurement cannot be taken on the whole porous media. Thus, we have to consider the porosity in a statistical way. This can be done by a sampling process. The sampling is carried out in an arbitrary way and different values of porosity are obtained when different sampling sizes and positions are considered. Since each chosen sample is within a compartment of regular shape, it has a deterministic porosity value. Average of these values will tend to the true porosity of the whole porous media, guaranteed by the so-called large number theorem. In fact, we have two choices: to approximate the porosity of the whole porous structure by the porosity of the samples chosen large enough (which is generally difficult as explained above) or, alternatively, to take a large number of samples.

For a porous structure represented by a spatial signal, it turns out that porosity have a very close relationship to the autocorrelation function. This connection is not perceptible in a deterministic case (one sample), but will be clear when explained in the following statistical synthetic models.

2.1 A Homogeneous Model of a Single Component Porous Media with Point-to-point Independence

For the sake of simplicity, we consider the one-dimensional case. We also assume the homogeneous porous set F to be hosted in the unit interval and defined as follows:

At the finest resolution, the proposed porous media F can be considered to be composed of n equal parts $[(i-1)/n, i/n]$, $i = 1, \cdots, n$ of the unit interval, where each part is associated with a binary random variable X_i indicating the membership of the subinterval $[(i-1)/n, i/n]$. We can write $F = \bigcup_{i:X_i=1} [(i-1)/n, i/n]$ and X_i is defined as:

$$\mathbb{P}(X_i = 1) = p, \quad \mathbb{P}(X_i = 0) = 1 - p \tag{2.2}$$

where $\mathbb{P}$ denotes the probability.

Suppose that any part $[(i-1)/n, i/n]$ is chosen in a random way, we further assume that the variables X_i are i.i.d. for all $1 \leq i \leq n$.

Consider the normalized version $I(t)$ of the autocorrelation function of F:

$$I(t) = \frac{\int \mathcal{X}_F(x)\mathcal{X}_F(x+t)dx}{\int \mathcal{X}_F(x)^2 dx} \tag{2.3}$$

where $\mathcal{X}_F$ is the indicator function of F: $\mathcal{X}_F(x) = \sum_{i=1}^{n} X_i \mathcal{X}_{[(i-1)/n, i/n]}(x)$ (the value at each i/n is treated in a natural way, i.e. it can only be 0 or 1). The normalization is applied to assure that $I(0) = 1$, which will facilitate our comparisons below.

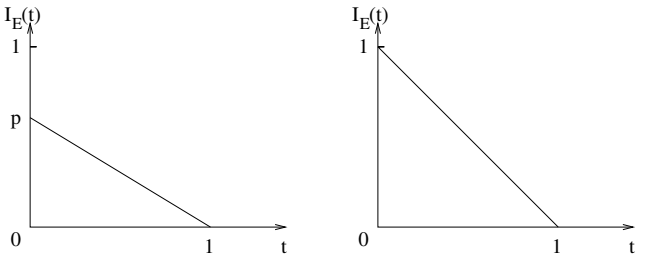


Fig. 1. The autocorrelation of: (a) the homogeneous model of single component porous media with point-to-point independence; (b) the unit segment.

In this example, we have

$$\int \mathcal{X}_F(x)^2 dx = \frac{\sum_{k=1}^{n} X_k^2}{n}$$

and

$$\int \mathcal{X}_F(x)\mathcal{X}_F(x+t)dx = \frac{\sum_{k=i}^{n} X_k X_{k-i+1}}{n}$$

for $t = (i-1)/n$, $i = 1, \cdots, n$.

Since we have assumed a statistical definition of F, it is suitable to redefine the autocorrelation function of F in the following way:

$$I_{\mathbb{E}}(t) = \frac{\mathbb{E}\int \mathcal{X}_F(x)\mathcal{X}_F(x+t)dx}{\mathbb{E}\int \mathcal{X}_F(x)^2 dx} \tag{2.4}$$

where $\mathbb{E}$ denotes the expectation.

Then, we have $\mathbb{E}\int \mathcal{X}_F(x)\mathcal{X}_F(x+t)dx = (1-\frac{i-1}{n})p^2$ and $\mathbb{E}\int \mathcal{X}_F(x)^2 dx = p$. On the other hand, variances of these two expressions involved in $I_{\mathbb{E}}(t)$, are

of order $1/n$, i.e. $O(1/n)$, which guarantees their convergence when n is large enough.

It is easy to see that:

$$I_{\mathbb{E}}(t) = \begin{cases} 1, & t = 0; \\ (1-(i-1)/n)p, & t = (i-1)/n, \quad i = 2, \cdots, n. \end{cases} \tag{2.5}$$

Let n go to infinity, we obtain:

$$I_{\mathbb{E}}(t) = \begin{cases} 1, & t = 0; \\ (1-t)p, & 0 < t < 1. \end{cases} \tag{2.6}$$

Fig. 1(a) illustrates $I_{\mathbb{E}}(t)$.

The porosity of F is $\frac{1}{n}\sum_{i=1}^{n} X_i$; statistically, it is: $\frac{1}{n}\mathbb{E}\sum_{i=1}^{n} X_i = p$.

Then p is obtained as the right limit of $I_{\mathbb{E}}(t)$ at $t = 0$ or the slope of the $I_{\mathbb{E}}(t)$. In this way, the porosity of a set F can be related to its autocorrelation function $I_{\mathbb{E}}(t)$ by considering an immediate neighborhood of $t = 0$ or root mean square slope of the autocorrelation function.

The simple case of the autocorrelation function of the segment $[0, 1]$, a limit case of Eq.(2.6), is illustrated in Fig. 1(b). In this case, there's no jump from $t = 0$ to $t = 0+$, meaning that the similarity is carried out by all the points of $[0, 1]$ and that $I_{\mathbb{E}}(t)$ smoothly decreases as t increases. However, for $p < 1$, from $t = 0$ to $t = 0+$, the similarity is carried out by only p percent of points of $[0, 1]$, which equals to the porosity of the set, Fig. 1(a).

2.2 A Homogeneous Model of Double Components Porous Media with Point-to-point Independence

To verify the relationship between porosity and the autocorrelation function for a more complex structure, we study another model of porous media composed of two components.

Consider F as composed from $2n$ equal parts $[\frac{i-1}{2n}, \frac{i}{2n}]$, $i = 1, \cdots, 2n$ of the unit interval, where each part is associated with a binary random variable X_i indicating the membership of the subinterval $[\frac{i-1}{2n}, \frac{i}{2n}]$. Clearly, we can write $F = \bigcup_{i:X_i=1} [\frac{i-1}{2n}, \frac{i}{2n}]$ and X_i is defined as:

$$\mathbb{P}(X_i = 1) = p_j, \quad \mathbb{P}(X_i = 0) = 1 - p_j, \tag{2.7}$$

where $j = 1$ for i even, $j = 2$ for i odd, and $\mathbb{P}$ denotes the probability.

We further assume that the variables X_i are independent between them, for all $1 \le i \le 2n$. Then we have $\mathbb{E}\int \mathcal{X}_F(x)^2 dx = (p_1 + p_2)/2$, which is in fact the statistical porosity of F, and we note $p = (p_1 + p_2)/2$. Moreover, we have:

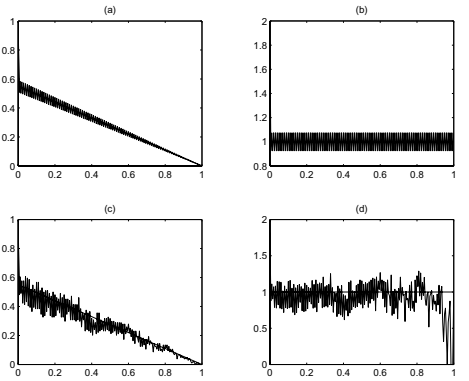


Fig. 2. The autocorrelation of porous media where $p_1 = 0.7$, $p_2 = 0.4$. Theoretical expressions: (a) $I_{\mathbb{E}}(t)$, (b)$I_{\mathbb{E}}(t)/I_{\mathbb{E},p}(t)$; simulation results (10,000 samples): (c)$I_{\mathbb{E}}(t)$, (d)$I_{\mathbb{E}}(t)/I_{\mathbb{E},p}(t)$.

$$\mathbb{E}\int \mathcal{X}_F(x)\mathcal{X}_F(x+t)dx = (1-\frac{k}{n})(p_1^2+p_2^2) \tag{2.8}$$

for $t = \dfrac{2k}{2n}$, and

$$\mathbb{E}\int \mathcal{X}_F(x)\mathcal{X}_F(x+t)dx = (1-\frac{2k-1}{2n})p_1p_2 \tag{2.9}$$

for $t = \dfrac{2k-1}{2n}$. On the other hand, variances of these two expressions, involved in $I_{\mathbb{E}}(t)$, are $O(1/n)$, which guarantees their convergence when n is large enough. Then:

$$I_{\mathbb{E}}(t) = \begin{cases} 1, & t = 0; \\ (1-\dfrac{2k}{2n})\dfrac{p_1^2+p_2^2}{p_1+p_2}, & t = \dfrac{2k}{2n}, \quad k = 1,\cdots,n; \\ (1-\dfrac{2k+1}{2n})\dfrac{2p_1p_2}{p_1+p_2}, & t = \dfrac{2k+1}{2n}, \quad k = 1,\cdots,n. \end{cases} \tag{2.10}$$

Consider the line of porosity defined by:

$$I_{\mathbb{E},p}(t) = (1-t)p, \quad 0 < t < 1. \tag{2.11}$$

Then, for $t = \dfrac{2k}{2n}$, $\quad k = 1,\cdots,n$, we have: $I_{\mathbb{E}}(t) = (1-t)p \cdot \dfrac{4p_1p_2}{(p_1+p_2)^2} < I_{\mathbb{E},p}(t)$, and for $t = \dfrac{2k+1}{2n}$, $\quad k = 1,\cdots,n$, we have: $I_{\mathbb{E}}(t) = (1-t)p \cdot \dfrac{2(p_1^2+p_2^2)}{(p_1+p_2)^2} > I_{\mathbb{E},p}(t)$.

Clearly, $I_{\mathbb{E}}(t)$ oscillates around $I_{\mathbb{E},p}(t)$, making the latter as its line of average. In Fig. 2, we show the theoretical and simulation results for $I_{\mathbb{E}}(t)$

and $I_{\mathbb{E}}(t)/I_{\mathbb{E},p}(t)$. We observe that the oscillations of the autocorrelation curve around its line of porosity will not vanish, reflecting the dependency between points. Also, the simulations indicate that the part of $I_{\mathbb{E}}(t)/I_{\mathbb{E},p}(t)$ that can be used should not be close to $t = 1$ since the divider will be too small.

2.3 A Homogeneous Model of Porous Media with Point-to-point Correlation in Terms of Distance

A more realistic model is to suppose some relationships between the points, i.e. a certain dependency between components. This dependency can be a function of the distance between points, and logically, it should decrease as the distance increases.

In this model, we consider the porous media as composed of one component, and we will keep all notations used in Section 2.1. We also have:

$$\mathbb{E}(X_i X_j) = \begin{cases} p, & i = j; \\ R(i,j)p, & i \neq j \end{cases} \tag{2.12}$$

where $R(i,j)$ is a function of $|i-j|$ that can be redefined as $R(r)$, $r = |i-j|/n$.

$R(r)$ is positive and its largest value is $R(0) = 1$. Moreover, $R(r)$ will oscillate around p and converges to p when r increases and $R(+\infty) = p$.

We have

$$\mathbb{E}\int \mathcal{X}_F(x)^2 dx = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}X_i^2 = p. \tag{2.13}$$

and for $t = (i-1)/n$,

$$\begin{aligned} \mathbb{E}\int \mathcal{X}_F(x)\mathcal{X}_F(x+t)dx &= \frac{1}{n}\sum_{k=i}^{n} \mathbb{E}(X_k X_{k-i+1}) \\ &= (1-t)R(t)p. \end{aligned} \tag{2.14}$$

Then

$$I_{\mathbb{E}}(t) = (1-t)R(t). \tag{2.15}$$

Rewrite $I_{\mathbb{E}}(t) = R_p(t)R_o(t)$, where $R_p(t) = p(1-t)$ and $R_o(t) = R(t)/p$. This decomposition describes the two components that are present in the autocorrelation function. Indeed, R_p is a straight line with slope $-p$ and cutting point p on the y-axis. This line, that we called the line of porosity in Section 2.2, corresponds to the autocorrelation function of the point-to-point independent homogeneous porous media. The curve R_o is the pure correlation component which contains the correlation information between points. In other terms, R_o gives the information about the average relationship between any two points of the set in terms of their distance. Thus, we can redefine R_o as the pure correlation function of the set:

$$R_o(t) = \frac{\mathbb{E}\int \mathcal{X}_F(x)\mathcal{X}_F(x+t)dx}{(1-t)\mathbb{E}\int \mathcal{X}_F(x)^2 dx}. \tag{2.16}$$

Remark: The following autocorrelation function is commonly used for stationary signals S:

$$C(t) = \frac{E[(S(x)-m)(S(x+t)-m)]}{E(S(x)-m)^2} \tag{2.17}$$

where $m = E(S(x))$ is independent of x.

However, for a spatial signal defined on a complex support, it is difficult or even impossible to define its mean value m, when taken in the usual meaning [5]. In order to use the last mentioned definition of the autocorrelation function, some ergodic condition on average has to be verified. To avoid this average problem, we chose to use the general form [4]. Moreover, as shown above, the general form can be decomposed, in a clear way, into two components that we call here the line of porosity and the pure correlation curve.

3 Complementary Parameters to Porosity

Obviously, porosity alone is far from enough to reflect the irregular morphology of micro-porous structures. Properties of porous media are highly dependent on the morphology of the pore space as well as those of its complementary part. To go one step further in the characterization process and extract complementary parameters, various strategies have to be developed for different degrees of complexity.

3.1 Low-complexity Texture: A Measure of the Departure from Homogeneity

Simply organized structures are generally composed of sub-parts having a narrow range of size distribution. Thread-like textures are a particular case (Fig. 3, left and middle parts). In this part, we use typical thread-like structures to show how, our newly developed analysis approach, based on the autocorrelation function, allows to extract pore frequency and extent, which are complementary parameters to porosity. From the previous result, we know that the LMS of the autocorrelation function of a perfectly homogeneous structure coincides with its porosity line. This ideal homogeneous structure will serve to measure the departure of other structures from homogeneity. Comparing images based on this departure from homogeneity, based on the autocorrelation function, is more feasible than a direct pixel-by-pixel comparison. To do so, we suggest to subtract, form the autocorrelation function, its porosity line, which is, as stated above, its LMS slope. This allows to keep only the information hidden in the remaining part, $I_{\mathbb{E}}(t) - R_p$. In fact, this departure from homogeneity can be related to the concept of lacunarity [12]. For this remaining curve, two different approaches can be adapted. The first one

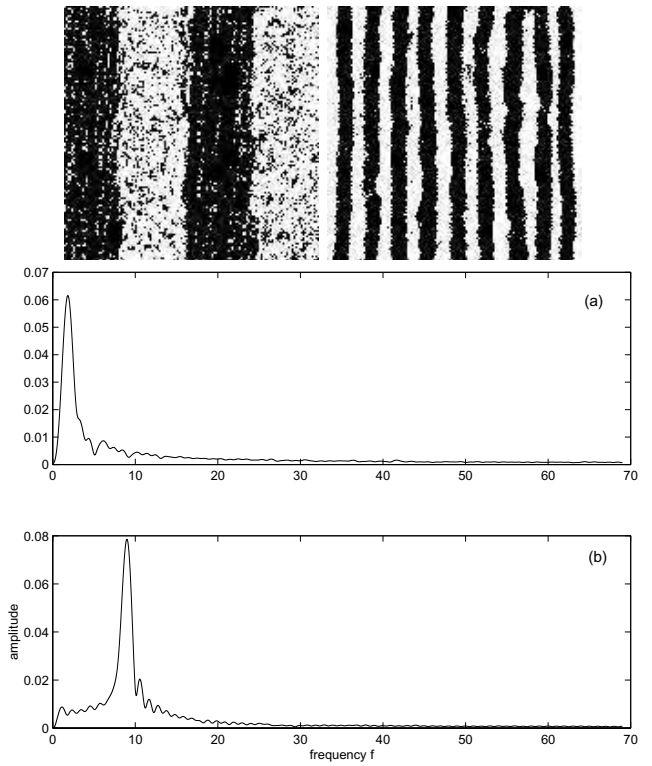


Fig. 3. Up left and right: two synthetic thread-like porous structures of same porosity; once porosity line cleared, the Fourier transform of the autocorrelation functions, (a) for the left image, (b) for the middle one.

is through the variance of $I_{\mathbb{E}}(t) - R_p$, which can be defined as a new formula for lacunarity as a departure from homogeneity, in terms of the autocorrelation function. This approach is not developed in this paper. Since we are concerned, in this section, by low-complexity images, it is possible to apply the Fourier transform to analyze the frequency components of $I_{\mathbb{E}}(t) - R_p$.

For example, for the two periodic patterns of the same porosity value: 0.4633, shown by two images on the left and middle places in Fig. 3, the Fourier transform of their corresponding $I_{\mathbb{E}}(t) - R_p$ is given on the right of Fig. 3. The results on frequency f as well as on $1/f$ give two parameters, independent from porosity, representing the number of threads as well as their sizes, respectively. In fact, the Fourier transform of $I_{\mathbb{E}}(t) - R_p$ gives a frequency expression of the departure of the signal from homogeneity (Wiener-Khinchin theorem). The main frequency component corresponds to periodic pattern once the porosity information is dropped.

We have also processed images where two frequencies have been superposed. To extract the average frequency f_m and the corresponding average length $1/\lambda_m$, we experimentally found that the following formula gives the nearest values for f_m compared to $1/\lambda_m$. If we note by $\mathcal{D}(f)$ the absolute value of Fourier transform of $I_{\mathbb{E}}(t) - R_p$, then, we have:

$$f_m = \left(\frac{\sum \mathcal{D}(f_i) f_i^{1/2}}{\sum \mathcal{D}(f_i)}\right)^2 \tag{3.18}$$

$$\lambda_m = \left(\frac{\sum \mathcal{D}(f_i) \lambda_i^{1/2}}{\sum \mathcal{D}(f_i)}\right)^2 \tag{3.19}$$

where $\lambda_i = 1/f_i$.

We have to stress that the above Fourier analysis method works well for non-trivial cases, i.e. porosity is not too small or not near 1 (its complementary is small). For complex structures of porosity near zero, we have to use the approaches we present below.

Remark: Relying on the theoretical proof given in Section 2, which gives a statistical meaning to porosity through autocorrelation function, we suggest to extract images porosity using the least mean square (LMS) slope of the autocorrelation function. This porosity line gives a statistical information on a given (deterministic) image, which, in fact, has been generated through a random way. This statistical meaning of porosity is more appropriate than using the porosity computed from a single image, the latter introducing a variation in the porosity of the whole images collection. As a practical example, we use a set of porous images of decreasing porosities to illustrate this correspondence of porosity to the LMS slopes of the autocorrelation function. The images are taken from K. Zhao et al. [13], Fig. 2 on page 118, which represent blending films of porous scaffold and show decreasing porosities in terms of the concentration of the two used polymers, PHB and PHBHHx. The left plot in Fig. 4 reports the equivalence (matching) between the LMS slopes of the autocorrelation of figures (a-f)in Fig. 2 in K. Zhao et al. [13] and the porosity values calculated by the porosity formula.

3.2 Fractal-like Texture

Fat Fractals: Synthetic Porous Media

When complexity of a structure increases further while it has a positive (non-zero) porosity, we propose a different approach to the autocorrelation function, combining fractal tools. A typical case for those more complex structures of positive porosity are known as fat fractals. Their main peculiarity, compared to thin fractals, consists in a finite and non-zero Lebesgue measure of their support. The empty holes of fat fractals have size-dependant power distribution similar to porous materials [3, 14] and they have been proposed as realistic models of micro-porous media [3], Let us recall how a 1-D regular

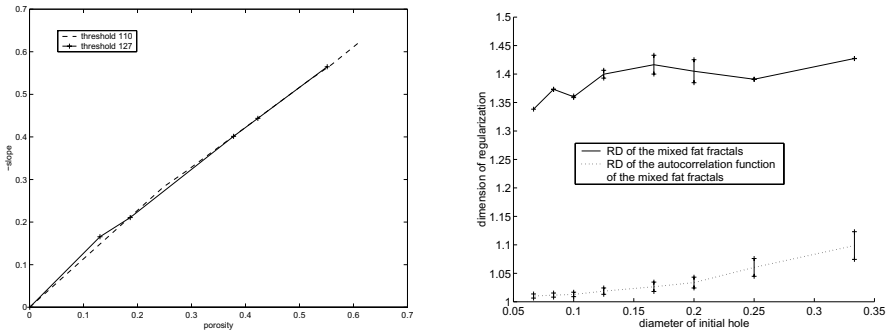


Fig. 4. Left: The matching between estimated porosity using the porosity formula and the slope of the linear part of the autocorrelation function of images in [13], page 118; Right: RD directly applied on the mixed fat fractals (top) and on their autocorrelation function (bottom).

fat fractal is generated. For sake of simplicity, we usually use the unit interval as initiator. A regular fat fractal can be obtained through the following simple iterative rule: from one step to the next, we just drop the open middle intervals of length l_n from each of the remaining intervals. If the length of the interval at step $n-1$ is L_{n-1}, then this removed length is L_{n-1} divided by a^n. After n iterations, the obtained set is composed of 2^n intervals of lengths L_n and n subsets which are composed of $N_n(k) = 2^{k-1}$ empty holes of lengths l_k, $(k = 1, \cdots, n)$, respectively. For a 1-D regular fat fractal, a is the only parameter involved in its definition. It is clear that the empty holes are symmetrically distributed in a 1-D regular fat fractal. To describe a porous media having asymmetric features, we have to use the mixed fat fractal which, intuitively, is a redistribution of the parts of the regular one, obtained by rearranging alternatively its voids and occupied intervals. We have considered eight fat fractals having the following values for the parameter a : 3, 4, 5, 6, 8, 10, 12 and 15. The autocorrelation function of these fat fractals is a more irregular curve, preventing thus a valuable application of the Fourier transform as was the case above [9, 15]. The irregularity of those curves suggests the utility of the fractal dimension usually used to quantify complexity. Thus, we use the regularization dimension (RD), introduced by J. Levy-vehel and F. Roueff [11], since it proved to be more sensitive to variations. However, one can also raise the question of applying the RD directly to the fat fractal instead of applying it to the autocorrelation function of the set. Comparison of the obtained results shows that, when RD is directly applied to the sets, a differentiation based on this dimension is less convincing compared to when it is applied to their autocorrelation function, the right plot of Fig. 4. In fact, RD of the autocorrelation function is a strictly increasing function of the size of the initial hole. This difference can be explained by the fact that autocorrelation function has a smoothing "action" since it attenuates the irregularities

of a signal (spatial signal representing the set in this case) and produces more uniform ones: sparse parts intersecting with themselves will still be sparse and when intersecting with denser parts, will again produce sparse parts. Also, the autocorrelation accumulates similarities that are dispersed all along the set. Once the autocorrelation function is applied, The similarity amount is globally decreasing as the translation value t increases.

Thin Fractals: Structures of Porosity Zero

As already mentioned in Section 3.1, images of neglected porosity cannot be processed by the Fourier analysis approach that we suggested above. We hereafter propose an alternative method based on the Hausdorff measure of these sets. In fact, typical non-trivial examples are sets having zero-Lebesgue measure, as it is the case for Cantor sets or, more generally thin fractals. For these sets, we have to use what we previously named the Hausdorff Measure Spectrum Function (HMSF) [7], which involves integration according to Hausdorff measure instead of Lebesgue measure used in the autocorrelation function, Eq.(2.1). In fact, we define the HMSF as follows:

$$I^{s_H}(t) = \int_{x \in F} \mathcal{X}_F(x)\mathcal{X}_F(x+t)d\mathcal{H}^{s_H}(x), \tag{3.20}$$

where s_H is the Hausdorff dimension of the set F [16]. Fig. 5 shows the HMSF of the Cantor fractal set as well as of a geometrical multifractal set. The latter example is generated on the unit interval using two similitudes, with the scaling factors equal to $1/2$ and $1/4$, on left and right extremities, respectively. This two scale-Cantor set gives rise to what is called a geometrical multifractal [17, 18]. Other typical examples can be found in [6, 7].

Distinguishing Sets of the Same Fractal Dimension Using HMSF

The fractal dimension is used to quantify the complexity of a structure. However, there is a need to further set apart, in a quantitative way, different fractal structures sharing the same fractal dimension. In [7, 8], we have proposed the HMSF as a potential method to distinguish structures. By the way it is constructed, the HMSF contains information on two-point statistics relationship between parts in terms of their distances. This is the reason that the HMSF provides additional information on the organization of the structure (what is generically called lacunarity). We proposed two different ways to exploit the HMSF in order to distinguish between sets having the same fractal dimension. The first one is based on what we called the translation invariance based method (TIBM). The second one consists in comparing the measure values of this function for different sets at a fixed level. We call it the fixed level based method (FLBM). The latter one is necessary only when the first step in not enough. FLBM associates an index to sets which allows for their

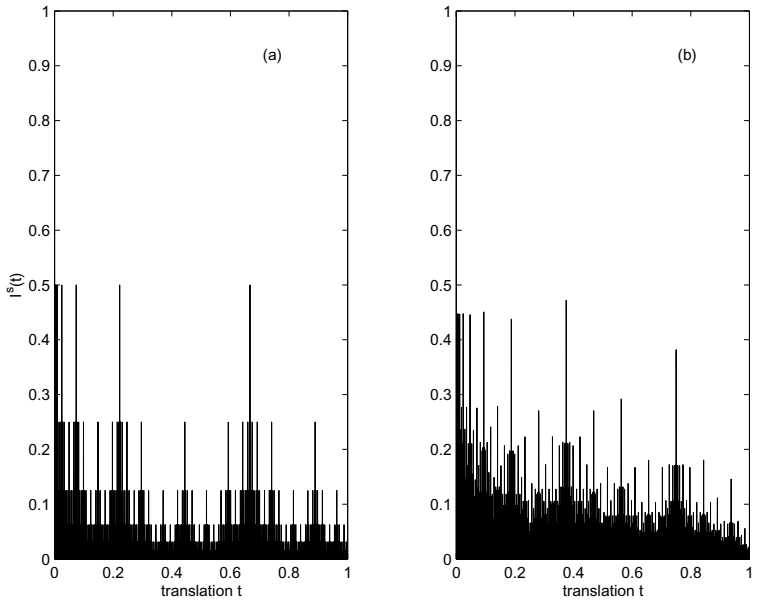


Fig. 5. The HMSF of the triadic Cantor set (left) and of the considered multifractal set (right).

differentiation as well as for evaluation of their degree of homogeneity. For geometrical multifractals, peaks of the HMSF have variation of different heights. However, they still converge towards several level values, from which we can extract geometrical information related to the homogeneity of the structure. Further analysis in this direction will be done in a future publication.

Estimation of the Hausdorff Dimension Using HMSF

Our HMSF method is based on the Hausdorff measure and dimension, which are generally hard to be determined. This a priori knowledge can be an obstacle towards the application of the HMSF. We proposed a method that allows to approximate the Hausdorff measure and the Hausdorff dimension. The method we propose here does not rely on the exact Hausdorff measure definition but substitutes for it, an appropriate approximative quantity. Indeed, except for uniform Cantor-type sets whose HMSF can theoretically be determined [6,7,9,16], it is necessary to use as a substitute, the measure spectrum function, which can be defined in the following way: given a fractal set F, we associate it with a family of covering sets $\{F_{i,n}\}$, where $n = 0, 1, 2, \cdots$ and $i \in I_n$. The set $i \in I_n$ is the set of indexes whose corresponding intervals $F_{i,n}$ are nonempty. We also let the size $|F_{i,n}| \to 0$ when n goes to infinity. If we write $F_n = \cup_{i \in I_n} F_{i,n}$, then $F = \cap_{n=1}^{\infty} F_n$. For the shift element t, the intersection $F_n \cap (F_n + t) = \cup F'_{k,n}$, where $F'_{k,n}$ corresponds to the intersec-

tion parts of n-th covering sets $\{F_{i,n}\}$ with their translates. Then, we can define the measure spectrum function as: $I_n^s(t) = \sum_k |F'_{k,n}|^s$. This substitute is a generic approximation of the Hausdorff measure. When dealing with a given fractal set, one uses a coarse-graining process to practically represent the set at different resolutions. If we use a coarse-graining whose elements are all of size ϵ_n at the coarse-graining level n, we can define different quantities to approximate the Hausdorff measure [16]. A first approximating quantity is similar to the one involved in the box-counting dimension: $Q_1(n, s) = N(n)\epsilon_n^s$, where $N(n)$ is the number of nonempty boxes of coarse-graining elements of size ϵ_n. A better approximating quantity is what we have previously named the adaptive coverings, noted Q_2 [16]. Indeed, we have successfully used this quantity to practically evaluate the Hausdorff dimension. Q_2 gathers into one element each group of joined coarse-graining elements, and redefines in this way another covering family (adaptive). It is usable when the condition that the maximum length of its adaptive covering goes to zero when the coarse-graining level n goes to infinity [16]. We believe that for advanced needs of practical problems, this approximation exercise should be continued in order to reach better estimations. This quantity shows a transition property around its value at Hausdorff dimension s_H, similar to that exhibited by the Hausdorff measure. However, instead of relying upon the critical behaviour of $I_s(0)$ (which is related to the Hausdorff measure of the set) when s is around s_H, we used the whole spectrum of measure function $I_s(t)$ for all values of t varying from 0 to the total length of the set. In other words, it is a "global" measure phase transition criterion that we use instead of a transition criterion around a single point. This method provides an interesting alternative to the usual scaling forms involved in the box-counting dimension, where convergence problems are encountered.

Uniqueness of HMSF

The classical definition of autocorrelation function degenerates when applied to signals having support of zero Lebesgue measure. Though the power law expression can be used to quantify this degeneracy, the latter can have severe consequences since the usual autocorrelation function can be dependent on the dynamical process. In other words, one can get two different autocorrelation functions for the same signal if there exist two different dynamical processes, usually named cascades (also artificially obtained by coarse-graining), corresponding to this same final set. If we get interested in the asymptotic or infinite state of a physical phenomena, this degeneracy becomes a major drawback in signal characterization. Indeed, the coarse-graining procedure is the main way that we apply to represent a geometric set. However, this somehow artificial way can associate various dynamical systems to the same set, implying divergence in the autocorrelation power law exponent. Being concerned with this issue, the HMSF that we proposed seems to be a suitable form to drop the limitations of the classical autocorrelation function. Indeed, we showed that

HMSF remedies to this situation and makes the characterization of the geometric information (at the static state) independent of the dynamic process involved [10].

4 Conclusion

In this paper, we established a statistical relationship between porosity and the autocorrelation function. This serves as a new definition, through the autocorrelation function, of the homogeneity and opens a new way to measure the departure of a given texture from its homogeneous ideal counterpart (the homogeneous image of the same porosity as the given texture). We also propose several new ways to investigate texture of porous media. These different approaches allow us to deal with a broad spectrum of texture complexity. Indeed, one cannot hope for an almighty solution for all textures. For various texture complexities, we have to develop different strategies aimed to different purposes. The new descriptions of texture obtained through these strategies define the various parameters which can be used for classification purposes or as reconstruction criteria. In fact, in this work, we have combined some largely used mathematical tools, namely the autocorrelation function and the Fourier transform, with more recent complexity analysis tools, i.e., the HMSF that we previously introduced along with the regularization dimension. This combination has been put in practice using computerized numerical means. This paper proposes some possible ways to probe the fine details in texture. However, for each of the proposed strategies, further work is going on theoretically and numerically. Parameter definitions can indeed be refined and more advanced algorithms have to be developed.

Acknowledgements. The authors would like to thank Professor Antoine Saucier for his valuable comments. This work has been supported by NSERC UFA and NSERC grant (RGPIN-227118), hold by F. Nekka as well as by Faculté de Pharmacie, Univrsité de Montréal (fonds de démarrage) and by CRM (Centre de Recherches Mathématiques) group funds.

References

1. W. E. Wallace, C. M. Guttman (2002) J. Res. Natl. Inst. Stand. Technol 107: 1–17
2. C. Allain, M. Cloitre (1987) Phys. Rev. A 36 no.12: 5751–5757
3. S. A. Bulgakov (1992) Phys. Rev. A 46 no.12:8024–8027
4. D. C. Champeney, (1973) Fourier Transforms and their Physical Applications. Academic Press, New York
5. L. S. Leibovitch (1998) Fractals and Chaos Samplified for the Life Sciences. Oxford University Press, New York

6. F. Nekka, J. Li (2002) Chaos, Solitons and Fractals 13, no.9:1807–1817
7. J. Li, F. Nekka (2004) Chaos, Solitons and Fractals 19, no.1:35–46
8. F. Nekka, J. Li (2004) Characterization of Fractal Structures Through a Hausdorff Measure Based Method. In: M. M. Novak (eds) Thinking in Pattern - Fractals and Related Phenomena in Nature. World Scientific, Singapore
9. J. Li, F. Nekka (2003) Pattern Recognition Letters 24 Issue 15:2723–2730
10. J. Li and F. Nekka, Is the Classical Autocorrelation Function Appropriate for Complex Signals? – The Necessity of the Generalized Autocorrelation Function. Submitted to Int. J. of Chaos and Bifurcation 2004
11. F. Roueff, J. Levy-vehel (1998) A Regularization Approach to Fractional Dimension Estimation. In: M. M. Novak (eds) Fractals and Beyond - Complexities in the Sciences. World Scientific, Singapore
12. T.G. Smith Jr. et al. (1996) J. Neurosc. Methods 69:123-136
13. K. Zhao, Y. Deng, G. Q. Chen (2003) Biochemical Engineering Journal 16:115-123
14. D. K. Umberger, J. D. Farmer (1985) Physical Review Letters 55 no.7:661-664
15. J. Li, C. Dubois, F. Nekka (2004) The Colligation of the Autocorrelation Function and the Regularization Dimension - A New Characterization of Porous Media. In: M. M. Novak (eds) Fractals 2004, Poster Abstracts, Complexity and Fractals in Nature, Vancouver
16. J. Li, A. Arneodo, F. Nekka (2004) CHAOS, An Interdisciplinary Journal of Nonlinear Science 14 no.4:1004–1017
17. J. Feder (1998) Fractals. Plenum Press, New York
18. T. Vicsek (1989) Fractal Growth Phenomena. World Scientific, Singapore

Fractal Inverse Problem: Approximation Formulation and Differential Methods

Eric Guérin and Eric Tosan

LIRIS - Université Claude Bernard - Bâtiment Nautibus - 43, Bd du 11 Novembre - 69622 Villeurbanne Cedex - France
eric.guerin@liris.cnrs.fr, eric.tosan@liris.cnrs.fr

Summary. An analytical approach to fractal inverse problem is presented in this paper. We recall the construction of an Hilbert space with address functions, that constitute a general framework for fractal modeling, since it includes IFS made fractals. A large scale of applications is shown, ranging from scalar function approximation to image compression.

1 Introduction

1.1 Fractal Inverse Problem

The fractal inverse problem is an important research area with a great number of potential application fields. It consists in finding a fractal model or code that generates a given object. This concept has been introduced by Barnsley with the well known *collage theorem* [2]. When the considered object is an image, we often speak about fractal image compression. A method has been proposed by Jacquin to solve this kind of inverse problem [13].

This problem has been studied by much authors. Generally speaking, inverse methods can be classified in two types:

- Direct methods: model characteristics are found directly. In the fractal case, very few direct methods have been proposed. In general, we have to deal with synthetic data entries. Some authors use wavelet decomposition to find frequency structures and extract IFS coefficients [3, 16]. A method using complex moment has been experienced to work for fractal images [1].
- Indirect methods: model characteristics are found indirectly. In general an optimization algorithm is used. These methods allows to deal with more complex models and less synthetic data entries. Inverse problem for mixed IFS has been performed with genetic methods [14].

Optimization methods used in indirect methods are generally stochastic, because it's not possible to calculate any derivative with respect to the model parameters.

1.2 Fractal approximation

In [17], VRSCAY and SAUPE introduced a derivative property in the numerical functions involved in fractal image coding with respect to their affine IFS parameters. This allows the use of a gradient descent method to solve the inverse problem and gives better results than the standard collage theorem.

In [8,9], we developed a method based on this property for fractal approximation of curves and surfaces. In [10,11], we have extended this method to surfaces. In [6,12], we introduced a *projected IFS tree model* to obtain a better approximation of natural surfaces and greyscale images.

In this paper we develop a differential approach of the fractal approximation problem based on formal multiresolution.

1.3 Formal multiresolution

In [7], we present an analytical approach of the approximation problem based on address functions. BARNSLEY introduces these functions to define a formal parameterization of attractors [2]. Address functions are functions that takes address arguments rather than numerical ones. They map from infinite words Σ^ω to a modelisation space $\mathcal{X} = \mathbb{R}^m$:

$$\begin{aligned} \phi : \Sigma^\omega &\to \mathcal{X} \\ \rho &\mapsto \phi(\rho) \end{aligned}$$

Address functions give a natural multiresolution formulation [7]: by selecting infinite words of Σ that have the form αk^ω, we define a finite family of points that can be viewed as the tabulation of ϕ at a given level n:

$$(\phi(\alpha k^\omega))_{\alpha \in \Sigma^n, k \in \Omega}$$

where Ω is a subset of Σ such that $\phi(k^\omega)$ represents the "boundaries" of the figure.

The modelisation space $\mathcal{X} = \mathbb{R}^m$ is Hilbertian, we define a new Hilbert space of address functions based on the following dot product [7]:

$$< \phi, \phi' >= \lim_{n \to \infty} \frac{1}{N^n} \sum_{\alpha \in \Sigma^n} \frac{1}{M} \sum_{k \in \Omega} < \phi(\alpha k^\omega), \phi'(\alpha k^\omega) >$$

with $M = |\Omega|$ and $N = |\Sigma|$. This constitutes a general frame of geometric fractal modeling and approximation.

2 IFS approximation

One convenient way to provide address functions is the use of IFS (Iterated Function Systems). Furthermore, we will see that these functions verify a decreasing condition and then belong to $L^2(\Sigma^\omega, \mathcal{X})$.

2.1 IFS model

Introduced by Barnsley [2] in 1988, the IFS (Iterated Function Systems) model generates a geometrical shape or an image [13] with an iterative process. An IFS-based modeling system is defined by a triple $(\mathcal{X}, d, \mathcal{S})$ where [18, 19]:

- $(\mathcal{X}, d)$ is a complete metric space, $\mathcal{X}$ is called *iteration space*;
- $\mathcal{S}$ is a semigroup acting on points of $\mathcal{X}$ such that: $\lambda \in \mathcal{X} \mapsto T\lambda \in \mathcal{X}$ where T is a contractive operator, $\mathcal{S}$ is called *iteration semigroup*.

An *IFS* $\mathbb{T}$ *(Iterative Function System)* is a finite subset of $\mathcal{S}$: $\mathbb{T} = \{T_0, ..., T_{N-1}\}$ with operators $T_i \in \mathcal{S}$. We note $\mathcal{H}(\mathcal{X})$ the set of non-empty compacts of $\mathcal{X}$. $\mathcal{H}(\mathcal{X})$ is a complete metric space with the Hausdorff distance. The associated Hutchinson operator is:

$$K \in \mathcal{H}(\mathcal{X}) \mapsto \mathbb{T}K = T_0K \cup ... \cup T_{N-1}K \ .$$

This operator is contractive in the complete metric space $\mathcal{H}(\mathcal{X})$ and admits a fixed point, called *attractor* [2]:

$$\mathcal{A}(\mathbb{T}) = \lim_{n\to\infty} \mathbb{T}^n K \text{ with } K \in \mathcal{H}(\mathcal{X}) \, .$$

By introducing a finite set Σ, the IFS can be indexed $\mathbb{T} = (T_i)_{i\in\Sigma}$ and the attractor $\mathcal{A}(\mathbb{T})$ has an *address function* [2,4] defined on Σ^ω, the set of infinite words of Σ:

$$\rho \in \Sigma^\omega \mapsto \phi(\rho) = \lim_{n\to\infty} T_{\rho_1}...T_{\rho_n}\lambda \in \mathcal{X} \text{ with } \lambda \in \mathcal{X} \, . \tag{1}$$

2.2 Approximation formulation

Corollary 1. *Every address function associated with an IFS is in* $L^2(\Sigma^\omega, \mathcal{X})$.

Proof. See [7].

In the Hilbert space of address functions, optimization problem can be expressed with the following formulation. Let φ be an address function, find $\mathbb{T}$ that minimizes the error function:

$$\mathbb{T} \in \mathcal{S}^\Sigma \to g(\mathbb{T}) = \|\Psi(\mathbb{T}) - \varphi\|_2^2 \in \mathbb{R}_+$$

with $\Psi(\mathbb{T})$ the address function associated with $\mathbb{T}$.

To apply standard non-linear fitting methods, the function g needs to have good properties. This function is a quadratic form of Ψ. In the following, we will expose these properties.

2.3 Affine IFS

We now deal with affine IFS, that means IFS defined with affine contractions in $\mathcal{X} = \mathbb{R}^m$. In this case, the contractive semigroup can be characterized. An affine operator is defined by a couple (u, L) with $u \in \mathbb{R}^m$ and L a $m \times m$ matrix:

$$Tp = u + Lp$$

The set of affine operators acting on $\mathbb{R}^m$ is a complete metric space with the following distance:

$$d(T, T') = \|u - u'\| + \|L - L'\|$$

where

$$\|L\| = \max_{\|u\|=1} \|Lu\| .$$

Proposition 1. *The affine contractive semigroup is an open set* $\mathcal{S} = \mathbb{R}^m \times \mathcal{B}_1$ *where* $\mathcal{B}_1 = \{L / \|L\| < 1\}$.

Proof. One can easily verify that $T \in \mathcal{S}$ implies its contraction:

$$\begin{aligned} \exists r \in [0,1[,\, \forall p, q \in \mathbb{R}^m,\, & d(Tp, Tq) \leq r d(p, q) \\ & \Leftrightarrow \exists r \in [0,1[,\, \forall u \in \mathbb{R}^m,\, \|Lu\| \leq r\|u\| \\ & \Leftrightarrow \exists r \in [0,1[,\, \forall u \in \mathbb{R}^m,\, \|u\| = 1, \|Lu\| \leq r \\ & \Leftrightarrow \max_{\|u\|=1} \|Lu\| < 1 \end{aligned}$$

2.4 Analyticity

In this section, we precise property of the function:

$$\psi : \mathbb{T} \in \mathcal{S}^{\Sigma} \rightarrow \psi(\mathbb{T}) \in C^0(\Sigma^{\omega}, \mathcal{X})$$

Definition 1. *Let* $\mathcal{S}^{\Sigma}$ *be the set of indexed IFS* $\mathbb{T} = (T_i)_{i \in \Sigma}$. *Let* ψ_ρ *be the following function, where* $\rho \in \Sigma^{\omega}$ *is fixed:*

$$\begin{aligned} \psi_\rho : \mathcal{S}^{\Sigma} &\rightarrow \mathcal{X} \\ \mathbb{T} &\mapsto \psi_\rho(\mathbb{T}) = \lim_{n \rightarrow \infty} T_{\rho_1} \dots T_{\rho_n} p \end{aligned}$$

As T_i is affine, we may decompose it in a translation vector u_i and a linear part L_i:

$$T_i p = u_i + L_i p$$

In this case, the product $T_i T_j$ gives $u_i + L_i u_j$ as translation vector and $L_i L_j$ as linear part. Then, we expand the matrix product:

$$\begin{aligned} T_{\rho_1} \dots T_{\rho_n} p = \quad & u_{\rho_1} \\ & + L_{\rho_1} u_{\rho_2} \\ & + L_{\rho_1} L_{\rho_2} u_{\rho_3} \\ & + \dots \\ & + L_{\rho_1} \dots L_{\rho_{n-1}} u_{\rho_n} \\ & + L_{\rho_1} \dots L_{\rho_n} p \end{aligned} \tag{2}$$

When n tends to infinity, p has no influence on this formula:

$$\lim_{n\to\infty}(L_{\rho_1}\dots L_{\rho_n}p)=0$$

because L_i are linear contractions. Then $\psi_\rho(\mathbb{T})$ can be written as a summation:

$$\psi_\rho(\mathbb{T})=\lim_{n\to\infty}\sum_{k=1}^{n}L_{\rho_1}\dots L_{\rho_{k-1}}u_{\rho_k}$$

Proposition 2. *For every ρ in Σ^ω, the function ψ_ρ is analytical on $\mathcal{S}^\Sigma$.*

Proof. See [7].

Proposition 3. *The function:*

$$\psi:\mathcal{S}^\Sigma\to L^2(\Sigma^\omega,\mathcal{X})$$

is analytical.

Proof. The function ψ is a family of functions:

$$\psi(\mathbb{T})=(\psi_\rho(\mathbb{T}))_{\rho\in\Sigma^\omega}.$$

The proof of analyticity of ψ_ρ based on differentials is valid with ψ when introducing ψ as a function of both $\mathbb{T}$ and ρ:

$$d^k\psi(\mathbb{T})(\rho)=d^k\psi_\rho(\mathbb{T}).$$

2.5 Error Estimation

In practical, this error function is approximated on samples, that means on a finite number of values:

$$g(\mathbb{T})\approx g_n(\mathbb{T})=\frac{1}{N^n}\sum_{\alpha\in\Sigma^n}\frac{1}{M}\sum_{k\in\Omega}\|\psi_{\alpha k^\omega}(\mathbb{T})-\varphi(\alpha k^\omega)\|^2$$

We toggle from a functional distance to a tabulation distance.

To perform finite exact computations, we take advantage of the fact that each transformation has a fixed point:

$$T_kc_k=c_k$$

We evaluate the function at a deep n, with $|\alpha|=n$. Then, the function has the form:

$$\begin{aligned}\psi_{\alpha k^\omega}(\mathbb{T})&=T_{\alpha_1}\dots T_{\alpha_n}\phi(k^\omega),\\&=T_{\alpha_1}\dots T_{\alpha_n}c_k.\end{aligned}$$

In this case, only polynomial computations have to be performed, g_n is a polynomial function:

$$g_n(\mathbb{T})=\frac{1}{N^n}\sum_{\alpha\in\Sigma^n}\frac{1}{M}\sum_{k\in\Omega}\|T_{\alpha_1}\dots T_{\alpha_n}c_k-\varphi(\alpha k^\omega)\|^2$$

2.6 Resolution

We proved the analyticity of affine IFS functions with respect to their matrix coefficients. We can now use a differential method to solve our problem. The literal derivative is more complex to evaluate than a numerical approximation with a perturbation. The optimization algorithm used is Levenberg-Marquardt, an improved gradient method [15].

3 Function Approximation

This section will show a very simple example of numerical optimization using affine IFS defined in $\mathbb{R}$.

3.1 Model overview

Let $\Sigma = \{0, \ldots, N-1\}$. Transformations operate on $\mathbb{R}$:

$$\begin{aligned} T_i : \mathbb{R} &\to \mathbb{R} \\ x &\mapsto a_i x + b_i \end{aligned}$$

Each transformation is defined by two scalars. In this case, the address function is:

$$\phi(\rho) = b_{\rho_1} + a_{\rho_1} b_{\rho_2} + a_{\rho_1} a_{\rho_2} b_{\rho_3} + \ldots$$

A simple series converge to this value:

$$\phi(\rho) = \lim_{n \to \infty} B_n$$

where

$$\begin{cases} B_1 = b_{\rho_1} \\ B_{i+1} = B_i + A_i b_{\rho_{i+1}} \text{ for } i \geq 1 \end{cases}$$

and

$$\begin{cases} A_1 = a_{\rho_1} \\ A_{i+1} = A_i a_{\rho_{i+1}} \text{ for } i \geq 1 \end{cases}$$

Remark 1. This kind of IFS is not Fractal Interpolation Functions since they are defined in $\mathbb{R}$ (FIF are defined in $\mathbb{R}^2$).

3.2 Approximation formulation

When dealing with approximation, a common data type is an ordered list of points $(x_i, y_i)_{i=1,\ldots,p}$. The value of x_i will be used to extract an address associated to the sample, whereas the value of y_i will be the target value of the address function. Let $\alpha^{(i)} = \alpha_1^{(i)} \ldots \alpha_n^{(i)}$ be the N-adic expansion of $\bar{x}_i$ with $x_i = \bar{x}_i + \epsilon_i$ and $\epsilon_i < \frac{1}{N^{n+1}}$:

$$\bar{x}_i = \sum_{j=1}^{n} \frac{1}{N^j} \alpha_j^{(i)}$$

Then, the approximation problem with affine IFS in $\mathbb{R}$ can be formulated. Given data entries $(x_i, y_i)_{i=1,\ldots,p}$ where $x_{i+1} > x_i$, and a number of transformations N, find the IFS that minimizes the error:

$$\begin{aligned} \mathbb{T}_{opt} &= \underset{\mathbb{T}\in\mathcal{S}^{\Sigma}}{\operatorname{argmin}}\, g_n(\mathbb{T}) \\ &= \underset{\mathbb{T}\in\mathcal{S}^{\Sigma}}{\operatorname{argmin}} \frac{1}{p} \sum_{i=1\ldots p} \left(\psi_{\alpha_1^{(i)}\ldots\alpha_n^{(i)}0^{\omega}}(\mathbb{T}) - y_i\right)^2 \end{aligned}$$

3.3 Results

We have tested our approximation method on several data sets, ranging from smooth curves to random data. As expected, the approximation quality depends on the number of transformations N taken.

Figure 1 shows the approximation of a cubic curve $y = 6(x - \frac{1}{2})^3$ with the method described previously. In these graphs, x-coordinates represents the address values and y-coordinates the values of the address function. The original curve contains 1000 points. When approximating with only 2 transformations, the fitting is not good. When the number of transformations becomes larger, the quality of approximation is better.

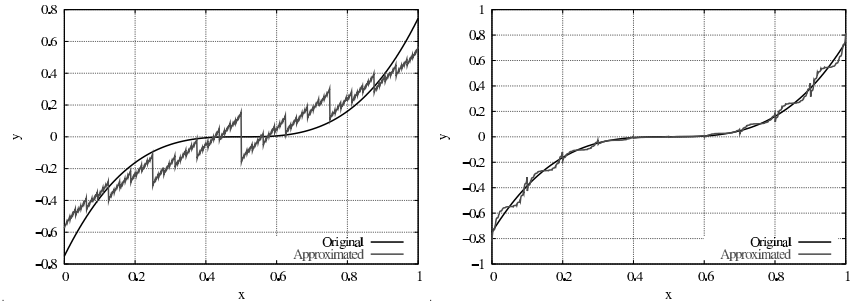
Figure 2 shows the approximation of a random function that contains 100 points. With only 5 transformations, the result is not so bad. Increasing the number of transformations leads to a better approximation. The upper limit of N is when we reach the number of data points: $N = p$. In this case, the exact reconstruction is possible. The method used to solve the approximation problem is not global. It means that the result can be a local minimum.

4 Modelisation of rough shapes

In order to propose an efficient solution to the problem of rough surface approximation, we have used a parametric model based on a fractal model. In [18,20], we have proposed a *projected IFS model* for fractal curve and surfaces. This model combines a fractal classical approach – Iterative Function Systems – and CAGD classical approach – free form based on control points. These points allow an easy and flexible control of the fractal shape generated by the IFS model and provide a high quality fitting.

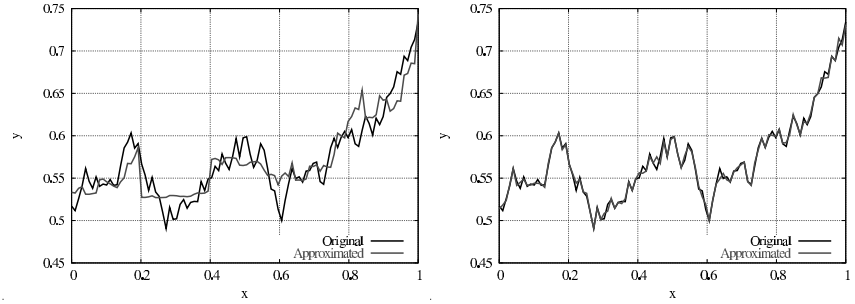
4.1 Projected IFS model

To allow more flexible modeling, we introduced and used a projected IFS model [18,19]. The way to obtain projected IFS attractors is to use a barycentric metric space $\mathcal{X} = \mathcal{B}^J$:



(a) With $N = 2$ (b) With $N = 10$

Fig. 1. Approximation of a cubic polynomial curve (1000 points)



(a) With $N = 5$ (b) With $N = 40$

Fig. 2. Approximation of a random function (100 points)

$$\mathcal{B}^J = \{(\lambda_j)_{j\in J} \,|\, \sum_{j\in J} \lambda_j = 1\}$$

Then, the iteration semigroup is constituted of matrices with barycentric columns:

$$S_J = \{T \,|\, \sum_{j\in J} T_{ij} = 1,\, \forall i \in J\}$$

This choice leads to the generalization of IFS attractors named *projected IFS attractors*:

$$P\mathcal{A}(\mathbb{T}) = \{P\lambda \,|\, \lambda \in \mathcal{A}(\mathbb{T})\}$$

where P is a polygon or grid of control points $P = (p_j)_{j\in J}$ and $P\lambda = \sum_{j\in J} \lambda_j p_j$. The associated address function is:

$$\varphi(\rho) = P\phi(\rho) = \sum_{i \in J} p_i \phi_i(\rho)$$

As shown in figure 3, a two-dimensional addressing can be easily calculated by a PÉANO code mapping.

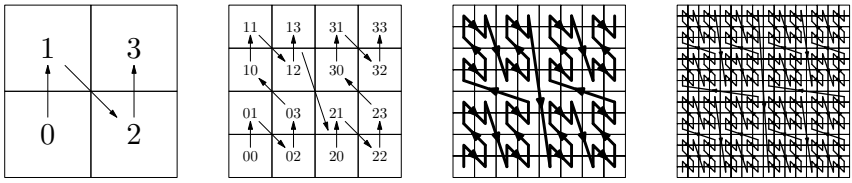


Fig. 3. PÉANO code with $\Sigma = \{0, 1, 2, 3\}$

In figure 4, an example of bivariate function generated by projected IFS is shown. This function defines a surface, projected through a 4×4 control points grid. Here, control points are scalars $p_i = z_i \in \mathbb{R}$:

$$\begin{aligned} \varphi : \Sigma^\omega &\to \mathbb{R} \\ \rho &\mapsto \varphi(\rho) = \textstyle\sum_{i \in J} z_i \phi_i(\rho) \end{aligned}$$

The construction of the projected attractor is determinist: it only requires recursive subdivisions as shown in figure 4.

4.2 Projected IFS tree model

Natural objects are composed of heterogeneous parts. To cope with this problem, we introduced another generalization: projected IFS trees model [6, 12].

Let Γ be a cut of the tree $(\Sigma^\infty, \leq)$, that means a finite part of Σ^* such that each word $\rho \in \Sigma^\omega$ admits a unique decomposition on $\Gamma \times \Sigma^\omega$:

$$\rho = \gamma\tau \text{ with } \gamma \in \Gamma \text{ and } \tau \in \Sigma^\omega.$$

If we denote $m = \max_{\gamma \in \Gamma} |\gamma|$, then we have the following decomposition:

$$\forall n \geq m \quad \Sigma^n = \bigcup_{\gamma \in \Gamma} \gamma \Sigma^{n-|\gamma|} \text{ and } \Sigma^\omega = \bigcup_{\gamma \in \Gamma} \gamma \Sigma^\omega$$

Drawn from the families:

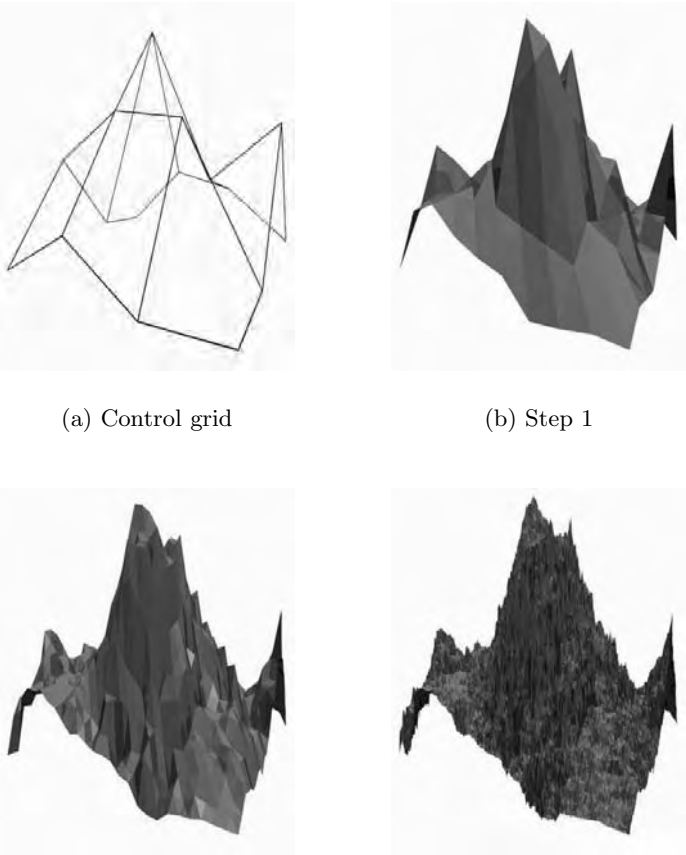
- of address functions $\phi^\gamma \in C^0(\Sigma^\omega, \mathcal{X}_\gamma)$,
- of affine functions $P^\gamma : \mathcal{X}_\gamma \to \mathcal{X}$,

we use the following address function to modelize surfaces:

$$\phi(\gamma\tau) = P^\gamma \phi^\gamma(\tau)$$

and:

$$\forall \gamma \in \Gamma, \forall i \in \Sigma, \; \phi^\gamma(i\tau) = T_i^\gamma \phi^\gamma(\tau)\,.$$



(a) Control grid (b) Step 1

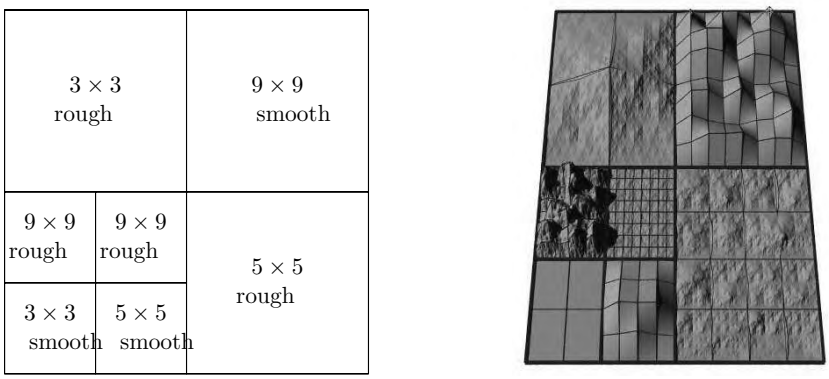
(c) Step 2 (d) Step 4

Fig. 4. Example of a projected IFS surface construction

Proposition 4. *Every address function* ϕ^γ *built on address functions associated with IFS* T^γ *is in* $L^2(\Sigma^\omega, \mathcal{X})$ *and* $\phi \in L^2(\Sigma^\omega, \mathcal{X})$.

Proof. The functions ϕ^γ are associated with IFS, that means that they verify the decreasing condition, and ϕ too (see [7]).

An example of heterogeneous surface is given in figure 5. Each patch of the surface can have different properties. In this example, we have mixed rough and smooth modeling together.



(a) Modeling quadtree (b) Associated surface

Fig. 5. Surface modeling with projected IFS quadtree

4.3 Approximation formulation

We want to approximate data entries arranged in grids $(z_{i,j})_{i,j\in 0\ldots 2^n}$ with a projected IFS tree $(P,\mathbb{T}) = (P^\gamma,\mathbb{T}^\gamma)_{\gamma\in\Gamma}$. Given a tree cut Γ, the model is described by two families of parameters: $(P^\gamma)_{\gamma\in\Gamma}$ and $(\mathbb{T}^\gamma)_{\gamma\in\Gamma}$. The address is split into two parts: the leaf $\gamma\in\Gamma$, address of the projected IFS model, and $\tau\in\Sigma^\omega$ address of the point in the projected IFS model:

$$\psi^\Gamma_{\gamma\tau}(P,\mathbb{T}) = P^\gamma\psi_\tau(\mathbb{T}^\gamma)$$

ψ^Γ is analytical with respect to $\mathbb{T} = (\mathbb{T}^\gamma)_{\gamma\in\Gamma}$ and affine with respect to $P = (P^\gamma)_{\gamma\in\Gamma}$.

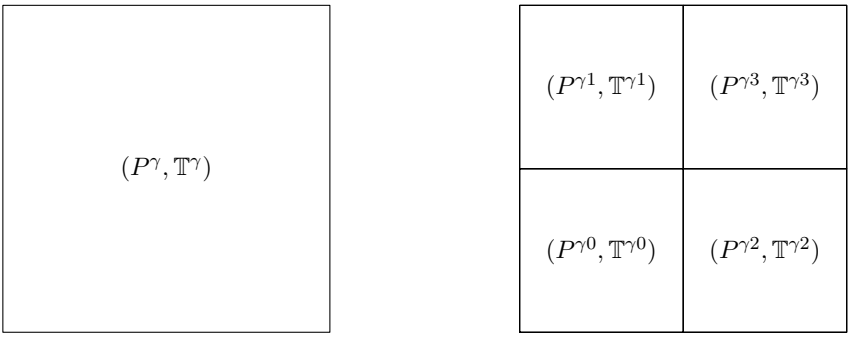
The approximation algorithm has to perform simultaneously two tasks: find the tree cut Γ and the associated projected IFS models $(P^\gamma,\mathbb{T}^\gamma)$. To satisfy this constraint, we have constructed another norm combining a maximum through the tree cut with a quadratic norm:

$$\|\psi^\Gamma(P,\mathbb{T}) - \varphi\| = \max_{\gamma\in\Gamma}\|P^\gamma\psi(\mathbb{T}^\gamma) - \varphi^\gamma\|$$

with $\varphi^\gamma(\tau) = \varphi(\gamma\tau)$.

The algorithm is then implemented using a threshold ϵ that indicates the maximum value allowed in a leaf. If this constraint is not satisfied, the leaf is split into four, recursively (see figure 6). Then, the whole norm is smaller or equal to the threshold:

$$\|\psi^\Gamma(P,\mathbb{T}) - \varphi\| \le \epsilon \quad\Leftrightarrow\quad \forall\gamma\in\Gamma,\ \|P^\gamma\psi(\mathbb{T}^\gamma) - \varphi^\gamma\| \le \epsilon$$



(a) Before refinement (b) After refinement

Fig. 6. Refinement of a projected IFS tree

The algorithm is recursive and uses only locally the analyticity property of the error function. The goal is to find the minimal cut that satisfies the constraint:

$$\Gamma_\epsilon = \min \left\{ \Gamma / \| \psi^\Gamma(P, \mathbb{T}) - \varphi \| \leq \epsilon \right\}$$

We construct an address function tabulation associated to the data entries:

$$\begin{aligned} \varphi(\alpha k^\omega) &= \varphi((\alpha' k'^\omega) \bullet (\alpha'' k''^\omega)), \\ &= z_{j',j''}. \end{aligned}$$

with

$$\begin{aligned} j' &= \textstyle\sum_{l=1\ldots n} 2^{n-l} \alpha'_l \\ j'' &= \textstyle\sum_{l=1\ldots n} 2^{n-l} \alpha''_l \end{aligned}$$

and α', α'' verifies $\alpha' \bullet \alpha'' = \alpha'_1 \bullet \alpha''_1 \ldots \alpha'_n \bullet \alpha''_n$, with $\alpha'_i \bullet \alpha''_i = 2\alpha'_i + \alpha''_i$ and $k = k' \bullet k''$.

Error estimation for a given leaf $\gamma \in \Gamma$ is:

$$g_n(P^\gamma, \mathbb{T}^\gamma) = \frac{1}{4^{n-|\gamma|}} \sum_{\alpha \in \Sigma^{n-|\gamma|}} \frac{1}{M} \sum_{k \in \Omega} \left(P^\gamma \psi_{\alpha k^\omega}(\mathbb{T}^\gamma) - \varphi(\gamma \alpha k^\omega) \right)^2$$

4.4 Surface reconstruction

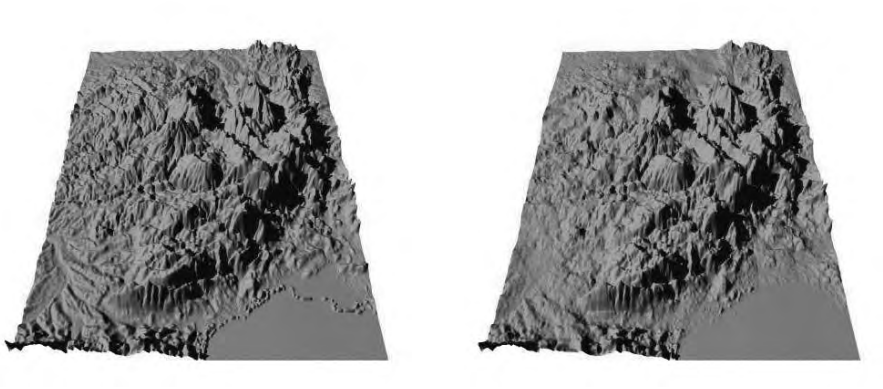
Figure 7 represents the result of a surface approximation. The data is an elevation grid of size 257×257 extracted from Digital Terrain Elevation Data (DTED) Level 0 [1]. In this example, the approximation method has been

[1] Data available at http://data.geocomm.com/catalog/FR/group121.html

applied with an error threshold based on a minimum local PSNR value. PSNR is directly related to the definition of $g_n(P^\gamma, \mathbb{T}^\gamma)$:

$$\mathrm{PSNR}(P^\gamma, \mathbb{T}^\gamma) = 10 \log_{10}\left(\frac{\max}{g_n(P^\gamma, \mathbb{T}^\gamma)}\right)$$

where max is the range of input data.



(a) Original surface (b) Approximated surface

Fig. 7. Approximation of the French "Massif central" mountain

In this example, the threshold ϵ is such that for all γ in Γ the value of $\mathrm{PSNR}(P^\gamma, \mathbb{T}^\gamma)$ is greater than $40dB$.

4.5 Image Compression

By using the same model, we are able to perform image compression. The input data is a greyscale grid of size 257×257. The difference is in the approximation method, that optimizes the rate/distortion ratio. Figure 8 shows an example of image compression. For a bit rate of $0.12bpp$, the corresponding error is PSNR=$28.3dB$, with the following classical definition of PSNR:

$$\mathrm{PSNR}(P, \mathbb{T}) = 10 \log_{10}\left(\frac{255}{g_n(P, \mathbb{T})}\right)$$

where $g_n(P, \mathbb{T}) = \sum_{\gamma \in \Gamma} g_n(P^\gamma, \mathbb{T}^\gamma)$ corresponds to the estimation of quadratic distance:

$$g_n(P, \mathbb{T}) \approx \|\psi^\Gamma(P, \mathbb{T}) - \varphi\|_2^2$$

Detailed method is available in [6, 12].

(a) Original image: portion of peppers

(b) Image compressed at $0.12bpp$, PSNR=$28.3dB$

Fig. 8. Image compression example

5 Conclusion

We showed that analytical approach and methods using derivation properties can be used to perform the fractal inverse problem. This problem can be formulated as an optimization problem in an Hilbert space. For a useful family of fractal model based on affine IFS, the error function is analytical. Hence, the optimization problem has a non-linear classical formulation. Methods based on non-linear optimization algorithms can be applied with interesting numerical results in surface reconstruction and image compression.

References

1. Toshimizu Abiko, Masayuki Kawamata, and Tatsuo Higuchi. An efficient algorithm for solving inverse problems of fractal images using the complex moment method. In *Proceedings of IEEE International Workshop on Intelligent Signal Processing and Communication Systems*, volume 1, pages S12.4.1–S12.4.6. November 1997.
2. Michael Barnsley. *Fractals everywhere.* Academic Press, 1988.
3. K Berkner. A wavelet-based solution to the inverse problem for fractal interpolation functions. In Tricot Lévy-Véhel, Lutton, editor, *Fractals in engineering'97*, pages 81–92. Springer Verlag, 1997.
4. Gerald A. Edgar. *Measure, Topology, and Fractal Geometry.* Springer Verlag, 1990.
5. Zhigang Feng and Heping Xie. On Stability of Fractal Interpolation. *Fractals*, 6(3):269–273, 1998.

6. Eric Guérin. *Approximation fractale de courbes et de surfaces*. Thèse de doctorat, Université Claude Bernard Lyon 1, December 2002.
7. Eric Guérin and Eric Tosan. Fractal inverse problem: an analytical approach. Research report RR-2004-005, LIRIS, January 2004. submitted to Fractals.
8. Eric Guérin, Eric Tosan, and Atilla Baskurt. Fractal coding of shapes based on a projected IFS model. In *ICIP 2000*, volume II, pages 203–206, September 2000.
9. Eric Guérin, Eric Tosan, and Atilla Baskurt. A fractal approximation of curves. *Fractals*, 9(1):95–103, March 2001.
10. Eric Guérin, Eric Tosan, and Atilla Baskurt. Fractal Approximation of Surfaces based on projected IFS attractors. In *Proceedings of EUROGRAPHICS'2001, short presentations*, 2001.
11. Eric Guérin, Eric Tosan, and Atilla Baskurt. Modeling and approximation of fractal surfaces with projected IFS attractors. In M. M. Novak, editor, *Emergent Nature*. World Scientific, 2002.
12. Eric Guérin, Eric Tosan, and Atilla Baskurt. Fractal Compression of Images with Projected IFS. In *PCS'2003, Picture Coding Symposium, St Malo*, April 2003.
13. A E Jacquin. Image coding based on a fractal theory of iterated contractive image transformations. *IEEE Trans. on Image Processing*, 1:18–30, January 1992.
14. Evelyne Lutton, Jacques Lévy-Véhel, Guillaume Cretin, Philippe Glevarec, and Cédric Roll. Mixed IFS : resolution of the inverse problem using genetic programming. *Complex Systems*, 9(5):375–398, 1995.
15. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C : The Art of Scientific Computing*, chapter Nonlinear Models. Cambridge University Press, 1993.
16. Z R Struzik, E H Dooijes, and F C A Groen. The solution of the inverse fractal problem with the help of wavelet decomposition. In M M Novak, editor, *Fractals reviews in the natural and applied sciences*, pages 332–343. Chapman and Hall, February 1995.
17. Edward R. Vrscay and Dietmar Saupe. Can one break the collage barrier in fractal image coding. In Dekking, Vehel, Lutton, and Tricot, editors, *Fractals : theory and applications in engineering*, pages 307–323. Springer, 1999.
18. Chems Eddine Zair and Eric Tosan. Fractal modeling using free form techniques. *Computer Graphics Forum*, 15(3):269–278, August 1996. EUROGRAPHICS'96 Conference issue.
19. Chems Eddine Zair and Eric Tosan. Computer Aided Geometric Design with IFS techniques. In M M Novak and T G Dewey, editors, *Fractals Frontiers*, pages 443–452. World Scientific Publishing, April 1997.
20. Chems Eddine Zair and Eric Tosan. Unified IFS-based Model to Generate Smooth or Fractal Forms. In A. Le Méhauté, C. Rabut, and L. L. Schumaker, editors, *Surface Fitting and Multiresolution Methods*, pages 335–344. Vanderbilt University Press, Nashville, TN, 1997.

Index