Anne-Maria Ernvall-Hytönen
Camilla Hollanti
(Eds.)

Proceedings of the 3rd Nordic EWM Summer School for PhD Students in Mathematics

Turku Centre *for* Computer Science

TUCS General Publication
No 53, June 2009

Proceedings of the 3rd Nordic EWM Summer School for PhD Students in Mathematics

Editors:

Anne-Maria Ernvall-Hytönen

Camilla Hollanti

TUCS General Publication
No 53, June 2009

Proceedings of the 3rd EWM Summer School for PhD Students in Mathematics

Preface

When the first preparations for the third EWM summer school for PhD students in mathematics were made, nobody really knew if anybody would attend. When the contribution deadline got nearer and nearer, we still did not know if we would get any contributions.

We were pleasantly surprised when our email inboxes started to fill with articles from around the world and from different mathematical areas. We were even happier when we learned that the quality of the contributions was very high. Although, truth to be told, sometimes we were completely at a loss at first when we needed to find a suitable referee.

As in all cases, only a part of the contributions could be published. Those are found in this book. We hope you enjoy reading the result.

The financial help from our sponsors cannot be thanked enough. Without their support this event would still be just a dream. Hence, sincere thanks go to the Finnish Cultural Fund, Otto A. Malm Foundation, Nokia, TUCS, Oskar Öflund Foundation, Finnish Centre of Excellence in Analysis and Dynamics Research, University of Turku, Google, European Women in Mathematics, and European Mathematical Society.

We would like to thank the Scientific committee for valuable opinions and scientific help. The help from organizing committee, local committee and other volunteers is gratefully acknowledged. Especially, we would like to thank Eeva Suvitie, Aasa Feragen, Anna Laine and Toni Ernvall.

Anne-Maria Ernvall-Hytönen and Camilla Hollanti,
editors of the proceedings

Scientific committee

Contents

The determination of the optimal sensors' location using genetic algorithm **5**
Astrakova, A.S; Bannikov, D.V; Cherny, S.G. and Lavrentiev, M.M. Jr.

Gender and financial mathematics: evidence from mutual fund performance in one of the emerging markets **23**
Azmi, R.

Introduction to algebraic number theory **34**
Bayer-Fluckiger, E.

Mathematics and gender studies: an overview **41**
Blunck, A.

Teaching mathematics and gender at the university **42**
Blunck, A.

On the evolutionary dynamics of virulence **43**
Boldin, B.

Local and global class field theory **61**
Coates, J. and Ramdorai, S.

Function theories in higher dimensions **62**
Eriksson, S.-L.

Some generalizations of trigonometric functional equations **63**
Fechner, Ż.

Women in mathematics in France **69**
Guillopé, C.

Modeling invasions and calculating establishment success chances **72**
Haccou, P.

What should maths teachers know about girls and boys? **73**
Hannula, M.

Curves of genus 2 on rational normal scrolls **75**
Hofmann, A.

On inequalities in borderline cases **84**
Hurri-Syrjänen, R.

On Sobolev spaces **85**
Hurri-Syrjänen, R.

Hausdorff measures and dimensions **115**
Järvenpää, M.

Continued fractions **123**
Lorentzen, L.

The Yamabe problem with singularities **156**
Madani, F.

Hausdorff dimensions of Good sets and strict Jarník sets for Fuchsian groups with parabolic elements **168**
Munday, S.

Methods for symmetric key cryptography and cryptanalysis **186**
Nyberg, K.

A glance at hyperbolic function theory in the context of geometric algebras: hypergenic operators **187**
Eriksson, S.-L. and Orelma, H.

Interaction of two charges in a uniform magnetic field: symmetries, reduction and non-integrability of the planar problem **198**
Pinheiro, D. and MacKay, R. S.

On an algorithm for factoring natural numbers **204**
Rubtsova, R.

Girls and boys and equity in mathematics: teachers' beliefs **207**
Soro, R.

On entire solutions of some inhomogeneous linear differential equations in a Banach space **210**
Gefter, S. and Stulova, T.

Collective animal behaviour: coming together **214**
Sumpter, D.

Collective animal behaviour: moving together **229**
Sumpter, D.

Evolution of body condition-dependent dispersal under kin competition **246**
Utz, M.

Implementation models of ICT in teaching of complex numbers **251**
Vrdoljak, A.

The determination of the optimal sensors' location using genetic algorithm

A.S. Astrakova [*], *D.V.Bannikov*[†], *S.G. Cherny*[‡] *and M.M. Lavrentiev, Jr.*[§]

Abstract

We consider the problem of placing sensors optimally for the earliest detection of tsunami waves. It is necessary to record a tsunami wave from an arbitrary point of a subduction zone as soon as possible, using a given number of sensors. We use a genetic algorithm to solve this problem. Wave travelling times were calculated from the linear approximation of a shallow water model. The proposed computational algorithm was verified on problems that can be solved also analyticaly. The real-life problem of optimal sensor placement on the Alaska-Aleutan subduction zone is solved.

1 Introduction

At the time being, mankind is not able to avoid most of natural disasters. However, problem of risk assessment and damage mitigation could be effectively solved. Catastrophic tsunami waves are definitely among the most dramatic disasters. Each additional minute in tsunami wave warning may save many lives in inundation coastal areas. Most of tsunamis (up to 85%) are caused by the strong underwater earthquakes. Majority of these earthquakes take place in subduction zones, where the Earth crust is penetrated into the mantle. One of the modern technologies for reporting tsunami is based on deep water bottom pressure recorders (BPR). Obviously, time required for tsunami wave detection depends on number and location of these BPR's. In this paper problem to find both optimal locations and minimal necessary number of such BPR's is solved with the help of so-called genetic optimization algorithm. Celebrated shallow water model is used for wave propagation simulation.

2 General Problem Statement

2.1 Basic notions

Let $\mathbf{\Omega}$ be domain with a parts of aquatoria (negative depth parameter $h < 0$, dry land ($h > 0$), and subduction zone $\mathbf{P}$. Domain $\mathbf{D}$ stands for the part of $\mathbf{\Omega}$, where pressure recorders could be displayed. The problem is: to determine locations for the given number L of BPR's such that seismic event from any point of $\mathbf{P}$ will be detected (wave is passed over at least one of BPR's) after minimal possible time. Without loss of generality we suppose $\mathbf{\Omega}$ rectangular (fig. 1), approximated by the mesh $\omega_h = \left\{(n,m); n = \overline{1,N}, m = \overline{1,M}\right\}$ with the given depths distribution h. Subduction zone $\mathbf{P}$ is determined at the same mesh by points $\{\mathbf{p}_j\}_{j=1}^{P}$, where $\mathbf{p}_j = \{x_j, y_j\} \in \omega_h$. Coordinates of each from L BPR's $\mathbf{q}_i = \{x_i, y_i\}$ are belonging to $\mathbf{D}$. Domain $\mathbf{D}$ is arbitrary subset, possibly coinciding with $\mathbf{\Omega}$. By configuration $\mathbf{Q} = \{\mathbf{q}_1, ..., \mathbf{q}_L\}$ we understand selected set of L BPR's, representing suggested solution to considered problem.

Suppose that disturbance (tsunami source) arises at point $\mathbf{p}_j \in \mathbf{P}$. This disturbance propagates over "water" part of domain $\mathbf{\Omega}$ with certain speed (depending on point). Velocity distribution determines traveling time from source to each point (fig. 2). We are interested in minimal time, required for disturbance (wave) to approach any "water" point $\mathbf{x}$ from the given source $\mathbf{p}_j$. Let γ be one of the ways, connecting points $\mathbf{p}_j$ and $\mathbf{x}$, and τ_γ be wave traveling time along this way. As $\tau(\mathbf{p}_j, \mathbf{x})$ we denote minimal time to approach $\mathbf{x}$ from $\mathbf{p}_j$**:**

$$\tau(\mathbf{p}_j, \mathbf{x}) = \min_{\gamma} \tau_\gamma. \tag{1}$$

Fig. 3 displays trajectories, along which wave is approaching points by minimal time, started from selected point at subduction zone.

[*]Novosibirsk State University, Novosibirsk, Russia
[†]Novosibirsk State University, Novosibirsk, Russia
[‡]Institute of Computational Technologies, Siberian Branch of Russian Academy of Science
[§]Novosibirsk State University, Novosibirsk, Russia

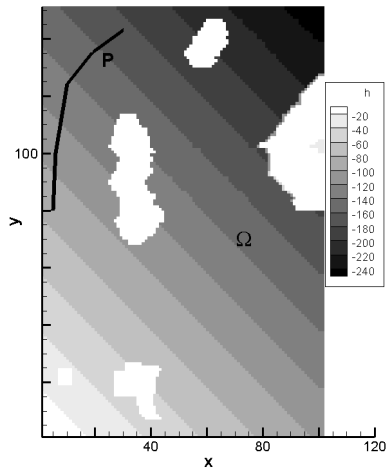


Fig. 1. Domain $\mathbf{\Omega}$ under consideration, which includes water area with depth $h < 0$, dry land ($h > 0$), and subduction zone $\mathbf{P}$. Intensity of grey color is proportional to water depth.

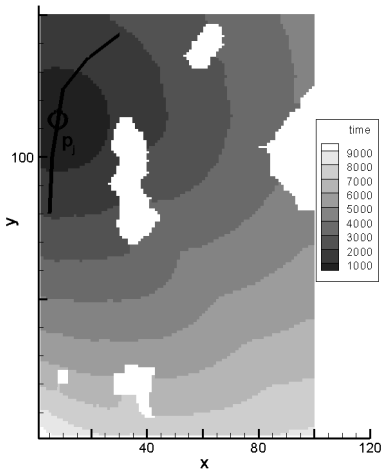


Fig. 2. Level lines of equal time wave propagation, caused by event at p_j.

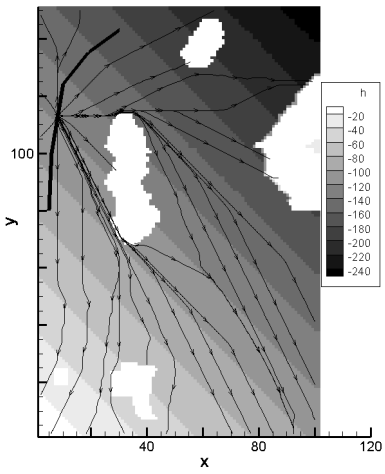


Fig. 3. Trajectories of disturbance propagation.

Time required for determination of tsunami wave (source at point $\mathbf{p}_j$) by configuration $\mathbf{Q}$ of L BPR's is calculated as follows

$$t(\mathbf{p}_j, \mathbf{Q}) = \min_{1 \leqslant i \leqslant L} \tau(\mathbf{p}_j, \mathbf{q}_i). \tag{2}$$

Guaranteed tsunami registration time from any point $\mathbf{p}_j \in \mathbf{P}$by configuration $\mathbf{Q}$ is nothing but

$$T(\mathbf{Q}) = \max_{1 \leqslant j \leqslant P} t(\mathbf{p}_j, \mathbf{Q}). \tag{3}$$

2.2 Statement of Optimization Problem

Mathematical statement of problem to determine optimal configuration $\mathbf{Q}$ (we speak about minimization of guaranteed time for wave registration from any point of $\mathbf{P}$) is formulated as follows:

Find configuration $\mathbf{Q} = \{\mathbf{q}_1, ..., \mathbf{q}_L\}$, which bring minimal value to goal functional $T(\mathbf{Q})$ in (3):

$$\min T(\mathbf{Q}) \tag{4}$$

where number L is given and phase restrictions are

$$\mathbf{Q} \in \mathbf{D}. \tag{5}$$

2.3 Wave propagation model

We consider linear approximation of shallow water equations [2]. In this case wave velocity is proportional to square root of depth h at each point:

$$v = \begin{cases} \sqrt{-gh}, \ h < 0, \\ 0 \ , \ h \geqslant 0. \end{cases} \tag{6}$$

Software application for fast calculation of traveling time $\tau(\mathbf{p}_j, \mathbf{x})$ between any two given points of aquatoria has been kindly provided by Prof. An.G. Marchuk (ICM&MG SB RAS, Novosibirsk) [3].

Examples of application for both shallow water model and traveling time calculation for selected bathymetry are given in fig.'s 2, 3.

3 Method of Solution

Suggested method of solution is based on Genetic Algorithm (GA). This traces evolution of population of individuals according to recombination and mutation, being selected under certain selection criteria. In

application to problem of determination of configuration $\mathbf{Q}$ from L BPR's, particular configuration $\mathbf{Q}$ stands for the individual. We will also refer as individual the set of coordinates

$$(x_1, y_1, x_2, y_2, ..., x_i, y_i, ..., x_L, y_L), \tag{7}$$

determining all BPR's from $\mathbf{Q}$.

3.1 Scheme of Algorithm

General scheme for GA is given in fig. 4. This consists from the following steps.

1. Initial population $\mathbf{Q}_1^s, ..., \mathbf{Q}_k^s, ..., \mathbf{Q}_p^s$ $(s = 0)$ is composed from $p = p_0$ individuals. Each individual is nothing but the set of numbers (7), which determine configuration being optimized under restrictions $\{x_{ki}^s, y_{ki}^s\} \in \mathbf{D}$, $k = 1, ..., p$, $i = 1, ..., L$. Initial population is generated randomly with respect to all parameters, satisfying (5). For better convergence is very favorable to have in initial population "nearly optimal" individual. Later we discuss how to fulfill this requirement.

2. Value of goal function $T(\mathbf{Q}_k^s)$ is calculated for each individual from configuration.

3. Selection of $Tr \cdot p$ best individuals (with smaller values of the goal function T). Then, by recombination and mutation new $\mathbf{Q}_1^{s+1}, ..., \mathbf{Q}_k^{s+1}, ..., \mathbf{Q}_p^{s+1}$ is constructed. Size of population does not increase$p \leqslant p_0$.

4. Return to the step 2, unless the desired number of generations N_{gen} is already reached.

Number of generations N_{gen} is determined by approaching of global minimum, that is absence of valuable change in minimal value of goal function from generation to generation.

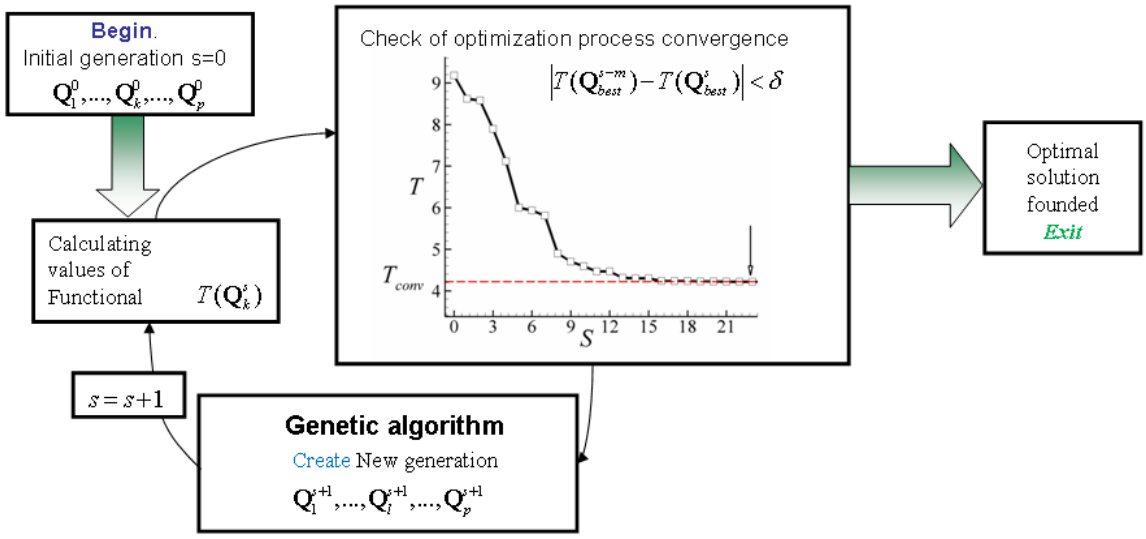


Fig. 4. Scheme of optimization process.

3.2 Operations of Genetic Algorithm

Let describe operations of GA in better details.

3.2.1 Selection

Certain fixed number $Tr \cdot p$ $(0 < Tr < 1)$ of individuals, possessing the smallest values of goal function, is selected. Rest of generation excluded from account.

3.2.2 Recombination

After selection, two parent individuals are randomly chosen, $\mathbf{Q}' = (x'_1, y'_1, x'_2, y'_2, ..., x'_L, y'_L)$ and $\mathbf{Q}'' = (x''_1, y''_1, x''_2, y''_2, ..., x''_L, y''_L)$. They "produce" new individual $\mathbf{Q}^{new} = (x_1^{new}, y_1^{new}, ..., x_L^{new}, y_L^{new})$ by recombination. This consists of sorting out BPR's $\mathbf{q'}_i$ and $\mathbf{q''}_i$, $i = 1, ..., L$ from both configurations $\mathbf{Q}'$ and $\mathbf{Q}''$ in order to determine new BPR $\mathbf{q}_i^{new} = (x_i^{new}, y_i^{new})$. Coordinates of this new one are (fig. 5):

$$x_i^{new} = \alpha_{x,i} x'_i + (1 - \alpha_{x,i}) x''_i, \;\; y_i^{new} = \alpha_{y,i} y'_i + (1 - \alpha_{y,i}) y''_i \tag{8}$$

where $\alpha_{x,i} \in RAND(-d,\ 1{+}d)$, $\alpha_{y,i} \in RAND(-d,\ 1{+}d)$. Quantity $d > 0$ is called *recombination parameter*. Note that both $\alpha_{x,i}$ è $\alpha_{y,i}$ could be greater than unit and less than zero. Thus, recombination extrapolates some individuals. Operation is applied $(1 - Tr) \cdot p$ times to obtain generation of p individuals. Scheme of BPR creation is presented in fig. 5.

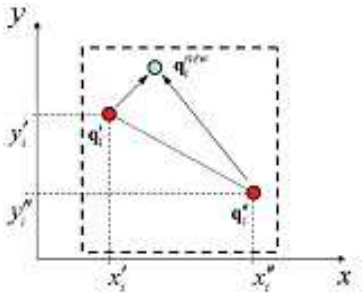


Fig. 5. Recombination.

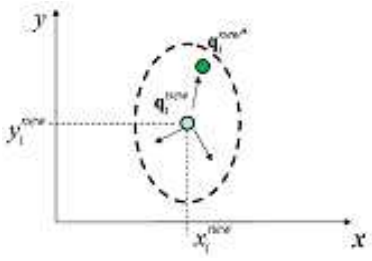


Fig. 6. Mutation.

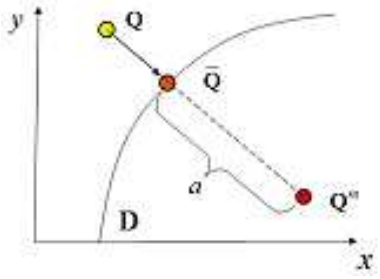


Fig. 7. Projection to domain **D** of configuration from one BPR.

3.2.3 Mutation

All individuals, obtained by selection and recombination (except the best one), are subject of slight variation according to:

$$z_i^{new*} = z_i^{new} \pm \mu (z_{R,i} - z_{L,i}) \delta, \;\; i = 1, ..., L. \tag{9}$$

By symbol z in (9) is understood any of coordinates x or y of i-th BPR from "mutating" configuration; $\mu \in [0, 1]$ - mutation parameter, $\delta = 2^{-16\gamma}$, γ - random value from $[0, 1]$ interval, values $z_{R,i}$ and $z_{L,i}$ are determined by phase restrictions. Mutation stretches individuals to wider domain. This increases the possibility to obtain global minimum. Scheme of mutation is given in fig. 6.

Preserving the best individual from each generation is called *cloning*. This is required as the best one could not be lost by recombination and mutation.

3.2.4 Projecting

New generation is constructed in such manner that restrictions in (5) are fulfilled. In case individual $\mathbf{Q}$ does not satisfy them, it is substituted with the new one $\bar{\mathbf{Q}}$, obtained by projection of $\mathbf{Q}$ to domain $\mathbf{D}$. This operation is executed as follows: $\bar{\mathbf{Q}} = \mathbf{Q} + a(\mathbf{Q}^m - \mathbf{Q})$, where $\mathbf{Q}^m$ stands for best individual from previous generation (having minimal value of goal function). Constant $a \in [0, 1]$ is picked up such that the worse BPR from configuration $\mathbf{Q}$ match the boundary of $\mathbf{D}$, while all the rest be inside it.

Schematically projection (for configuration from one BPR) is shown in fig. 7.

3.2.5 Stop Criteria

New generation are constructed until N_{gen} ones will be calculated. Stop criteria is simple – absence of appreciable change in goal function (3) from generation to generation. Using preliminary runs number of generations N_{gen} is determined. Individual from the last generation, processing minimal value of goal function, is regarded as problem solution.

In this study we did use the following parameters in GA: $d = 0.7$, $\mu = 0.1$, $Tr = 0.3$. These values have been determined through numerical experiments with test problems (see section 4 for details). The best performance of GA is observed for these values of parameters.

4 Numerical Experiments

In order to verify and calibrate implementation of the proposed numerical algorithm, several experiments have been done with simulated domain geometries, bathymetric and subduction zones shapes. Flat and variable depth profiles have been tested, areas with positive h (islands) were considered. Subduction zones were shaped as circle segments and intervals to compare numerical results with analytic solutions.

4.1 Constant depth and subduction zone is semicircle

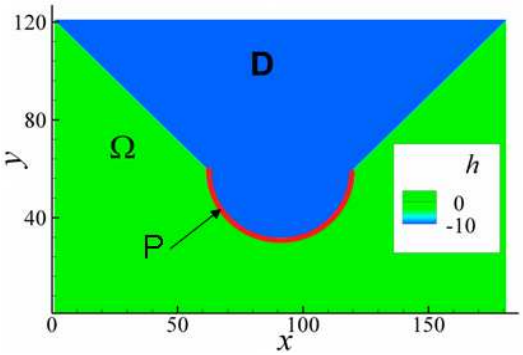


Fig. 8. Aquatoria geometry and depth profile.

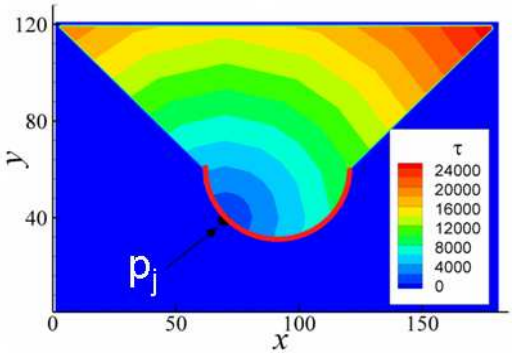


Fig. 9. Isochrones $\tau = const$.

Rectangular aquatoria $\mathbf{\Omega}$, subduction zone $\mathbf{P}$ as semicircle, and phase restrictions $\mathbf{D}$ from the first series of numerical tests is shown in fig. 8. Water area with constant depth $h = -10$ m takes exactly domain $\mathbf{D}$. Rest of domain $\mathbf{\Omega}$ is supposed to be dry land. Time isochrones for wave (caused by seismic event at point propagation $\mathbf{p}_j$ from subduction zone $\mathbf{P}$) over $\mathbf{D}$ is depicted in fig. 9.

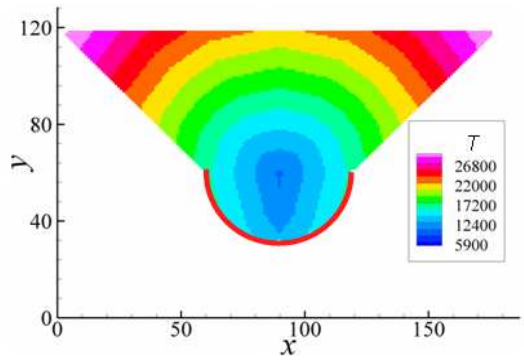


Fig 10. Traveling time $T(\mathbf{q})$ distribution for one station.

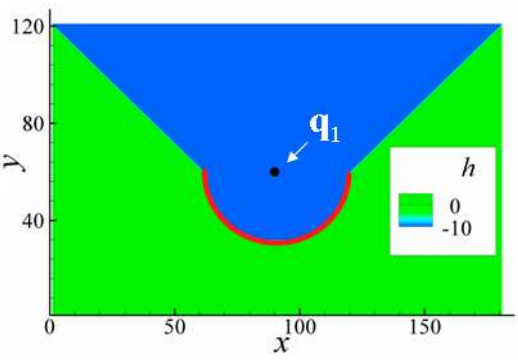


Fig. 11. Problem solution for one BPR.

4.1.1 Optimal location for one station

For almost trivial case then configuration $\mathbf{Q}$ consists of one BPR ($L = 1$), value of goal function T in (3) is easily calculated for all points of "water" domain $\mathbf{D}$. Fig. 10 presents distribution of values of Ó in $\mathbf{D}$. Obvious optimal position of BPR, providing minimal value for goal function T, is nothing but center of the circle. This result was numerically obtained for (4), (5) (see fig.11).

4.1.2 Optimal location for several stations

Numerical tests with two BPR's.

Evolution of Generations

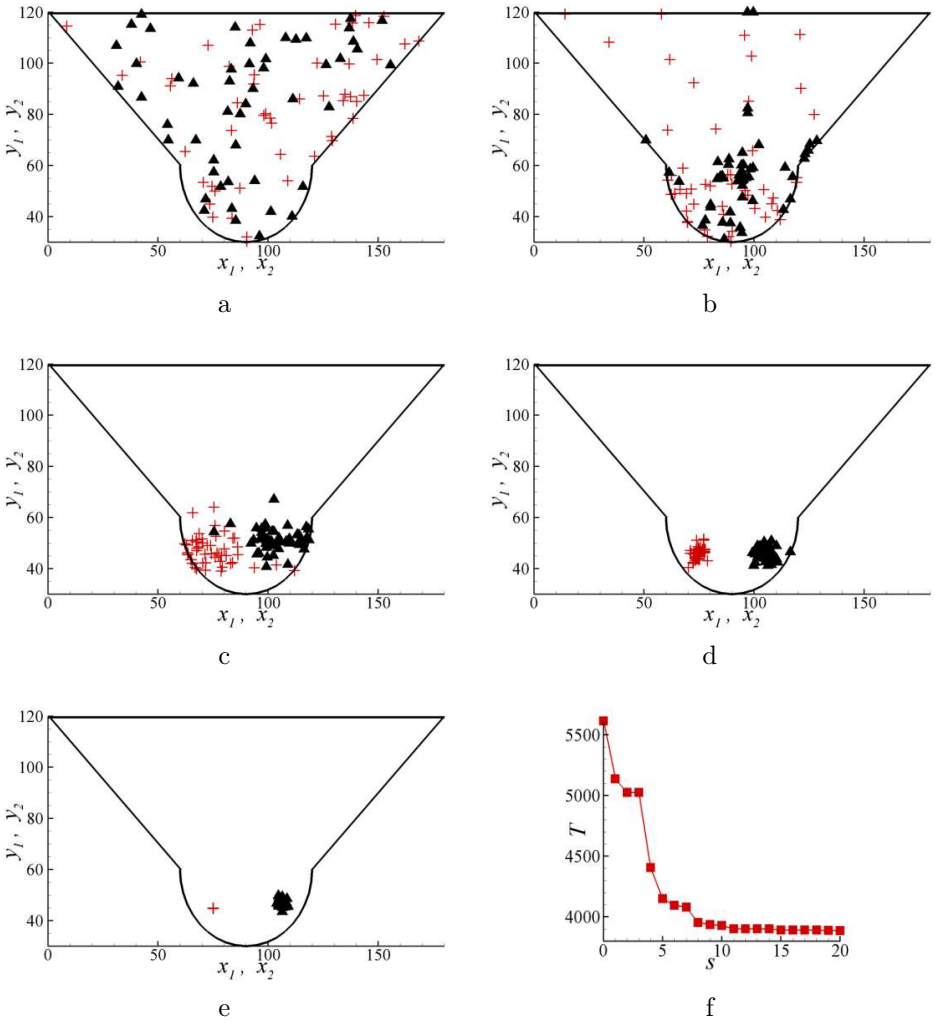


Fig. 12. Evolution of configurations from two BPR's (▲ - first and + - second BPR, respectively). Number of generation s is presented: (a) – s =1; (b) – s =3; (c) – s =5; (d) – s =11; (e) – s =20). Figure (f) visualizes convergence of the goal function T.

Configurations convergence for the case of two BPR's, along with dynamics of goal function T are given in fig. 12. Number of individuals in generation p=50 was used. All individuals are shown in fig. 12.

Impact p – size of generation – on problem solution

For the small size p of generation individuals with low values of T (however, being far from global minimum) may provoke absence of alternative configurations. This may prevent appearance of better solutions in the sequel generations. As a result, evolution will lead to local minimum, which is shown in fig. 13, where results for generations of different size are displayed. For each size five numerical tests were performed (this is also indicated in fig. 13). For large size individuals from new generation have wider distribution of parameters, compared to smaller ones. However, larger size of generation causes proportional increase in CPU resources (time required for calculation). Thus, number of individuals should be balanced to achieve both converge with global minimum and reasonable performance. This population-sizing problem attracts many attention in literature [4, 5, 6].

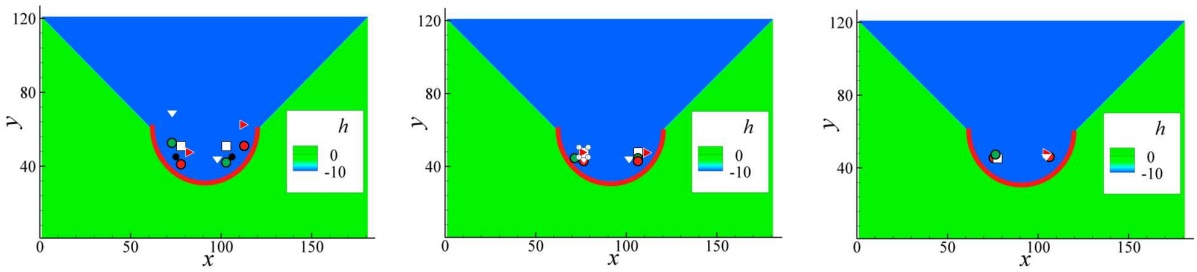


Fig. 13. Locations of obtained optimal positions for 5 different runs of optimization code: (a) - p=20; (b) - p=40; (c) - p=80.

Convergence histories for generations of the same sizes for the best (with respect to value of T) individuals of each generation is shown in fig. 14. Similarly as it has been done in fig. 13, for each p convergence of five numerical runs are presented. In all cases obtained solutions provide slightly higher values of T, compared to exact solution. This difference goes down then the size $p$increases. The larger is size p, the more generations are needed for convergence.

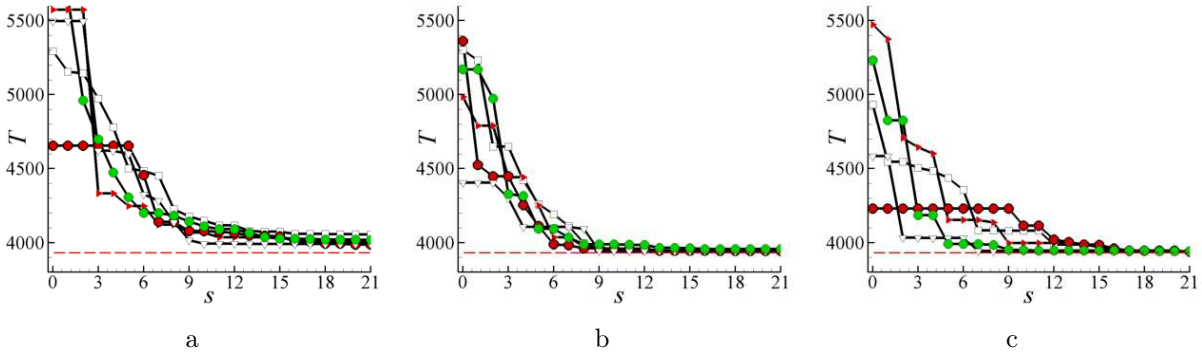


Fig. 14. Convergence histories for different generation sizes: (a) - p =20; (b) - p=40; (c) - p=80, 5 optimization restarts have been done in each case.

Numerical results for two and more BPR's

Fig. 15 demonstrates solutions to optimization problems for various values of BPR's number L. For each value of L five restarts have been done, best of the obtained solutions was chosen. Generation size was fixed at the level p =100.

For this flat bottom case exact solution is symmetric and therefore, could be easily calculated in analytic form. Comparison of calculated values of goal function T with its values for exact solution is given in fig. 16. Horizontal axis shows number of BPR's.

Parameters of GA – impact on convergence and precision

Calibration of the GA parameters has been obtained by numeric optimization of configuration with $L = 6$. Data of obtained numerical solution for various ratios between the number of individuals in initial generation

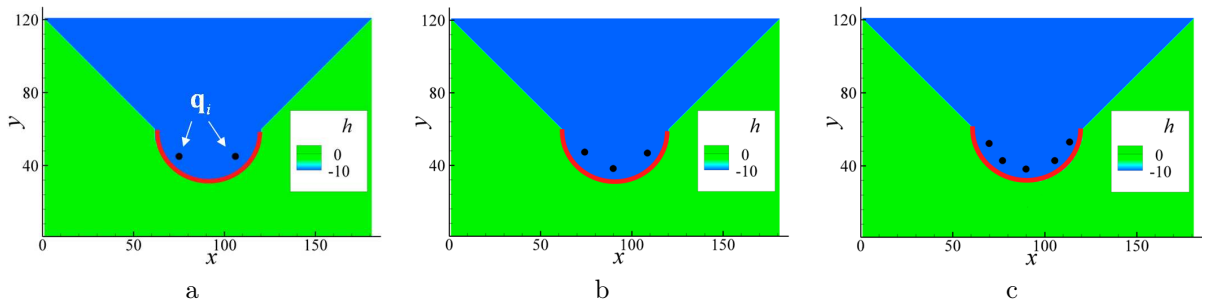


Fig. 15. Numerically obtained configurations for: (a) - L=2; (b) - L=3; (c) - L=5.

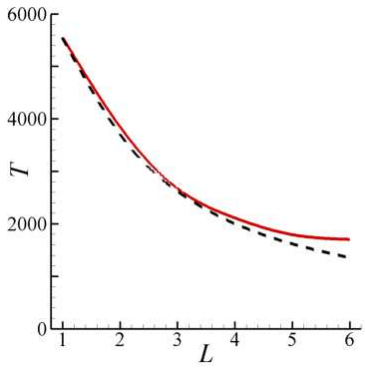


Fig. 16. Values of goal function T vs number of BPR's: solid line – numerical solution, dashed line – exact solution.

p_0 and number p of individual in the sequel generations are summarized in Table 1. Parameter N_{gen} indicates number of calculated generations, such that problem solution does not change for the following generations. Average time was calculated for five optimization restarts according to:

$$T_{average} = \frac{1}{n}\sum_{i=1}^{n} T_i,$$

where n – represents the number of optimization restarts, executed for each of cases considered; T_i- solution obtained after i-th restart; $T_{\min} = \min(T_1, ..., T_n)$. Parameter n takes the value ***5*** for p=100, 1000, 3000 and takes the value ***3*** for p=10000, 30000. Exact solution (time) in with case is T_{exact}=1355.

Table 1

Quantity	Values				
p_0 / p	300/100	3000/1000	30000/10000	10000/100	30000/100
N_{gen}	50	100	120	60	85
$T_{average}$	2458	2462	1419	2088	2288
T_{min}	2177	2167	1355	1683	2112

Convergence history for the best optimization restart with p_0=30000 and p=10000 is given in fig. 17. Dependence of GE performance on mutation parameter μ (9) has been studied, see Table 2. Here

$$\delta T = 0.5(T_{\max} - T_{\min}),$$

with T_{max}=max(T_1,...,T_n). Rest of GA parameters have been fixed as $d = 0.7$, $Tr = 0.3$, $p_0 = p = 100$. Better values of tsunami detection time was observed for small values of μ. Both these numbers $\mu = 0.2$ and $\mu = 0.1$ have been selected for the further numerical experiments. Corresponding columns in Table 2 are marked in bold.

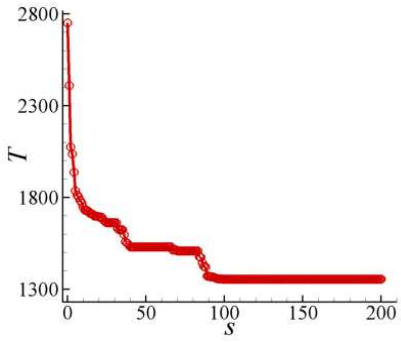


Fig. 17. Goal function convergence history for the best obtained solution with p_0=30000 and p=10000.

Table 2

Quantity	Values				
μ	**0.1**	**0.2**	0.3	0.5	1
N_{gen}	**50**	**70**	75	80	85
Taverage $\pm\, \delta T$	**2279±895**	**2203±786**	2320±774.5	2399±794	2305±578
T_{min}	**1355**	**1399**	1400	1696	1723

Solution dependence on recombination parameter d (8) for selected values of mutation μ is given in Table 3 for selection parameter $Tr = 0.1$ (section 3.2.1) and in Table 4 for $Tr = 0.3$. All these data were obtained with $p_0 = p = 100$ by 16 optimization restarts, $n = 16$.

Table 3

Quantity	Values for μ=0.1 and $Tr = 0.1$			Values for μ=0.2 and $Tr = 0.1$		
d	0.3	0.7	0.9	0.3	0.7	0.9
N_{gen}	50	70	70	50	95	100
Taverage $\pm\, \delta T$	2600±593	2434±620.5	2159±747	2570±466.5	2204±732	2238±597.5
T_{min}	2178	1858	1423	2112	1735	1723

As the value $\mu = 0.1$ corresponds to better numerical results, numerical study for $Tr = 0.5$ has been done only with $\mu = 0.1$, see Table 5.

There are almost equally effective sets of parameters $d = 0.7$, $\mu = 0.1$, $Tr = 0.3$ and $d = 0.9$, $\mu = 0.1$, $Tr = 0.3$ in table 4. But the first one has an advantage in the convergence rate: in this case 60 generations are necessary for convergence, the second one requires 70 generations.

Summing up we conclude that GA implementation is the most effective with the following parameters: $d = 0.7$, $\mu = 0.1$, $Tr = 0.3$. Corresponding column in Table 4 is marked in bold. Observation that increase in value of p supports global minima determination is effective.

4.1.3 Some properties of genetic optimization algorithm

For the cases of large enough number of stations, say $L \geqslant 5$, too small size of generation (e.g. $p \sim 100$) convergence with local minima can take place. Thus, solution obtained for $L = 8$ and $p = 100$ is presented in fig. 18. It is clear that only four of eight BPR's are displaced in correct positions, while the rest four ones are far from the exact solution.

The following three schemes have been proposed to avoid this phenomenon.

Increase the number p of individuals in generation

Result of this obvious factor on global minima approximation is given in fig.'s 13-14 for $L = 2$ and in Table 1 for $L = 6$.

Consequent problem solution increasing number of BPR's from 1 to L

Illustration of this scheme is given in fig. 19. For the case of L=8, p=100 transition from $L = 7$ to $L = 8$ is demonstrated. This transition is made by introducing into initial population for L=8 an individual, which

Table 4

Quantity	Values for $\mu = 0.1$ and $Tr = 0.3$			Values for $\mu = 0.2$ and $Tr = 0.3$		
d	0.5	**0.7**	0.9	0.5	0.7	0.9
N_{gen}	50	**60**	70	50	70	80
Taverage $\pm\, \delta T$	2259±789	**2279±895**	2108±832	2280±1016	2203±786	2181±949.5
T_{min}	1723	**1355**	1356	1399	1399	1399

Table 5

Quantity	Values for $\mu = 0.1$ è $Tr = 0.5$		
d	0.5	0.7	0.9
N_{gen}	300	400	60
Taverage $\pm\, \delta T$	2342±775	2089±792	2223±944
T_{min}	1399	1399	1416

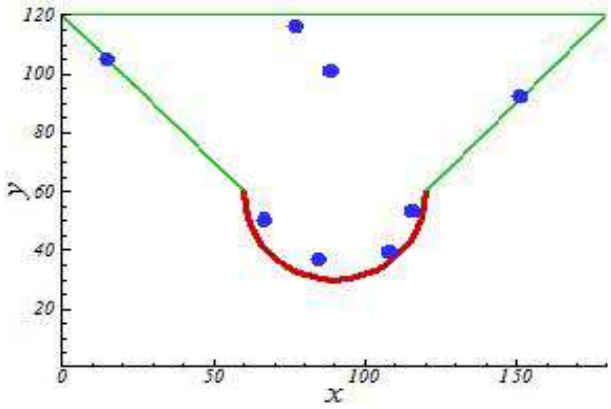


Fig. 18. Local minima obtained by GA with $L = 8$, $p = 100$.

is supposed to be close to global minima. Such individual was chosen by adding eighth BPR to optimal configuration, obtained for $L = 7$. One of possible addition is shown in fig. 19 (b).

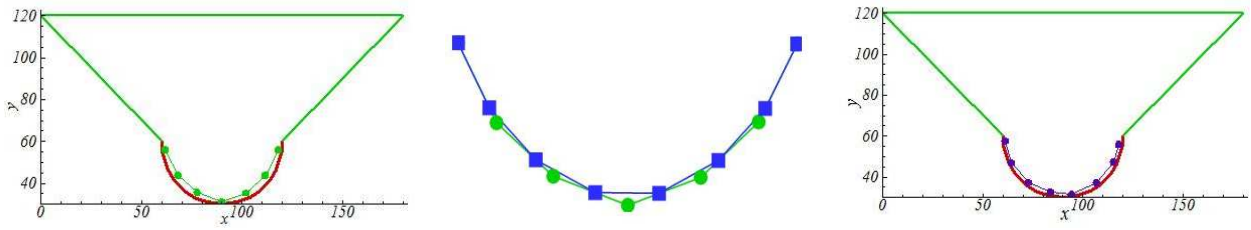


Fig. 19. Consequent problem solution for L=8: (a) optimal solution for 7 BPR's; (b) - Construction of "good" individual of 8 BPR's (squares) from the optimal individual of 7 BPR's (circles); (c) - obtained optimal solution (global minima) for 8 BPR's.

Insertion of "good" configuration to initial generation

In some cases "good" individual (close to global minima) for initial generation could be obtained for large L directly (no consequent solution for different L) by projection of subduction zone $\mathbf{P}$ onto the nearest boundary of domain $\mathbf{D}$. This idea is illustrated in fig. 20 for the case of Alaska-Aleutan subduction zone.

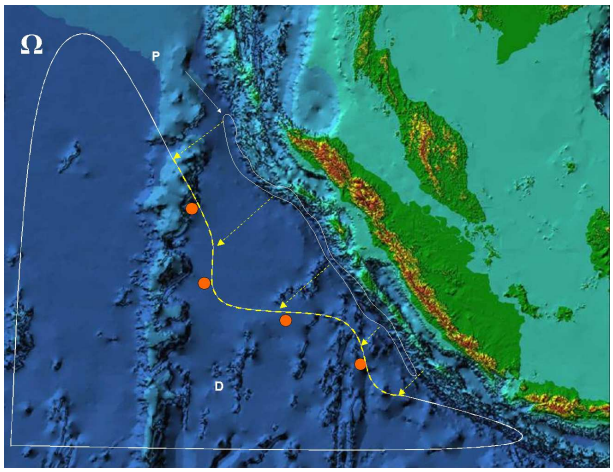


Fig. 20. Composition of "good" configuration for initial generation: BPR's are uniformly distributed along projection of zone $\mathbf{Đ}$ to the closest boundary of $\mathbf{D}$.

4.2 Constant depth, subduction zone is semicircle, presence of island in D

Now we modify the problem considered in section 4.1 by introducing the circle island into domain $\mathbf{D}$, see fig. 21. Fig. 21 displays the distribution of our goal function T for $\mathbf{D}$ in case of one BPR. Clearly, exact solution in this case, providing global minimal value of T, is the lowest point of the island. This exact solution has been reproduced in numerical experiment, see fig. 23 (a). Fig. 23 (b) displays numerical solution for 2 BPR's.

4.3 Constant depth, subduction zone is combination of two semicircles

Geometry of aquatoria $\mathbf{\Omega}$ with subduction zone $\mathbf{P}$, being combination of two semicircles is given in fig. 24. Distribution of traveling times for tsunami source at point $\mathbf{p}_j$ is displaced in fig. 25.

Fig. 26 shows the distribution of values of goal function T in case of one BPR. Location of global minima is easily identified. This solution of optimization problem (4), (5) has been calculated, see fig. 27.

Fig. 28 demonstrates solutions of optimization problems for various value of L. Optimal detection time versus the BPR's number is drawn in fig. 29.

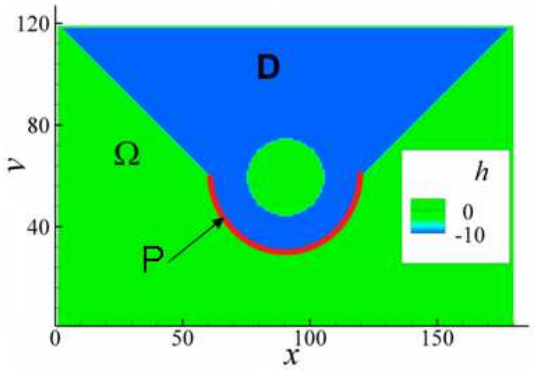


Fig. 21. Aquatoria geometry and relief of bottom.

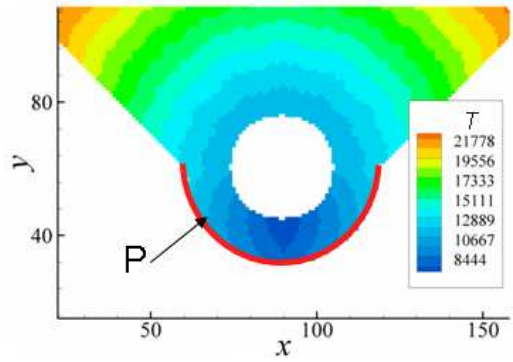
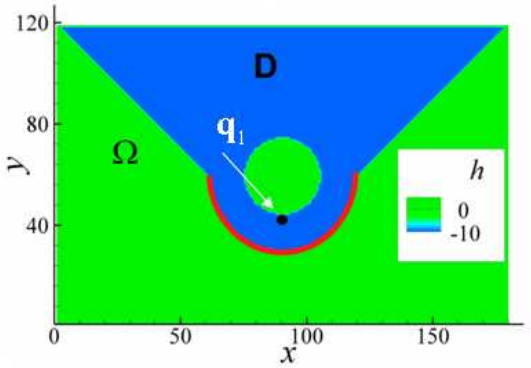
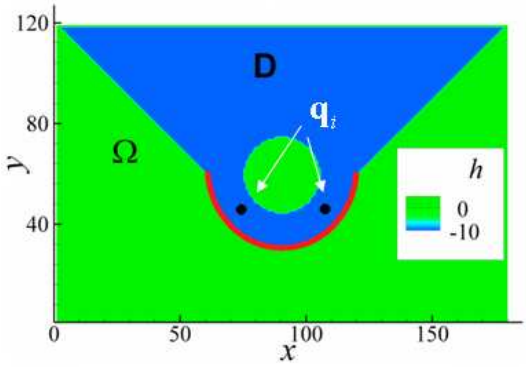


Fig. 22. Traveling time $T(\mathbf{q})$ distribution for one BPR.



a



b

Fig. 23. Optimal configurations numerically obtained for L=1 (a) and L=2 (b).

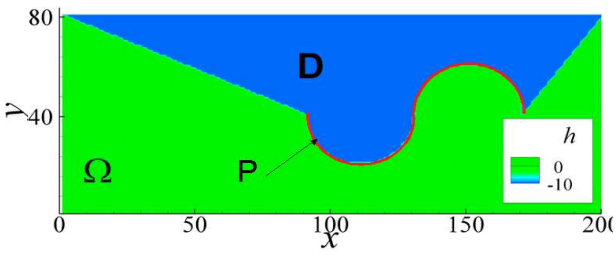


Fig. 24. Aquatoria geometry, uniform depth profile.

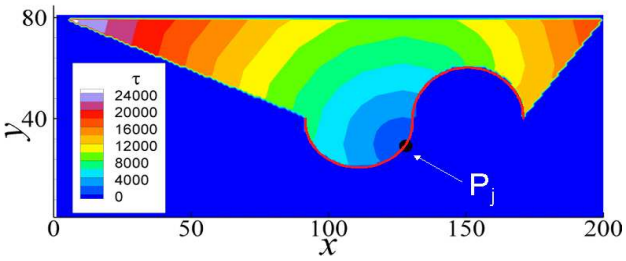


Fig. 25. Isochrones $\tau = const$.

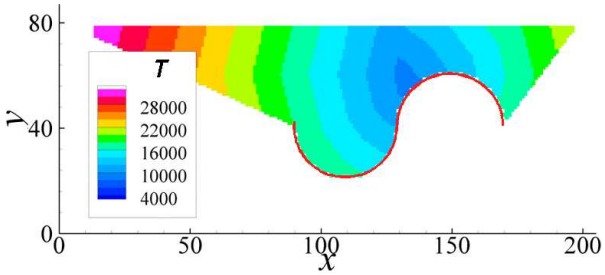


Fig. 26. Distribution of traveling time $T(\mathbf{q})$ for one BPR.

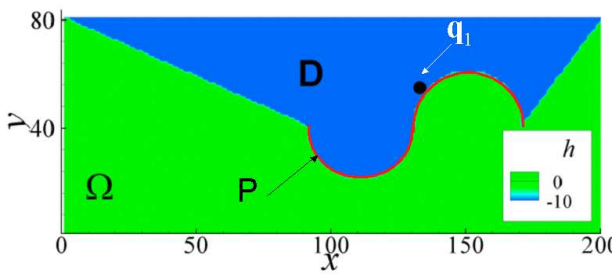
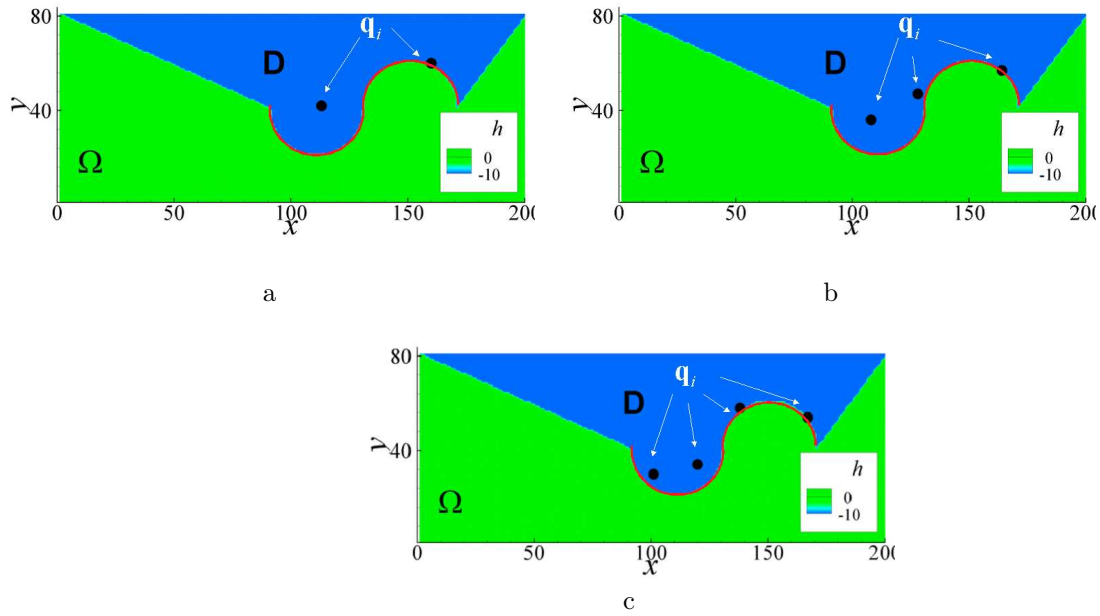


Fig. 27. Problem solution, numerically obtained for one BPR.



a b

c

Fig. 28. Numerically obtained optimal configuration for the cases: L=2 (a); L=3 (b); L=4 (c).

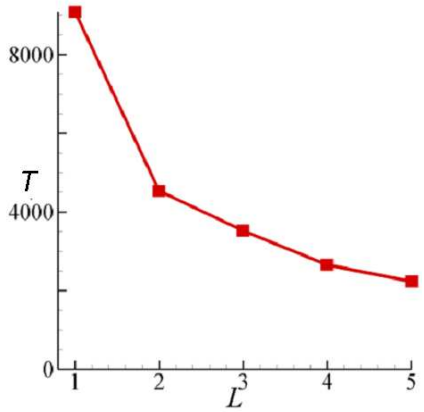


Fig. 29. Time of tsunami detection in dependence on the number of BPR's.

4.4 Variable depth, subduction zone is horizontal interval

Rectangular domain $\mathbf{\Omega}$ with variable depth profile, subduction zone coincides with "lower" boundary of $\mathbf{\Omega}$, and admissible for BPR's area $\mathbf{D}$ (shown by rectangular frame) are given in fig. 30. Traveling times from the source point $\mathbf{p}_j$ and trajectories to obtain given point by minimal time are shown in fig. 31.

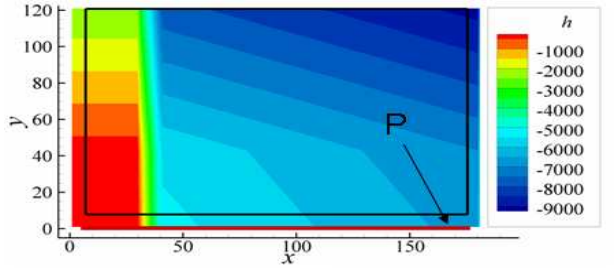


Fig. 30. Aquatoria geometry and depth profile.

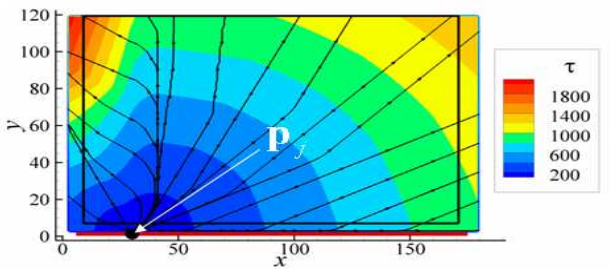


Fig. 31. Isochrones $\tau = const$ and trajectories of wave first arrival in given points.

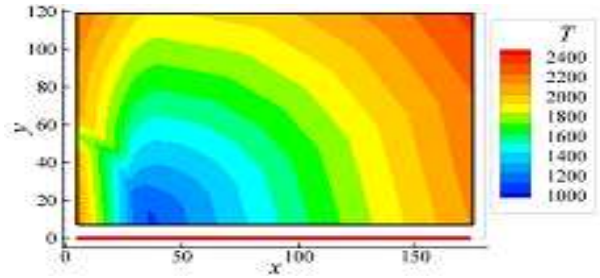


Fig. 32. Traveling time $T(\mathbf{q})$ distribution for one BPR.

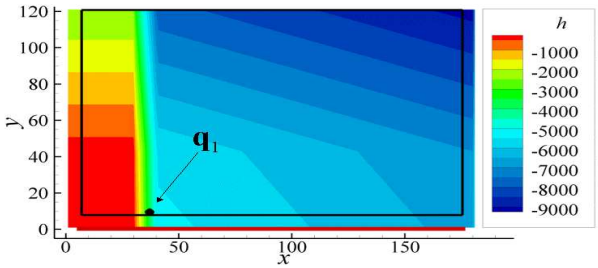
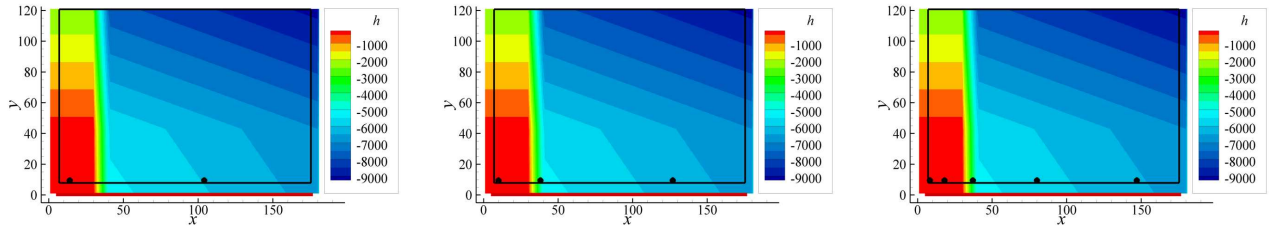


Fig. 33. Problem numerical solution for L=1.

Distribution of the goal function T for one BPR in domain $\mathbf{D}$ in presented in fig. 32. Optimal deployment of one BPR, which provide global minima for T, is clear. Result of numerical solution of optimization problem (4), (5), coinciding with the one of fig. 32, is shown in fig. 33.

Numerically obtained solutions to optimization problem for different number of BPR's are depicted in fig. 34. In each case five optimization restarts were executed. Best solutions have been chosen. Size of generation were fixed as $p = 100$. Fig. 35 demonstrates minimal (from five restarts) obtained values of goal function T versus number of BPR's.



a b c

Fig. 34. Numerically obtained configurations: $L = 2$ (a); $L = 3$ (b); $L = 5$ (c).

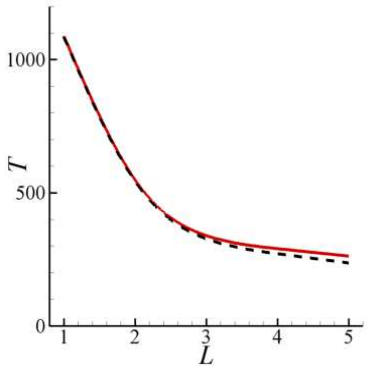


Fig. 35. Value of the goal function T vs number of BPR's: numerically obtained solution – solid line, exact solution – dashed line.

5 Algorithm application to real bathymetry at Alaska-Aleutan subduction zone

Real bathymetry around Alaska-Aleutan subduction zone **P** could be observed in fig. 36. White lines show the borders of domain **D** for BPR's deployment. Traveling times for tsunami wave are given in fig. 37.

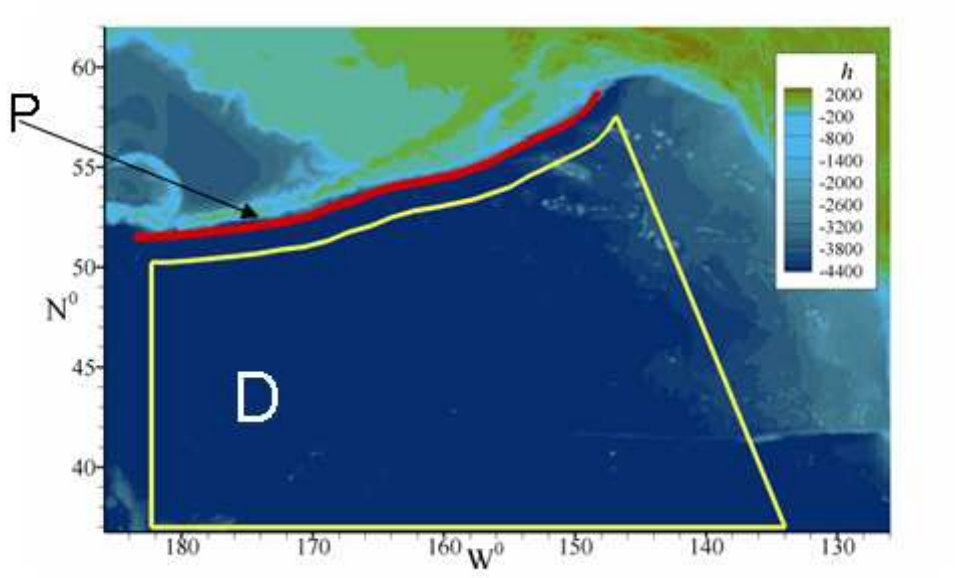


Fig. 36. Aquatoria geometry and depth profile.

Fig. 38 demonstrates solutions to optimization problem with different number L of BPR's. For each value of L five optimization restarts were executed, solutions with the best (minimal) values of time were plotted. Size of generations were fixed p=100. Locations of BPR's of NOAA DART buoys are given in fig. 39, marked with small white circles. In the same figure numerically obtained optimal solution is presented by black framed larger circles **O**. Decrease of time required for tsunami wave detection for larger number of BPR's is illustrated in fig. 40, obtained results are shown with squares. Black cross indicates this detection time for the current position of DART buoys.

Note that presented software application is able to solve the problem of optimal deployment of certain number of additional BPR's provided that locations of some buoys are already fixed.

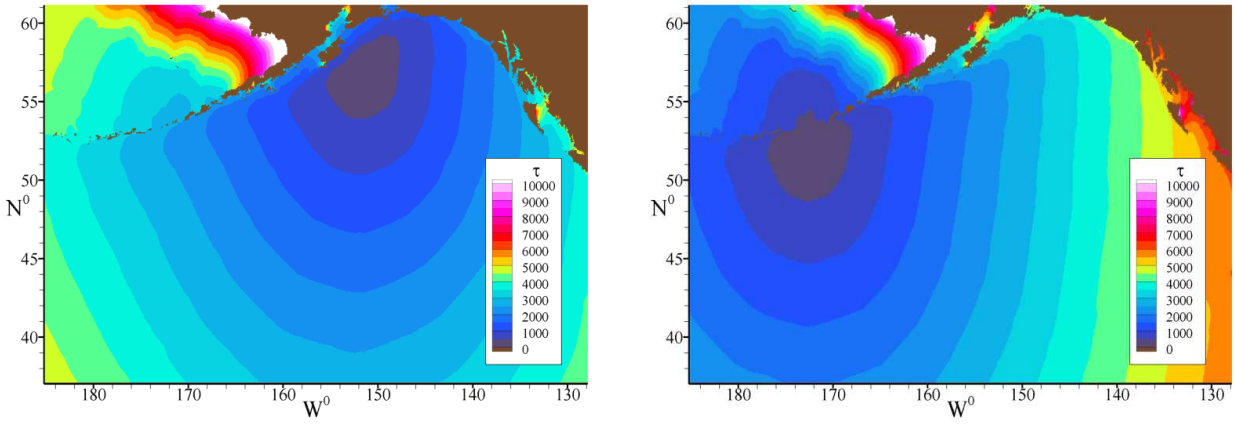


Fig. 37. Level lines $\tau = const$ for different sources.

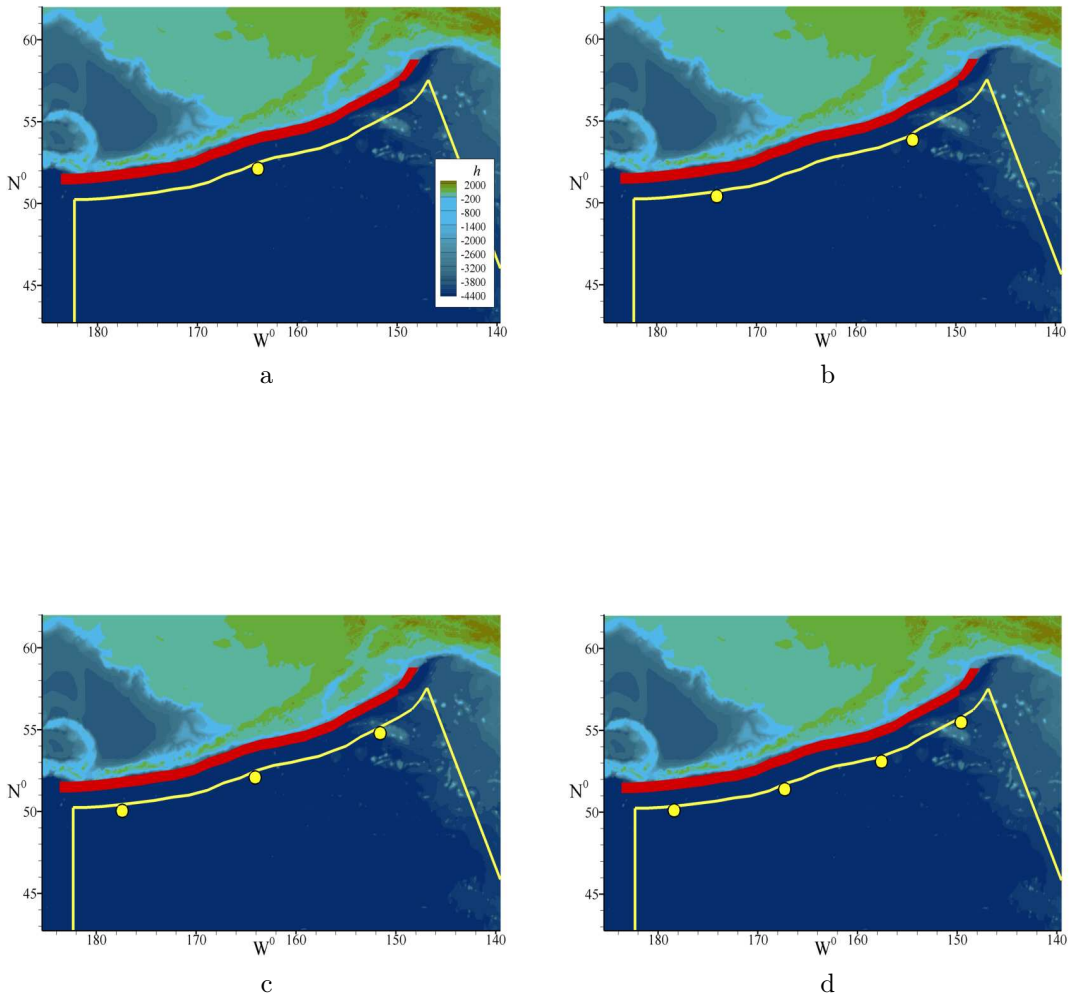


Fig. 38. Numerically obtained configurations for: $L = 1$ (a); $L = 2$ (b); $L = 3$ (c); $L = 4$ (d).

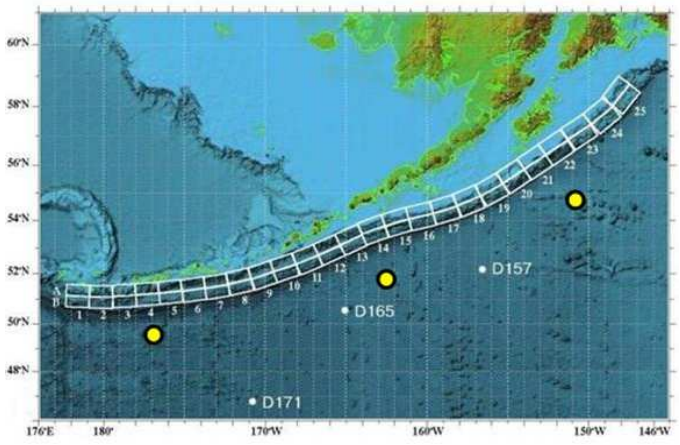


Fig. 39. Positions of existing BPR's – DART buoys of NOAA – are indicated by small white circles, numerically obtained optimal locations are shown with larger circles, framed in black (**O**).

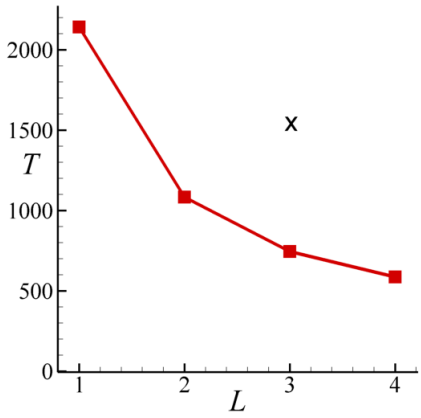


Fig. 40. Guaranteed time of tsunami wave detection for various number of BPR's: ■ – numerically optimized; **x** – for the deployed system.

6 Some data about developed software application

All program modules have been written as FORTRAN codes (GA and application to calculate traveling time between any two given points of aquatoria). Typical time for optimization is compared to 2 min on Pentium IV 3 GHz, compaq visual fortran 6.6 (N_{gen}=50, $p = 200$, $L = 6$).

There exists parallel implementation of GA FORTRAN code for multicomputer systems. All communications are implemented using MPI. Preliminary test calculations show near linear speedup up to 64 processors.

7 Discussion and conclusion

One of the methods of the optimal sensors location for the earliest warning of tsunami wave in defended centers of the Pacific region is considering in [1]. It consists in the consideration of six possible positions of BPR's and choosing among them those which in case of putting there the sensors will allow to save maximum of people in zones concerned. Braddock uses 18 representative tsunami generation regions and 27 representative population centers with known population sizes. For each generation region it's determined quantity of people that will be rescued for the certain set of BPR's. There for we compare tsunami travel time from the generation region to population center with the sum of tsunami travel time to BPR's, response time of the sensor (to signal and confirm generation of tsunami) and the reaction warning period at population center. This procedure takes place for all generation regions and all centers. Thus author introduces the functional expressing the ratio of part of people that will be rescue to the general number of population. Optimal location of minimum number of BPR's providing maximal value of the functional is the solution of this problem. This approach is associated with fixed location of BPR', so it can't guarantee minimal possible tsunami wave detection time.

In our case we consider solution of the minimal time detection problem for free BPR's location. It allows to ensure maximal possible time which is necessary for warning and evacuation and thereby it assures people's rescue from the disaster. Method of the genetic optimization have allowed to formulate and solve problem for whole searching domain.

Math statement of problem to optimize sensor network location in order to detect earlier tsunami wave, caused by earthquake within the given subduction zone, is proposed. Efficient numerical algorithm, based on GA, has been developed. Algorithm parameters have been calibrated in numerical tests. Problem of DART buoys optimization around Alaska-Aleutan subduction zone has been studied. The method can be applied to other real-life problems.

References

[1] Braddock R.D. Network properties of DEEP-SEA tsunami detectors. *22nd IUGG Intern. Tsunami Symposium, Chania*, 43–47, 2005.

[2] Pelinovsky E.N. Gydrodynamics of Tsunami Waves. *Nizhnii Novgorod, Institute of Applied Physics RAS*, 1996, 276 p.

[3] Marchuk An.G. A method for determination of wave rays in non-homogeneous media. *Bulletin of the Novosibirsk Computing Center. Series: Mathematical Problems in Geophysics*, (10):51–58, 2005.

[4] Deb K. Multi-Objective Optimization using Evolutionary Algorithms. *John Wiley & Sons.*, 2002, 497 p.

[5] Cantu-Paz E. Efficient and accurate parallel genetic algorithms. *Boston: Kluwer Academic Publishers*, 2000.

[6] Harik G., Cantu-Paz E., Goldberg D., Miller B. L. The gambler's ruin problem, genetic algorithms, and the sizing of populations. *Proc. Of the 4^{th} Intern.Conf. on Evolutionary Computation, New York*, 7–12, 1997.

Gender and Financial Mathematics: Evidence from Mutual Fund Performance in one of the Emerging Markets

*Rania A. Azmi**
University of Portsmouth, UK
rania.azmi@port.ac.uk, rania.a.azmi@gmail.com

Abstract

One of the key issues in financial mathematics is portfolio selection, in which fund managers are responsible for the performance of their mutual funds. Historically, there has been a claim that women perform worse, on average, than men do on mathematical tasks. In contrary to this claim, this paper finds the performance of women outperforming the men performance in mutual funds business, where financial mathematics models are crucial and highly utilized. In particular, this paper examines the determinants of mutual fund performance in one of the emerging markets (Egypt), with an emphasis on the factor of fund manager gender. This paper offers new insights into the Egyptian mutual fund industry. The results regarding the determinants examined (fund manager gender, in addition to fund age, size, objective, total risk, systematic risk, expenses ratio and type) show significant relation between fund's manager gender, expenses ratio, objective, type and total risk, and fund performance. For the common investor who wanted to invest in Mutual Funds in an emerging market like Egypt's market during a five-year period (January 1999 -December 2003), the selection criteria that could have provided the best results in selecting the fund are: a fund managed by a woman, an open end fund, with a growth objective and low expenses ratio. In contrast, the traditional selection criteria of size of the fund and its age appeared to be statistically irrelevant in this study. The result of having a relation between fund manager gender and fund performance in an emerging market clearly warrant future studies.

Key Words: Financial Mathematics, Gender, Mutual Funds.

1 Introduction to the Study of Gender, Financial Mathematics and Mutual Funds

Simon (2000) emphasises that the notion that mathematics is a masculine pursuit persists, however, he supports the commonsense view that no inherent characteristics of mathematics warrant excluding women and the same statement holds true for the sub-disciplines of mathematics. As times have changed, and more women are entering what were once exclusively masculine professions of mathematics and science; women are thriving.

Roth (2003) argues that there is a claim of significant gender differences in participation rates in mathematics education studies, and in related careers, besides the claim that women perform significantly worse, on average, than men do on mathematical tasks and activities.

One clear example of a profession which was traditionally dominated by men, which requires mathematical proficiency, is mutual fund management. The main duties of portfolio or mutual funds manager are to select securities to purchase, determine which ones to sell, and rebalance the portfolio in conjunction with the fund's buying, selling, contributions and redemption activity. Performing such duties relies in most of the cases on financial mathematics models. At the core of portfolio management activities is portfolio optimisation.

Portfolio optimisation is mainly focused on the determination of an optimal investment strategy on a financial market. The fund manager must determine how many shares of which securities to hold when, in order to maximise his/her utility of wealth at the end of the planning period. Portfolio optimisation involves modern methods of financial mathematics.

Beazer (1975) explains that optimisation models are models designed to be used as operational devices to improve portfolio (or mutual funds) performance. Dyer et al. (1992) argue that many of the topics in Multi-Criteria Decision Making have been optimisation-related, in which Goal Programming, conceived by Charnes, Cooper & Ferguson (1955), was an early contribution. Goal Programming is the most widely used

*The author would like to thank Dr. Dylan Jones, senior lecturer in the Department of Mathematics at the University of Portsmouth. Dr. Jones provided invaluable comments on this manuscript.

approach in the field of Multiple Criteria Decision Making that enables the decision maker to incorporate numerous variations of constraints and goals, particularly in the field of Portfolio Selection.

The old model of Markowitz (1952), including several variations, is still representing the foundation of fund managers' investment decision. The development of modern, time-continuous portfolio optimisation, which is one of the growing research areas in the field of financial mathematics, had advanced so far that many algorithms now suggest themselves for practical application and implementation. The basic objective of a mutual fund is to provide a diversified portfolio so as to reduce the risk in investments at a lower cost. Markowitz suggests that investors should consider risk and return together and determine the allocation of funds among investment alternatives on the basis of the trade-off between them.

Mankert (2006) emphasises that portfolio (mutual fund) models are tools intended to help portfolio managers decide on weights of the assets within a portfolio. The ideas of Markowitz (1952) have had a great impact on portfolio theory. However, in practical portfolio management the use of Markowitz model has not had the same impact as it has had in academia. Many fund managers consider developing new models, using their financial mathematics capabilities, which build on Markowitz model and aims at handing some of its practical problems.

Over and above, Modern Financial Theory is based, amongst the others, on the Portfolio Theory of Markowitz (1952), the Arbitrage Principles of Modigliani & Miller (1958), the Capital Asset Pricing Model of Sharpe (1964), Lintner (1965) and Black (1972), and the Options Pricing Theory of Black & Scholes (1973). Hromis (2004) argues that the Modern Financial Theory is based on a heavy use of mathematical models.

In a dynamic market the need for professional fund managers increase enormously. The complexity involved in analyzing individual securities makes it extremely difficult for an investor to take the investment decision on their own, particularly under current circumstances of the financial crisis. So it makes lots of sense to depend on those who are good at this job of managing funds.

Beckmann & Menkhoff (2008) mention that fund managers are not only worth a detailed analysis because they are experts in managing risks but also because they work in a field of financial decision making.

This paper examines the performance of female and male mutual funds mangers, as one of the important careers in the field of financial mathematics, in order to investigate the claim that women perform lower, on average, than men do on financial mathematics activities, particularly in the field of mutual funds management.

The reminder of the paper is organized as follows. Section (2) provides an overview of mutual fund industry in an emerging market. Section (3) discusses the relevant literature review. Section (4) outlines the research data and relevant analysis. Empirical results are summarised in section (5), while section (6) gives some implications and areas for future research and finally section (7) provides concluding remarks.

2 Mutual Fund Industry in Emerging Markets

The important role of mutual funds in terms of stock markets efficiency, liquidity and transparency in emerging markets raises the need for studying the factors behind the performance of mutual funds.

While there is an extensive collection of literature on emerging markets, these mainly focus on the US funds investing in the emerging markets (for example Aggarwal, Klapper & Wysocki, 2004 and Gottesman & Morey, 2006), there is very limited work that has been done on mutual funds that exist in emerging markets, particularly when it comes to studying such factors as gender. This could be due to difficulties in portfolio evaluation in theses markets.

Nevertheless, the size and return of available funds in emerging markets and their growth prospect warrant in-depth study into these markets (Gottesman & Morey, 2006).

There is a need for mutual fund performance attribution in emerging markets, particularly with factors such as fund manager gender. A growing number of literatures in developed market investigate gender differences amongst mutual fund managers. Based on findings from the existing literature on gender differences (for example, Beckmann, Lutje & Rebeggiani, 2007; Bliss & Potter, 2001; Niessen & Ruenzi, 2005 and Veleva, 2005), it is hypothesized that female fund managers are well educated, in financial mathematics, take less risk and follow less extreme investment styles that are more consistent over time. Furthermore, female fund managers are expected to be less overconfident and therefore to trade less.

This paper is distinct from other gender and financial mathematics related papers in that it provides imperical evidence on emerging markets, particularly Egypt, where the mutual fund industry is emerging and where the gender role is increasing. Mutual fund industry started in Egypt on 1994 with the establishment of the first mutual fund by the National Bank of Egypt (Azmi, 2005).

Although the number of funds in Egypt is very small compared to established markets, the growth is high with the increasing openness of the Egyptian economy, together with the active implementation of the

privatization program, although this would change with the current financial crisis.

This paper also contributes to the growing literature on gender and mutual fund performance evaluation and attribution. In particular, this paper is designed to provide evidence on relation between different fund factors, especially the gender of the fund manager, and its performance in an emerging market.

3 Literature Review

Treynor (1965), Sharpe (1966), and Jensen (1968) develop the standard indices to measure risk adjusted mutual fund returns. Numerous studies have tested the performance of mutual funds compared to a certain benchmark, usually market index (Artikis, 2002, Cresson, Cudd and Lipscomb, 2002, Daniel, Grinblatt, Titman and Wermers, 1997, Lehmann and Modest, 1987, Matallin and Nieto, 2002, Otten and Schweitzer, 2002, Persson, 1998, Raj, Forsyth and Tomini, 2003, and Zheng, 1999).

Lehmann and Modest (1987), and Daniel et al. (1997) find the performance of mutual funds to underperform the market index in the US financial market, consistent with a study by Persson (1998) who finds the performance of the Swedish mutual funds to underperform the market index, and in contrast to a later study by Artikis (2002) who argues that the performance of the Greek mutual funds outperforms the market index.

Otten and Schweitzer (2002) compare the European mutual fund industry with the United States using risk-adjusted measures of performance. They find that Europe is still lagging the US mutual fund industry where it comes to total asset size or average fund size.

Cresson, Cudd and Lipscomb (2002) show that fund performance outperforms the market index in the short-term, whereas it underperforms the market index in the long-term. Matallin and Nieto (2002) claim that the performance of most of the Spanish mutual funds underperforms the Spanish financial market index.

Previous studies on the evaluation of the mutual fund's performance in developed countries varied in their results. In addition, there are some factors (like: the volatility of markets, the size of government involvement and the extent of regulations) which distinguish mutual funds in emerging markets from their counterparts in more established markets. Studies examining the relation between mutual fund performance and factors such as fund manager gender, fund expenses, size, age, and objective report conflicting results.

For example, Barber & Odean (2001) find that the average portfolio turnover rate for men is significantly higher than for women, and mutual fund performance losses are significantly more pronounced for men. While Niessen & Ruenzi (2005) hypothesise in their research that female fund manager take less risk and follow less extreme investment styles that are more consistent over time. Their empirical results support these hypotheses, but they find no evidence that behavioural differences between female and male fund managers are reflected in fund performance.

Beckmann, Lutje & Rebeggiani (2007) argue that Italian female professionals do not only assess themselves as more risk averse than their male colleagues, they also prefer a more passive portfolio management compared to the level they are allowed to. Besides, in a competitive tournament scenario near the end of the investment period, female asset managers do not try to become the ultimate top performer when they have outperformed the peer group. However in case of underperformance, the risk of deviating from the benchmark makes female professionals more willing than their male colleagues to seize a chance of catching up.

Aside from fund manger gender relation to fund performance, Volkman & Wohar (1995) and Gallagher (2002) claim positive relations between fund performance as a dependent variable in their studies and fund's objective as well as fund's systematic risk as independent variables. Whereas Peterson, Pietranico, Riepe and Xu (2001) find negative relation between fund performance and fund's systematic risk. Other studies (example Carhart, 1997) find no relation between fund performance and fund's age.

Table (1) lists the literatures' results on the factors determining the performance of the mutual funds.

Literature in the area of finance and financial mathematics, particularly on mutual funds, provides a range of factors that contribute to the performance of a particular fund. Although the direction and extent to which these factors influence performance varies among the developed countries funds, evaluating developing countries funds according to such factors has scarcely been investigated, particularly for fund manger gender factor.

This paper attempts to accomplish this by empirically tests the relation between mutual fund performance and fund manager gender, in addition to other seven factors, which are: fund age, size, objective, expenses ratio, systematic risk, type and total risk in one of the emerging markets, that is Egypt's market.

Determinants	Authors	Results (Relevance to Mutual Fund Performance)
Fund Manager Gender	Atkinson, Baird and Frye (2003)	There is a Relation
	Barber and Odean (2001)	There is a Relation
	Beckmann, Lutje and Rebeggiani (2007)	There is a Relation
	Bliss and Potter (2001)	There is a Relation
	Niessen and Ruenzi (2005)	No Evidence of Behavioural Differences
	Veleva (2005)	There is a Correlation
Expense Ratio	Peterson, Pietranico, Riepe and Xu (2001)	There is a direct relation
	Gottesman & Morey (2006)	Negative Relation
Fund Systematic Risk	Gallagher (2002)	Positive Relation
	Peterson, Pietranico, Riepe and Xu (2001)	Negative Relation
Fund Total Risk	Das, Kish, Muething and Taylor (2002)	Positive Relation
Fund Age	Carhart (1997)	No Relation
	Gallagher (2002)	Negative Relation
Fund Size	Grinblatt and Titman (1994)	No Relation
	Volkman and Wohar (1995)	No Relation
	Carhart (1997)	No Relation
	Israelsen (1998)	Positive Relation
	Ramasamy and Yeung (2003)	Positive Relation
Fund Objective (Income- Growth-Income/Growth)	Volkman and Wohar (1995)	Positive Relation (Growth)
	Bauman (1968)	No Clear Relation
Fund Type (Open/ Closed End)	Glenn (2004)	Positive Relation (Closed End)

Table (1): Review of Different Research Results on the Determinants of Mutual Fund Performance

4 Methodology and Data

4.1 Research Model

The research model is based on the relevant literature reviewed in the previous section. The research model depicts the relations between fund manager gender, in addition to other seven factors of mutual funds, and the fund performance.

Other factors (variables) were excluded because of the lack of their information in the Egyptian emerging market (examples: fund turnover, liquidity, etc.) or to avoid the multicolinearity issue.

4.2 Research Variables and Measurement

a. Mutual fund performance

Mutual fund performance is the dependent variable in the research model. There are three measures of mutual fund performance based on the literature reviewed in the previous section. But for the purpose of this research, Sharpe's Index will be used to measure mutual funds performance as it is the recommended measure of mutual fund performance in the Egyptian emerging market where diversification opportunities locally are not good enough to eliminate entirely the unsystematic risk and active stocks are limited (Azab, 2002; Azmi, 2005).

The Sharpe's index is computed by applying the following:-

$$SI_p = (R_p - R_{rf})/D_p$$

Where:

SI_p = Sharpe's index for portfolio (mutual fund) p.

R_p = Return on portfolio p.

R_{rf} = Return on risk-free asset.

D_p = Standard deviation of portfolio p.

The numerator is the excess return above the risk-free return on a portfolio, and Dp is the measure of total risk of the portfolio. A portfolio performs better than the benchmark if its Sharpe's index is greater than that of the benchmark.

b. Fund manager gender and other mutual fund factors

Fund manager gender and other seven factors are examined in terms of their relation with mutual fund performance as follows:-

- Fund manager's gender (Male/ Female) is examined in terms of its relation with mutual fund performance. It is measured by one dummy variable of gender (Male= 1, Female= 0).
- Mutual fund age is computed on a quarterly base during the study period of 5 years (Jan. 1999- Dec. 2003).
- Mutual fund size is computed by applying the following: Total assets value of the fund/ Number of mutual fund's shares outstanding.
- Mutual fund objective is measured by two dummy variables of income and growth objectives (Income: (1,0) ; Growth: (0,1); Income/Growth: (0,0)).
- Fund's total risk is measured by the standard deviation (the square root of the variance) of the fund's returns.
- Fund systematic risk is measured by beta coefficient (Miller, 2001) as follows:-

$$\beta_i = \frac{\mathrm{Cov}(X_i, X_m)}{\sigma_m^2} = \frac{\sum_{t=1}^{n}(X_{it} - \overline{X}_i)(X_{mt} - \overline{X}_m)}{\sum_{t=1}^{n}(X_{mt} - \overline{X}_m)^2}$$

Where:-

β_i: The Beta coefficient of mutual fund i.

Cov (X_i, X_m): Covariance between the return of the mutual fund i and the return of the market portfolio (m).

σ_m^2: Variance in market portfolio return.

X_{it}: The return of mutual fund i in the period t.

$\overline{X}_i$: The average returns of fund i during the period.

X_{mt}: Market return in the period t.

$\overline{X}_m$: The average returns of the market portfolio during the period.

- Fund's expenses ratio: computed by applying the following: Expenses/ Net assets value of the mutual fund.
- Mutual fund's type (Open/ closed end): measured by one dummy variable of fund's type (Open= 1, Closed= 0).

These variables are considered to be the determinants of mutual funds performance according to the literature review. VIF is calculated for them to identify multicolinearity. VIF, variance inflation factor, if highly collinear a high value is calculated, higher than 5 (Levine et al., 2005, p.632).

In this study, the calculated VIF values indicate no existence of the multicolinearity issue with the current independent variables as the calculated VIF values are less than 5 as shown in table (2).

Table (2): VIF Values for the Independent Variables of the Research

The Independent Variables	VIF Values
Fund Manager Gender	1.532
Fund Size	1.396
Fund Total Risk	1.067
Fund Systematic Risk	1.042
Fund Expense Ratio	1.055
Fund Age	1.811
Fund Type	1.246
Fund Objective (Income)	2.077
Fund Objective (Growth)	2.653

4.3 Research Hypotheses

1. Fund manger gender and mutual fund performance (hypothesis H1)

   Atkinson, Baird and Frye (2003) show that fund manager gender is related to fund performance. Bliss and Potter (2001) argue that female fund managers outperform their male counterparts, consistent with later study by Veleva (2005) who finds a correlation between the percentage of female representation and total (and average) annual returns. Bliss and Potter (2001) further compare data from domestic and international US equity funds and expected women to hold less risky portfolios than men. Assuming them to be less overconfident, female asset managers are expected to trade less than their male counterparts, and thus to perform better (Bliss & Potter, 2001; Barber & Odean, 2001).
   H1: There is a relation between fund manager gender and fund performance.

2. Other factors of mutual fund performance (hypotheses H2: H8)

   Carhart (1997) show that fund age is not related to performance, in contrast to a later study by Gallagher (2002) who finds balanced mutual fund performance to be negatively related to fund age.
   H2: There is a relation between fund age and fund performance.

   Grinblatt and Titman (1994), Volkman and Wohar (1995), and Carhart (1997) find no relation between fund performance and its size, in contrast to Israelsen (1998), and Ramasamy and Yeung (2003) who find fund performance to be positively related to fund size.
   H3: There is a relation between fund size and fund performance.

   Volkman and Wohar (1995) compare the fund performance of growth, income, growth/income objectives of a fund. They find fund performance to be positively related to fund objective when it is the growth, in contrast to Bauman (1968) who find no clear relation between fund performance and its objective.
   H4: There is a relation between fund objective and fund performance.

   Glenn (2004) examines the relation between fund performance and its type (open/ closed end). He finds a significantly positive relation between performance and fund type when it is closed end fund, in contrast to Kacperczyk, Sialm & Zheng (2005) who find a positive relation between fund performance and fund type when it is open end fund.
   H5: There is a relation between fund type and fund performance.

   Peterson, Pietranico, Riepe and Xu (2001) show that expense ratios are directly related to the variability of mutual fund returns, consistent with a study by Goettesman & Morey (2006) who find a strong relation between fund performance and fund's expenses ratio.
   H6: There is a relation between fund's expense ratio and fund performance.

   Gallagher (2002) find positive relation between fund performance and systematic risk, whereas Peterson et al. (2001) argue fund performance to be negatively related to fund's systematic risk.
   H7: There is a relation between fund systematic risk and fund performance.

   Das, Kish, Muething and Taylor (2002) find fund performance to be positively related to the total risk (as measured by the standard deviation).
   H8: There is a relation between fund total risk and fund performance.

4.4 Data and Sampling

Sharpe's Index is used to evaluate the risk-adjusted performance of the mutual funds operating in the Egyptian stock exchange during the period from January 1999 to December 2003 using quarterly data. Then fund manager gender and other seven factors are examined in terms of their relation with mutual fund performance using a multiple regression.

The data for the estimation of Sharpe's index as well as fund's manager gender, age, size, type, total risk, objective, systematic risk and expenses ratio are collected from the Capital Market Authority of Egypt and the Cairo and Alexandria Stock Exchange, in addition to the investment management companies of the Egyptian mutual funds.

The benchmark used to compare the risk-adjusted performance of the Egyptian mutual funds is the CASE 30 index[1]. The risk free return necessary to compute the Sharpe's index is the reported 3-months Egyptian Treasury bill yield.

In order to avoid survivorship bias, the sampling period is chosen to include all mutual funds during 5 years including those funds which did not survive after the study period. The research population consists

[1]CASE 30 Index: Cairo and Alexandria Stock Exchange Index for the top 30 companies, CASE 30, is the benchmark of the Egyptian emerging market. http://www.egyptse.com/index.asp

of 21 mutual funds in which 19 funds are included in the research based on their inception dates during 5 years from Jan. 1999 to Dec. 2003, in which Delta fund, a closed end fund for example, is included although it is no longer available in the Egyptian stock market since January 2004.

The study examines multiple regression model using the risk-adjusted measure of mutual fund performance with the fund manger gender, in addition to other seven determinants of mutual fund performance in the Egyptian emerging market as follows:-

$$SI_j = a + b_1 G_j + b_2 R_{S_j} + b_3 R_{\beta_j} + b_4 E_j + b_5 S_j + b_6 A_j + b_7 T_j + b_8 O_{I_j} + b_9 O_{G_j} + e$$

Where:

- SI_j: The performance of the mutual funds measured by Sharpe's index during the period j.
- G_j: The gender of the mutual fund manger during the period j.
- R_{S_j}: The total risk of the mutual fund, measured by the standard deviation, during the period j.
- R_{β_j}: The systematic risk of the mutual fund, measured by Beta coefficient, during the period j.
- E_j: The expenses ratio of the mutual fund during the period j.
- S_j: The size of the mutual fund during the period j.
- A_j: The mutual fund age during the period j.
- T_j: The mutual fund type during the period j.
- O_{I_j}: The mutual fund objective as a dummy variable representing income fund during the period j.
- O_{G_j}: The mutual fund objective as a dummy variable representing growth fund during the period j.
- $a, b1, b2, \ldots, b9$: The multiple regression coefficients with fund performance.
- e: The random error of the multiple regression model.

5 Results

The results of the multiple regressions processed show that the performance of the Egyptian mutual funds is related to five factors as shown in table (3).

Table (3): The Factors Determining the Performance of the Mutual Funds Operating in the Egyptian Emerging Market

Independents Variables	Regression Coefficients	t-test	Standard Error	Significance level
Manager Gender	-1.449	-5.988	0.242	0.000
Expenses Ratio	-4.050	-4.258	0.951	0.000
Fund Objective (Growth)	0.519	2.455	0.211	0.015
Total Risk	-0.107	-2.271	0.047	0.024
Fund Type	0.683	2.014	0.339	0.045

Table (3) reports the regression coefficients, t-test, standard errors and the significance level for five independent variables, in which a regression coefficient with positive sign indicating a positive relation between the fund's determinant and the fund performance, and with a negative sign indicating reverse or negative relation. **Therefore, the results support the acceptance of the hypotheses: H1, H4, H5, H6 and H8, whereas the results do not support the acceptance of the hypotheses: H2, H3 and H7**.

The results suggest that mutual fund performance is inversely related to fund's manager gender (when it is a male), expenses ratio and total risk, and positively related to fund's objective (when it is growth) and fund's type (when it is an open-end fund).

The results of the relation between fund performance and fund's manager gender implies that male fund managers perform, on average, 1.449 points more poorly than do female fund managers.

Years (study period)	Total Mutual Funds		Mutual Funds Managed by Women		Mutual Funds Managed by Men	
	Number	Percentage	Number	Percentage	Number	Percentage
1999	19	100%	4	21.05%	15	78.95%
2000			4	21.05%	15	78.95%
2001			4	21.05%	15	78.95%
2002			5	26.32%	14	73.68%
2003			8	42.11%	11	57.89%

Table (4): Number and Percentage of Mutual Funds Managed by Women vs. Men

6 Implications and Areas for Future Research

Several studies either on fund's performance evaluation (Cresson et al., 2002, Daniel et al., 1997, Lehmann and Modest, 1987, Otten and Schweitzer, 2002, Raj et al., 2003, etc.) or fund's performance attribution (Atkinson et al., 2003, Carhart, 1997, Grinblatt and Titman, 1994, Volkman and Wohar, 1995, etc.) report conflicting results.

The result of fund's manager gender is interesting as existing studies (Bliss and Potter, 2001, Schubert, Gysler, Brown, and Brachinger, 2000, Veleva, 2005, etc.) show that men and women view money, risk, and investing differently.

There is also anecdotal evidence and research suggesting that women might actually be better investors than men. However, none of this has historically mattered in the mutual fund industry in the United States because the number of women fund managers was negligible. But this is changing as women represent 11 percent of the fund managers in the USA and in Egypt women represent, surprisingly, on average 27 percent of the fund managers (Azmi, 2005).

This result of having a relation between fund performance and fund's manager gender is somewhat surprising particularly in an emerging market like Egypt's Stock Market, and clearly warrant future studies. But such result could be attributed or interpreted in light of the trend of increasing number of funds managed by women vs. men by the end of study's period. Table (4) shows the percentages of women and men managing mutual funds in the Egyptian emerging market during the study period of 5 years.

Also this result of fund manager gender influence over fund performance could be attributed to the high impact of manager characteristics in fund performance particularly in an emerging market. Therefore, other studies in management aspects of mutual funds in Egypt or other emerging countries are encouraged as they could reveal more determinants of mutual funds performance in these markets.

The Egyptian mutual funds categorized into 3 groups based on their objectives (income, growth, income/growth funds) in which the results reveal a relation between fund performance and fund's objective when its come to growth objective. This result is in line with what has been found in literature (example Volkman & Wohar, 1995) and can be attributed to the basic relationship between return and risk.

The results show a negative relation between fund performance and its expense ratio and a negative relation also between fund performance and fund's total risk as measured by the standard deviation. This is consistent with the international fund literatures which imply that those literatures do indeed apply to emerging market funds.

Although the results of this study regarding the relation between fund performance and fund type are in line with some literatures (example Rao, 2001), they are contradicting with other studies like Glenn study (2004). And this could be attributed to the very small number of existed closed end funds (only 2 closed-end funds) in the Egyptian emerging market.

Fund age, size and systematic risk found to be of no significant relation with the Egyptian mutual funds in this study which contradicts with what has been found on the international literature of funds in the US and other developed markets, but it is justifiable with the circumstances of emerging markets as they are characterised by limited numbers of active stocks, etc. Therefore, mutual funds manager usually find limited opportunities in emerging markets to diversify or eliminate the unsystematic risk and that why the results suggest that unsystematic risk proportion in the fund's total risk influence more fund performance than the systematic risk proportion (although this is changing due to the current financial and economic turmoil).

To summarise, there is no significant evidence of relation between mutual fund performance in the Egyptian emerging market and fund's systematic risk, age and size during the period from Jan. 1999 to Dec. 2003. The results indicate significantly positive relation between fund performance and fund's type (open) and objective (growth), during the same period. These results imply positive performance of a mutual fund when its type is an open end fund and also when its objective is growth over income or income/ growth. The results also indicate negative relation between fund performance and fund's expenses ratio, fund's total risk, and fund's manager gender (when it is a male).

Therefore other studies are warrant for validating or finding out in line or different results concerning the set of factors of influence on the mutual fund performance and whether women are outperforming men in this career.

7 Concluding Remarks

This paper finds significant relation between mutual fund performance and fund's manager gender, suggesting that women perform well in financial mathematics field but equality is yet to be realised in terms of more women joining this career.

The main limitation in this paper is the small number of the available mutual funds in the Egyptian emerging market and the lack of information for other factors that could be of influence over mutual funds performance in the Egyptian emerging market (example: the factors related to management characteristics).

Future researches are warrant in the area of fund performance attribution in emerging markets; perhaps with larger emphasis on manager specific factors and gender differences.

References

[1] Artikis, Panayiotis (2002). Evaluation of Equity Mutual Funds Operating in the Greek Financial Market, Journal of Managerial Finance, Vol. 28, pp. 27-42.

[2] Aggarwal, R., Klapper, L., and Wysocki, P. (2004). Disclosure Quality and Emerging Market Mutual Fund Investment, The World Bank Research Paper, pp. 1-38.

[3] Atkinson, S., Baird, S., and Frye, A. (2003). Do Female Mutual Funds Managers Manage Differently?, Journal of Financial Research, pp. 27-32.

[4] Azab, Bassam (2002). The Performance of the Egyptian Stock Market, International Banking and Finance, The Birmingham Business School.

[5] Azmi, Rania (2005). Analytical Study of the Egyptian Mutual Funds Performance and its Determinants, Unpublished Master Thesis, Department of Business Administration, Faculty of Commerce, Alexandria University, Egypt.

[6] Barber, B., and Odean, T. (2001). Gender, Overconfidence, and Common Stock Investment, Quarterly Journal of Economics, pp. 261- 292.

[7] Bauman, Scott (1968). Evaluation of Prospective Investment Performance, the Journal of Finance, pp. 276-295.

[8] Beazer, William (1975). Optimization of Bank Portfolios, Lexington Books, D.C. Heath and Company.

[9] Beckmann, D., Lutje, T., and Rebeggiani, L. (2007). Italian Asset Managers' Behavior: Evidence on Overconfidence, Risk Taking, and Gender, Discussion Paper No. 358, Leibniz Universitat Hannover, Department of Economics, ISSN: 0949-9962.

[10] Beckmann, D. and Menkhoff, L. (2008). Will Women Be Women? Analyzing the Gender Difference among Financial Experts, KYKLOS, Vol. 61, pp. 364-384.

[11] Black, Fischer (1972). Capital Market Equilibrium with Restricted Borrowing, the Journal of Business, Vol. 45, pp. 444-455.

[12] Black, F. and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities, the Journal of Portfolio Economy, Vol. 81, pp. 637-654.

[13] Bliss, R., and Potter, M. (2001). Mutual Fund Managers: Does Gender Matter?, Babson College, U.S.A.

[14] Carhart, Mark (1997). On Persistence in Mutual Fund Performance, the Journal of Finance, pp. 57-81.

[15] Charnes, A., Cooper, W. and Ferguson, R. (1955). Optimal estimation of executive Compensation by Linear Programming, Journal of Management Science, Vol. 1, No. 1, pp. 138-151.

[16] Cresson, J., Cudd, R., and Lipscomb, T. (2002). The Early Attraction of S & P 500 Index Funds: Is Perfect Tracking Performance an Illusion? , Journal of Managerial Finance, Vol. 28, pp. 1-8.

[17] Daniel, K., Grinblatt, M., Titman, S., and Wermers, R. (1997). Measuring Mutual Fund Performance with Characteristic-Based Benchmarks, the Journal of Finance, Vol. 52, pp. 1035-1058.

[18] Das, N., Kish, R., Muething, D., and Taylor, L. (2002). Literature on Hedge Funds, Discussion Paper, Lehigh University, Pennsylvania, pp. 1-88.

[19] Dyer, J., Fishburn, P., Steuer, R., Wallenius, J. and Zionts, S. (1992). Multiple Criteria Decision Making, Multiattribute Utility Theory: The Next Ten Years , Management Science, Vol. 38, pp. 645-654.

[20] Gallagher, David (2002). Investment Manager Characteristics, Strategy and Fund Performance, Unpublished Doctoral Dissertation, School of Business, Faculty of Economics and Business, The University of Sydney, Australia.

[21] Glenn, Brian (2004). The Mechanics Behind Investment Funds: Why Closed-End Funds Provide Superior Returns, Journal of Managerial Finance, Vol. 30, pp. 86-102.

[22] Gottesman, A., and Morey, M. (2006). Predicting Emerging Market Mutual Fund Performance, http://ssrn.com/abstract=897184.

[23] Grinblatt, M., and Titman, S. (1994). A Study of Monthly Mutual Fund Returns and Performance Evaluation Techniques, Journal of Financial and Quantitative Analysis, Vol. 29, pp. 419-444.

[24] Hromis, Gabriela (2004). The Possibility of Significant Change in Financial Theory, Bachelor Thesis, Ekonomska Fakulteta, Univerza V Ljubljani, pp. 1-46.

[25] Israelsen, Craig (1998). Characteristics of Winning Mutual Funds, Journal of Financial Planning, pp. 20-28.

[26] Jensen, Michael (1968). The Performance of Mutual Funds in the Period 1945-1964, the Journal of Finance, Vol. 23, pp. 389-416.

[27] Kacperczyk, M., Sialm, C., and Zheng, L. (2005). Unobserved Actions of Mutual Funds, the Sauder School of Business, University of British Columbia.

[28] Lehmann, B., and Modest, D. (1987). Mutual Fund Performance Evaluation: A Comparison of Benchmarks and Benchmarks Comparison, the Journal of Finance, Vol. 42, pp. 233-265.

[29] Levine, D., Stephan, D., Krehbiel, T., and Berenson, M. (2005). Statistics for Managers, International edition, Pearson Prentice Hall.

[30] Lintner, John (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets, the Journal of Review of Economics and Statistics, Vol. 47, pp. 13-37.

[31] Mankert, Charlotta (2006). The Black-Litterman Model: Mathematical and Behavioral Finance Approaches Towards its Use in Practice , Licentiate Thesis, the Royal Institute of Technology, Stockholm, Sweden, pp. 1-111.

[32] Markowitz, Harry (1952). Portfolio Selection, the Journal of Finance, Vol. 7, pp. 77-91.

[33] Matallin, J., and Nieto, L. (2002). Mutual Funds as an Alternative to Direct Stock Investment, the Journal of Applied Financial Economics, pp. 743-750.

[34] Modigliani, F. and Miller, M. (1958). The Cost of Capital, Corporation Finance and the Theory of Investment, the American Economic Review, Vol. 48, pp. 261-297.

[35] Niessen, A., and Ruenzi, S. (2005). Sex Matters: Gender and Mutual Funds, Department of Corporate Finance, University of Cologne.

[36] Otten, R., and Schweitzer, M. (2002). A Comparison between the European and the U.S. Mutual Fund Industry, Journal of Managerial Finance, Vol. 28, pp. 14-34.

[37] Persson, Mattias (1998). Performance of Swedish Mutual Funds, Department of Economics, Lund University, Sweden.

[38] Peterson, J., Pietranico, P., Riepe, M., and Xu, F. (2001). Explaining the performance of Domestic Equity Mutual Funds, Journal of Investing, Institutional Investor Inc., pp. 1-14.

[39] Raj, M., Forsyth, M., and Tomini, O. (2003). Fund Performance in a Downside Context, Journal of Investing, Institutional Investor Inc., pp. 50-63.

[40] Ramasamy, B., and Yeung, M. (2003). Evaluating Mutual Funds in Emerging Markets: Factors that Matter to Financial Advisors, The International Journal of Bank Marketing, Vol. 21, pp. 122-136.

[41] Rao, Umamaheswar (2001). Mutual Fund Performance during Up and Down Market Conditions, Review of Business, Summer issue, pp. 62-75.

[42] Roth, Louise (2003). A Research Note on Gender Differences in Compensation on Wall Street, Social Forces, Vol. 82, pp. 783-802.

[43] Schubert, R., Gysler, M., Brown, M., and Brachinger, H. (2000). Gender Specific Attitudes Towards Risk and Ambiguity, Center for Economic Research, Swiss Federal Institute of Technology.

[44] Sharpe, William (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk, the Journal of Finance, Vol. 19, pp. 425-442.

[45] Sharpe, William (1966). Mutual Fund Performance, Journal of Business, Vol. 39, pp. 119-138.

[46] Simon, Marilyn (2000). The Evolving Role of Women in Mathematics, the National Council of Teachers of Mathematics, Journal of Mathematics Teacher, Vol. 93, pp. 782-786.

[47] Treynor, Jack (1965). How to Rate Management of Investment Funds, Harvard Business Review, Vol. 43, pp. 63-73.

[48] Veleva, Vesela (2005). Gender Diversity and Financial Performance, Citizens Advisers Inc., pp. 1-4.

[49] Volkman, D., and Wohar, M. (1995). Determinants of Persistence in Performance of Mutual Funds, the Journal of Financial Research, pp. 30-52.

[50] Zheng, Lu (1999). Is Money Smart? A Study of Mutual Fund Investor's Fund Selection Ability, The Journal of Finance, pp. 901-933.

Introduction to algebraic number theory

Eva Bayer-Fluckiger
Ecole Polytechnique Fédérale de Lausanne, Switzerland

1 An example – sums of squares

The origin of the theory lies in concrete and beautiful problems. Some of these are easy to solve, but others are inaccessible with elementary means, and lead to the introduction of new objects and methods that can then be used to deal with them. The aim of this section is to illustrate this through one example, namely sums of two squares of integers.

Problem 1.1. Which positive integers can be written as sums of two squares?

Experimentally, one sees that – for instance – 5 is a sum of two squares, namely $5 = 1 + 4$, but 3 cannot be expressed as a sum of two squares. In the 17th century, Pierre de Fermat proved the following beautiful theorem:

Theorem 1.2. *Let p be an odd prime number. Then p is a sum of two squares if and only if $p \equiv 1 \pmod 4$.*

One direction is obvious: one sees very easily that if p is a sum of two squares then $p \equiv 1 \pmod 4$. The converse is much more difficult. Fermat's original proof is different from the one presented here.

The main idea of the proof we discuss here is to consider a new object, the ring $\mathbf{Z}[i]$, where i is the imaginary number, $i^2 = -1$. In other words, we consider the subset of the complex numbers consisting of the expressions $a + bi$ where a and b are integers. This is closed by addition and multiplication, hence it is a subring of the complex numbers. It is called the *ring of Gaussian integers*. Note that Gauss lived much later than Fermat – this notion was not used in Fermat's original proof.

Let us denote by $x \to \overline{x}$ the complex conjugation. This clearly preserves the ring of Gaussian integers – if $x = a + bi \in \mathbf{Z}[i]$, then $\overline{x} = a - bi \in \mathbf{Z}[i]$. Let us look at the complex norm : $N : \mathbf{Z}[i] \to \mathbf{N}$ defined by $N(x) = x\overline{x}$. If $x = a + bi$, then $N(x) = a^2 + b^2$. This leads us to the following basic observation :

Remark 1.3. A positive integer n is a sum of two squares if and only if $n \in N(\mathbf{Z}[i])$.

This motivates a study of the ring of Gaussian integers and of the norm, with the hope that it will lead to the solution of our problem. It is natural to look at the Gaussian integers as a generalization of the ordinary integers, and investigate whether some of the basic properties of $\mathbf{Z}$ are also true for $\mathbf{Z}[i]$. One of these is the Euclidean division :

Euclidean division Let α and β be two integers, $\beta \neq 0$. Then there exist two integers γ and δ such that $\alpha = \beta\gamma + \delta$ and that the absolute value of δ less than the absolute value of β.

This can ideed be generalized to the Gaussian integers, as follows :

Proposition 1.4. *(Euclidean division for the Gaussian integers) : Let $\alpha, \beta \in \mathbf{Z}[i]$ with $\beta \neq 0$. Then there exist γ and δ in $\mathbf{Z}[i]$ such that $\alpha = \beta\gamma + \delta$ and that $N(\delta) < N(\beta)$.*

Note that here we use the norm N to measure the size of the elements of the ring of Gaussian integers.

Proof. Set $x = \frac{\alpha}{\beta}$. Note that there exists $\gamma \in \mathbf{Z}[i]$ such that $N(x - \gamma) < 1$. We conclude by setting $\delta = \alpha\beta - \gamma$. □

Recall that an ideal is said to be *principal* if it is generated by a single element. For instance, every ideal of the ring $\mathbf{Z}$ is principal. This can be proved by Euclidean division – if I is a non–zero ideal, then take an element of I such that its absolute value is minimal. Then Euclidean division shows that this element generates the ideal. The same method leads to the following :

Corollary 1.5. *Every ideal of $\mathbf{Z}[i]$ is principal.*

Proof. Let I be a non–zero ideal of $\mathbf{Z}[i]$, and let $\beta \in I$ be a non–zero element such that $N(\beta) < N(\beta')$ for all non–zero $\beta' \in I$. Let $\alpha \in I$. Then there exist $\gamma, \delta \in \mathbf{Z}[i]$ such that $\alpha = \beta\gamma + \delta$ and that $N(\delta) < N(\beta)$. As $\beta \in I$ and I is an ideal, we have $\beta\gamma \in I$. Therefore $\delta \in I$, and by minimality of $N(\beta)$ this implies that $\delta = 0$. □

In other words, the ring $\mathbf{Z}[i]$ is a *principal ideal domain.*

We also need a lemma concerning the invertible elements of $\mathbf{Z}[i]$:

Lemma 1.6. *Let $x \in \mathbf{Z}[i]$. Then there exists $y \in \mathbf{Z}[i]$ with $xy = 1$ if and only if $N(x) = 1$.*

Proof. Suppose that there exists $y \in \mathbf{Z}[i]$ with $xy = 1$. Then $N(xy) = N(x)N(y) = N(1) = 1$, therefore $N(x) = 1$ or -1. As $N(x) \geq 0$, we must have $N(x) = 1$. Conversely, suppose that $N(x) = 1$. We have $N(x) = x\overline{x} = 1$, hence $\overline{x}$ is the inverse of x. □

The final ingredient in our proof of Fermat's theorem concerns the behavior of prime numbers in the ring $\mathbf{Z}[i]$. Let p be a prime number. Then the ideal $p\mathbf{Z}$ is a maximal ideal of the ring $\mathbf{Z}$. Of course $p\mathbf{Z}[i]$ is an ideal of the ring $\mathbf{Z}[i]$, but is it always maximal ? The following proposition shows that this is not the case, and gives a complete answer to our question.

Proposition 1.7. *Let p be an odd prime number. Then the ideal $p\mathbf{Z}[i]$ is maximal if and only if $p \equiv 3 \pmod 4$.*

Proof. The ideal $p\mathbf{Z}[i]$ is maximal if and only if $\mathbf{Z}[i]/p\mathbf{Z}[i]$ is a field. Note that $\mathbf{Z}[i] = \mathbf{Z}[X]/(X^2+1)$, hence we have $\mathbf{Z}[i]/p\mathbf{Z}[i] \simeq \mathbf{F}_p[X]/(X^2+1)$, where $\mathbf{F}_p$ is the finite field with p elements. On the other hand, $\mathbf{F}_p[X]/(X^2+1)$ is a field if and only if the polynomial $X^2+1 \in \mathbf{F}_p[X]$ is irreducible, and this happens if and only if -1 is not a square modulo p, or equivalently $p \equiv 3 \pmod 4$. □

We can now prove Fermat's theorem. Recall that it is sufficient to show that every prime number p with $p \equiv 1 \pmod 4$ is the norm of an element of $\mathbf{Z}[i]$.

Proof of Fermat's theorem. Let p be a prime number such that $p \equiv 1 \pmod 4$. Then by the previous proposition, the ideal $p\mathbf{Z}[i]$ is not maximal. Hence there exists an ideal I of the ring $\mathbf{Z}[i]$ strictly containing the ideal $p\mathbf{Z}[i]$. As every ideal of $\mathbf{Z}[i]$ is principal, there exists $\alpha \in \mathbf{Z}[i]$ such that $I = \alpha\mathbf{Z}[i]$. This α is not an invertible element of $\mathbf{Z}[i]$ because $I \neq \mathbf{Z}[i]$. Therefore by the lemma $N(\alpha) \neq 1$. As $p\mathbf{Z}[i]$ is contained in I, there exists $\beta \in \mathbf{Z}[i]$ with $p = \alpha\beta$. The element β is not invertible because $p\mathbf{Z}[i]$ is not equal to I, hence $N(\beta) \neq 1$. We have $p^2 = N(p) = N(\alpha\beta) = N(\alpha)N(\beta)$, hence $N(\alpha) = p$. This implies that p is a sum of two squares. □

This gives us a motivation to study rings such as $\mathbf{Z}[i]$, in particular their ideals and invertible elements. We will do this in the following.

2 Algebraic number fields and rings of integers

The previous section shows that it is worth while to study rings of the type $\mathbf{Z}[i]$, and of course also their fields of fractions, such as $\mathbf{Q}[i]$. This leads us to the notion of *algebraic number field.*

Definition 2.1. An *algebraic number field* K is a field extension of finite degree of the field of rational numbers $\mathbf{Q}$.

Recall that if F is a field, then a field extension of F is a field K containing F. Then K is a vector space over F, and the dimension of this vector space is called the *degree* of the extension. The degree of K over F is denoted by $[K : F]$.

The simplest example of an algebraic number field is $\mathbf{Q}$, the only number field of degree 1. The next simplest ones are number fields of degree 2, the so–called *quadratic extensions.* It is easy to see that these are of the form $K = \mathbf{Q}(\sqrt{d})$, where d is a square–free integer. We say that K is a *real quadratic field* if $d > 0$, and that K is an it imaginary quadratic field if $d < 0$.

We would also like a generalization of the rings $\mathbf{Z}$ and $\mathbf{Z}[i]$. This is based on the following observation: if K is an algebraic number field and $\alpha \in K$, then there exists a monic polynomial $f \in \mathbf{Q}[X]$ such that $f(\alpha) = 0$. Indeed, if $n = [K : \mathbf{Q}]$, then the elements $1, \alpha, \ldots, \alpha^n$ are linearly dependent over $\mathbf{Q}$.

Definition 2.2. Let K be an algebraic number field and let $\alpha \in K$. Then we say that α is an *integer of K* if there exists a monic polynomial $f \in \mathbf{Z}[X]$ such as $f(\alpha) = 0$.

Let us denote by O_K the set of elements that are integers of K. We have the following

Theorem 2.3. *O_K is a subring of K.*

It seems hard to prove directly that if α and β are integers of K then so is their sum and their product. One uses the following lemma :

Lemma 2.4. *Let K be an algebraic number field, and let $\alpha \in K$. Then $\alpha \in O_K$ if and only if $\mathbf{Z}[\alpha]$ is a finitely generated abelian group.*

The proof of this lemma can be found (for instance) in [2], 2.1. Let us show how to use the lemma to prove the theorem :

Proof of theorem. Let $\alpha, \beta \in O_K$. Then by the lemma $\mathbf{Z}[\alpha]$ and $\mathbf{Z}[\beta]$ are finitely generated abelian groups. Hence $\mathbf{Z}[\alpha, \beta]$ is also a finitely generated abelian group. As this group contains $\mathbf{Z}[\alpha + \beta]$, $\mathbf{Z}[\alpha - \beta]$, $\mathbf{Z}[\alpha\beta]$, these are also finitely generated abelian groups, and hence another application of the lemma shows that $\alpha + \beta, \alpha - \beta, \alpha\beta \in O_K$. □

The ring O_K is called the *ring of integers* of K. It has the following properties :

Proposition 2.5. *Let K be an algebraic number field of degree n. We have the following :*

(i) $O_K \cap \mathbf{Q} = \mathbf{Z}$

(ii) $O_K \mathbf{Q} = K$

(iii) O_K *is a free abelian group of rank n.*

See for instance [2], Chap.II.

It is not easy in general to determine the ring of integers of an algebraic number field. However, this can be done for quadratic fields, and one gets the following result :

Theorem 2.6. *Let d be a square–free integer, and let $K = \mathbf{Q}(\sqrt{d})$ be the corresponding quadratic field. Then*

$$O_K = \mathbf{Z}[\sqrt{d}] \text{ if } d \equiv 3 \pmod 4.$$

$$O_K = \mathbf{Z}[\frac{\sqrt{d}+1}{2}] \text{ if } d \equiv 1 \pmod 4.$$

See for instance [2], 2.5.

In order to prove more properties of rings of integers, we need some more algebraic notions which will be developed in the next section.

3 Trace, norm and discriminant

Let F be a field, and let K be an extension of degree n of F. Let $x \in K$. Then the multiplication by x

$$m_x : K \to K$$

$$m_x(y) = xy$$

is an F–linear map. Let us define the *trace* of x by $Tr(x) = Tr(m_x)$ and the *norm* of x by $N(x) = det(m_x)$.

Let $x_1, \ldots, x_n \in K$. Let us denote by $D(x_1, \ldots, x_n)$ the determinant of the matrix $Tr(x_i x_j)$.

The following is well–known (see for instance [2], 2.6).

Proposition 3.1. *Let $x \in K$, and let $x_1, \ldots, x_n$ be the roots of the minimal polynomial of K over F counted with multiplicity $[K : F(x)]$. Then $Tr(x) = x_1 + \cdots + x_n$, and $N(x) = x_1 \cdots x_n$.*

Let K be an algebraic number field. Then we have

Proposition 3.2. *Let $x \in O_K$. Then $Tr(x), N(x) \in \mathbf{Z}$.*

Proof. This follows from the previous proposition, and from the fact that $O_K \cap \mathbf{Q} = \mathbf{Z}$. □

Definition 3.3. The *discriminant* of the number field K is by definition

$$disc(K) = D_K(e_1, \ldots, e_n)$$

where $\{e_1, \ldots, e_n\}$ is a basis of the free abelian group O_K.

It is easy to see that this does not depend on the choice of the basis $\{e_1, \ldots, e_n\}$. Indeed, if $\{f_1, \ldots, f_n\}$ is another basis, then $D_K(e_1, \ldots, e_n)$ and $D_K(f_1, \ldots, f_n)$ differ by the square of the determinant of the change of basis matrix A. As A is integral and invertible, $det(A) = \pm 1$, so $det(A)^2 = 1$. Hence $D_K(e_1, \ldots, e_n) = D_K(f_1, \ldots, f_n)$.

For more details concerning discriminants, see for instance [2], 2.7.

Example 3.4. Let $K = \mathbf{Q}(\sqrt{d})$ be a quadratic field. Let $x = a + b\sqrt{d}$ with $a, b \in \mathbf{Q}$. Then $Tr(x) = 2a$ and $N(x) = a^2 + db^2$.

If $d \equiv 3 \pmod 4$, then $disc(K) = 4d$, and if $d \equiv 1 \pmod 4$ then $disc(K) = d$.

We also need the notion of norm for ideals :

Definition 3.5. Let I be an ideal of O_K. Then we define the norm of I as being the number of elements of O_K/I.

Proposition 3.6. *If $I = \alpha O_K$ for some $\alpha \in O_K$, then $N(I) = |N(\alpha)|$.*

For a proof, see for instance [2], 3.5.

4 Euclidean division and principal ideal domains

We proved that $\mathbf{Z}[i]$ is a principal ideal domain by showing the existence of Euclidean division in this ring. Both notions make sense for arbitrary rings of integers, but they don't always hold, nor are they always equivalent.

Definition 4.1. Let K be an algebraic number field and let O_K be its ring of integers. We say that K is Euclidean (or that O_K is Euclidean) if for all $\alpha, \beta \in O_K$ with $\beta \neq 0$ there exist γ and δ in O_K such that $\alpha = \beta\gamma + \delta$ and that $|N(\delta)| < |N(\beta)|$.

It is easy to see that if K is Euclidean, then O_K is a principal ideal domain – the proof we saw for the Gaussian integers works in general. However, the converse is not true in general : for instance, if $K = \mathbf{Q}(\sqrt{-19})$, then O_K is a principal ideal domain but is not Euclidean.

On the other hand, being a principal ideal domain is equivalent to having unique factorization :

Proposition 4.2. *Let K be an algebraic number field and let O_K be its ring of integers. Then O_K is a principal ideal domain if and only if O_K has unique factorization.*

These equivalent properties do not always hold, as shown by the following example :

Example 4.3. Let $K = \mathbf{Q}(\sqrt{-6})$. Then $O_K = \mathbf{Z}[\sqrt{-6})]$. The element $6 \in O_K$ has two different decompositions into products of irreductible elements of O_K, namely

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

Let us check that $2, 3$ and $\sqrt{-6}$ are irreducible elements of O_K. If for instance we had $2 = \alpha\beta$ with $\alpha, \beta \in O_K$ and neither α nor β invertible in O_K then $N(2) = 4 = N(\alpha)N(\beta)$ and $N(\alpha) \neq 1$, $N(\beta) \neq 1$. This imples that $N(\alpha) = 2$. If $\alpha = a + b\sqrt{-6}$, $a, b \in \mathbf{Z}$, then $N(\alpha) = a^2 + 6b^2$, and clearly this cannot be equal to 2. Hence 2 is irreducible. Similarly, we show that 3 and $\sqrt{-6}$ are irreducible.

It is a basic problem in number theory to decide which algebraic number fields K have principal rings of integers.

5 Ideal class groups

As we saw in the previous section, unique factorization does not always hold in rings of integers of algebraic number fields. However, such a property exists at the level of ideals. The results below can be found in most books on algebraic number theory, for instance [2], Chap.III.

Let K be an algebraic number field and let O_K be the ring of integers of K. Recall that if I and J are ideals of O_K, then the product IJ is the set of finite sums of products ab with $a \in I$ and $b \in J$.

Theorem 5.1. *Let I be an ideal of O_K. Then there exist distinct prime ideals $P_1, \ldots, P_r$ and positive integers e_i such that*

$$I = P_1^{e_1} \ldots P_r^{e_r}.$$

Moreover, this decomposition is unique up to permutation.

It is also useful to note the following :

Proposition 5.2. *An ideal of O_K is maximal if and only if it is prime.*

In order to go further, we need the notion of *fractional ideal* :

Definition 5.3. A *fractional ideal* of K is a subset I of K such that there exist $\alpha \in O_K$ and an ideal J of O_K such that $\alpha I = J$.

Proposition 5.4. *For any prime ideal P of O_K, there exist a fractional ideal P' such that $PP' = O_K$.*

The ideal P' is called the *inverse* of P and is denoted by P^{-1}. With this notation, we get the following

Theorem 5.5. *Let I be an ideal of O_K. Then there exist distinct prime ideals $P_1, \ldots, P_r$ and integers e_i such that*

$$I = P_1^{e_1} \ldots P_r^{e_r}.$$

Moreover, this decomposition is unique up to permutation.

This shows that all fractional ideals of K are invertible, and hence the set of fractional ideals of K is a group. Let us denote this group by $\mathcal{I}_K$, and let $\mathcal{P}_K$ be the subgroup of principal fractional ideals.

Definition 5.6. The *ideal class group* of K is by definition the quotient group $C_K = \mathcal{I}_K/\mathcal{P}_K$.

Theorem 5.7. *The group C_K is finite.*

Definition 5.8. The *class number* of K is by definition the cardinal of the finite group C_K. It is denoted by h_K.

Note that $h_K = 1$ if and only if the ring O_K is a principal ideal domain. It is an important open question whether or not there are infinitely many algebraic number fields K with $h_K = 1$. The conjectured answer is "yes".

6 Decomposition of prime numbers in algebraic number fields

In our discussion of sums of two squares, we needed to know which prime numbers remained prime (equivalently, maximal) when extended to the ring of Gaussian integers. We will now examine this question in general. For the proofs of the results given below, see for instance [2], 5.2.

Let K be an algebraic number field of degree n, and let O_K be the ring of integers of K.

Let p be a prime number. Then there exist distinct prime ideals P_i and positive integers e_i such that

$$pO_K = P_1^{e_1} \ldots P_r^{e_r}.$$

We say that p *ramifies in* K (or that it is ramified in K) if there exists an i with $e_i > 1$. The integer e_i is called the *ramification index*. Otherwise, we say that p is *unramified* in K. This notion is related to the discriminant

Theorem 6.1. *A prime number p ramifies in K if and only if p divides $disc(K)$.*

This immediately implies that a number field has only finitely many ramified primes.

Example 6.2. Let $K = \mathbf{Q}(\sqrt{d})$ be a quadratic field, and let p be a prime number. If $p \equiv 1 \pmod 4$, then $disc(K) = d$, hence p ramifies in K if and only it divides d. If $p \equiv 3 \pmod 4$, then $disc(K) = 4d$, hence 2 and the prime divisors of d are ramified.

Definition 6.3. Let P be a prime ideal of O_K. The number $f_P = N(P)$ is called the *residual degree* of P.

Set $f_i = f_{P_i}$. Then we have

Theorem 6.4. *For any prime number p, we have $e_1 f_1 + \cdots + e_r f_r = n$.*

Let us denote by $G(K/\mathbf{Q})$ the set of field automorphisms of K that are the identity on $\mathbf{Q}$. Recall that $K/\mathbf{Q}$ is *Galois* if the number of elements of $G(K/\mathbf{Q})$ is equal to n, the degree of the extension $K/\mathbf{Q}$. If this is the case then we say that K is a *Galois number field*.

Proposition 6.5. *Suppose that K is a Galois number field. Then $e_i = e$ and $f_i = f$ for all $i = 1, \ldots, r$.*

In particular, the formula of the previous theorem becomes $efr = n$.

Definition 6.6. Let p be a prime number. We say that p is *inert* in K if $pO_K = P$ is a prime ideal.

In other words, we have $r = e = 1$. Note that this implies that $f = n$.

7 An example – cyclotomic fields

Let p be a prime number, and let ζ be a primitive p–th root of unity (that is, $\zeta^p = 1$ and $\zeta \neq 1$). Set $K = \mathbf{Q}(\zeta)$. This field is called the pth *cyclotomic field*.

Set $\phi_p(X) = X^{p-1} + \cdots + X + 1 \in \mathbf{Z}[X]$. Note that $X^p - 1 = \phi_p(X)(X-1)$, hence ζ is a root of ϕ_p. We have

Proposition 7.1. *The polynomial ϕ_p is irreducible.*

This follows from Eisenstein's criterion, see for instance [2], 2.9.

Corollary 7.2. *We have $[K : \mathbf{Q}] = p - 1$.*

Proof. Indeed, as ϕ is irreducible, we have $K \simeq \mathbf{Q}[X]/(\phi(X))$. But the degree of ϕ is $p-1$, hence $[K : \mathbf{Q}] = p - 1$. □

The following properties are proved for instance in [2], 2.9.

Proposition 7.3. *The ring of integers of K is $O_K = \mathbf{Z}[\zeta]$.*

Proposition 7.4. *The discriminant of K is p^{p-2}.*

This implies that the only ramified prime is p. We have $pO_K = p^{p-1}$, where the prime ideal P is principal generated by $\zeta - 1$. We have $r = f = 1$ and $e = p - 1$.

8 The canonical embedding of an algebraic number field

Let K be an algebraic number field of degree n, and let O_K be its ring of integers. It is interesting to study the $\mathbf{Q}$–linear embeddings $\sigma : K \to \mathbf{C}$. Some of these have their image contained in $\mathbf{R}$, and these are called *real embeddings of K*, the others are called *imaginary embeddings of K*. If σ is an imaginary embedding, then so is its complex conjugate $\overline{\sigma}$, hence imaginary embeddings come in pairs. Let r_1 be the number of real embeddings and $2r_2$ the number of imaginary embeddings of K. Then we have $r_1 + 2r_2 = n$ (see for instance [2], 4.2.).

Example 8.1. Let $K = \mathbf{Q}(\sqrt{d})$ be a quadratic field. If K is real, then $r_1 = 2$ and $r_2 = 0$, whereas if K is imaginary then $r_1 = 0$ and $r_2 = 1$.

Let

$$\sigma_1, \ldots, \sigma_{r_1} : K \to \mathbf{R}$$

be the real embeddings of K, and let

$$\sigma_{r_1+1}, \ldots, \sigma_{r_1+r_2}, \overline{\sigma}_{r_1+1}, \ldots, \overline{\sigma}_{r_1+r_2} : K \to \mathbf{C}$$

be the imaginary embeddings of K. The *canonical embedding* of K is by definition

$$\sigma : K \to \mathbf{R}^n$$

defined by $\sigma(x) =$

$$(\sigma_1(x), \ldots, \sigma_{r_1}(x), Re(\sigma_{r_1+1}(x)), Im(\sigma_{r_1+1}(x)), \ldots, Re(\sigma_{r_1+r_2}(x)), Im(\sigma_{r_1+r_2}(x))).$$

Let us recall that a *lattice* of rank n is a discrete subgroup of $\mathbf{R}^n$ that contains a basis of $\mathbf{R}^n$. The *covolume* of a lattice is by definition the volume of any fundamental domain (cf. for instance [2]. 4.2). Then for any ideal I of O_K, $\sigma(I)$ is a lattice of rank n.

Using these observations, we can apply geometric methods to obtain arithmetic results. One of the basic relations between geometry and number theory is the following

Proposition 8.2. *The covolume of $\sigma(O_K)$ is the absolute value of the discriminant of K.*

Using geometric methods, one obtains information concerning class numbers, for instance

Theorem 8.3. *(Minkowski bound) Every ideal class contains an integral ideal I with*

$$N(I) \leq (\frac{4}{\pi})^{r_2} \frac{n!}{n^n} |disc(K)|^{1/2}.$$

Cf. [2], 1.3. Using the same method, one obtains other finiteness results, for instance that there are only finitely many algebraic number fields with a given discriminant.

References

[1] A. Fröhlich, M. Taylor, *Algebraic Number Theory*, Cambridge studies in advanced mathematics **27**, Cambridge University Press (1991).

[2] P. Samuel, *Theorie algébrique des nombres*, Hermann, Collection Méthodes (1967).

[3] J-P. Serre, *A course in arithmetic*, Springer–Verlag, Graduate Texts in Mathematics (1993).

[4] P. Swinnerton-Dyer, *A Brief Guide to Algebraic Number Theory*. Cambridge University Press (2001).

[5] E. Weiss, *Algebraic Number Theory*, McGraw–Hill Book Company, Inc. (1963).

Mathematics and Gender Studies: an Overview

Andrea Blunck
University of Hamburg

Abstract

Mathematics and gender studies seem to be two disciplines that are far away from each other. The category "gender" does not seem to play any role in mathematics. So it is difficult to access mathematics from a gender studies perspective, and research on mathematics and gender is mostly concerned with meta-mathematical topics. I subdivide the research on mathematics and gender into four areas:

1. history of mathematics,
2. didactics of mathematics,
3. mathematics as field of study or work,
4. feminist science studies on mathematics.

In areas 1 - 3 there has been done quite a lot of work. For each of these areas I will present some typical questions and some interesting results. Research in area 4, however, is only at the beginning. Possible questions are: Does gender matter when a mathematical theory comes into being? Does mathematics participate in the construction of gender? I will present some more possible questions and some approaches how to tackle them.

References

[1] **Main source:** Andrea Blunck, Irene Pieper-Seier: Mathematik: Genderforschung auf schwierigem Terrain, in: Ruth Becker, Beate Kortendiek (eds.): *Handbuch Frauen- und Geschlechterforschung. Theorie, Methoden, Empirie, VS-Verlag*, Wiesbaden 2008, 812-820.

[2] Andrea Blunck: Research on Mathematics and Gender: Implications for Teaching, in: Maria Chionidou-Moskofoglou, Andrea Blunck, Renata Siemienska, Yvette Solomon, Renate Tanzberger (eds.): *Promoting Equity in Maths Achievement. The Current Discussion. Publicacions i Edicions Universitat de Barcelona* 2008, 127-132.

[3] Louise S. Grinstein, Paul J. Campbell (eds.): Women of Mathematics. A Biobibliographic Sourcebook. Greenwood Press, New York, 1987.

[4] Andrea Lenzner: Women in Mathematics. A Cross-Cultural Comparison. *Waxmann*, Münster, 2006.

[5] Heather Mendick: Masculinities in Mathematics. *Open University Press*, Maidenhead, 2006.

[6] Margaret A.M. Murray: Women Becoming Mathematicians. Creating a Professional Identity in Post-World War II America. *The MIT Press*, Cambridge Mass., 2001.

[7] Pat Rogers, Gabriele Kaiser (eds.): Equity in Mathematics Education. Influences of Feminism and Culture. *The Falmer Press*, London, 1995.

Teaching Mathematics and Gender at the University

Andrea Blunck
Hamburg

Abstract

In this talk I will explain what it means to teach "mathematics and gender studies" at a German university mathematics department. I will present the gender-related courses I give, namely:

- Women in the History of Mathematics,
- Gender and MIN (Mathematics, Informatics, Natural Sciences; joint course with Ingrid Schirmer, Dept. of Informatics),
- various seminars.

Moreover, I will explain why I think it is useful for students of mathematics to learn something about gender and mathematics.

References

[1] Andrea Blunck: Research on Mathematics and Gender: Implications for Teaching, in: Maria Chionidou-Moskofoglou, Andrea Blunck, Renata Siemienska, Yvette Solomon, Renate Tanzberger (eds.): *Promoting Equity in Maths Achievement. The Current Discussion. Publicacions i Edicions Universitat de Barcelona* 2008, 127-132.

On the evolutionary dynamics of virulence

*Barbara Boldin**
Department of Mathematics and Statistics
FIN-00014 University of Helsinki, Finland
barbara.boldin@helsinki.fi, barbara.boldin@gmail.com

Abstract

The aim of these notes is to demonstrate, by way of examples, how the techniques of Adaptive Dynamics can be used to study the evolutionary dynamics of infectious diseases. We focus on evolution of a single phenotypic trait, namely the disease induced death rate, or virulence. In a series of worked-out examples, we introduce the basic notions of Adaptive Dynamics and follow (some of) the development of evolutionary epidemiology through the years. We begin with the so called single infection model, discuss the conventional evolutionary wisdom and the trade-off hypothesis. Later on, we focus on the role of multiple infections in the evolution of infectious diseases. We investigate in more detail a superinfection model and discuss how the details of the superinfection process shape the course of evolution. In the last part of these notes, we introduce an example of a nested model that explicitly links the epidemiological dynamics at the host population level to the dynamics of infection in a single infected host. Such a nested model allows us to derive the precise form of the superinfection probability from the underlying mechanistic submodel of within-host dynamics.

1 Introduction

In 1973, the Russian evolutionary biologist Theodosius Dobzhansky wrote in one of his essays: "*Nothing in biology makes sense except in the light of evolution*" [21]. This is especially true for microorganisms, such as bacteria and viruses, for which it is now clear that evolution occurs not only on ecological time scales, but even during a course of an infection of a single infected host. For HIV-1 virus, for instance, the mutation rate per base pair is of the order of 10^{-5} to 10^{-4}. It is estimated that 10^9 replication cycles occur per day within a single infected individual, which means that a tremendeous selection pressure is exerted on the virus even during the course of a single infection [45, 46, 50]. Another example of the rapid evolution of pathogens are bacteria which have developed resistance to antibiotics, e.g. Methicillin-resistant *Staphylococcus aureus* (MRSA), Vancomycin-resistant *Staphylococcus aureus* (VRSA) or Vancomycin-resistant *enterococcus* (VRE) [6, 9, 12].

Natural selection acts on pathogens on several different levels. At the host population level, for instance, pathogens compete for susceptible hosts. These can either be uninfected, or, if multiple infections are possible, can also include already infected hosts. At the within-host level, pathogens compete for, for instance, uninfected target cells and even for resources within a single cell. These different levels of selection are interrelated and ideally, evolution of pathogens should be investigated by taking the different levels into account. However, this very quickly leads to complicated models. Traditionally, mathematical models explore selection pressures only at the host population level [5, 8, 13, 14, 26, 33, 34, 37, 40, 42, 47, 48, 51], but the importance of the so called nested (or embedded) models has increasingly been realized in the recent years [1, 2, 7, 11, 29, 39]. In these notes, we start simple by investigating the selection pressures at the host population level (i.e., by ignoring within-host dynamics). Later on, we present an example of a nested model that explicitly links the within-host pathogen dynamics to the traits that determine the spread of the infection at the host population level (e.g. virulence and transmissibility).

It is well documented that selection acts not only on different levels but also on different pathogen traits, such as for instance, the rate of pathogen reproduction within a host, the rate of evasion from the immune system, the infection induced death rate [32, 43, 52, 54]. In the first part of these notes we focus only on evolution of the disease induced death rate (virulence). When an explicit model of within-host dynamics is embedded into the epidemiological model, virulence will naturally be related to within-host production rate of the pathogen and we will thus focus on evolution of intra-host pruduction rate.

Clearly, some pathogens (such as the virus causing the common cold) are virtually avirulent, while others (for instance, the *ebola* virus) are almost always lethal. What are the factors that determine the levels of

*Acknowledgement. This work was supported by the Academy of Finland (Finnish Centre of Excellence in Analysis and Dynamics Research)

virulence of various pathogens? Extensive studies of different aspects of pathogen dynamics have shown that several mechanisms may explain the very different evolutionary paths of pathogens. Transmission mode (i.e., the way in which the pathogen is transmitted from one host to another) and host population regulation, for example, are known to play an important role in the evolution of pathogens [8, 18, 23, 24]. Another factor that plays a significant role, and one that we shall investigate in more detail in these notes, are multiple infections of the host. In general, a host that is already infected by some strain is not completely immune to infections by different strains. It seems reasonable to expect that the success of a reinfecting strain depends on several things, for instance, the difference in within-host competitive ability with the resident strain, the reinfection dose or the host susceptibility to another infection (this can either be reduced due to some partial immunity the first strain confers or increased because the host's immune system is weakened by the first infection). It is thus important to understand how different assumptions regarding the reinfection process shape the course of pathogen evolution.

Apart from a few introductory examples we mainly focus on the role of reinfection (in particular superinfection). We refer the reader to [18, 8, 23, 24, 47] and the references therein for some studies of other aspects of the evolutionary dynamics of infectious diseases.

Throughout the notes we use the tools of Adaptive Dynamics [27, 28, 19]. The reader who is not familiar with the terminology of Adaptive Dynamics can consult the boxes, where the basic notions are explained.

2 The basic model

We base the examples on a simple SI (Susceptible - Infected) model. Our basic assumptions are:

(i) The population birth rate is constant and is denoted by b. All newborns are susceptible.

(ii) Susceptible individuals die at a constant per capita rate d.

(iii) Infected individuals die at an increased per capita rate $d + \alpha$.

(iv) New infections occur according to the Law of Mass Action. That is, the rate at which an infected individual infects susceptible hosts is proportional to the abundance of susceptible hosts in the population, βS.

(v) Infected hosts become infectious at the moment of infection.

The disease induced death rate α is often called *virulence* (see however [10, 53] for other meanings of the term virulence). Note that assumptions (i) and (iv) imply that the pathogen is transmitted only horizontally (i.e. from one host to another) and not vertically (from the mother to a newborn child). Note also that we did not include any recovery which means that we limit ourselves to chronic pathogens (see [47] for a comparable study in the context of an SIR model).

If we denote by S and I the abundance of, respectively, susceptible and infected hosts, we can translate the above assumptions into the following system of ODEs,

$$\begin{aligned} \frac{dS}{dt} &= b - \beta SI - dS, \\ \frac{dI}{dt} &= \beta SI - (d+\alpha)I. \end{aligned} \tag{1}$$

System (1) has two equilibria: the infection free steady state,

$$\bar{S} = \frac{b}{d}, \quad \bar{I} = 0,$$

and the endemic steady state given by

$$\hat{S} = \frac{d+\alpha}{\beta}, \quad \hat{I} = \frac{b}{d+\alpha} - \frac{d}{\beta}. \tag{2}$$

The endemic equilibrium is biologically meaningful only when it is positive. This is the case precisely when the *basic reproduction ratio*, $\mathcal{R}_0$, is larger than 1. The basic reproduction ratio is defined as the expected number of new infections caused by a single infected host in an otherwise uninfected population [20]. In this case, $\mathcal{R}_0$ can easily be determined: since each infected individual is expected to live $\frac{1}{d+\alpha}$ units of time and is in that time expected to infect $\beta\frac{b}{d}$ new individuals, we find that

$$\mathcal{R}_0 = \frac{b\beta}{d(d+\alpha)}.$$

The basic reproduction ratio also determines stability of the two equilibria. If $\mathcal{R}_0 < 1$, the infection free steady state is the only biologically meaningful steady state and it is globally stable: when every infected individual produces, on average, less than one new infection, then every introduction of the infection will inevitably die out. If, on the other hand, $\mathcal{R}_0 > 1$, the endemic steady state is globally stable, while the infection free steady state is unstable (cf. [20], Exercise 3.11).

Our aim now is to investigate how virulence α changes in the course of evolution. To keep things simple, we shall assume that the host does not coevolve with the evolving pathogen. That is, the host parameters b and d are assumed to be fixed.

3 Evolution of virulence in the context of a Single Infection Model

In order to study the competition of multiple pathogen strains (characterized by different values of α) in a populations of hosts, we have to specify assumptions about how multiple strains are handled within a single infected host. We begin with the simplest possible assumption, namely that a host infected by one strain is completely protected from further infections (in other words, we assume complete cross immunity). This yields the so called *Single Infection Model*. In a special case where only two strains (characterized by virulence values α_1 and α_2) circulate in the population, we can describe the dynamics by

$$\begin{aligned} \frac{dS}{dt} &= b - \beta S I_1 - \beta S I_2 - dS, \\ \frac{dI_1}{dt} &= \beta S I_1 - (d + \alpha_1) I_1, \\ \frac{dI_2}{dt} &= \beta S I_2 - (d + \alpha_2) I_2. \end{aligned} \tag{3}$$

Suppose that a mutant strain α_m is introduced into a population in which the resident strain α_r is endemic. The mutant strain grows (or declines) according to

$$\frac{dI_m}{dt} = \big(\beta S - (d + \alpha_m)\big) I_m.$$

Box 1. Some basic notions of Adaptive Dynamics

The **invasion exponent** $r(x, y)$ is defined as the growth rate of a mutant population with trait y in the environment set by the resident population with trait x. If the invasion exponent is differentiable as a function of y in the point $y = x$, then the sign of the **selection gradient**

$$\frac{\partial r}{\partial y}\Big|_{y=x}$$

determines the direction of evolution from the resident trait x. If it is positive, the trait will (at least locally) increase in the course of evolution and if it is negative, the trait will (locally) decrease in the course of evolution. **Singular strategies** are trait values in which the selection gradient vanishes, i.e.

$$\frac{\partial r}{\partial y}\Big|_{y=x} = 0.$$

A singular trait x^* is called **convergence stable**, if a nearby strategy can be invaded (only) by traits that are nearer to x^*. That is, if $x < x^*$, then $r(x, y) > 0$ for $x < y < x^*$, while for $x > x^*$ the invasion exponent is positive when $x^* < y < x$. Convergence stable strategies are thus (local) attractors for monomorphic evolutionary dynamics.
An **evolutionarily stable strategy** (**ESS**) is a strategy that cannot be invaded by neighbouring traits. That is, x^* is an ESS if $r(x^*, y) < 0$ for $y \in (x^* - \varepsilon, x^* + \varepsilon)$ with some $\varepsilon > 0$. Despite the enticing 'stable' in its name, an ESS may not be an evolutionary attractor. If it is, it is called a **continuously stable strategy** (**CSS**).
An **evolutionary branching point** is a singular strategy that is convergent stable, but not an ESS.
If the invasion exponent is differentiable twice, then the second partial derivatives allow us to classify the singular points. In particular, if

$$\frac{\partial^2 r}{\partial y^2}\Big|_{y=x=x^*} < 0,$$

the point x^* is an ESS. If

$$\frac{\partial^2 r}{\partial y^2}\Big|_{y=x=x^*} < \frac{\partial^2 r}{\partial x^2}\Big|_{y=x=x^*},$$

then x^* is convergence stable. A singular point for which

$$\frac{\partial^2 r}{\partial x^2}\Big|_{y=x=x^*} > \frac{\partial^2 r}{\partial y^2}\Big|_{y=x=x^*} > 0$$

is a branching point.

Two basic assumptions of Adaptive Dynamics are that (i) mutants are introduced in small numbers and (ii) mutations are rare on the ecological time scale. The first assumption allows us to view the abundance of susceptibles S as depending only on the resident strain, α_r. That is, the mutant is so rare that it initially doesn't influence the environment into which it is introduced. The second assumption allows us to presume that the resident population has reached an equilibrium, $\hat{S}(\alpha_r)$.

The invasion criterion can thus be formulated in terms of the *invasion exponent*: if the per capita growth rate of a mutant α_m that is introduced into the resident population infected with α_r,

$$s(\alpha_r, \alpha_m) = \beta \hat{S}(\alpha_r) - (d + \alpha_m) \tag{4}$$

is positive, the mutant will invade, while the invasion fails if $s(\alpha_r, \alpha_m)$ is negative. This is indeed the case when we model invasions deterministically. If the model was stochastic, the invasion would still fail when the growth rate of the mutant is negative. With a positive growth rate, however, the mutant succeeds only with some positive (but smaller than 1) probability since the mutant may go extinct because of demographic stochasticity while it is still rare.

Box 2. Pairwise invasibility plots

Pairwise invasibility plots (**PIPs**) are a handy way of representing graphically the ability of a mutant trait to grow in the resident community. A PIP is constructed by plotting the sign of the invasion exponent $r(x, y)$ for all feasible pairs (x, y) of (resident, mutant) trait values.

If the resident population is at a stable equilibrium, then $r(x, x) = 0$ and so the zero contour lines contain at least the main diagonal. The shapes of other zero contour lines, if there are any, contain important information about the course of evolution. In particular, singular points are found as intersections of zero contour lines with the main diagonal. If we now imagine that black and white regions in the PIP represent the regions where the invasion exponent is, respectively, negative and positive, then the 'character' of a singular strategy can easily be recognized from a pairwise invasibility plot. Namely, if x^* is to be an ESS, and hence uninvadable by the neighbouring strategies, the straight vertical line through (x^*, x^*) must lie, at least locally, in the region where the invasion exponent is negative, i.e. in the black region. The singular trait x^* is convergence stable when the regions left of (x^*, x^*) are, at least close to the diagonal, white above the diagonal and black below the diagonal (i.e. x^* is locally attracting from the left), while the regions right of the point (x^*, x^*) are (at least close to the diagonal) black above the diagonal and white below the diagonal (in other words, x^* is locally attracting from the right).

We can reformulate the invasion criterion in terms of the basic reproduction ratio: the mutant α_m invades if the basic reproduction ratio of the mutant in the environment set by the resident,

$$\mathcal{R}_0(\hat{S}(\alpha_r), \alpha_m) = \frac{\beta \hat{S}(\alpha_r)}{d + \alpha_m} \tag{5}$$

exceeds 1, while the invasion fails if $\mathcal{R}_0(\hat{S}(\alpha_r), \alpha_m) < 1$.

3.1 Conventional wisdom.

It was believed for a long time that all pathogens would eventually evolve to be benign to their hosts. The words of the French-American microbiologist René Dubos (1965) reflect this, in that time widely accepted, idea: "*Given enough time, a state of peaceful coexistence eventually becomes established between any host and parasite.*"

In the context of our model, evolution to avirulence is certain if we assume that transmisibility β and virulence α are independent of one another. Using (2) and (4) we can then write the invasion exponent as

$$s(\alpha_r, \alpha_m) = \beta \hat{S}(\alpha_r) - (d + \alpha_m) = \alpha_r - \alpha_m. \tag{6}$$

Hence, mutants that decrease virulence are successful, while those that increase it are not. Assuming that mutualism is not possible (that is, α is always nonnegative), we conclude that evolution indeed drives virulence towards zero. We thus recover the so called *conventional evolutionary wisdom*: pathogens evolve to become avirulent. Figure 1 shows the (very trivial) corresponding pairwise invasibility plot.

3.2 The trade-off hypothesis

Supporters of the avirulence hypothesis argued that the reason we still observe virulent microorganisms today is simply that the process of pathogen adaptation to their hosts has not been long enough. However, there are many examples of host-parasite systems with very long coevolutionary history in which pathogens have not evolved towards avirulence [25, 31].

The first breaktrough in our understanding of why this could be the case came in the early 1980's with the work of Anderson and May [3, 4, 36, 37] and Levin and Pimentel [34]. What they suggested was a *trade-off* between virulence and transmissibility, essentially reflecting the idea that 'you don't get something for nothing': pathogens aim to increase transmission to new hosts, but cannot do so without simultaneously harming the host, i.e. increasing the host's death rate.

The trade-off hypothesis was doubted at first, mainly due to the lack of empirical support. However, there is now good experimental evidence that such trade-offs exist [16, 17, 22, 35, 41, 55]. Note, however,

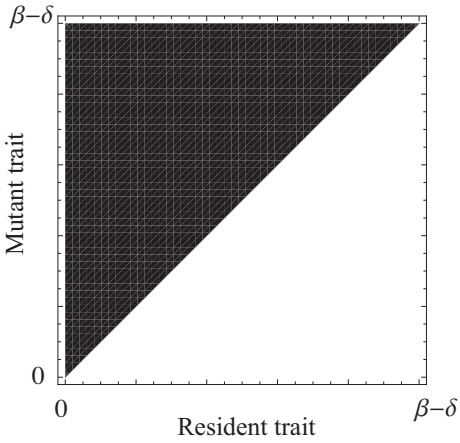


Figure 1: Pairwise invasibility plot corresponding to the invasion exponent in (6). White represents the regions where the mutant can invade, while the invasion fails in black regions. Pathogens can persist in the population when $\mathcal{R}_0 > 1$, which is equivalent to $\alpha < \beta - \delta$. Evolution drives the pathogens towards avirulence.

that the lack of empirical support for the existence of a trade-off between transmissibility and virulence in any particular case may simply be due to the fact that the harmful effects of the pathogen on the host manifest themselves in some other form, for instance in decreasing host's fecundity [44, 49].

So let us suppose that there exists a trade-off between virulence and transmission rate and let us see what this means for the evolution of the pathogen. A mutant strain α_m can invade when

$$s(\alpha_r, \alpha_m) = \beta(\alpha_m)\hat{S}(\alpha_r) - (d + \alpha_m) > 0 \tag{7}$$

where

$$\hat{S}(\alpha) = \frac{d + \alpha}{\beta(\alpha)}.$$

Thus, the mutant succeeds if it decreases $\hat{S}(\alpha)$ set by the resident. Since $s(\alpha_r, \alpha_m) > 0$ implies that $s(\alpha_r, \alpha_m) < 0$ (i.e., the resident cannot invade back), the evolution proceeds, in a series of trait substitutions, towards a local minimum of $\hat{S}(\alpha)$. The traits that (locally) minimize the steady state abundance of susceptible hosts are necessarily uninvadable and thus represent the possible end points of evolution, i.e., the continuously stable strategies.

The ultimate evolutionary winner is thus the strain that is able to persist in the worst possible environment, i.e. with the least amount of susceptible hosts. This is sometimes called the *pessimization principle* [19]. Note, incidentally, that minimization of $\hat{S}$ is equivalent to maximization of the basic reproduction ratio. The evolutionary winner is therefore the trait that (locally) maximizes $\mathcal{R}_0$.

The precise conclusions about the outcome of evolution will depend on the shape of the trade-off function $\beta(\alpha)$. If the trade-off is concave then there exists a single maximum of $\mathcal{R}_0$. If $\beta(\alpha)$ is convex, then there are no maxima of $\mathcal{R}_0$. There may, however, exist a single minimum. This minimum is an evolutionary repeller and represents a separating point for evolutionary outcomes: if the starting virulence is below the value that minimizes $\mathcal{R}_0$, evolution will drive virulence towards zero, while a starting point above the threshold virulence level means that virulence increases indefinitely. In Figure 2 we present the pairwise invasibility plots for two choices of $\beta(\alpha)$.

Remark 3.1. Evolution acts as an optimization only in a very special case, when the dimension of the environment equals one [38]. This is the case here where the only environmental variable for the pathogens is the abundance of susceptible hosts, $\hat{S}$. This simplicity comes not only because of the assumption of complete cross-immunity between strains but also because of the very simple demography of the host population. Once more realistic assumptions of density dependence in birth or death rates are included, the evolutionary dynamics becomes richer and we no longer find optimization. We refer the reader to the studies of Pugliese [48] and Svennungsen and Kisdi [51], where density dependence was taken into account. This shows that demography plays an important role in the evolution of pathogens.

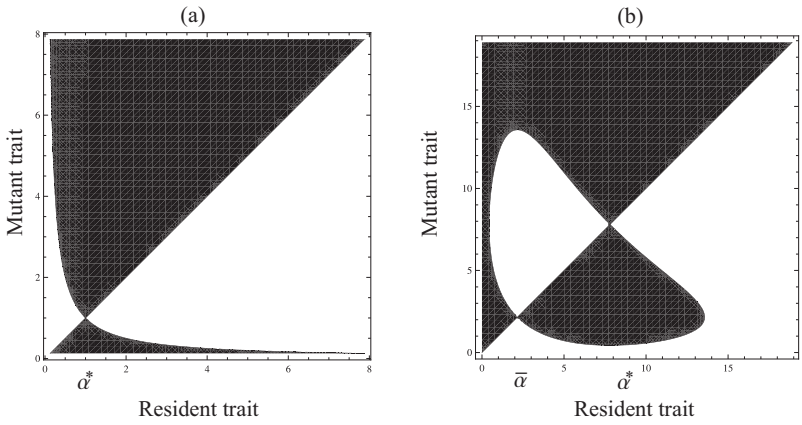


Figure 2: PIPs corresponding to (7) and (a) a concave trade-off function $\beta(\alpha) = \frac{10\alpha}{\alpha+1}$, (b) a convex-concave trade-off function $\beta(\alpha) = \frac{A}{B+(1-B)e^{-C\alpha}}$ with $A = 2, B = 0.1, C = 0.4$. In (a), evolution drives virulence to some intermediate value, denoted by α^*. In (b), the point $\bar{\alpha}$ is an evolutionary repeller: traits starting below this value evolve towards zero, traits above evolve towards the continuously stable strategy α^*.

4 The superinfection model

In this section we relax the assumption of complete cross-immunity and consider the possibility of reinfections.

Multiple infections within a single host can be modeled in different ways. *Superinfection models* assume that within-host dynamics is fast compared to processes at the host population level. If an individual infected by strain α_1 is reinfected by another strain α_2, then the better within-host competitor immediately ousts the other strain and takes over the host (here we have in mind a very simplistic within-host scenario where, in the long run, only one strain can persist inside a host). *Coinfection models*, on the other hand, incorporate also the transient dynamics where the host harbors both strains α_1 and α_2.

Since there will always be a period in which a reinfected host harbours more than one strain, coinfection models may be argued to be more realistic than superinfection model. The added realism, however, does not come for free and the models very quickly become untractable when the number of strains increases. In these notes we limit ourselves to studying the superinfection models and refer the reader to [40, 18] for a study of virulence evolution in the context of a coinfection model and for a derivation of superinfection as the limiting case of a coinfection process.

To include the possibility of superinfections we extend the single infection model with two strains to

$$\begin{aligned}
\frac{dS}{dt} &= b - \beta(\alpha_1)SI_1 - \beta(\alpha_2)SI_2 - dS,\\
\frac{dI_1}{dt} &= \beta(\alpha_1)SI_1 + \beta(\alpha_1)\psi(\alpha_2,\alpha_1)I_1I_2 - \beta(\alpha_2)\psi(\alpha_1,\alpha_2)I_1I_2 - (d+\alpha_1)I_1,\\
\frac{dI_2}{dt} &= \beta(\alpha_2)SI_2 + \beta(\alpha_2)\psi(\alpha_1,\alpha_2)I_1I_2 - \beta(\alpha_1)\psi(\alpha_2,\alpha_1)I_1I_2 - (d+\alpha_2)I_2,
\end{aligned} \tag{8}$$

where ψ denotes the *superinfection function*. More precisely, we define

$$\begin{aligned}
\psi(\alpha_1,\alpha_2) := {} & \textit{the probability that, upon reinfection, the newly infecting}\\
& \textit{strain } \alpha_2 \textit{ takes over the host that is already infected with } \alpha_1.
\end{aligned}$$

Remark 4.1. The model could in principle include a rather more general description of a superinfection. For instance, instead of $\beta(\alpha_2)\psi(\alpha_1,\alpha_2)$, one could write $\beta(\alpha_2)\psi(\alpha_1,\alpha_2)\sigma(\alpha_1)$ to take into account the fact that already infected individuals may differ in their susceptibility to an infection from uninfected individuals. If $0 < \sigma(\alpha_1) < 1$, then infection with α_1 has conferred some partial immunity and the individual is less susceptible to infection with α_2 than an uninfected host. If $\sigma(\alpha_1) > 1$, on the other hand, the host resistance to infection by α_2 has decreased because of the existing infection by α_1. Such modifications would be easy to include, however, to keep the presentation simple, we choose not to do so.

We shall in fact assume that the probability of superinfection depends only on the difference of the two strains. That is, we shall write

$$\psi(\alpha_1,\alpha_2) = \phi(\alpha_2 - \alpha_1)$$

and, to shorten the notation, we define

$$\Phi(\alpha_1, \alpha_2) := \beta(\alpha_2)\phi(\alpha_2 - \alpha_1) - \beta(\alpha_1)\phi(\alpha_1 - \alpha_2). \tag{9}$$

The invasion exponent now takes the form

$$\begin{aligned} r(\alpha_r, \alpha_m) &= \beta(\alpha_m)\hat{S}(\alpha_r) - (d + \alpha_m) + \Phi(\alpha_r, \alpha_m)\hat{I}(\alpha_r) \\ &= s(\alpha_r, \alpha_m) + \Phi(\alpha_r, \alpha_m)\hat{I}(\alpha_r), \end{aligned} \tag{10}$$

where s is the invasion exponent from the single infection model. Note that the resident equilibrium is the same as in the basic model since the resident is assumed to consist of one strain only and thus no superinfections take place.

The assumptions regarding the superinfection function are now crucial and, as we shall see below, different choices can lead to very different evolutionary outcomes. Note that, since mutations are assumed to be small, the outcome of invasion relies only on the behaviour of ϕ in the vicinity of zero (the shape of ϕ away from zero, however, plays a role in global and in polymorphic dynamics; see [8]).

We assume that the superinfection function is a nonnegative, increasing function and consider the following three classes of superinfection functions:

(A) $\phi(\alpha) = 0$ for $\alpha \leq 0$, ϕ has a jump discontinuity in $\alpha = 0$,

(B) $\phi(\alpha) = 0$ for $\alpha \leq 0$, ϕ is continuous in $\alpha = 0$ and is differentiable twice in zero from the right with $\phi'_+(0) > 0$,

(C) $\phi(0) > 0$, ϕ is differentiable.

Selection gradient exists in cases (B) and (C). We find that a singular strategy α^* satisfies

$$\frac{\partial r}{\partial \alpha_m}\Big|_{\alpha_m=\alpha_r=\alpha^*} = \frac{\partial s}{\partial \alpha_m}\Big|_{\alpha_m=\alpha_r=\alpha^*} + \Phi_0\hat{I}(\alpha^*) = \beta'(\alpha^*)\hat{S}(\alpha^*) - 1 + \Phi_0\hat{I}(\alpha^*) = 0, \tag{11}$$

where

$$\Phi_0 = \begin{cases} \beta(\alpha^*)\phi'_+(0), & \text{in case (B)} \\ \beta'(\alpha^*)\phi(0) + 2\beta(\alpha^*)\phi'(0), & \text{in case (C)}, \end{cases}$$

Since $\beta(\alpha)$ is assumed to be increasing we see that $\Phi_0 > 0$ in both case (B) and case (C). It is then clear from (11) that superinfections drive virulence beyond the point that maximizes $\mathcal{R}_0$. In other words, when a host can be superinfected, pathogens evolve to be more virulent than when there is complete cross-immunity between strains.

The three classes of superinfection functions give very different evolutionary outcomes. In the examples that follow we give some biological motivation for a particular class and study the corresponding adaptive dynamics.

4.1 Case A

We first consider the case where ϕ is zero on $(-\infty, 0]$ and moreover has a jump discontinuity at zero. One may, for instance, have in mind

$$\phi(\alpha) = \begin{cases} 1, & \alpha > 0 \\ 0, & \alpha \leq 0, \end{cases} \tag{12}$$

which corresponds to the deterministic description of an invasion: reinfection with a more virulent strain succeeds with probability 1, while a reinfection with a less (or equally) virulent strain always fails.

The discontinuity at the origin implies that even a slightly larger mutant strain will successfully invade the resident strain in the population and that the less virulent strain can never invade back. Indeed, for $\alpha_m = \alpha_r \pm \varepsilon$, the term $\Phi(\alpha_r, \alpha_m)$ in (10) is $\mathcal{O}(1)$, while $s(\alpha_r, \alpha_m)$ is of the order $\mathcal{O}(\varepsilon)$, and so it is the sign of $\Phi(\alpha_r, \alpha_m)$ that determines the outcome of an invasion. Evolution thus increases virulence with every successful mutation. In the long run, therefore, virulence increases either indefinitely or to the maximum virulence level that still allows the infection to persist in the population. As we have seen in the previous section, the convexity of the trade-off determines which of the two scenarios applies.

The evolution of virulence in the context of a superinfection model with a discontinuous superinfection function has been studied already in the 1990's. In their 1994 paper [42], Nowak and May consider the superinfection functions of the form $c\phi$ with ϕ in (12) and some $c > 0$. They find that, assuming that mutations are generated uniformly on some interval $[\alpha_{\min}, \alpha_{\max}]$, a continuum of strains can coexist on the evolutionary time scale.

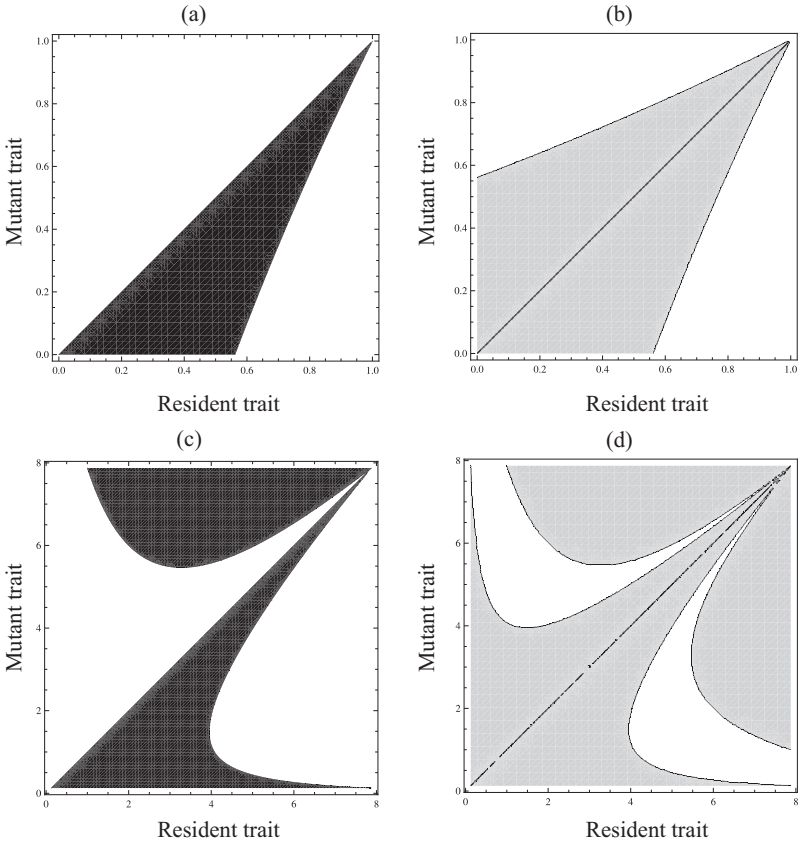


Figure 3: Case (A), superinfection function is 2ϕ with ϕ in (12): (a) PIP corresponding to $\beta = 2$, (b) regions of mutual invadability in (a) are depicted in white, (c) PIP corresponding to $\beta(\alpha) = \frac{10\alpha}{\alpha+1}$, (d) regions of coexistence in (c) are depicted in white.

Adaptive Dynamics, on the other hand, assumes more realistically, that the mutants arise locally around the strains that are already present in the population. In Figure 3, we present the pairwise invasibility plots corresponding to the superinfection function 2ϕ with ϕ as in (12). We moreover consider two choices of β: in the top row (Figures 3a and 3b) we consider a constant transmissibility β while in the bottom row (Figures 3c and 3d) we take $\beta = \frac{10\alpha}{\alpha+1}$.

As predicted, we observe increase of virulence in the course of evolution via a series of trait substitutions. For comparison, we note that in the case of a constant transmission rate, the single infection model predicts evolution towards avirulence (i.e. $\alpha = 0$; see Figure 1), while the second choice of β would lead to some intermediate level of virulence (cf. Figure 2a). When virulence evolves into vicinity of $\alpha_{\max}$ (the maximum virulence value that still allows persistence of infection in the population), however, some pairs of strains (α_1, α_2) are *mutually invadable*, which means that both $r(\alpha_1, \alpha_2)$ and $r(\alpha_2, \alpha_1)$ are positive (i.e., both (α_1, α_2) and (α_2, α_1) fall into the white region of the PIP; see Figures 3b and 3d). In the case of constant β, this region is more easily accessible by small mutations around the boundary $\alpha_{\max}$.

Thus, we find coexistence of two strains. What happens with dimorphisms on the evolutionary time scale? Since the environment the pathogens experience has now become three dimensional (set by $\hat{S}, \hat{I}_{\alpha_1}, \hat{I}_{\alpha_2}$) it is now in principle possible that a third strain could coexist with α_1 and α_2. The dimension of the environment then increases to four, so there is a possibility of the fourth strain, and so on. However, the dynamics of polymorphisms has, to our knowledge, never been investigated in the context of this model and we thus end here the discussion of Case A.

4.2 Case B

We now assume that the superinfection function is continuous in $\alpha = 0$. As we shall see in the following section (when the within-host dynamics is taken into account), the continuity of the superinfection function at the origin arises naturally when we consider the reinfection process as a stochastic event.

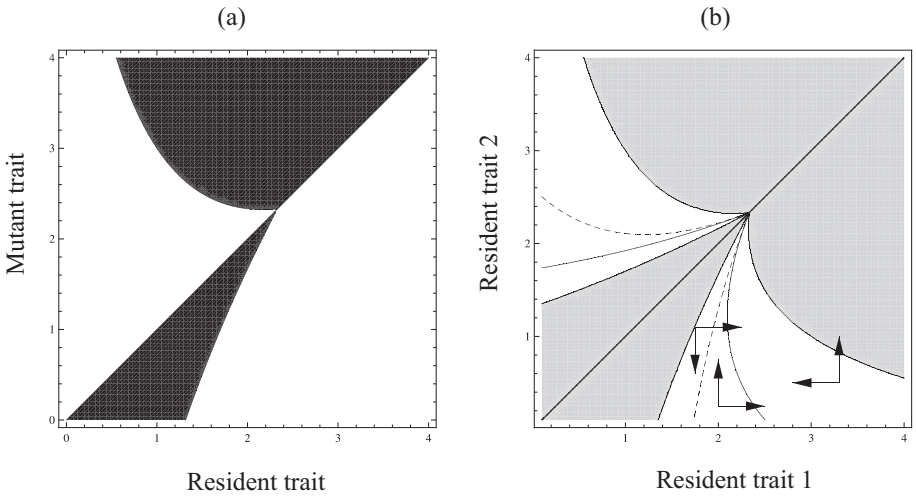


Figure 4: (a) Pairwise invasibility plots corresponding to Example 4.2: $\beta(\alpha) = 10$ and $\phi(\alpha) = \frac{\alpha}{\alpha+2}$ for $\alpha > 0$, (b) The regions of coexistence (depicted in white), along with isoclines that depict the evolution of dimorphisms. The isocline $\frac{\partial R(\alpha_1,\alpha_2,\alpha_m)}{\partial \alpha_m}|_{\alpha_m=\alpha_1} = 0$ is depicted with a full line (in the interior of the white region), while the dashed line represents the isocline $\frac{\partial R(\alpha_1,\alpha_2,\alpha_m)}{\partial \alpha_m}|_{\alpha_m=\alpha_2} = 0$.

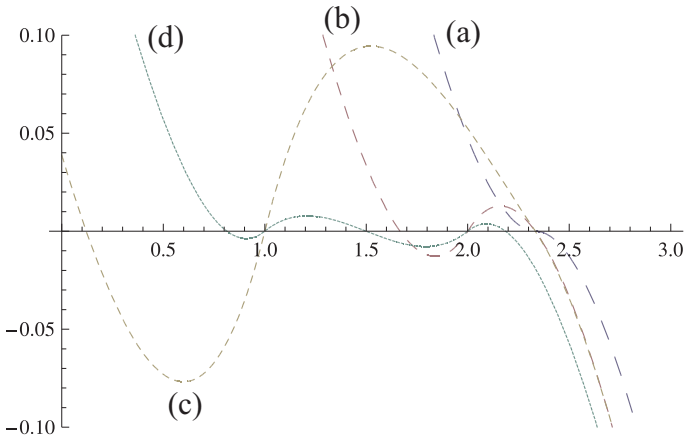


Figure 5: Plots of (a) $r(\alpha^*, \alpha_m)$, (b) $R(\alpha^*, 2, \alpha_m)$, (c) $R(\alpha^*, 1, \alpha_m)$ and (d) $R(1, 2, \alpha_m)$.

Even though the superinfection function may not be differentiable, the selection gradient exists and is given by

$$\frac{\partial r}{\partial \alpha_m}\Big|_{\alpha_m=\alpha_r} = \frac{\partial s}{\partial \alpha_m}\Big|_{\alpha_m=\alpha_r} + \beta(\alpha_r)\phi'_+(0)\hat{I}(\alpha_r), \tag{13}$$

which means that the singular strategies can be determined by calculating the points in which the selection gradient vanishes. Note that, if $\phi'_+(0) = 0$, the singular strategies coincide with the ones obtained from the single infection model.

Since r isn't differentiable twice we cannot characterize the singular strategies in the usual way, however we can write

$$\frac{\partial^2 r}{\partial \alpha_m^2}\Big|_{\alpha_m=\alpha_r=\alpha^*} = \begin{cases} \beta''(\alpha^*)\hat{S}(\alpha^*) + 2\beta'(\alpha^*)\hat{I}(\alpha^*)\phi'_+(0) + \hat{I}(\alpha^*)\beta(\alpha^*)\phi''_+(0), & \alpha > \alpha^* \\ \beta''(\alpha^*)\hat{S}(\alpha^*) - \hat{I}(\alpha^*)\beta(\alpha^*)\phi''_+(0), & \alpha < \alpha^*, \end{cases} \tag{14}$$

Expression in (14) allows us to determine whether the singular strategies are invadable or not. The discontinuity of ϕ' at the origin implies that the non-generic type of singular points, which are invadable from one side but uninvadable from the other, may now be the rule rather than the exception. We demonstrate this on two examples.

Example 4.2.

Suppose that the transmission rate β is constant. Singular strategies are then given by

$$\alpha^* = \frac{\beta b}{\frac{1}{\phi'_+(0)} + d} - d.$$

Furthermore, expression (14) simplifies to

$$\frac{\partial^2 r}{\partial \alpha_m^2}\Big|_{\alpha_m=\alpha_r=\alpha^*} = \begin{cases} \beta \hat{I}(\alpha^*)\phi''_+(0), & \alpha > \alpha^* \\ -\beta \hat{I}(\alpha^*)\phi''_+(0), & \alpha < \alpha^*, \end{cases} \tag{15}$$

In this case, the singularity will always be invadable from one side but not from the other. The curvature of the superinfection function at the origin determines which of the two sides is invadable: if $\phi''_+(0) > 0$, then the singularity is invadable from above and uninvadable from below, and vice versa if $\phi''_+(0) < 0$.

In Figure 4a we show the pairwise invasibility plot corresponding to $\phi(\alpha) = \frac{0.5\alpha}{0.5\alpha+1}$ (for $\alpha > 0$ and $\phi(\alpha) = 0$ otherwise). The fact that ϕ is concave implies that the singular strategy $\alpha^* = \frac{7}{3}$ is invadable from below and uninvadable from above. Simple geometric arguments show that there must exist a region of mutual invadability close the singular strategy. Hence, after evolution has brought virulence in the vicinity of α^* the population becomes dimorphic. To decide whether any such dimorphism would be converging or diverging, we calculate the invasion exponent with two resident strategies, α_1 and α_2, $R(\alpha_1, \alpha_2, \alpha_m)$. Because of continuity of invasion exponent, the graph of R will be (for small perturbations) similar to the graph of r and R will thus in a generic case have three roots. In Figure 5 we show the graphs of R for a few choices of dimorphic residents. The nature of dimorphisms can, however, most easily be determined using the isoclines

$$\frac{\partial R(\alpha_1, \alpha_2, \alpha_m)}{\partial \alpha_m}\Big|_{\alpha_m=\alpha_1} = 0 \quad \text{and} \quad \frac{\partial R(\alpha_1, \alpha_2, \alpha_m)}{\partial \alpha_m}\Big|_{\alpha_m=\alpha_2} = 0,$$

which we show in Figure 4b. The isoclines, along with the arrows showing the direction of dimorphic evolution, reveal that, in this case, dimorphisms are only of transient nature and eventually all converge to the monomorphic singularity. Further numerical experiments (not shown here) reveal, however, that divergent dimorphisms are possible as well in the context of this model (see also [7]).

Example 4.3.

Let us now consider the trade-off $\beta(\alpha) = \frac{10\alpha}{\alpha+1}$ and the superinfection functions of the form

$$\phi_a(\alpha) = \begin{cases} \dfrac{a\alpha}{a\alpha + 1}, & \alpha > 0 \\ 0, & \alpha \leq 0, \end{cases} \tag{16}$$

for some $a \geq 0$. Note that $\phi'_{a+}(0) = a$ and $\phi''_{a+}(0) = -2a^2$. The parameter a therefore represents the slope of the superinfection function at the origin (to the right). As we shall see in the next section, the slope of ϕ at the origin is related to the reinfection dose, i.e. the number of pathogens of the superinfecting strain. So how do the singular strategies depend on the value of a?

In Figure 6 we show a series of pairwise invasibility plots corresponding to increasing values of a. Note that the case $a = 0$ corresponds to the single infection model (superinfections are not possible) and we thus recover the PIP in Figure 2a. The singular strategy is a CSS in this case.

When a increases, the singular point increases and moves towards to boundary $\alpha_{\max}$. In the limit $a \to \infty$ we should recover the PIP from case (A) since for $a \to \infty$ the superinfection functions converge to the step function in (12).

Note that, both the trade-off function and the superinfection functions are concave. This implies that the second derivative in (14) is always negative for $\alpha > \alpha^*$, which means that the singular strategy is never invadable from above. For small values of a, the singular strategy is uninvadable also from below and is thus a CSS. At some critical value $\bar{a}$ a bifurcation occurs and the singular strategy becomes invadable from below (see Figure 7).

4.3 Case C

The case where $\phi(0) > 0$ corresponds to the situation where both the resident and the invading population are assumed to be finite and subject to demographic stochasticity. Indeed, if N denotes the abundance of the resident strain and n the number of newly introduced pathogens of identical virulence, then the new

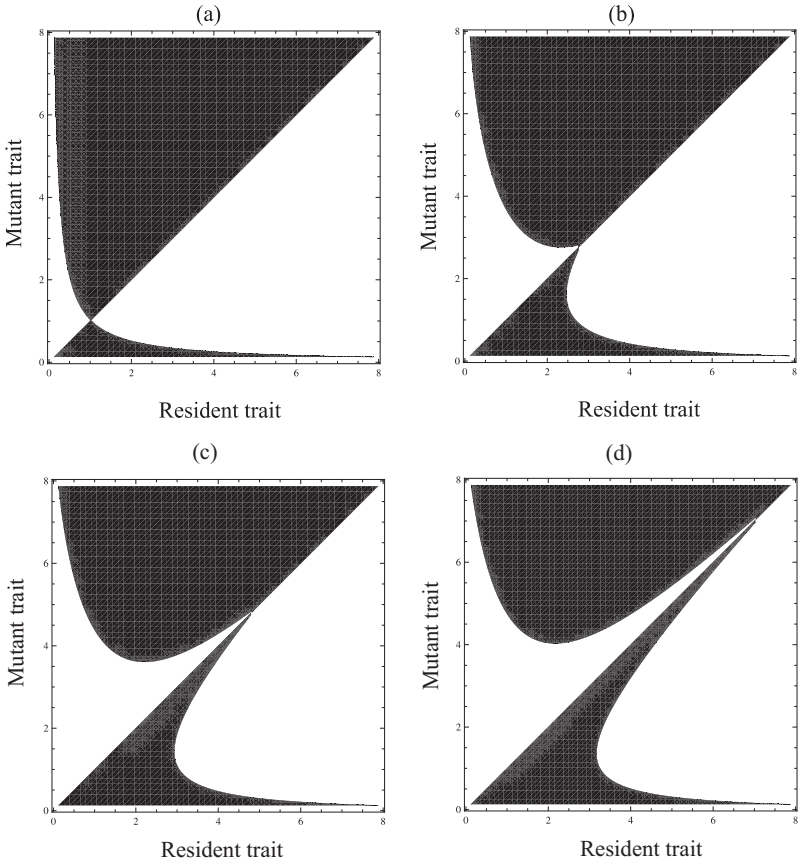


Figure 6: Pairwise invasibility plots corresponding to Example 4.3: $\beta(\alpha) = \frac{10\alpha}{\alpha+1}$ and ϕ_a in (16) with (a) a = 0.01, (b) a = 0.7, (c) a = 2, (d) a =20.

infection settles with probability $\phi(0) = \frac{n}{n+N} > 0$. In reality, n will typically be very small. On the other hand, N is large and in the limiting case, where N is considered to infinite we obtain $\phi(0) = 0$, as was the case in previous examples. The additional assumption of differentiability in Case (C) is made purely to simplify the analysis.

The case where $\phi(0) > 0$ was studied extensively by Pugliese in [48] and more recently by Boldin et al. in [8] (where, in addition, different assumptions about the host population regulation were investigated). We refer the reader to the papers for more details, here we only summarize the main findings.

As in Case (A) and (B), we no longer have optimization in the course of evolution (which can easily be recognized by the loss of skew symmetry in PIPs). However, superinfections do not imply evolutionary coexistence per se. By writing out the second derivative

$$\frac{\partial^2 r}{\partial \alpha_m^2}\Big|_{\alpha_m=\alpha_r=\alpha^*} = \beta''(\alpha^*)(\hat{S}(\alpha^*) + \phi(0)) + 2\beta'(\alpha^*)\phi'(0)$$

we observe that the curvature of the trade-off plays a role in the characterization of singular strategies. It was shown in [48] that, if the trade-off function belongs to a certain family of concave functions, the singular strategy is unique and it is always a CSS, which means that dimorphisms, if they occur, are only of transient nature and are eventually resolved in a CSS. As was shown by Boldin et al. in [8], branching points can be found, even among the concave trade-offs. In [8] it was furthermore investigated how the assumptions about the host population regulation influence the occurence of branching. This was done by investigating in detail three population dynamics regimes, (i) the constant population birth rate (as we assume here), (ii) constant population size and (iii) logistic population growth. Case (i) appears to be the most conducive to branching. Branching is found also in other models, however, the convexity ranges of the trade-offs that yield branching are narrower. Moreover, we found mutual exclusion, which is contrary to the common belief that superinfections promote coexistence. We refer the reader to [8] for some more detailed examples and for a discussion on the dynamics after branching.

Remark 4.4. In these notes we investigated the adaptive dynamics by choosing a particular trade-off. Using the so called *critical function analysis*, one can turn the question around and ask: which trade-offs lead to

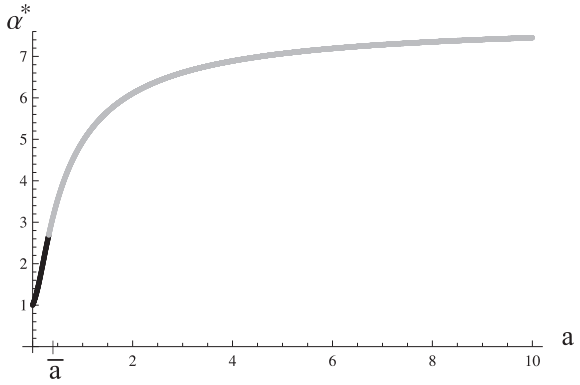


Figure 7: Singular strategy α^* as a function of a (Example 4.3). The black line shows the region where α^* is a CSS. The grey line shows the region where the singular strategy invadable from below, but not from above.

a particular outcome (for instance, branching)? We shall not go into the details of critical function analysis here but refer the interested reader to [8, 51] for examples.

5 Linking population dynamics to the dynamics within the host

Even though transmissibility β as well as virulence α are likely to be related to the individual's within host state of infection (such as for instance, the amount of viruses the individual harbors), our modeling thus far ignored a detailed descripition of the infection within a host. As a consequence, we had to settle with some phenomenological trade-off $\beta = \beta(\alpha)$ and superinfection function ϕ.

In this secion we introduce an explicit model of within-host pathogen dynamics that follows the time evolution of target cells and free pathogens. This will allow us to make more natural (and ultimately more easily tested by experiments) assumptions about how transmissibility and virulence depend on individual's infection state. Moreover, such description will allow us to derive the superinfection function from the mechanistic intra-host submodel.

This section is based on the modeling and analysis presented in [7].

5.1 A model of within-host pathogen dynamics

We describe the dynamics inside a single host using three variables: T, T^* and V represent, respectively, the number of uninfected and infected target cells and the number of free pathogens. We assume that

(i) In the absence of infection, target cells are produced at a constant rate λ and die at a constant per capita rate δ.

(ii) Free pathogens inside a host die at per capita rate c.

(iii) Infections of uninfected target cells are described by the mass action term kVT. That is, the rate at which pathogens find uninfected target cells, successfully bind to the surface of the cell and/or enter the target cell, is proportional to the product of the numbers of uninfected target cells and free pathogens. Upon infection, the uninfected target cell and the pathogen that infected it, form an infected cell.

(iv) Infected target cells produce free pathogens at a rate p. This production comes at a cost, namely, it increases the death rate of infected target cells by $\mu(p)$. We assume that μ is a nonnegative, increasing function of the production rate p.

These assumptions yield the following system of ODEs,

$$\begin{aligned} \frac{dT}{dt} &= \lambda - kVT - \delta T \\ \frac{dT^*}{dt} &= kVT - (\delta + \mu(p))T^* \\ \frac{dV}{dt} &= pT^* - kVT - cV. \end{aligned} \tag{17}$$

System (17) has two equilibria: the infection free steady state,

$$\bar{V} = \bar{T}^* = 0 \quad \text{and} \quad \bar{T} = \frac{\lambda}{\delta} \tag{18}$$

and the nontrivial equilibrium given by

$$\begin{aligned} \hat{T} &= \frac{c}{k(\mathcal{B}_0(p)-1)} \\ \hat{T}^* &= \frac{\mathcal{B}_0(p)}{p}\Big(\lambda - \frac{cd}{k(\mathcal{B}_0(p)-1)}\Big) \\ \hat{V} &= \frac{\lambda}{c}\Big(\mathcal{B}_0(p)-1\Big) - \frac{\delta}{k}. \end{aligned} \tag{19}$$

Here, $\mathcal{B}_0$ stands for the so called *burst size*, i.e. the expected number of pathogens produced by one infected target cell. If pathogens are produced at rate p, then

$$\mathcal{B}_0(p) = \frac{p}{\delta + \mu(p)}.$$

The nontrivial steady state given by (19) is biologically meaningful only when all three components in (19) are positive. This is the case when the within-host basic reproduction ratio of a pathogen, $\mathcal{R}_0^w$ (the superscript w serves to distinguish it from the pathogen's basic reproduction ratio at the host population level) exceeds one. $\mathcal{R}_0^w$ is defined as the expected number of new pathogens produced by a single pathogen introduced into a virgin cell environment. Since free pathogens need to enter uninfected target cells in order to reproduce and since the probability with which the pathogen enters a target cell in a virgin environment equals $\frac{k\lambda}{k\lambda+dc}$, the within-host basic reproduction ratio of a pathogen with trait p equals

$$\mathcal{R}_0^w(p) = \frac{k\lambda}{k\lambda + \delta c}\mathcal{B}_0(p).$$

When the nontrivial equilibrium exists, it is locally asymptotically stable, while the infection free steady state is unstable in that case (see [15] for a global stability result).

We now consider the rate of pathogen production p as the (only) trait that is subject to natural selection. All the other parameters in the within-host model will be kept constant throughout. We furthermore assume for simplicity that, if a target cell is infected with one trait, it is protected from further infections. In other words, we do not consider superinfections or coinfections at the cell level. Since this assumption implies that the pathogens compete within a host for only one resource, i.e. uninfected target cells, the evolutionary dynamics at the within-host level is very simple. Namely, when a mutant trait, say q, is introduced into a host where the trait p is resident, the mutant is successful (according to the deterministic model) if and only if it exploits the resource better than the resident, i.e. when $\hat{T}(q) < \hat{T}(p)$. Note, incidentally, that minimization of $\hat{T}$ is equivalent to maximization of $\mathcal{R}_0^w$ and also to maximization of $\mathcal{B}_0$.

5.2 The superinfection model revisited

We can now rewrite the superinfection model in the form

$$\begin{aligned} \frac{dS}{dt} &= b - \beta(p)SI_p - \beta(q)SI_q - dS \\ \frac{dI_p}{dt} &= \beta(p)SI_p + \Phi(q,p)I_pI_q - (\alpha(p)+d)I_p \\ \frac{dI_q}{dt} &= \beta(q)SI_q + \Phi(p,q)I_pI_q - (\alpha(q)+d)I_q, \end{aligned} \tag{20}$$

where now

$$\Phi(p,q) = \beta(q)\phi(p,q) - \beta(p)\phi(q,p)$$

and the superinfection function $\phi(p,q)$ is defined as

$$\begin{aligned} \phi(p,q) := \; &\textit{probability that the reinfecting strain } q \textit{ wins the within-host competition} \\ &\textit{with strain } p \textit{ and takes over the host that is already infected by } p. \end{aligned}$$

As before, we could now investigate how different choices of superinfection function shape the course of evolution. However, the mechanistic submodel of within-host dynamics now allows us to derive explicit expression for the superinfection probability directly from the underlying branching process. We now present the derivation.

5.3 The dynamics in the initial stages of a superinfection

In the initial stages of a superinfection, the invading trait q is likely to be present only in small quantities. Hence, even when $\hat{T}(q) < \hat{T}(p)$ (and so the newly introduced trait has the potential to outcompete the resident trait), trait q may go extinct due to demographic stochasticity in the initial stages of a superinfection, when it is still rare. We thus describe the initial stages of an invasion as a stochastic birth-and-death process. At this point, lytic viruses have to be distinguished from the non-lytic (or budding) viruses. Here we present the derivation of the superinfection probability only for non-lytic viruses and refer the reader to [47] for a similar analysis of lytic viruses.

Suppose first that only one free pathogen with trait q is introduced into a host that is already infected by trait p. If we assume that the trait p resides at a stable equilibrium, then the new trait q is introduced into an environment given by the steady state value of $\hat{T}(p)$,

$$\hat{T}(p) = \frac{c}{k(\mathcal{B}_0(p) - 1)}. \tag{21}$$

The probability that the clan of this initially introduced pathogen survives in an already infected host, is given as the smallest fixed point of a generating function [30]. In order to compute it, we must first derive the probabilities π_n with which one free pathogen with trait q will produce n new pathogens.

In order to reproduce, a pathogen must bind to an uninfected target cell. This happens with probability

$$\frac{k\hat{T}(p)}{k\hat{T}(p) + c}.$$

When the pathogen enters a target cell, its survival relies on the survival of the target cell that hosts it. The life span of a target cell infected with trait q is exponentially distributed with parameter $(\mu(q) + d)$. The infected target cell produces free pathogens according to a Poisson process with parameter q. So the probability density that an infected target cell lives t units of time and in that time produces n offspring equals

$$(\delta + \mu(q))e^{-(\delta+\mu(q))t}e^{-qt}\frac{q^n t^n}{n!}.$$

Accounting for all possible times t, we arrive at the following expression for π_n,

$$\pi_n = \frac{k\hat{T}(p)}{k\hat{T}(p)+c}\int_0^\infty (\delta + \mu(q))e^{-(\delta+\mu(q))t}e^{-qt}\frac{q^n t^n}{n!}dt, \tag{22}$$

which is valid for $n \geq 1$. For the probability of having no offspring at all, however, we have to take into account that the pathogen may never reproduce simply because it never enters an uninfected target cell. Since the probability with which the pathogen dies before it binds to a cell equals $\frac{c}{k\hat{T}(p)+c}$, we obtain the following generating function $G(z)$,

$$G(z) = \frac{c}{k\hat{T}(p) + c} + \sum_{n=0}^{\infty} \pi_n z^n,$$

which, by using (22) and interchanging the order of summation and integration, can be written as

$$G(z) = \frac{c}{k\hat{T}(p) + c} + \frac{k\hat{T}(p)}{k\hat{T}(p) + c} \cdot \frac{1}{1 + \mathcal{B}_0(q)(1 - z)}. \tag{23}$$

The probability with which the clan of the invading pathogen goes extinct is given as the smallest solution of $G(z) = z$ (cf. [30]). Whether this solution lies in $[0, 1]$, depends on the value of $G'(1)$, which equals the invaders within-host reproduction ratio in the environment set by the resident,

$$\mathcal{R}_0^w(\hat{T}(p), q) = \frac{k\hat{T}(p)}{k\hat{T}(p) + c}\mathcal{B}_0(q).$$

If $\mathcal{R}_0^w(\hat{T}(p), q) \leq 1$, the clan will go extinct with certainty. If, on the other hand, $\mathcal{R}_0^w(\hat{T}(p), q) > 1$, the invasion will be successful with nonzero probability.

Let $P(p, q)$ denote the probability of extinction of trait q, following an introduction of a single free pathogen into an environment set by the resident trait p. Using (23) and (21) we obtain

$$P(p, q) = \min\left\{1, \frac{c}{c + k\hat{T}(p)} + \frac{1}{\mathcal{B}_0(q)}\right\} = \min\left\{1, 1 - \frac{1}{\mathcal{B}_0(p)} + \frac{1}{\mathcal{B}_0(q)}\right\}.$$

Thus, the invading trait has a nonzero probability of success only when its burst size exceeds the burst size of the resident trait p. We also observe that (i) $P(p,p) = 1$, as it should be since the resident trait resides at a stable equilibrium, and (ii) when $\mathcal{B}_0(q) \to \infty$, the invading trait will survive with certainty, provided that the pathogen initially introduced makes it to an uninfected target cell. The probability of extinction must therefore equal the probability with which the pathogen dies before it enters a target cell. And indeed,

$$\lim_{\mathcal{B}_0(q)\to\infty} P(p,q) = \frac{c}{k\hat{T}(p)+c}.$$

The complementary probability

$$\phi_1(p,q) = 1 - P(p,q)$$

is the probability that the clan of one free pathogen with trait q survives in the environment set by the resident trait.

When n pathogens are introduced, therefore, the probability of survival equals

$$\phi_n(p,q) = \begin{cases} 1 - P^n(p,q), & \mathcal{B}_0(p) < \mathcal{B}_0(q) \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

or, rewritten in terms of $\hat{T}(p)$ and $\hat{T}(q)$,

$$\phi_n(p,q) = \begin{cases} 1 - \Big(1 - \dfrac{k\hat{T}(p)}{c+k\hat{T}(p)} + \dfrac{k\hat{T}(q)}{c+k\hat{T}(q)}\Big)^n, & \hat{T}(q) < \hat{T}(p) \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

We observe that the superinfection functions are continuous, but not differentiable in $q = p$. The fact that they are increasing as functions of $\mathcal{B}_0(q)$, implies that the traits that significantly increase the burst size also have a better chance of surviving in the host than the traits which are only slightly better within-host competitors than the resident.

Note also that $\{\phi_n\}$ is an increasing sequence for every resident strategy, that is,

$$\phi_0(p,q) < \phi_1(p,q) < \phi_2(p,q) < ...$$

for every p. Thus, if p and q are given, the larger the reinfection dose is, the better the chances of survival of the mutant are.

In the limit, when the number of initially introduced pathogens goes to infinity, we have

$$\lim_{n\to\infty} \phi_n(p,q) = \begin{cases} 1, & \hat{T}(q) < \hat{T}(p) \\ 0, & \text{otherwise,} \end{cases}$$

i.e., the superinfection function is a discontinuous function which furthermore doesn't discriminate among the winning strategies: every trait that reduces the steady state level of target cells to a lower level than the resident trait succeeds with probability one. Hence, when infinitely many pathogens with trait q are introduced, the deterministic description gives the full story: if the newly introduced trait goes extinct, it is because it loses the competition within the host and not due to bad luck while still rare. In this deterministic description, an even slighly better within-host competitor will outcompete the resident strain and evolution will drive p towards the within-host optimum. Assuming small mutational steps, therefore, the outcome of evolution at the population level will be the same as in a single infected host. Contrary to the within-host evolution, however, we do not have an optimization model at the population level (we refer to [7] for details). For comparison, we note that while the basic superinfection model predicted ever increasing virulence, the superinfection model with a nested within-host submodel predicts evolution towards the within-host optimum.

If, on the other hand, n approaches zero, the chance of a successful invasion becomes virtually zero. In this case, therefore, superinfections play a negligible role. In the limit $n = 0$ we end up with the single infection model.

For intermediate levels of n, the singular strategy lies inbetween the within-host optimum and the optimum of the single infection model. With increasing n, the singular strategy moves from the optimum of the single infection model towards the optimum of the within-host model. Depending on the trade-off and the reinfection dose n, the convergence stable strategies can either be uninvadable or invadable. However, because of the fact that the superinfection function is merely continuous, we may again get singular points that are invadable from one side only. If branching occurs, one of the two strains has very little room to

evolve and remains virtually constant through the course of evolution. We refer the reader to [7] for examples and more details.

References

[1] S. Alizon and M. van Baalen. Emergence of a convex trade-off between transmission and virulence. *The American Naturalist*, 165(6):155–167, 2005.

[2] S. Alizon and M. van Baalen. Transmission–virulence trade-offs in vector-borne diseases. *Theoretical Population Biology*, 74(1):6–15, 2008.

[3] R. Anderson and R. May. Population biology of infectious diseases: Part i. *Nature*, 280(5721):361–367, 1979.

[4] R. M. Anderson and R. M. May. Coevolution of hosts and parasites. *Parasitology*, 85:411–426, 1982.

[5] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, UK, 1991.

[6] F. Baquero and J. Blázquez. Evolution of antibiotic resistance. *Trends in Ecology & Evolution*, 12(12):482–487, 1997.

[7] B. Boldin and O. Diekmann. Superinfections can induce evolutionarily stable coexistence of pathogens. *J. Math. Biol.*, 56(5):635–672, 2008.

[8] B. Boldin, S. A. H. Geritz, and É. Kisdi. Superinfections and adaptive dynamics of pathogen virulence revisited: A critical function analysis. *Evolutionary Ecology Research*, 11(2):153–175, 2009.

[9] M. Boni and M. Feldman. Evolution of antibiotic resistance by human and bacterial niche construction. *Evolution*, 59(3):477–491, 2005.

[10] J. J. Bull. Virulence. *Evolution*, 48:11423–1437, 1994.

[11] D. Coombs, M. Gilchrist, and C. Ball. Evaluating the importance of within-and between-host selection pressures on the evolution of chronic pathogens. *Theoretical Population Biology*, 72(4):576–591, 2007.

[12] J. Davies. Origins and evolution of antibiotic resistance. *Microbiologia(Madrid)*, 12(1):9–16, 1996.

[13] T. Day. On the evolution of virulence and the relationship between various measures of mortality. *Proceedings of the Royal Society B: Biological Sciences*, 269(1498):1317–1323, 2002.

[14] T. Day and S. R. Proulx. A general theory for the evolutionary dynamics of virulence. *Am. Nat.*, 163:40–63, 2004.

[15] P. De Leenheer and H. L. Smith. Virus dynamics: a global analysis. *SIAM J. Appl. Math.*, 63(4):1313–1327, 2003.

[16] J. C. de Roode, R. Pansini, S. J. Cheesman, M. E. H. Helinski, S. Huijben, A. R. Wargo, A. S. Bell, B. H. K. Chan, D. Walliker, and A. F. Read. Virulence and competitive ability in genetically diverse malaria infections. *Proceedings of the National Academy of Sciences*, 102(21):7624–7628, 2005.

[17] J. C. de Roode, A. J. Yates, and S. Altizer. Virulence-transmission trade-offs and population divergence in virulence in a naturally occurring butterfly parasite. *Proceedings of the National Academy of Sciences*, 105(21):7489, 2008.

[18] U. Dieckmann, J. A. J. Metz, M. W. Sabelis, and K. Sigmund. *Adaptive Dynamics of Infectious Diseases: In Pursuit of Virulence Management*. Cambridge studies in Adaptive Dynamics, Cambridge University Press, 2002.

[19] O. Diekmann. A beginner's guide to adaptive dynamics. In *Mathematical modelling of population dynamics*, volume 63 of *Banach Center Publ.*, pages 47–86. Polish Acad. Sci., Warsaw, 2004.

[20] O. Diekmann and J. A. P. Heesterbeek. *Mathematical epidemiology of infectious diseases*. Wiley Series in Mathematical and Computational Biology. John Wiley & Sons Ltd., Chichester, 2000. Model building, analysis and interpretation.

[21] T. Dobzhansky. Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, 35:125–129, 1973.

[22] G. Dwyer, S. A. Levin, and L. Buttel. A simulation model of the population dynamics and evolution myxomatosis. *Ecol. Monogr.*, 60:423–447, 1990.

[23] P. W. Ewald. Host-parasite relations, vectors, and the evolution of disease severity. *Ann. Rev. Ecol. Syst.*, 14:465–485, 1983.

[24] P. W. Ewald. *Evolution of infectious disease*. Oxford University Press, Oxford, UK, 1994.

[25] F. Fenner and F. Ratcliffe. Myxomatosis. Cambridge, 1965.

[26] V. V. Ganusov and R. Antia. Trade-offs and the evolution of virulence of microparasites: do details matter? *Theor. Popul. Biol.*, 64(2):211–220, 2003.

[27] S. Geritz, J. Metz, E. Kisdi, and G. Meszena. The dynamics of adaptation and evolutionary branching. *Phys. Rev. Letters*, 78:2024–2027, 1997.

[28] S. A. H. Geritz, E. Kisdi, G. Meszena, and J. A. J. Metz. Evolutionary singular strategies and the adaptive growth and branching of the evolutionary tree. *Evol. Ecol.*, 12:35–57, 1998.

[29] M. A. Gilchrist and D. Coombs. Evolution of virulence: Interdependence, constraints and selection using nested models. *Theor. Pop. Biol.*, 69:145–153, 2006.

[30] P. Haccou, P. Jagers, and V. Vatutin. *Branching processes: variation, growth, and extinction of populations.* Cambridge studies in Adaptive Dynamics, Cambridge University Press, 2005.

[31] E. Herre. Population structure and the evolution of virulence in nematode parasites of fig wasps. *Science*, 259(5100):1442–1445, 1993.

[32] Y. Iwasa, F. Michor, and M. Nowak. Virus evolution within patients increases pathogenicity. *Journal of theoretical biology*, 232(1):17–26, 2005.

[33] R. E. Lenski and R. M. May. The evolution of virulence in parasites and pathogens: reconciliation between the competing hypotheses. *J. Theor. Biol.*, 169:253–265, 1994.

[34] S. A. Levin and D. Pimentel. Selection of intermediate rates of increase in parasite-host systems. *Am. Nat.*, 117:308–315, 1981.

[35] M. J. Mackinnon and A. F. Read. Genetic relationships between parasite virulence and transmission in the rodent malaria Plasmodium chabaudi. *Evolution*, pages 689–703, 1999.

[36] R. May and R. Anderson. Population biology of infectious diseases: Part II. *Nature*, 280(5722):455–461, 1979.

[37] R. M. May and R. M. Anderson. Epidemiology and genetics in the coevolution of parasites and hosts. *Proc. Roy. Soc. Lond. B Bio*, 219:281–313, 1983.

[38] J. A. J. Metz, S. D. Mylius, and O. Diekmann. When does evolution optimize. *Evol. Ecol. Res.*, 10:629–654, 2008.

[39] N. Mideo, S. Alizon, and T. Day. Linking within-and between-host dynamics in the evolutionary epidemiology of infectious diseases. *Trends in Ecology & Evolution*, 23(9):511–517, 2008.

[40] J. Mosquera and F. R. Adler. Evolution of virulence: a unified framework for coinfection and superinfection. *J. Theor. Biol.*, 195:293–313, 1998.

[41] J. Muskett, N. Reed, and D. Thornton. Increased virulence of an infectious bursal disease live virus vaccine after passage in chicks. *Vaccine*, 3(3):309–12, 1985.

[42] M. A. Nowak and R. M. May. Superinfection and the evolution of parasite virulence. *Proc. Roy. Soc. Lond. B*, 255:81–89, 1994.

[43] M. A. Nowak and R. M. May. *Virus Dynamics: Mathematical principles of immunology and virology.* Oxford University Press, Oxford, UK, 2000.

[44] K. O'Keefe and J. Antonovics. Playing by different rules: the evolution of virulence in sterilizing pathogens. *The American Naturalist*, 159(6):597–605, 2002.

[45] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271:1582–1586, 1996.

[46] D. B. Preston, B. J. Poiesz, and L. A. Loeb. Fidelity of HIV-l reverse transcriptase. *Science*, 242:1168–1171, 1988.

[47] A. Pugliese. Evolutionary dynamics of virulence. To appear in 'Elements of adaptive dynamics, U. Dieckmann and J.A.J. Metz (eds.)., Cambridge Univ. Press'.

[48] A. Pugliese. On the evolutionary coexistence of parasite strains. *Math. Biosci.*, 177/178:355–375, 2002.

[49] O. Restif and J. Koella. Concurrent evolution of resistance and tolerance to pathogens. *The American Naturalist*, 164(4):90–102, 2004.

[50] J. D. Roberts, K. Bebenek, and T. A. Kunkel. The accuracy of reverse transcriptase from HIV-1. *Science*, 242:1171–1173, 1988.

[51] T. O. Svennungsen and É. Kisdi. Evolutionary branching of virulence in a single-infection model. *J. Theor. Biol.*, 257(3):408–418, 2009.

[52] D. Tebit, I. Nankya, E. Arts, and Y. Gao. Hiv diversity, recombination and disease progression: how does fitness" fit" into the puzzle. *AIDS Rev*, 9(2):75–87, 2007.

[53] S. Thomas and J. Elkinton. Pathogenicity and virulence. *Journal of Invertebrate Pathology*, 85(3):146–151, 2004.

[54] R. Troyer, K. Collins, A. Abraha, E. Fraundorf, D. Moore, R. Krizan, Z. Toossi, R. Colebunders, M. Jensen, J. Mullins, et al. Changes in human immunodeficiency virus type 1 fitness and genetic diversity during disease progression supplemental material for this article may be found at http://jvi. asm. org/. *Journal of Virology*, 79(14):9006–9018, 2005.

[55] M. E. Wickham, N. F. Brown, E. C. Boyle, B. K. Coombes, and B. B. Finlay. Virulence is positively selected by transmission success between mammalian hosts. *Current Biology*, 17(9):783–788, 2007.

Local and global class field theory

John Coates [*] *and Sujatha Ramdorai* [†]

Abstract

These lectures will give an account of Local and Global Class field Theory as covered by Serre and Tate in the classic book 'Algebraic Number Theory' by Cassels and Fröhlich.

[*]University of Cambridge, United Kingdom
[†]TATA Institute of Fundamental Research, Mumbai, India

Function Theories in Higher dimensions

*Sirkka-Liisa Eriksson**

Abstract

There is a rich interplay between potential theory in the plane and complex function theory. In higher dimensions potential theory is very well developed but extensions of one variable complex function theory to higher dimensions has plenty of open problems. A higher dimensional generalization of the algebra of complex numbers is a Clifford algebra which is the smallest extension of the Euclidean space $\mathbb{R}^n$ to an associative algebra that inherits the algebraic, geometric and metric properties of the Euclidean space $\mathbb{R}^n$. It is a generalization of the algebra of quaternions introduced by Hamilton around 1843.

In 1935 Fueter defined regular functions in the algebra of quaternions and was able to prove the main theorems as Cauchy theorem, Cauchy formula and Laurent power series expansions. The key idea is that a regular function is a conjugate gradient of a harmonic function. Delanghe in 1970 was able to prove similar results in $\mathbb{R}^n$ using Clifford algebras.

A generalization of the Laplacian is the Laplace-Beltrami operator defined on manifolds. We consider hyperbolic harmonic functions with respect to the Laplace-Beltrami operator of the hyperbolic metric $ds^2 = x_n^{-2} \sum_{i=0}^{n} dx_i^2$. An important fact is this hyperbolic distance is Möbius invariant without any conformal factor. Leutwiler noticed around 1990 that if the usual Euclidean metric is changed to a hyperbolic one then the power function, calculated using Clifford algebra, is the conjugate gradient of the a hyperbolic harmonic functions. We study generalized holomorphic functions, called hypermonogenic functions, connected to the hyperbolic metric. They satisfy many similar properties as holomorphic functions.

*Tampere University of Technology, Department of Mathematics, P.O.Box 553, FI-33101 Tampere, Finland

Some generalizations of trigonometric functional equations

Żywilla Fechner
Silesian University, Institute of Mathematics, Bankowa 14, PL-40-007 Katowice, Poland
e-mail: zfechner@math.us.edu.pl

Abstract

We discuss some trigonometric functional equations for mappings defined on a group and taking values in the field of complex numbers. Moreover, we will present some methods which are helpful in solving these equations. First, we give some examples of applying spectral analysis to solve the cosine functional equation, next we show that under certain assumptions Bochner Theorem can be used. Our main goal is to introduce some integral-type generalizations of the cosine equation. The main tool in solving this equation is a version of Wiener's tauberian theorem. Finally, we give certain modifications of these equations and pose some possible directions of further investigations.

1 Introduction

Observe that the function $f(x) = \cos x$ for $x \in \mathbb{R}$ satisfies the following equation:

$$f(x+y) + f(x-y) = 2f(x)f(y), \tag{1}$$

for all $x, y \in \mathbb{R}$, which justifies to call (1) the *cosine functional equation*. It is also known as *d'Alembert functional equation*, since it was introduced by J. d'Alembert in [2]. Let us notice that we may consider functions to be defined not only on the real line $\mathbb{R}$, but also on an arbitrary group G. In the target space we require addition and multiplication, thus we need a field or, in vector case, a ring or an algebra. In the next section we discuss certain generalizations of (1), our aim is to present different methods which are useful in solving the cosine equation.

Now observe that the pair $(g(x), f(x)) = (\sin x, \cos x)$ for $x \in \mathbb{R}$ satisfies

$$g(x+y) + g(x-y) = 2g(x)f(y) \tag{2}$$

for all $x, y \in \mathbb{R}$. This equation is known as the *cosine-sine equation* or *Wilson's functional equation*. Of course, the foregoing remarks about the domain and the range of solutions remain valid in case of Wilson's equation. We discuss the form of solutions of equation (2) in Section 2.

There exist a number of generalizations of equations (1) and (2) in different directions. Pl. Kannappan [9] investigated (1) for mappings defined on a group and taking values in the field of complex numbers, see Theorem 2.1 below. Some results concerning d'Alembert functional equation on step 2 nilpotent groups are due to H. Stetkær in [15] and [16]. There are also known some generalizations for vector-valued functions defined on abelian groups (cf. L. Rejtő [13], S. Kurepa [11]) and for Wilson's-type generalizations by S. Kurepa [10] and by the author [6].

Another generalizations are due to W. Chojnacki in [4]. He has considered a d'Alembert-type equation:

$$\int_K f(x + h \cdot y)d\nu(h) = f(x)f(y), \quad x, y \in G \tag{3}$$

and a Wilson's-type equation:

$$\int_K g(x + h \cdot y)d\nu(h) = g(x)f(y), \quad x, y \in G, \tag{4}$$

where K is a compact group with a Haar measure ν acting on a locally compact abelian group G with a Haar measure m and functions $f, g\colon G \to \mathbb{C}$ are m-measurable and essentially bounded. By the use the Fourier analysis he found the explicit formula of f in (3) and (4) without giving a description of g in (4). If we consider a locally compact abelian group, then equations (1) and (2) are special cases of (3) and (4); it is enough to take $K = \mathbb{Z}_2$ with a normalized counting measure and K acting on G by the rule $0 \cdot x = x$ and $1 \cdot x = -x$ for all $x \in G$.

A one more generalization of Wilson's equation has been considered by H. Stetkær in [14] in connection with spherical functions:

$$\int_K g(x + k \cdot y)dk = g(x)f(y), \quad x, y \in G,$$

where K is a compact group acting by automorphisms on an abelian group G and dk denotes the normalized Haar measure on K.

Now, let us focus on the following generalization of d'Alembert functional equation:

$$(f * \mu_y)(x) + (f * (\mu_y)^-)(x) = f(x)f(y), \quad x, y \in G, \tag{5}$$

where G is a locally compact abelian group, $f \in L^\infty(G)$, $\mathcal{B}(G)$ is a family of all Borel subsets of G, $\mu\colon \mathcal{B}(G) \to \mathbb{C}$ is a regular bounded measure and

$$\mu^-(A) := \mu(-A), \quad \mu_y(A) := \mu(A + y), \quad A \in \mathcal{B}(G).$$

Equation (5) was introduced and solved by Z. Gajda in [7]. It is easy to see that the cosine equation is a special case of (5). Indeed, it is enough to take μ given by

$$\mu(A) := \begin{cases} \frac{1}{2}, & 0 \in A; \\ 0, & 0 \notin A; \end{cases} \quad A \in \mathcal{B}(G). \tag{6}$$

In Section 3 we are going to present some results concerning Wilson's-type extensions of equation (5) and we will suggest some possible directions of further research.

All terminology concerning the field of harmonic analysis is in accordance with the monograph of E. Hewitt, K. A. Ross [8].

One more notion may be useful: for any function $F\colon G \to \mathbb{C}$ defined on a group G the functions

$$F_o(x) := \frac{F(x) - F(-x)}{2}, \quad F_e(x) := \frac{F(x) + F(-x)}{2}, \quad x \in G$$

are the *odd* and the *even parts* of F, respectively.

2 D'Alembert and Wilson's equations

2.1 The cosine equation

In this section we discuss some possible solutions of cosine functional equation (1) in different settings. Classical results for d'Alembert functional equation have been obtained by Pl. Kannappan [9].

Theorem 2.1 (Pl. Kannappan, [9]). *Let $(H, +)$ be a group (not necessery abelian) and $f\colon H \to \mathbb{C}$ satisfies*

$$f(x + y + z) = f(x + z + y), \quad x, y, z \in H. \tag{7}$$

The function f satisfies equation (1) *iff there exists a homomorphism $m\colon H \to \mathbb{C}$ such that*

$$f(x) = \frac{m(x) + m(-x)}{2}, \quad x \in H. \tag{8}$$

Note that we do not impose any topology on H; in particular, we do not assume any regularity conditions. The proof of this theorem is divided into two complementary parts: first it is assumed that $f(H) \subset \{-1, 1\}$ and then that $f(x_1)^2 \neq 1$ for some $x_1 \in G$. First part is a straightforward calculation and in the second part it is shown that the function

$$m(x) := f(x) + \frac{f(x + x_1) - f(x)f(x_1)}{f(x_1)^2 - 1}, \quad x \in H,$$

is a homomorphism. Therefore there is the explicit formula for the homomorphism m from (8).

Now we give a rough idea about solutions of d'Alembert functional equation obtained by means of spectral analysis. The following version of Wiener's tauberian theorem can be useful (see Székelyhidi [17], p. 9):

Theorem 2.2 (Wiener). *If G is a locally compact abelian group, then any nonzero closed invariant subspace of $L^\infty(G)$ contains a character.*

We give two examples of solutions of d'Alembert functional equation for groups. Both examples may be derived from a general theorem due to L. Székelyhidi [17], which tells us that each solution of (1) on a locally compact abelian group is a real part of some character. We present special case of a reasoning from [17].

Example 2.3. Let $(\mathbb{Z}_m, +)$ be the additive group of all remainders from division by m equipped with the discrete topology. If $f\colon \mathbb{Z}_m \to \mathbb{C}$ is a bounded function satisfying d'Alembert functional equation for all $x, y \in \mathbb{Z}_m$, then there exists $k \in \mathbb{Z}$ such that

$$f(y) = \cos\left(\frac{2\pi y}{m}\right), \quad y \in \mathbb{Z}_m.$$

Indeed, let $\tau(f)$ denote the minimal proper closed invariant subspace containing f. First, observe that if f is a solution of equation (1), then each function $g \in \tau(f)$ satisfies (2). To show this fix $x, z \in \mathbb{Z}_m$. From (1) applied for $x - z$ instead of x we have

$$f_z(x+y) + f_z(x-y) = 2f_z(x)f(y), \quad y \in \mathbb{Z}_m.$$

Since a linear combination of solutions of (2) is a solution of (2) we obtain

$$g(x+y) + g(x-y) = 2g(x)f(y), \quad y \in \mathbb{Z}_m$$

for any function $g \in \tau(f)$.

By Wiener's theorem the space $\tau(f)$ contains a character. Therefore (cf. E. Hewitt and K. A. Ross, [8], p. 367) there exists an $l \in \{0, 1, \ldots, m-1\}$ such that the character

$$\chi(k) = e^{\frac{2\pi i k}{m}}, \quad k \in \mathbb{Z}_m$$

is an element of $\tau(f)$, i.e.

$$\chi(x+y) + \chi(x-y) = 2\chi(x)f(y), \quad y \in \mathbb{Z}_m.$$

Dividing by $\chi(x) \neq 0$ we arrive at

$$f(y) = \frac{\chi(y) + \chi(-y)}{2} = \frac{1}{2}\left[e^{\frac{2\pi i y}{m}} + e^{\frac{-2\pi i y}{m}}\right] = \cos\left(\frac{2\pi y}{m}\right), \quad y \in \mathbb{Z}_m.$$

We give an example of a solution of (1) on a multiplicative group $\mathbb{Z}_p \setminus \{0\}$ for p being a prime number.

Example 2.4. Let χ be a Dirichlet character (cf. e.g. T. Apostol [3], Chapter 6) given by

$$\chi(n) = \left(\frac{n}{p}\right) = \begin{cases} 1, & n = x^2 \mod p \quad \text{for some } x, \\ -1, & n = x^2 \mod p \quad \text{for no } x. \end{cases}$$

If f is a solution of (1), then $f(\mathbb{Z}_p) \subset \{-1, 1\}$. It is easily seen that for each character χ we have $\chi\left(x^{-1}\right) = \overline{\chi(x)}$ for all $x \in G$. Therefore $f = \chi$.

The following result is due to T. A. O'Connor [12]:

Theorem 2.5 (T. A. O'Connor)**.** *Let $(H, +)$ be a connected, separable, locally compact abelian group and let $f\colon H \to \mathbb{R}$ be a bounded continuous function such that $f(0) = 1$. The function f satisfies* (1) *iff there exists a character χ_0 such that*

$$f(x) = \Re\chi_0(x), \quad x \in H.$$

The main tool to prove Theorem 5 is Bochner's characterization of positive definite functions; namely, let $\check{H}$ denote the set of all characters of H. It is shown that a solution of d'Alembert equation is positive definite and thus by Bochner theorem there exists a regular measure $P\colon \mathcal{B}(\check{H}) \to [0, 1]$ such that

$$f(x) = \int_{\check{H}} \Re\chi(x)dP(\chi), \quad x \in H.$$

For each $x \in H$ we define a random variable $\ell_x\colon \check{H} \to \mathbb{R}$ by the formula $\ell_x(\chi) = \Re(\chi)$ for $\chi \in \check{H}$. The variance $\mathrm{Var}\ell_x = 0$ and thus $\ell_x = 0$ a.e. Therefore, if $(x_n)_{n\in\mathbb{N}}$ is a countable and dense subset of H and

$$E := \left(\left\{\chi \in \check{H} : \ell_{x_n}(\chi) = const, \quad n \in \mathbb{N}\right\}\right),$$

then $P(E) = 1$ and thus there exists a character χ_0 such that $E = \{\chi_0, \overline{\chi_0}\}$. Now we have the desired representation.

One should be aware that O'Connor's results is far from being general, since it deals only with real bounded solutions, however it is a nice example of using Bochner theorem combined with some basic probability notions.

2.2 The cosine-sine equation

Now we discuss solutions of Wilson's functional equation. Typical method used in solving Wilson's equation is to show that under some additional assumptions f satisfies (1). In order to do this one may try to use results from previous section. However, this method may not be useful. It is worth to underline that the main difficulty in solving Wilson's-type equations is to give a description of g. This is not obvious even in the real case (cf. J. Aczél [1]). For complex-valued mappings defined on an abelian group one of possible methods to find the form of g is spectral synthesis (for details see the monograph [17] of L. Székelyhidi). Namely, we have the following result:

Theorem 2.6 (L. Székelyhidi, [17], p. 109)**.** *Let G be an abelian group and let $f, g\colon G \to \mathbb{C}$ be in $L^\infty(G)$. The pair (g, f) satisfies Wilson's functional equation for all $x, y \in G$ iff either*

(i) $f = 0$ and g is arbitrary,

or

(ii) there exist a character $\chi_1 \in \Gamma$ and a constant $\alpha \in \mathbb{C}$ such that $\chi_1^2 = 1$ and

$$f(x) = \chi_1(x), \quad g(x) = \alpha\chi_1(x), \quad x \in G, \tag{9}$$

or

(iii) there exist a character $\chi \in \Gamma$ and constants $K, L \in \mathbb{C}$ such that $\chi^2 \neq 1$ and

$$f(x) = \frac{\chi(x) + \chi(-x)}{2}, \quad g(x) = K\chi(x) + L\chi(-x), \quad x \in G. \tag{10}$$

As we observed it in the Introduction f is a generalization of the cosine and g is a generalization of the sine. To obtain this it is enough to take $\chi(x) := \exp(ix)$ and $K = -L = \frac{1}{2}$. It is worth to notice that in the general case g need not to be odd. More precisely, g is odd iff $K = -L$.

3 Some integral generalizations

In the Introduction we have mentioned some possible integral generalizations of d'Alembert and Wilson's equations. The purpose of the present part is to give a rough idea how to solve (5) and the following equation:

$$(g * \mu_y)(x) + (g * (\mu_y)^-)(x) = g(x)f(y), \quad x, y \in G. \tag{11}$$

Taking a measure μ given by (6) equation (11) becomes Wilson's functional equation (2).

We will need some notations concerning algebra $L^1(G)$. Namely, Fourier transforms of a function $f \in L^1(G)$ and a bounded regular measure $\mu\colon \mathcal{B}(G) \to \mathbb{C}$ are defined in the following way:

$$\widehat{f}(\gamma) := \int_G f(x)\gamma(-x)dm(x), \quad \widehat{\mu}(\gamma) := \int_G \gamma(-x)d\mu(x), \quad \gamma \in \Gamma.$$

For any ideal $\mathcal{I}$ of the algebra $L^1(G)$ let

$$Z(\mathcal{I}) := \{\gamma \in \Gamma : \widehat{g}(\gamma) = 0 \text{ for all } g \in \mathcal{I}\}.$$

Now we cite a one more version of Wiener's theorem (see e.g. Gajda, [7]).

Theorem 3.1 (Wiener)**.** *Assume that $\mathcal{I}$ is a closed ideal in $L^1(G)$. If $Z(\mathcal{I}) = \varnothing$, then $\mathcal{I} = L^1(G)$.*

In the following theorem we have a description of solutions of (5).

Theorem 3.2 (Z. Gajda, [7])**.** *Let $\mu\colon \mathcal{B}(G) \to \mathbb{C}$ be a bounded regular measure. Then a function $f \in L^\infty(G)$ which does not vanish m- locally almost everywhere (m-l.a.e.) satisfies equation* (5) *if and only if there exists a character $\gamma \in \Gamma$ such that*

$$f(y) = (\gamma * \mu_y)(0) + (\gamma * (\mu_y)^-)(0) = \int_G \{\gamma(y - s) + \gamma(s - y)\}\, d\mu(s) \tag{12}$$

Now we enumerate the steps which were used to solve equation (5). First, we define A as the set of all $\psi \in L^1(G)$ of the form

$$\psi = (\phi * (\mu_y)^-) + (\phi * \mu_y) - f(y)\phi,$$

where ϕ ranges over $L^1(G)$ and y over G. Next, after some calculations involving Fubini theorem and convolution properties we obtain

$$\int_G f(x)\psi(x)dm(x) = 0, \quad \psi \in A. \tag{13}$$

Let $\mathcal{I}$ be the linear space spanned by the set A. It is easy to verify that $\mathcal{I}$ is an ideal of $L^1(G)$. Let $\mathcal{F}$ denote the closure of $\mathcal{I}$ in the norm topology. Clearly, $\mathcal{F}$ is a closed ideal in $L^1(G)$. Suppose that for every $\gamma \in \Gamma$ there exists a $y \in G$ such that

$$((\mu_y)^-)\widehat{\ }(\gamma) + (\mu_y)\widehat{\ }(\gamma) \neq f(y). \tag{14}$$

It can be shown that $Z(\mathcal{F}) = \varnothing$. Thus, by Wiener's theorem we infer that $\mathcal{I} = L^1(G)$. Consequently, A is linearly dense subset of $L_1(G)$. From (13) we derive that $f = 0$ m-l.a.e., which is impossible. Therefore, there exists a character $\gamma \in \Gamma$ such that (12) holds.

Now, we are going to outline the reasonings from [5], where one can find the detailed proof of solution of (11). Using Gajda's methods presented above we show that if (g, f) satisfies (11), then f satisfies (5), hence, there exists a character γ such that f is given by (12). Next we show that (g, f) satisfies (11) iff (g_o, f) and (g_e, f) satisfy (11), therefore we may deal with the even and the odd part of g separately. In the even case we show that g_e is proportional to f. The odd case is a bit more complicated; we use the form of solution of classical Wilson equation (cf. Theorem 2.6), Theorem 3.2 and linear independence of the set of all characters of G to show that g_o is proportional to the odd part of a certain character γ. Finally, we obtain the following theorem:

Theorem 3.3. *Let $\mu\colon \mathcal{B}(G) \to \mathbb{C}$ be a bounded regular measure, $f, g \in L^\infty(G)$ and assume that f, g do not vanish m-l.a.e. If the pair (g, f) satisfies equation* (11)*, then there exist a character $\gamma \in \Gamma$ and constants $C_1, C_2 \in \mathbb{C}$ such that f is the form of* (12) *and*

$$g(x) = C_1\gamma(x) - C_2\gamma(-x), \quad x \in G. \tag{15}$$

Conversely, if $\gamma \in \Gamma$ is a character, $C_1, C_2 \in \mathbb{C}$ are constants and f is given by (12) *and g by* (15)*, then the pair (g, f) fulfills* (11)*.*

It is worth to notice that constants appearing in Theorem 3.3 depend on measure μ and values $f(0)$ and $g(0)$.

Problem 3.4. It could be interesting to find the general solution of

$$(f * \mu_y)(x) + (g * (\mu_y)^-)(x) = h(x)k(y), \quad x, y \in G, \tag{16}$$

for unknown mappings $f, g, h, k \in L^\infty(G)$ and a bounded regular measure $\mu\colon \mathcal{B}(G) \to \mathbb{C}$. Equation (16) is a generalization of a number of classical functional equations, which appear in the monograph of L. Székelyhidi, [17], Chapters $10 - 13$.

References

[1] J. Aczél, Lectures on functional equations and their applications, *Academic Press*, New York-San Francisco-London, 1966.

[2] J. d'Alembert, Mémoire sur les principes de mécanique, *Hist. Acad. Sci.*, Paris, 1769, 278-286.

[3] T. Apostol, Introduction to Analytic Number Theory, *Undergraduate Texts in Mathematics*, New York-Heidelberg-Berlin, 1976.

[4] W. Chojnacki, On some functional equation generalizing Cauchy's and d'Alembert's functional equations, *Colloq. Math.* 55 (1988), 170–178.

[5] Ż. Fechner, A generalization of Gajda's equation, *J. Math. Anal. Appl.*, (354): 584–593, 2009.

[6] Ż. Fechner, Wilson's functional equation in Banach algebras, *Acta Sci. Math. (Szeged)*: 75(2009), 131–142 .

[7] Z. Gajda. A generalization of D'Alembert's Functional Equation, *Funkc. Ekvacioj*, (33):69–77, 1990.

[8] E. Hewitt and K. A. Ross, Abstract harmonic analysis. Vol. I: Structure of topological groups. Integration theory, group representation *Die Grundlehren der mathematischen Wissenschaften, Bd. 115 Academic Press, Inc., Publishers, New York; Springer-Verlag, Berlin-Göttingen-Heidelberg*, 1963.

[9] Pl. Kannappan, The functional equation $f(xy) + f(xy^{-1}) = 2f(x)f(y)$ for groups, *Proc. Amer. Math. Soc.* 19 (1968), 69–74.

[10] S. Kurepa, On some functional equations in Banach spaces, *Studia Math.* 19 (1960), 149–158.

[11] S. Kurepa, A cosine functional equation in Banach algebras, *Acta Sci. Math. (Szeged)* 23 (1962), 255–267.

[12] T. A. O'Connor, A solution of d'Alembert functional equation on a locally compact Abelian group, *Aequationes Math.* 15 (1977), 235–238.

[13] L. Rejtő, B-algebra valued solution of the cosine equation, *Studia Sci. Math. Hungar.* 7 (1972), 331–336.

[14] H. Stetkær, Wilson's functional equations on groups, *Aequationes Math.* 49 (1995), 252–275.

[15] H. Stetkær, D'Alembert's functional equations on metabelian groups, *Aequationes Math.* 59 (2000), 306–320.

[16] H. Stetkær, D'Alembert's and Wilson's functional equations on step 2 nilpotent groups, *Aequationes Math.* 67 (2004), 241-262.

[17] L. Székelyhidi, Convolution Type Functional Equation on Topological Abelian Groups, *World Scientific, Singapore-New Jersey-London-Hong Kong,* 1991.

Women in Mathematics in France

Colette Guillopé[*†]
Université Paris Est, Laboratoire d'Analyse et de Mathématiques Appliquées, UMR-CNRS 8050, 61 avenue du Général de Gaulle, 94010 Créteil Cedex, France, guillope@univ-paris12.fr

Abstract

This paper discusses the current status of women mathematicians in France. Recent sex-disaggregated data for universities and research institutes reveal how the situation of women mathematicians is deteriorating in France. The paper describes the actions of the associations of women scientists or engineers in France to promote sciences towards young people and to help improve the situation of women. The paper ends by analysing the recent government programs being implemented in the past ten years.

1 Status of women mathematicians

Although girls in France have a slightly higher success rate than boys in high school and account for over 45% of students in the standard scientific track, attrition becomes significant in higher education. Throughout undergraduate and graduate university studies, the proportion of female students in fundamental sciences is constant, around 27% [1]. In the selective parallel track for entering Engineering Schools (a French peculiarity), women account for 25% of the students, but 18% in mathematics- and physics-oriented disciplines [2].

Other French distinctive features include a favorable –although not perfect– social situation (low-cost public daycares, school all day long for young children, paid maternity leave), and the fact that a large portion of women mathematicians are civil servants, hired for a permanent position in the public research system in their late twenties or early thirties, after a short postdoctoral period.

Women account for 21% of mathematics faculty at French universities and 16% of mathematics researchers at the French National Center for Scientific Research (CNRS, the major public research institution in France, and the largest in Europe) [3]. Although these numbers might seem high compared to other countries such as USA, Canada or Finland, they remain unsatisfying and there is no progression, actually a noticeable decrease (20% women in mathematics at CNRS in 1989, the percentage being about constant since 1992, 16 to 17%). The number of mathematicians employed by CNRS has increased from 250 to 350 in the past 20 years, though the number of women mathematicians has stayed constant, about 50. Women Phd account for about 26% of all Phd's in mathematics, which is comparable to the number of associate professors at universities: there is no male advantage at this entry level. Note that there are very few women entering the most prestigious institution CNRS (zero to two a year, which amounts to about 0 to 10%).

However, the so-called glass ceiling remains very real. In mathematics, at universities, 26% of associate professors, but only 10% of full professors are women (respectively 30% and 9,7% in 1996) and the male advantage (ratio of the proportion of senior researchers –or full professors– among men over the proportion of senior researchers –or full professors– among women) is as high as 2.65 (whereas it is 2.4 in physics). At CNRS, the situation is more favorable, with 17% and 15% women among junior and senior researchers respectively, and with a male advantage of 1.04. The male advantage for mathematics is much lower than the male advantage in physics (which is 1.4) and than the overall male advantage at CNRS including all disciplines, which is 1.55 [3]. These figures give an idea of the thickness of the glass ceiling women find when they look for a promotion: in mathematics, men have 2.65 more chances to be promoted than women. Note that the number of women at CNRS (55 in 2005) is much lower than the number of women at universities (696 in 2006) [4], where the male advantage is huge; moreover men mathematicians very often prefer to become full professors at university rather than continue their career at CNRS. A finer study of the population of mathematicians shows that this is in the 30–40 age category that the thickness of glass ceiling is the largest: in other words, men mathematicians tend to get promoted between 30 and 40, whereas women tend to stay blocked in their career at that age.

[*]Association *femmes et mathématiques*, Institut Henri Poincaré, 11 rue Pierre et Marie Curie, 75230 Paris Cedex 05, France, fetm@ihp.jussieu.fr.

[†]Association "Femmes et Sciences", 9 rue Vésale, 75005 Paris, France, femmes.sciences@orange.fr.

2 Encouraging Girls to Choose Scientific Careers

This is one of the main goals of the "Femmes et Sciences" (F & S, Women and Science) Association [5], working in close partnership with *femmes et mathématiques* (F & M, Women and Mathematics) [6] and the "Association Française des Femmes Ingénieurs" (FI, French Women Engineers). Their members visit high schools to meet students, particularly girls, and inform them about scientific studies and careers, and to bear testimony to how rewarding they find their profession. As an example, in 2007, an action towards Parisian pupils was organized with the participation of more than 100 scientific female students from universities or Engineering schools who were trained by the three associations, and with the support of the City of Paris.

Some other programs of these three associations are:

- the common website "elles-en-sciences" [7] (She In Sciences), aimed at girls as well as their parents and teachers, launched in 2005 with support from the Ministry for Higher Education and Research;
- a booklet to help teachers tackle stereotypes about girls in science, published in 2006 [8];
- a photograph exhibition of women trained in mathematics *Women In Math... Why Not You?*, created by F & M, which has been circulating everywhere in France in high schools, commercial centers,...
- mentoring for teenagers;
- a yearly colloquium organized by F & S, where high school students are invited to meet women scientists;
- regular public debates, where gender specialists, sociologists, philosophers, historians and scientific women talk about general thematics concerning women and sciences ;
- a two-day forum is organized every other year by F & M for young women mathematicians, where they can present their work and prepare their interviews for teaching and research positions.

The Mission for the Place of Women at CNRS [9], established in July 2001, has developed various educational outreach tools, including the *Women in Physics* exhibition, created for the World Year of Physics in 2005. Now coming with a DVD, the exhibition continues to travel through France (already over 70 showings and debates in high schools, science centers and museums, conferences, etc.) and is now touring abroad in its English translation through partnerships developed with the USA, Canada and South Africa.

In the private sector, some companies have recently started positive actions like the "Elles bougent!" (Women On The Move!) Association [10] created in 2005 with the financial support of aviation, rail transport and automobile industries, in collaboration with related Engineering Schools, to present the careers available to girls and to offer mentoring.

3 Promoting Women in Science and Gender Equality

In France, several institutional structures are now operational, such as the "Mission for Gender Equality in Science and Technology", later entitled "Mission for Parity in Research and Higher Education", created in September 2001 at the Ministry of Higher Education and Research. This bureau has been transformed recently in 2009 in the "Mission of Parity and for Fighting Against Discriminations" [11]. The yearly "Irène Joliot-Curie" Award created by the Mission, in collaboration with EADS, to promote women in research and technology, was awarded to several mathematicians and physicists since 2005.

At CNRS, the Mission for the Place of Women [9] has remained very active, collecting and analyzing sex-disaggregated data, producing surveys and studies, sponsoring various colloquia, nominating CNRS women researchers for awards, promoting gender research, organizing gender trainings sessions across the country, and fostering gender equality within CNRS.

France is also actively involved in the European Platform of Women Scientists [12] created in 2005, particularly through former F & S founder, physicist Claudine Hermann, now on the administration board of EPWS, and through CNRS.

The 14th International Conference of Women Engineers and Scientists, ICWES14, was hosted by France in July 2008, in partnership with the Mission for the place of women at CNRS and the F & S, F & M and FI associations [13].

The "Société Française de Physique" (SFP, French Physical Society) [14], as the three above-cited Associations, pays much attention to promotions and appointments in leading positions and governmental committees, often unfair to women, and lodges complaints when necessary. The SFP also lobbies for more

women as chairpersons and speakers in scientific conferences and colloquia. The SFP and F & S also participate every three years in the IUPAP International Conference on Women in Physics: the last one was held in Seoul in 2008 [15].

New awards for women scientists have been created in France in 2005.

- The Excellencia Trophy for high-tech women engineers (fundamental research, applied research, R & D, production, and students preparing to enter high-tech professions).;
- The City of Paris Award for a young (<35 years) female Parisian scientist, award which has been discontinued as a consequence of the change of the vice-mayor for universities, innovation and research, after the 2008 city elections.

In collaboration with the French Academy of Sciences, L'Oréal and UNESCO also have launched new national doctoral fellowships in 2007, which are awarded to women, every other year in exact sciences (mathematics, physics, chemistry, computer science, engineering sciences), or in natural sciences (biology, health sciences, Earth sciences). These fellowships help the doctoral student in her last year of doctorate to promote her work in view of applying for her future job.

Aknowledgement

C.G. wishes to thank C. Thibaut, A. Pépin, M. Ducloy, E. Giacobino and M. Leduc for inspiring this presentation. C.G. wishes also to thank L. Broze for providing the statistics on women mathematicians in France, that she collected in 2007 on the occasion of the 20th birthday of the association *femmes et mathématiques*. A more complete study about women in science in France, including a sketch of history of education and a section on gender studies in France, can be found in [16].

References

[1] http://www.eduscol.education.fr/D0234/filles_garcons_chiffres2008.pdf

[2] http://www.femmes-ingenieurs.org/offres/gestion/menu_82_perso_82_1363/statistiques.html

[3] A. Pépin. Status of women in physics in France, Internal CNRS Report (2008), including data from the French Ministry of Higher Education and Research; M. Crance, The place of women at CNRS: Key figures, IPAM-CNRS (2008) http://www.cnrs.fr/mpdf/IMG/pdf/2006_PlacedesFemmes.pdf

[4] L. Broze and V. Lizan. Mathématicienne, un genre en voie de disparition en France, *Matapli*, 89, 2009, under press.

[5] http://www.femmesetsciences.fr

[6] http://www.femmes-et-maths.fr

[7] http://www.elles-en-sciences.org/

[8] http://www.femmesetsciences.fr/ideesrecues.htm

[9] http://www.cnrs.fr/mission-femmes/

[10] http://www.ellesbougent.com

[11] http://www.enseignementsup-recherche.gouv.fr/pid20161/mission-parite.html

[12] http://www.epws.org

[13] http://www.icwes14.org/

[14] http://www.sfpnet.fr

[15] C. Thibault, A. Pépin, M. Ducloy, E. Giacobino and M. Leduc. French Women in Physics: Status and Actions, *3rd IUPAP International Conference on Women in Physics*, Seoul (Korea), 2008. http://www.icwip2008.org/

[16] C. Hermann and F. Cyrot-Lackmann, Women in Science in France, *Science in Context* 15, 529–556, 2002. Revised and update version to be published by the same authors plus J. Peiffer and H. Rouch, in *Gender and Science: Studies Across Cultures*, Cambridge U. Press, India.

Modeling invasions and calculating establishment success chances

Patsy Haccou
Leiden University, the Netherlands

Abstract

Invasions play a big role in many biological contexts. Since in most cases initial numbers of invaders are small, branching processes provide a good way to model and study such processes. I will discuss the basic Galton Watson Branching process, multitype GWBPs and the inhomogeneous branching process (Smith and Wilkinson). I will show how to calculate establishment success chances, and show how these models can be applied in a biological context.

References

[1] Athreya, K.B., Ney, P. 1972 Branching Processes, *Springer-Verlag*

[2] Mode, C. 1972 Multitype Branching Processes, *Elsevier*

[3] Jagers, P. 1975 Branching Processes with Biological Applications

[4] Haccou, P., Jagers, P., Vatutin, V.A. 2005 Branching Processes: Variation, Growth, and Extinction of Populations, *Cambridge University Press*

What should maths teachers know about girls and boys?

Markku Hannula
University of Turku, Finland

There used to be clear gender differences favouring males in large scale mathematics performance tests (Hyde, Fennema & Lamon, 1990), but these overall differences have disappeared in most countries (IEA / TIMSS, 1999; OECD / PISA, 2000). There are tasks in mathematics that still produce large gender differences in favour of boys. For example, some conceptual tasks about fractions (Hannula, 2003) and infinity (Hannula, Pehkonen, Maijala, Soro, 2006), when the issues have not been explicitly dealt in school. Problem solving has been among the topics that have produced the largest and most durable gender differences, and this was reflected also in PISA study, where boys scored better in 15 of the 27 OECD countries. There are also task types where female students tend to perform higher, such as when there is a straightforward or standard solution or the task is in shopping context (van den Heuvel Panhuizen, 1997). Overall, girls seem to perform well on tasks that they have been taught to solve.

Moreover, as soon as mathematics becomes optional in schools, there tends to be overrepresentation of male students over female students. In Finland, roughly one third of girls and two thirds of boys choose the more advanced mathematics course in upper secondary school (Välijärvi & Tuomi 1995). This is also reflected in their respective test performances, the more students have studied mathematics, the better they tend to perform in tests. This leads to a widening gender gap in performance as students get older. At university level, mathematics programs typically attract mainly male students, although mathematics teacher education programs typically attract more female students. The ratio of female students decreases the further the studies continue.

As numerous studies on achievement differences indicate, there is no reason to believe that female students are underrepresented due to inferior mathematics skills. Rather, female students tend to opt out mathematics more often than male students at equal performance level. Some studies have indicated that students tend to perceive mathematics as a male domain (Frost, Hyde and Fennema, 1994), but this belief is mainly held by male student's and hence does not give an appropriate explanation to why female students who perceive mathematics as gender neutral opt out mathematics.

Studies on student mathematical self-efficacy beliefs have produced very consistent results that indicate that across age and performance levels female students tend to have lower self-confidence in mathematics than male students (e.g. Hannula, Maijala, Pehkonen & Nurmi, 2005; Leder, 1995). Lower self-confidence among female students has been found even on level of individual tasks, in case of both correct and incorrect answers (Hannula , Maijala, Pehkonen & Soro, 2002). Related to low self-confidence, female students also suffer mathematics anxiety more often than male students (Frost, Hyde & Fennema, 1994; Hembree, 1990). These results in affect provide explanation to why female students choose usually not to study optional mathematics, especially when we consider that female students may have higher performance levels in arts and social sciences Lower self-efficacy is also likely to explain why female students rely on school-taught solutions methods and avoid non-standard or own solution methods that include an element of risk.

There is no reason to believe that the low level of female students' self-efficacy beliefs is a natural and permanent gender characteristic of female sex. The research has cumulated evidence for the hypothesis that female students' lack of confidence in mathematics is consistent with their teachers' beliefs (Li, 1999; Soro, 2002; Sumpter, 2009) and that teachers' typical interaction patterns with male and female students may thus attribute to the generation of gender differences. Mathematics teachers tend to believe that their male students often have hidden talent, but due to being lazy and careless they underperform, while female students tend to reach their performance due to diligence and hard work even if they are not very talented. These teacher beliefs are assumed to lead to different feedback to male and female students and thus contributes to the observed gender differences in self-efficacy beliefs. Another theory formulates the female beliefs as "learned helplessness" (Licht & Dweck, 1987). According to this theory, male students get typically their negative feedback due to misbehaviour and lack of effort, while well-behaving female students get more negative feedback on their cognitive performance. Moreover, female students get positive feedback on tidyness and behaviour while male students mainly on their performance. In summary, this pattern leads female students to attribute success to effort and failure to lack of talent while male students learn to attribute success to talent and failure to lack of effort.

According to the tendencies found when women have started working in the traditionally male fields Räsänen (1989b) distinguishes three stages also in teachers' attitudes towards girls studying mathematics or

physics: 1. These subjects are not suitable for girls. Girls do not belong to these lessons. 2. It is all right that also girls study these subjects but they are taught in the same way as used earlier. 3. The pedagogical methods have to be developed so that both girls and boys will benefit as much as possible from teaching. Although there seems to be a general belief in most cultures that mathematics is suitable for girls, few teachers are aware of the need for and have skills to implement gender sensitive teaching.

References

[1] Frost, L. A., Hyde, J. S. & Fennema, E. 1994. Gender, mathematics performance, and mathematics related attitudes and affect: a meta-analytic synthesis. International Journal of Educational Research 21 (4), 373-385.

[2] Hannula, M.S. 2003b. Locating fraction on a number line. In N.A. Pateman, B.J. Dougherty & J. Zilliox (eds.) Proceedings of the 27th Conference of the International Group for the Psychology of Mathematics Education, Vol 3, 17-24.

[3] Hannula, M.S. , Maijala, H., Pehkonen, E. & Nurmi, A. (2005). Gender comparisons of pupils' self-confidence in mathematics learning. Nordic Studies in Mathematics Education 10 (3-4), 29-42.

[4] Hannula, M. S., Pehkonen, E.; Maijala, H.; Soro, R. 2006. Levels of students' understanding on infinity. Teaching Mathematics and Computer Science 4 (2), 317-337.

[5] Hembree, R. (1990), "The Nature, Effects, and Relief of Mathematics Anxiety," Journal of Research in Mathematics Education, 21, 33-46

[6] Hyde, J. S., Fennema, E. & Lamon, S. J. 1990. Gender differences in mathematics performance: A meta-analysis. Psychological bulletin 107, 139-155.

[7] Leder, G. 1995. Equity inside the mathematics classroom: Fact or artifact? In Secada,W.G., Fennema, E. & Adajian L.B (eds.) New directions for equity in mathematics education. Cambridge University Press.

[8] Li, Q. (1999). Teachers' beliefs and gender differences in mathematics: A review. Educational Research, 41(1), 63-76.

[9] Licht, B. G. & Dweck, C. S. 1987. Sex differences in achievement orientations. In M. Arnot & G. Weiner (eds.): Gender and the politics of schooling. London: Unwin-Hyman.

[10] Räsänen, L. (1989b). Är lärarna intresserade av jämställdhetsförsök? [Are teachers interested in gender equality studies?] BRYT AVAA. Yhteispohjoismainen Bryt-projekti tiedottaa. Marraskuu 1989 [Nordic BRYT AVAA project informs].

[11] Soro, R. (2002). Opettajien uskomukset tytöistä, pojista ja tasa-arvosta matematiikassa. [Teachers' beliefs about girls, boys, and equity in mathematics.]: Annales universitatis Turkuensis C 191. Turku, Finland: University of Turku.

[12] Sumpter, L. 2009. On aspects of mathematical reasoning; Affect and gender. Umeå university.

[13] Välijärvi, J. & Tuomi, P. 1995. Lukio nuorten valintojen ja oppimisen ympäristönä. Jyväskylän yliopisto, kasvatustieteiden tutkimuslaitoksen julkaisusarja A. Tutkimuksia 60.

[14] van den Heuvel-Panhuizen, M. 1997. How equally suited is realistic mathematics education for boys and girls?-A first exploration. In. E. Pehkonen (ed.) Proceedings of the 21st Conference of the International Group for the Psychology of Mathematics Education. Vol 3, 65-72.

Curves of genus 2 on rational normal scrolls

*Andrea Hofmann**
Matematisk Institutt, Universitetet i Oslo, Postboks 1053, Blindern, N-0316 Oslo.

Abstract

The objects of study in this article are curves of genus 2 in projective space which lie on rational normal scrolls. It is known that a curve C of genus 2 has a unique g_2^1 which gives rise to a surface scroll that contains C. For curves of genus 2 and degree d, $6 \leq d \leq 8$, we find a threefold scroll that contains the curve C and whose ideal together with the ideal of the surface scroll generates the ideal of C. This result leads to the conjecture that for a genus 2 curve C of arbitrary degree $d \geq 6$ we can always find a threefold scroll whose ideal together with the ideal of the surface scroll generates the ideal of C. Furthermore we study the syzygies of genus 2 curves and ask whether the ith syzygies of the ideal of the surface scroll S together with the ith syzygies of the ideals of all threefold scrolls that contain C generate the ith syzygies of I_C.

1 Introduction

In this paper we study the ideal and syzygies of curves of genus 2 and degree d in $\mathbf{P}^{d-2}$. It is known that the ideal of such a curve C is generated by quadrics. One interesting issue is then to investigate if the ideal of C is generated by quadrics that generate the ideals of scrolls the curve lies on. One can pose the same question for the ith syzygies of I_C. Here we are only interested in two-dimensional and three-dimensional rational normal scrolls that contain C.

1.1 Main results

A curve of genus 2 and degree d in $\mathbf{P}^{d-2}$ lies on one rational normal surface scroll S which is generated by the unique $g_2^1(C)$ and on a two-dimensional family of rational normal three-fold scrolls where each scroll is generated by a $g_3^1(C)$.

Proposition 1.1. *For a curve C of genus 2 and degree $6 \leq d \leq 8$ the following holds:*

(a) $S \cap V = C$,

(b) $I_S + I_V = I_C$,

where S is the $g_2^1(C)$-scroll and V is a $g_3^1(C)$-scroll that does not contain S. In particular, the ideal I_C is generated by quadrics of rank 4 or less.

These results lead to the following conjecture:

Conjecture 1.2. For $d \geq 9$, the ideal of a curve C of genus 2 and degree d in $\mathbf{P}^{d-2}$ is generated by the quadrics in I_S and I_V where S is the $g_2^1(C)$-scroll and V is a $g_3^1(C)$-scroll that does not contain S.

Now one can ask the same question for higher syzygies:

Question 1.3. Let C be a curve of genus 2 and degree $d \geq 6$. For $1 \leq i \leq d-6$, are the ith syzygies of I_C generated by the ith syzygies of I_S, where S is the $g_2^1(C)$-scroll, and the ith syzygies of the ideals of all $g_3^1(C)$-scrolls that do not contain S?

Remark 1.4. As we will see in Example 3.2, the ith syzygies of the ideal I_S and of the ideal of only one such $g_3^1(C)$-scroll are not enough to generate all ith syzygies of I_C.

*I wish to thank my advisor Kristian Ranestad for useful discussions and proofreading.

1.2 Preliminaries

In algebraic geometry one important issue is to study properties of projective algebraic varieties and to classify these objects according to their properties. In this article we study curves embedded in the projective space $\mathbf{P}^n$.

Definition 1.5. The n-dimensional projective space $\mathbf{P}^n$ over the complex numbers $\mathbf{C}$ is $\mathbf{C}^{n+1} - \{0\}/\sim$ where two points $x = (x_0, x_1, \ldots, x_n)$ and $y = (y_0, y_1, \ldots, y_n) \in \mathbf{C}^{n+1}$ are equivalent (we write $x \sim y$), if there exists a $\lambda \in \mathbf{C} - \{0\}$ such that $x_i = \lambda y_i$ for all $i = 0, \ldots, n$.

From now on let $x_0, x_1, \ldots, x_n$ denote the coordinates in $\mathbf{P}^n$.

A projective variety $X \subseteq \mathbf{P}^n$ is an irreducible zero set of finitely many homogeneous polynomials $f_1, \ldots, f_r \in \mathbf{C}[x_0, x_1, \ldots, x_n]$:

$$X = \{P \in \mathbf{P}^n | f_i(P) = 0 \text{ for all } i = 1, \ldots, r\}.$$

The ideal of X is then defined to be the ideal generated by $f_1, \ldots, f_r$:

$$I_X = (f_1, \ldots, f_r).$$

In order to study and classify algebraic varieties, the attention was drawn to syzygies and minimal free resolutions in the past decades. One is interested in the connection between the geometry of projective varieties X and the minimal free resolution of its ideal I_X.

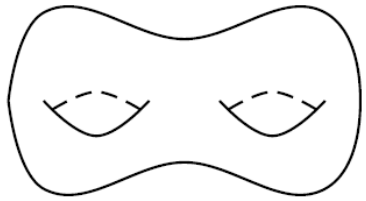
If Y is a variety which contains X, then $I_Y \subseteq I_X$ and one aim is to investigate which syzygies of I_X are generated by the syzygies of I_Y. We will call these syzygies geometric syzygies.

One natural question is then to ask whether the space of all ith syzygies of the ideal I_X is spanned by geometric syzygies.

This question was discussed for some varieties in [4], [7], [8] and [9].

We are studying curves of genus 2 and degree greater or equal to 6 which are lying in a natural way on rational normal scrolls and look at the syzygies of the ideals of the scrolls. Here we will focus on rational normal scrolls of dimension 2 and 3.

A curve of genus 2 over $\mathbf{C}$ looks topologically like a double torus when considered as a two-dimensional real manifold:



Now, if a a curve C is embedded in $\mathbf{P}^n$, then each linear space $H \subseteq \mathbf{P}^n$ of dimension $n-1$, which we also call a hyperplane, intersects the curve in a finite number of points. The number of these points is called the degree of C.

If C is a smooth curve of degree d in $\mathbf{P}^n$, we say that it has an embedding of degree d in $\mathbf{P}^n$.

Definition 1.6. A rational normal scroll Y of dimension k in $\mathbf{P}^n$ is the union of $(k-1)$-dimensional linear spaces, parametrized over $\mathbf{P}^1$, such that Y is linearly normal.

Here the notion of "linearly normal" is equivalent with $\deg(Y) + \dim(Y) = n + 1$. For the definition of linearly normal see also [1], Chapter III, Exercise appendix D. Each such $(k-1)$-dimensional linear space is also called a fiber of the scroll. See [2], Appendix A2H, or [3] for three different characterizations of a rational normal scroll.

On each scroll there exist rational curves of arbitrary degree which intersect each fiber in the scroll in exactly one point. Actually, there exist k rational curves on the scroll such that each of these curves intersects each fiber of the scroll in one point and such that the degrees of these curves are $d_1, \ldots, d_k$ with $d_1 + \cdots + d_k = n - k + 1$ and such that each fiber in the scroll is spanned by the k points in the intersection of the k curves and this fiber.

Definition 1.7. We call $(d_1, \ldots, d_k)$ the type of the scroll, and we will always order these numbers in such a way that $d_1 \geq d_2 \geq \cdots \geq d_k$. The degree of the scroll is given by $d_1 + d_2 + \cdots + d_k$.

Proposition 1.8. *By an appropriate choice of coordinates, the ideal I_Y of a rational normal scroll Y in $\mathbf{P}^n$ of scroll type $(d_1, \ldots, d_k)$ is generated by the 2×2 minors of the following $2 \times m$ matrix:*

$$\begin{pmatrix} x_0 & x_1 & \cdots & x_{d_1-1} & x_{d_1+1} & \cdots & x_{d_1+d_2} & \ldots & x_{n-d_k} & \ldots & x_{n-1} \\ x_1 & x_2 & \cdots & x_{d_1} & x_{d_1+2} & \cdots & x_{d_1+d_2+1} & \ldots & x_{n-d_k+1} & \ldots & x_n \end{pmatrix}$$

(Notice that m is the degree of the scroll.)

Proof. See [3]. □

Definition 1.9. A $g^1_k(C)$ is a family of k-tuples of points on C parametrized by a $\mathbf{P}^1$.
A point $P \in \mathbf{P}^n$ is a basepoint for a $g^1_k(C)$ if P is contained in all k-tuples of points in the $g^1_k(C)$. If the $g^1_k(C)$ has no basepoints, then we say it is basepoint-free.

Notice that a basepoint-free $g^1_k(C)$ gives rise to a $k:1$ map $C \to \mathbf{P}^1$.

The following is known (use the Riemann-Roch Theorem for curves in [6], Chapter IV):

Proposition 1.10. *A curve of genus 2 has an embedding of degree d in $\mathbf{P}^{d-2}$ when $d \geq 5$.*

Now let C be a curve of genus 2, embedded in $\mathbf{P}^{d-2}$.

A $g^1_k(C)$, $2 \leq k \leq d-2$, gives rise to a rational normal scroll of dimension k: Each fiber of such a scroll is given by the $(k-1)$-dimensional linear space spanned by a k-tuple in the $g^1_k(C)$. We call such a scroll a $g^1_k(C)$-scroll.

There exists exactly one $g^1_2(C)$ and the family of all $g^1_3(C)$ is an Abelian surface, which we denote by $Pic^3(C)$. This surface is isomorphic to $Jac(C)$, the Jacobian variety of C. You can read about the Jacobian variety and $Pic^d(C)$ in [1], Chapter I, §3.

We will denote a $g^1_3(C)$ by $|D|$ and a member of $|D|$ (i.e. three points on C) by D'.

Mainly we are interested in the two-dimensional scroll $S = \bigcup_{(P,P') \in g^1_2(C)} L_{P,P'}$ and in three-dimensional scrolls $V = V_{|D|} = \bigcup_{D' \in |D|} \overline{D}'$ where by $L_{P,P'}$ we mean the line through the points P and P' and by $\overline{D}'$ we mean the plane spanned by the three points in D'.
The degree of S is $d-2-2+1 = d-3$ and the degree of V is equal to $d-2-3+1 = d-4$.

Since by construction $C \subseteq S$ and $C \subseteq V$, it follows that $I_S \subseteq I_C$ and $I_V \subseteq I_C$. It is known that also the ideal I_C is generated by quadrics:

Proposition 1.11. *Let C be a curve of genus 2 and degree $d \geq 6$. The ideal I_C is generated by $\binom{d-3}{2} + d - 5$ quadrics; $\binom{d-3}{2}$ of these come already from I_S.*

Proof. By Theorem (4.a.1) in [5], C is projectively normal and the ideal I_C is generated by quadrics. Now we can apply the Riemann-Roch formula for curves (see e.g. [6], Chapter IV) to obtain that I_C is generated by $\binom{d-3}{2} + d - 5$ quadrics. Since $C \subseteq S$, $I_S \subseteq I_C$. By Proposition 1.8 the ideal I_S is generated by the 2×2 minors of a $2 \times (d-3)$ matrix, i.e. the number of generators of I_S is equal to $\binom{d-3}{2}$. □

Now one interesting question is to ask whether the quadrics in I_S and I_V for a general V are enough to generate I_C. This problem will be discussed in Section 2 for curves of degree $d = 6$, $d = 7$ and $d = 8$.

In the following, C will always denote an irreducible, smooth curve of genus 2 and we will denote the quadrics in its ideal by $I_C(2)$.

2 The ideal of genus 2 curves of degree $d \geq 6$

2.1 Genus 2 curves of degree 6

Let C be a curve of genus 2 and degree 6 in $\mathbf{P}^4$. Now we know that C is lying on a $g^1_2(C)$-scroll of type $(3,0)$ or $(2,1)$. Assume that S has scroll type $(2,1)$ (the case where the scroll type is $(3,0)$ is analogous). Then it is possible to choose coordinates of $\mathbf{P}^4$ in such a way that its ideal I_S is generated by the 2×2 minors of the following matrix:

$$\begin{pmatrix} x_0 & x_1 & x_3 \\ x_1 & x_2 & x_4 \end{pmatrix}.$$

Set $Q_1 = x_0x_2 - x_1^2$, $Q_2 = x_0x_4 - x_1x_3$ and $Q_3 = x_1x_4 - x_2x_3$.

Now we have an explicit description of the ideal I_C:

Proposition 2.1. *The four quadrics Q_1, Q_2, Q_3, Q where Q is a quadric in $\mathbf{P}^4$ that contains C but not S, generate the ideal I_C.*

On the other hand, the quadrics Q_1, Q_2, Q_3, Q where Q is a general quadric in $\mathbf{P}^4$, generate the ideal of a curve of genus 2 and degree 6 in $\mathbf{P}^4$.

Proof. For a general quadric Q in $\mathbf{P}^4$ which contains C and not S, we have $\dim(S \cap Q) = 1$ and $\deg(S \cap Q) = 6$. Hence the intersection is a curve C of degree 6. Moreover, with the adjunction formula in [6] (Proposition 1.5 in Chapter V) it can be calculated that the genus of C is 2. □

A quadric of rank 3 in $\mathbf{P}^4$ is a cone over a conic in $\mathbf{P}^2$ with a line as vertex. A special hyperplane intersection H_Q of this quadric splits into two $\mathbf{P}^2$, call them A_1 and A_2, which intersect in the quadric's vertex line L. The two planes A_1 and A_2 move in a one-dimensional family, parametrized by the conic, while the line L is fixed. Since C is of degree 6 the intersection $H_Q \cap C$ consists of 6 points which are distributed among $A_1 - L$, $A_2 - L$ and L. We say that a quadric of rank 3 in $\mathbf{P}^4$ is of type (a_1, a_2, a_3) if the six points in $H_Q \cap C$ are distributed in such a way that a_1 points lie in $A_1 - L$, a_2 points lie on L and a_3 points lie in $A_2 - L$.

Lemma 2.2. *We have the following possibilities for a_1, a_2 and a_3:*

a_1	a_2	a_3
2	2	2
3	0	3

Proof. Since A_1 and A_2 are lying in the same family of planes parametrized by a conic, $A_1 \cap C$ and $A_2 \cap C$ are parametrized by the same conic, i.e. $a_1 = a_3$. Since, for any curve C', a $g^1_1(C')$ gives an isomorphism $C' \cong \mathbf{P}^1$, it follows that C' has a g^1_1 if and only if the genus of C' is 0. So our genus 2 curve C cannot have a g^1_1. This implies that $a_1 \neq 1$ and $a_3 \neq 1$. Moreover, since each hyperplane in $\mathbf{P}^4$ intersects the curve in 6 points, it is impossible to have $a_1 + a_2 = 6$ or $a_2 + a_3 = 6$, in other words we must have $a_1 \geq 2$, $a_2 \geq 2$. □

Quadrics of type $(2, 2, 2)$ contain the scroll S since they contain the $g^1_2(C)$. On the other hand, the quadrics of type $(3, 0, 3)$ will not contain S since the $g^1_3(C)$ involved here has no basepoints. So we have seen that a quadric does not contain S if and only if for a special hyperplane intersection we can write $H_Q \cap C = 2D'$, where D' is lying in somebasepoint-free $g^1_3(C)$.

In a similar way we can look at special hyperplane intersections H_Q of quadrics of rank 4 in $\mathbf{P}^4$: A quadric of rank 4 in $\mathbf{P}^4$ is a cone over a quadric in $\mathbf{P}^3$ with a point P as vertex. A special hyperplane intersection H_Q of this quadric splits into two $\mathbf{P}^2$, A_1 and A_2, that intersect in a line l. A general such H_Q will intersect the curve C in such a way that no point of $H_Q \cap C$ is lying on l. That is, a quadric of type (a_1, a_2, a_3) will be a quadric where a_1 points of $H_Q \cap C$ lie on $A_1 - l$, a_3 points are lying on $A_2 - l$ and a_2 is 1 if $P \in H_Q \cap C$ or 0 if $P \notin H_Q \cap C$.

Lemma 2.3. *Here we have the following possibilities for a_1, a_2 and a_3:*

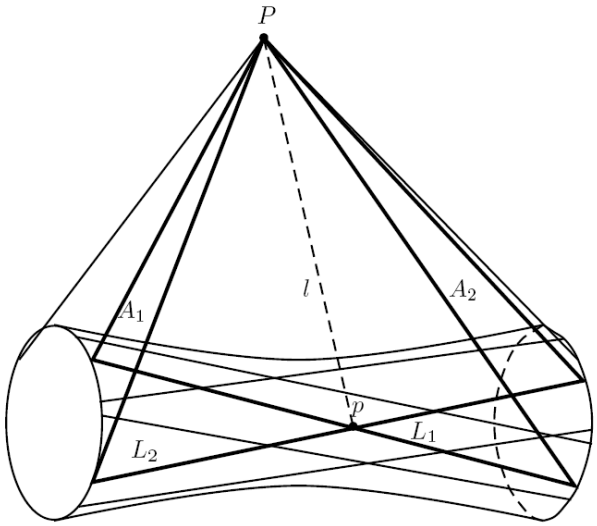
a_1	a_2	a_3
2	1	3
2	0	4
3	0	3

Proof. Since the two $\mathbf{P}^2$ are lying in two different families, we no longer need to have $a_1 = a_3$, but still, by the same reasons as in Lemma 2.2, $a_1 \geq 2$, $a_3 \geq 2$. □

Note that the quadrics of type $(2, 1, 3)$ and $(2, 0, 4)$ contain S while a quadric of type $(3, 0, 3)$ does not contain S. Moreover, the two families of planes in a quadric of type $(3, 0, 3)$ give us two rational normal scrolls $V_{|D_1|}$ and $V_{|D_2|}$, which are both equal to the quadric. We conclude that a quadric of rank 4 in $\mathbf{P}^4$ does not contain S if and only if we can write $H_Q \cap C = D'_1 + D'_2$ where D'_1 and D'_2 are members of some basepoint-free $g^1_3(C)$'s $|D_1|$ and $|D_2|$.

Example 2.4.

Below we have illustrated a quadric of rank 4 in $\mathbf{P}^4$. This is the cone over a smooth quadric in $\mathbf{P}^3$ with a point as vertex. There are two different families of lines on the quadric in $\mathbf{P}^3$, and a special hyperplane intersection H_Q of the quadric in $\mathbf{P}^4$ consists of two planes A_1 and A_2 which intersect in a line l. For each point p on the smooth quadric in $\mathbf{P}^3$, the tangent plane at this point will intersect the quadric in exactly the two lines L_1 and L_2. By moving the point p we obtain the two families of lines.



There is an involution

$$i : Pic^3(C) \to Pic^3(C),$$

$$|D| \mapsto |(H_{\mathcal{Q}} \cap C) - D|.$$

The quotient $Pic^3(C)/(i(|D|) \sim |D|)$, i.e. here we identify a $g^1_3(C)$ $|D|$ with $|(H_{\mathcal{Q}} \cap C) - D|$, is an algebraic surface, called Kummer's quartic surface.

All the quadrics in the ideal I_C span a linear space $\mathbf{P}^3$ while the quadrics in I_S span a $\mathbf{P}^2$. Call these linear spaces $\mathbf{P}^3_{\mathcal{C}}$ and $\mathbf{P}^2_{\mathcal{S}}$. The locus $\{Q \in I_C(2) | rank(Q) \leq 4\}$ splits into the $\mathbf{P}^2_{\mathcal{S}}$ and Kummer's quartic surface K in $\mathbf{P}^3_{\mathcal{Q}}$. Here the points on $K - \mathbf{P}^2_{\mathcal{S}}$ correspond to quadrics that contain C but not S. The Kummer surface has 16 simple nodes. They correspond to quadrics of rank 3, 15 of these lie on $K - \mathbf{P}^2_{\mathcal{S}}$ and they correspond to points $|D|$ in $Pic^3(C)$ which satisfy $|H_{\mathcal{Q}} \cap C| = 2|D|$. These again correspond to those quadrics of rank 3 that contain C and $V_{|D|}$ but not S. The remaining node lies in $\mathbf{P}^2_{\mathcal{S}}$ and corresponds to the only quadric of rank 3 that contains S. For a special hyperplane intersection of that quadric we can write $H_{\mathcal{Q}} \cap C = E_1 + E_2 + R_1 + R_2$ where E_1 and E_2 are elements in the $g^1_2(C)$ and R_1 and R_2 are points in $H_{\mathcal{Q}} \cap C$.

In particular we can take the extra quadric in Proposition 2.1 to be of rank 4:

Proposition 2.5. *The ideal of a curve C of genus 2 and degree 6 in $\mathbf{P}^4$ is generated by the quadrics in I_S and I_V where V is a $g^1_3(C)$-scroll that does not contain S.*

Proof. In this situation, a $g^1_3(C)$-scroll is a quadric of rank 4 and if we add an arbitrary quadric in $K - \mathbf{P}^2_{\mathcal{S}}$, then this quadric together with the quadrics that generate I_S generate the ideal I_C. □

2.2 Genus 2 curves of degree 7

Let C be a curve of genus 2 and degree 7 in $\mathbf{P}^5$. We know that the degree of any $g^1_2(C)$-scroll is 4, so C lies on a scrollar surface S of scroll type $(2,2)$, $(3,1)$ or $(4,0)$. Assume that S is of scroll type $(2,2)$ (the other two cases are analogous).

One can choose coordinates such that the ideal I_S is generated by the 2×2 minors of the following matrix:

$$\begin{pmatrix} x_0 & x_1 & x_3 & x_4 \\ x_1 & x_2 & x_4 & x_5 \end{pmatrix}. \tag{1}$$

We include the following proposition without proof in order to show that we can obtain an explicit description of the generators of I_C. Notice that by Proposition 1.11 we need 2 quadrics in addition to I_S to generate I_C.

Proposition 2.6. *The ideal I_C is generated by the 2×2 minors of the matrix (1) and two quadrics*

$$\begin{aligned} Q_1 &= l_1x_0 + l_2x_1 + l_3x_3 + l_4x_4, \\ Q_2 &= l_1x_1 + l_2x_2 + l_3x_4 + l_4x_5 \end{aligned}$$

where l_1, l_2, l_3 and l_4 are linear forms in $\mathbf{P}^5$.

Conversely, if we take

$$\begin{aligned} Q_1 &= l_1x_0 + l_2x_1 + l_3x_3 + l_4x_4, \\ Q_2 &= l_1x_1 + l_2x_2 + l_3x_4 + l_4x_5 \end{aligned}$$

with general linear forms $l_1, \ldots, l_4$ in $\mathbf{P}^5$, then the 2×2 minors of (1) together with Q_1 and Q_2 generate the ideal of a smooth curve of genus 2 and degree 7 in $\mathbf{P}^5$.

A special hyperplane intersection $H_{\mathcal{Q}}$ of a quadric of rank 4 in $\mathbf{P}^5$ splits into two $\mathbf{P}^3$ that intersect in a $\mathbf{P}^2$. Now we know that $H_{\mathcal{Q}} \cap C$ consists of seven points. Let Δ_1 and Δ_2 denote the two $\mathbf{P}^3$ and let $\Delta_{1,2}$ denote the intersection of Δ_1 and Δ_2. Then the seven points in $H_Q \cap C$ are divided between $\Delta_1 - \Delta_{1,2}$, $\Delta_2 - \Delta_{1,2}$ and $\Delta_{1,2}$. Note that as in the case of a curve of degree 6, Δ_1 and Δ_2 move in a one-dimensional family, while $\Delta_{1,2}$ is fixed. We say that a quadric has type (a_1, a_2, a_3) if a special hyperplane intersection of this quadric intersects C in such a way that a_1 points are in $\Delta_1 - \Delta_{1,2}$, a_2 points are in $\Delta_{1,2}$ and a_3 points are in $\Delta_2 - \Delta_{1,2}$.

Proposition 2.7. *Let $V = V_{|D|}$ be a $g^1_3(C)$-scroll that does not contain the $g^1_2(C)$-scroll S. Then $I_S(2) \cap I_V(2) = (Q)$, where Q is of type $(2, 2, 3)$, and this implies that $S \cap V = C$.*

Proof. The intersection of two quadrics Q_1 and Q_2 in $\mathbf{P}^5$ is of dimension 3 and degree 4. So if S and V are contained in Q_1 and Q_2, then it follows that $Q_1 \cap Q_2 = V \cup \mathbf{P}^3$. Since S is irreducible, it must be either contained in V or in $\mathbf{P}^3$. But by assumption, S is not contained in V and it cannot be contained in $\mathbf{P}^3$ either, since it spans all of $\mathbf{P}^5$. So we have seen that at most one quadric can contain both S and V.

On the other hand, if $D' \in |D|$ and $E \in g^1_2(C)$, then we can write $H_{\mathcal{Q}} \cap C = D' + E + B_1 + B_2$, where B_1 and B_2 are points on the curve. If we project from the line L that intersects the curve of degree 7 in $\mathbf{P}^5$ in the two points B_1 and B_2, we obtain a curve C' of genus 2 and degree 5 on a quadric Q' in $\mathbf{P}^3$. There are two families of lines on Q'. A line in one family will intersect the curve C' in an element in $|D|$ and a line in the other family will intersect the curve in an element in the $g^1_2(C)$. The cone over this quadric Q' with the line L as vertex is a quadric of type $(2, 2, 3)$ which contains V and S. □

Proposition 2.8. *For a three-dimensional rational normal scroll V that does not contain S we have $I_S + I_V = I_C$.*

Proof. Notice that by Proposition 1.8 and 1.11 the ideals I_S, I_V and I_C are generated by quadrics. Obviously, $I_S(2) + I_V(2) \subseteq I_C(2)$. Moreover we know that there are six generators in $I_S(2)$ and three generators in $I_V(2)$ and by Proposition the intersection $I_S(2) \cap I_V(2)$ consists of only one quadric, hence there are $6 + 3 - 1 = 8$ independent quadrics in $I_S(2) + I_V(2)$ which by Proposition 1.11 is exactly the number of generators of I_C. □

2.3 Genus 2 curves of degree 8

Let C be a curve of genus 2 and degree 8 in $\mathbf{P}^6$. Since the $g^1_2(C)$-scroll S has degree 5, the scroll type of S can be $(3, 2)$, $(4, 1)$ or $(5, 0)$. We assume that the scroll type is $(3, 2)$ (the other two cases are analogous) and then we know that after an appropriate choice of coordinates, the ideal I_S is generated by the 2×2 minors of the following matrix:

$$\begin{pmatrix} x_0 & x_1 & x_2 & x_4 & x_5 \\ x_1 & x_2 & x_3 & x_5 & x_6 \end{pmatrix}. \tag{2}$$

Note that by Proposition 1.11 we need 3 extra quadrics in addition to the 2×2 minors of the above matrix to generate the ideal I_C. As in the case $d = 7$ we found an explicit description of the 3 extra quadrics (also here no proof is stated):

Proposition 2.9. *There are three quadrics*

$$\begin{aligned} Q_1 &= l_1x_0 + l_2x_1 + l_3x_4, \\ Q_2 &= l_1x_1 + l_2x_2 + l_3x_5, \\ Q_3 &= l_1x_2 + l_2x_3 + l_3x_6 \end{aligned}$$

where l_1, l_2 and l_3 are linear forms in $\mathbf{P}^6$ which together with the 10 2×2 minors of the matrix (2) generate I_C.

On the other hand, if we set

$$\begin{aligned} Q_1 &= l_1x_0 + l_2x_1 + l_3x_4, \\ Q_2 &= l_1x_1 + l_2x_2 + l_3x_5, \\ Q_3 &= l_1x_2 + l_2x_3 + l_3x_6 \end{aligned}$$

with general linear forms l_1, l_2 and l_3 in $\mathbf{P}^6$, then Q_1, Q_2 and Q_3 together with the 2×2 minors of the matrix (2) generate the ideal of a smooth curve of genus 2 and degree 8 on S.

We also obtain a result analogous to Proposition 2.8 (also here the proof is omitted):

Proposition 2.10. *For a $g^1_3(C)$-scroll V that does not contain S the following holds:*

(a) $S \cap V = C$

(b) $I_S + I_V = I_C$.

2.4 Genus 2 curves of arbitrary degree d, $d \geq 9$

The results for curves of low degree presented in the previous section lead to the following conjecture:

Conjecture 2.11. Let C be a curve of genus 2 and degree $d \geq 9$ in $\mathbf{P}^{d-2}$. For a general $g^1_3(C)$-scroll V we have $I_S + I_V = I_C$, i.e. the quadrics in I_S and I_V generate I_C. In particular, I_C is generated by quadrics of rank 4 or less.

Remark 2.12. Examples for $d = 12$, $14 \leq d \leq 47$, can be calculated with the computeralgebrasystem Macaulay 2. With this system we can explicitly construct the ideal of a surface scroll S and the ideal of a threefold scroll V that does not contain S in such a way that they both contain a genus 2 curve C of degree d. The computations show then that the two ideals I_S and I_V generate the ideal I_C.

3 Resolutions and syzygies

Having investigated the ideal I_C, the next step is to look at so-called higher syzygies of I_C:

If X is a variety and I_X is generated by the homogeneous polynomials $f_1, \ldots, f_m$, then we can ask if there are any relations between $f_1, \ldots, f_m$, that is if there exist polynomials $g_1, \ldots, g_m$ in $R = \mathbf{C}[x_0, \ldots, x_n]$ such that $\sum_{i=1}^m g_i f_i = 0$. Any such relation is called a first syzygy of I_X. The degree of a syzygy is the degree of the coefficients g_i. We can continue in this way and ask if there are any relations between the relations and so on.

In this way we obtain an exact complex, also called a resolution of I_X:

$$0 \to M_l \stackrel{\phi_{l-1}}{\to} \cdots \stackrel{\phi_2}{\to} M_2 \stackrel{\phi_1}{\to} M_1 \stackrel{\phi_0}{\to} I_X \to 0,$$

where the $M_i = \oplus_j R(-j)^{\beta_{i,j}}$ are free R-modules and the maps ϕ_i are given by multiplication with matrices that have the coefficients g_k of the syzygies as entries.

The image of ϕ_i is called the ith syzygy module of I_X and the ideal I_X itself is called the 0th syzygy module.

Example 3.1. We have seen that, after choice of coordinates, the ideal of a rational normal surface scroll S in $\mathbf{P}^4$ is generated by the 2×2 minors of the following matrix:

$$\begin{pmatrix} x_0 & x_1 & x_3 \\ x_1 & x_2 & x_4 \end{pmatrix}.$$

Set $Q_1 = x_0x_2 - x_1^2$, $Q_2 = x_0x_4 - x_1x_3$ and $Q_3 = x_1x_4 - x_2x_3$. We obtain the first syzygies in the following way: Obviously,

$$0 = \det \begin{pmatrix} x_0 & x_1 & x_3 \\ x_0 & x_1 & x_3 \\ x_1 & x_2 & x_4 \end{pmatrix} = x_0Q_3 - x_1Q_2 + x_3Q_1.$$

Likewise,

$$x_1Q_3 - x_2Q_2 + x_4Q_1 = 0.$$

And it is easy to check that $x_0Q_3 - x_1Q_2 + x_3Q_1$ and $x_1Q_3 - x_2Q_2 + x_4Q_1$ are independent over $R = \mathbf{C}[x_0, x_1, x_2, x_3, x_4]$. That is, there are no second syzygies.

Thus, we obtain the following resolution:

$$0 \to R(-3)^2 \xrightarrow{\phi} R(-2)^3 \xrightarrow{\psi} I_S \to 0,$$

where ψ is given by multiplication with the matrix

$$\begin{pmatrix} x_0x_2 - x_1^2 & x_0x_4 - x_1x_3 & x_1x_4 - x_2x_3 \end{pmatrix}$$

and ϕ is given by multiplication with the matrix

$$\begin{pmatrix} x_3 & x_4 \\ -x_1 & -x_2 \\ x_0 & x_1 \end{pmatrix}.$$

A rational normal threefold-scroll in $\mathbf{P}^4$ is a quadric cone in $\mathbf{P}^4$. Let V be a threefold-scroll with ideal $I_V = (\tilde{Q})$. Then the resolution of I_V is given by

$$0 \to R(-2) \xrightarrow{g} I_V \to 0,$$

where the map g is given by multiplication with $\tilde{Q}$.

The resolution of the ideal I_C of a curve of genus 2 and degree 6 lying on S and the V from above is given by

$$0 \to R(-5)^2 \xrightarrow{f_3} R(-3)^2 \oplus R(-4)^3 \xrightarrow{f_2} R(-2)^4 \xrightarrow{f_1} I_C \to 0,$$

where (after choice) f_1 is given by multiplication with the matrix

$$\begin{pmatrix} x_0x_2 - x_1^2 & x_0x_4 - x_1x_3 & x_1x_4 - x_2x_3 & \tilde{Q} \end{pmatrix}$$

f_2 is given by multiplication with the matrix

$$\begin{pmatrix} x_3 & x_4 & \tilde{Q} & 0 & 0 \\ -x_1 & -x_2 & 0 & \tilde{Q} & 0 \\ x_0 & x_1 & 0 & 0 & \tilde{Q} \\ 0 & 0 & -x_0x_2 + x_1^2 & -x_0x_4 + x_1x_3 & -x_1x_4 + x_2x_3 \end{pmatrix}$$

and f_3 is given by multiplication with the matrix

$$\begin{pmatrix} \tilde{Q} & 0 \\ 0 & \tilde{Q} \\ -x_3 & -x_4 \\ x_1 & x_2 \\ -x_0 & -x_1 \end{pmatrix}.$$

Here we see that only the 0th syzygies of I_C can be generated by the 0th syzygies of I_S and I_V where S is the $g_2^1(C)$-scroll and V a $g_3^1(C)$-scroll with $I_V = (\tilde{Q})$.

Example 3.2. In $\mathbf{P}^5$, the resolutions of I_S, I_V and I_C where C is a curve of genus 2 and degree 7, S is the $g^1_2(C)$-scroll and V a $g^1_3(C)$-scroll, are the following:

$$0 \to R(-4)^3 \to R(-3)^8 \to R(-2)^6 \to I_S \to 0,$$

$$0 \to R(-3)^2 \to R(-2)^3 \to I_V \to 0$$

and

$$0 \to R(-6)^2 \to R(-4)^3 \oplus R(-5)^4 \to R(-3)^{12} \to R(-2)^8 \to I_C \to 0.$$

As we have already shown, the 0th syzygies of I_C are generated by the 0th syzygies of I_S and I_V where V is a $g^1_3(C)$-scroll that does not contain S. Now one can study the first syzygies of I_C and investigate if these are generated by the first syzygies of I_S and the first syzygies of the ideals of all $g^1_3(C)$-scrolls that do not contain S. Obviously, one such $g^1_3(C)$-scroll is not enough.

Remark 3.3. Now Examples 3.1 and 3.2 lead to the following question, analogous to the question for the generators of the ideal: For any $1 \leq i \leq d-6$, are the ith syzygies of I_C generated by the ith syzygies of I_S, where S is the $g^1_2(C)$-scroll, and the ith syzygies of the ideals of *all* $g^1_3(C)$-scrolls that do not contain S?

References

[1] Arbarello, Cornalba, Griffiths and Harris. Geometry of Algebraic Curves. Springer Verlag, 1985 .

[2] D. Eisenbud. The Geometry of Syzygies-A Second Course in Commutative Algebra and Algebraic Geometry. *Springer Verlag*, 2005.

[3] D. Eisenbud, J. Harris. On Varieties of Minimal Degree (A Centennial Account). *Proceedings of Symposia in Pure Mathematics*, (46): 3–13, 1987.

[4] F. Eusen. Der Ansatz von Paranjape und Ramanan bei der Vermutung von Green. *Ph.D. Thesis, Universität Hannover*, 1994.

[5] M.L. Green. Koszul Cohomology and the Geometry of Projective Varieties. *Journal of Differential Geometry*, (19): 125–171, 1984.

[6] R. Hartshorne. Algebraic Geometry. *Springer Verlag*, 1977.

[7] H.-C. Graf von Bothmer. Geometrische Syzygien von kanonischen Kurven. *Ph.D. Thesis, University of Bayreuth*, 2000.

[8] H.-C. Graf von Bothmer, K. Hulek. Geometric Syzygies of Elliptic Normal Curves and their Secant Varieties. *Manuscripta Mathematica*, (113):35–68, 2004.

[9] H.-C. Graf von Bothmer. Scrollar Syzygies of General Canonical Curves with Genus at most 8. *Trans. Amer. Math. Soc.*,(359):465–488, 2007.

On inequalities in borderline cases

Ritva Hurri-Syrjänen

Abstract

In the mini-course we studied the validity of classical Sobolev inequalities and embeddings also when the underlying bounded domain D in $\mathbb{R}^n$, $n \geq 2$, need not have a smooth boundary ∂D. The target space of the embedding was of Lebesgue or Hölder type. Our goal in this special lecture is consider the position in limiting cases. This is a survey talk. Among other things we consider the following:

If ∂D is smooth, it is a familiar fact that the Sobolev space $W^{1,n}(D)$ is embedded in the Orlicz space L_ϕ with Young function ϕ with values which behave like $\exp(t^{n/(n-1)})$ for large values of the argument t. We consider here, on one hand, spaces of functions that are larger than $W^{1,n}(D)$, but are contained in

$$\bigcap_{1<p<n} W^{1,p}(D)\,,$$

and on the other hand, also spaces of functions that are smaller than $W^{1,n}(D)$, but each $W^{1,p}(D)$ with $p > n$ is contained in them.

One of our objects is to give conditions on D which are sufficient to ensure Sobolev inequalities of exponential type yet allow the boundary of D to be quite rough. We show that if D is a bounded c_0-John domain and u is a function on D such that

$$I(a,D) := \left(\int_D |\nabla u(x)|^n \log^{an}(e + |\nabla u(x)|)\,dx\right)^{1/n} < \infty$$

for some $a < 1 - 1/n$, then there are constants $c_i = c_i(a, c_0, |D|, n)$, $i = 1, 2$, such that

$$\int_D \exp\left(\frac{|u(x) - u_D|}{c_1 I(a,D)}\right)^{\alpha} dx \leq c_2\,,$$

where $\alpha = n/(n - 1 - an)$ and $u_D = |D|^{-1}\int_D u(x)\,dx$. When ∂D is smooth, the special case $a = 0$ corresponds to the Trudinger embedding and $a < 0$ to the result obtained by Fusco, Lions and Sbordone. In a limiting case, corresponding to $a = 1 - 1/n$, we show that a double exponential inequality holds in smooth and certain nonsmooth domains D in $\mathbb{R}^n$: given any positive constants A_1 and $A_2 < 1$ such that $A_1 A_2 < 2e^{-1}$, $A_3 > 0$, which does not depend on $|\nabla u|$, such that

$$\int_D \exp\left(A_1 \exp\left(A_2 \frac{|u(x) - u_D|}{I(D)}\right)^{n/(n-1)}\right) dx \leq A_3\,,$$

where

$$I(D) := \left(\int_D |\nabla u(x)|^n \log^{n-1}(e + |\nabla u(x)|)\,dx\right)^{1/n} < \infty\,.$$

This limiting case as well as the previous John domain case is from the joint work of Edmunds and Hurri-Syrjänen.

The background information for this special lecture is found in the papers and books listed in the Bibliography.

References

[Ad] Robert A. Adams, *Sobolev Spaces*. Academic Press, Inc., Orlando, 1989.

[EdEv] D. E. Edmunds and W. D. Evans, *Hardy Operators, Function Spaces and Embeddings*. Springer, Springer-Verlag Berlin Heidelberg, 2004.

[EH-S] D. E. Edmunds and R. Hurri-Syrjänen, *Sobolev inequalities of exponential type*, Israel J. Math. **123**(2001), 61–92.

[FLS] N. Fusco, P.L. Lions and C. Sbordone, *Sobolev imbedding theorems in borderline cases*. Proc. Amer. Math. Soc. **124**(1996), 561–565.

[GiTu] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*. 2nd Edition. Revised 3rd Printing. Springer-Verlag, Berlin-Heidelberg-New York, 1998.

[Mo] J. Moser, *A sharp form of an inequality by N. Trudinger*. Indiana Univ. Math. J. **20**(1971), 1077–1092.

[Tru] N. S. Trudinger, *On imbeddings into Orlicz spaces and some applications*. J. Math. Mech. **17**(1967), 473–483.

[Str] R. S. Strichartz, *A note on Trudinger's extension of Sobolev inequality*. Indiana Univ. Math. J. **21**(1972), 841–842.

[Zi] William Ziemer, *Weakly Differentiable Functions*. Springer-Verlag, New York Inc., 1989.

On Sobolev spaces

Ritva Hurri-Syrjänen

Contents

1 Introduction **85**

2 The spaces $\mathcal{C}^k(G)$ and $\mathcal{C}_0^k(G)$ **86**

3 Partial differentiation **87**

4 Multi-indices **88**

5 Weak derivatives **88**

6 Sobolev spaces $W^{m,p}(G)$ **89**

7 Approximation by smooth functions **90**

8 Partition of unity **91**

9 The space $H^{m,p}(G)$ **91**

10 One version of the Poincaré inequality **93**

11 Sobolev spaces in $\mathbb{R}$ **95**

12 ACL-property **95**

13 Corollaries **96**

14 Embedding theorems **97**

15 Embedding theorems in the space $W_0^{m,p}(D)$ **103**

16 Embedding theorems in the space $W^{m,p}(D)$ **104**

17 Potential inequality **107**

18 On Poincaré domains **109**

18.1 Conditions for Poincaré domains 113

1 Introduction

These lecture notes contain the basic information on Sobolev spaces, so that we will be able to present the classical Sobolev embedding theorems for $W_0^{1,p}(D)$-functions and in certain classes of domains D in $\mathbb{R}^n$ for $W^{1,p}(D)$-functions, $1 \leq p < n$ and $p > n$, in the summer school on Monday, the 22nd of June. The case $p = n$ is to be considered on Friday, the 26th of June. My lecture notes are based on the lectures I gave on Sobolev spaces at the University of Helsinki in 2004. When I prepared the lectures back then, I mainly followed O. Martio's lectures on the classical results of Sobolev spaces at the University of Jyväskylä from the 1980's, [M]. Chapters 2–17 partly follow notes from those lectures. Chapter 18 is based on some parts of my doctoral thesis, [Hu]. I thank P. Harjulehto for some useful comments.

For further reading, there are several excellent books as well as lecture notes on Sobolev spaces listed in the Bibliography.

2 The spaces $\mathcal{C}^k(G)$ and $\mathcal{C}_0^k(G)$

Let G be an open set in Euclidean n-space $\mathbb{R}^n$, $n \geq 1$. Let $k = 0, 1, 2, \ldots$. We write

$$\mathcal{C}^k(G) = \{u : G \to \mathbb{R} | \text{ function } u \text{ has continuous partial derivatives of the } k\text{th order in } G\}.$$

Especially,

$$\mathcal{C}^0(G) = \mathcal{C}(G) = \{u : G \to \mathbb{R} | \text{ function } u \text{ is continuous in } G\},$$
$$\mathcal{C}^1(G) = \{u \in C(G) | \text{ function } u \text{ has continuous first order partial derivatives in } G\},$$

$$\mathcal{C}^\infty(G) = \bigcap_k \mathcal{C}^k(G) = \{u : G \to \mathbb{R} | \text{ function } u \text{ has continuous partial derivatives of all orders in } G\}.$$

Spaces $\mathcal{C}^k(G)$, $k = 0, 1, \cdots$, and $\mathcal{C}^\infty(G)$ are real linear vector spaces with natural operations. The addition is done by points as well as multiplication by a real number by points; if $f \in \mathcal{C}^k(G)$ and $g \in \mathcal{C}^k(G)$, $k = 0, 1, \cdots, \infty$, and $\lambda \in \mathbb{R}$ then

$$(f + g)(x) = f(x) + g(x) \text{ for all } x \in G$$
$$(\lambda f)(x) = \lambda f(x) \text{ for all } x \in G.$$

Note that

$$\mathcal{C}^\infty(G) \subset \mathcal{C}^{k+1}(G) \subset \mathcal{C}^k(G) \subset \mathcal{C}^1(G) \subset \mathcal{C}(G)$$

for all $k = 2, \cdots$.

Support 2.1. The closure of a set A is denoted by $\overline{A}$. Let G be an open set. Let $u : G \to \mathbb{R}$ be a function. The set $\overline{\{x \in G | u(x) \neq 0\}}$ which we write $\operatorname{spt} u$ is the support of a function u. Also notation $\operatorname{supp} u$ for $\operatorname{spt} u$ is used in literature.

Note that for a function $u : G \to \mathbb{R}^n$

1. $\operatorname{spt} u \subset \overline{G}$
2. The point $x \in G$ is also in $\operatorname{spt} u$ if and only if in every neighbourhood of x there is a point y with $u(y) \neq 0$.
3. The set $\operatorname{spt} u$ is a closed set in $\mathbb{R}^n$.

As an example, define $u : \mathbb{R} \to \mathbb{R}$ such that

$$u(x) = \begin{cases} |x| - 1 & , \text{ if } x \in [-1, 1], \\ 0 & , \text{ otherwise.} \end{cases}$$

Then $\{x \in G | u(x) \neq 0\} =]-1, 1[$ and $\operatorname{spt} u = [-1, 1]$.

If $u : G \to \mathbb{R}$ and $k = 0, 1, 2, \ldots, \infty$, we write

$$\mathcal{C}_0^k(G) = \{u \in \mathcal{C}^k(G) | \operatorname{spt} u \text{ is compact in } G\}.$$

The spaces $\mathcal{C}_0^k(G)$ are linear subspaces of $\mathcal{C}^k(G)$. Note that

$$\begin{array}{ccccccc} \mathcal{C}^0(G) & \supset & \mathcal{C}^1(G) & \supset & \mathcal{C}^2(G) \supset \cdots & \supset & \mathcal{C}^\infty(G) \\ \cup & & \cup & & \cup & & \cup \\ \mathcal{C}_0^0(G) & \supset & \mathcal{C}_0^1(G) & \supset & \mathcal{C}_0^2(G) \supset \cdots & \supset & \mathcal{C}_0^\infty(G) \end{array}$$

For our previous example function $u : \mathbb{R} \to \mathbb{R}$,

$$u(x) = \begin{cases} |x| - 1 & , \text{ if } x \in [-1, 1], \\ 0 & , \text{ otherwise,} \end{cases}$$

we have $u \in \mathcal{C}_0(\mathbb{R})$.

Remark 2.2. Let $i = 1, 2, \ldots, n$. If $u \in \mathcal{C}^k(G)$, then $D_i u$ is the partial derivative of u with respect to x_i. The mapping

$$D_i : \mathcal{C}^k(G) \to \mathcal{C}^{k-1}(G)$$

is a linear mapping with the properties

1. if $u \in \mathcal{C}^k(G)$ then $D_i u \in \mathcal{C}^{k-1}(G)$,

2. if $u \in \mathcal{C}_0^k(G)$ then $D_i u \in \mathcal{C}_0^{k-1}(G)$.

3 Partial differentiation

Lemma 3.1 (Partial differentiation in $\mathbb{R}$)**.** *Assume that Δ is an open subset of $\mathbb{R}$, $u \in \mathcal{C}^1(\Delta)$, and $v \in \mathcal{C}_0^1(\Delta)$. Then*

$$\int_\Delta u(t)v'(t)\,dt = -\int_\Delta u'(t)v(t)\,dt.$$

Lemma 3.2. *Let G be a domain in $\mathbb{R}^n$ that is G is an open and connected set in $\mathbb{R}^n$. If $u \in \mathcal{C}^1(G)$ and $v \in \mathcal{C}_0^1(G)$, then*

$$\int_G uD_i v\,dm = -\int_G vD_i u\,dm, \quad i = 1, \ldots, n.$$

Proof. Functions $uD_i v$ and $vD_i u$ are continuous functions in domain G and they differ from zero only in a compact set in G. Hence, the integrals $\int_G uD_i v\,dm$ and $\int_G vD_i u\,dm$ are well defined.

Let us set $u(x)D_i v(x) = 0$ and $v(x)D_i u(x) = 0$ for all x in $\mathbb{R}^n \setminus G$. Then we can integrate the integrals over $\mathbb{R}^n$,

$$\begin{aligned} \int_G uD_i v\,dm &= \int_{\mathbb{R}^n} uD_i v\,dm \\ &= \int_{\mathbb{R}^{n-1}} \left(\int_{\mathbb{R}} uD_i v\,dm_1 \right) dm_{n-1}. \end{aligned}$$

Let us fix $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in \mathbb{R}^{n-1}$. Let us denote

$$\tilde{u}(t) = u(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n), \quad t \in \mathbb{R},$$

and

$$\tilde{v}(t) = v(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n), \quad t \in \mathbb{R};$$

here $\tilde{u} \in \mathcal{C}^1(\Delta)$ and $\tilde{v} \in \mathcal{C}_0^1(\Delta)$.

Since Δ is an open set in $\mathbb{R}$,

$$\begin{aligned} \int_{\mathbb{R}} uD_i v\,dm_1 &= \int_\Delta \tilde{u}\tilde{v}'\,dt \\ &= -\int_\Delta \tilde{u}'\tilde{v}\,dt \\ &= -\int_{\mathbb{R}} vD_i u\,dm_1. \end{aligned}$$

The claim follows from Fubini's theorem. □

4 Multi-indices

Let $n = 1, 2, \dots$ and $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$ where $\mathbb{N} = \{0, 1, 2, \dots\}$. We call α as a multi-index. Its degree or norm is

$$|\alpha| = \sum_{i=1}^{n} \alpha_i, \quad |\alpha| \geq 0.$$

The αth partial derivative is

$$\mathrm{D}^\alpha = D_1^{\alpha_1} D_2^{\alpha_2} \cdot \cdots \cdot D_n^{\alpha_n},$$

where

$$\mathrm{D}_j^{\alpha_j} = D_j D_j \cdot \cdots \cdot D_j = \frac{\partial^{\alpha_j}}{(\partial x_j)^{\alpha_j}}, j = 1, \cdots, n.$$

Especially, D_j^0 of the function is the function itself.

If $u \in \mathcal{C}^{|\alpha|}(G)$, then $\mathrm{D}^\alpha u$ exists and

$$\begin{aligned} \mathrm{D}^\alpha u &= \frac{\partial^{\alpha_1} \partial^{\alpha_2} \dots \partial^{\alpha_n}}{(\partial x_1)^{\alpha_1} (\partial x_2)^{\alpha_2} \dots (\partial x_n)^{\alpha_n}} u \\ &= D_1^{\alpha_1} D_2^{\alpha_2} \cdot \cdots \cdot D_n^{\alpha_n} u. \end{aligned}$$

Especially, if $\alpha = (0, 0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{N}^n$, then $|\alpha| = 1$ and

$$\mathrm{D}^\alpha u = D_j u.$$

Lemma 4.1 (General partial derivatives rule in $\mathbb{R}^n$)**.** *Assume that $u \in \mathcal{C}^{|\alpha|}(G)$ and $v \in \mathcal{C}_0^{|\alpha|}(G)$. Then*

$$\int_G u \, \mathrm{D}^\alpha v \, dm = (-1)^{|\alpha|} \int_G v \, \mathrm{D}^\alpha u \, dm. \tag{1}$$

Proof. 1. If $|\alpha| = 0$ that is $\alpha = (0, \cdots, 0)$, then we have an identity.

2. If $|\alpha| = 1$, then the claim is Lemma 3.2.

3. If $|\alpha| > 1$, then the claim follows from Lemma 3.2 by induction.

□

5 Weak derivatives

Let G be an open set in $\mathbb{R}^n$, let u be in $L^1_{\mathrm{loc}}(G)$ and let α be a multi-index with $\alpha \in \mathbb{N}^n$. The function $v \in L^1_{\mathrm{loc}}(G)$ is called the α^{th} weak derivative of the function u, if

$$\int_G v\varphi \, dm = (-1)^{|\alpha|} \int_G u \, \mathrm{D}^\alpha \varphi \, dm, \quad \forall \varphi \in \mathcal{C}_0^{|\alpha|}(G). \tag{2}$$

Let us denote $v = \mathrm{D}^\alpha u$. The function $\mathrm{D}^\alpha u$ is called also the Sobolev derivative or the generalized partial derivative of u.

Remarks 5.1.

1. The function $\mathrm{D}^\alpha u$ is uniquely determined up to a set of measure zero.

2. If $u \in \mathcal{C}^{|\alpha|}(G)$, then $\mathrm{D}^\beta u \in L^p_{\mathrm{loc}}(G)$, for all β, $|\beta| \leq |\alpha|$ with all $p \in [1, \infty]$. The function $\mathrm{D}^\beta u$ is the β^{th} weak derivative, since by Lemma 4.1

$$\int_G u \, \mathrm{D}^\beta \varphi \, dm = (-1)^{|\beta|} \int_G \mathrm{D}^\beta u \varphi \, dm, \quad \forall \varphi \in C_0^\infty(G).$$

3. The integrals in (2) exist and are in $\mathbb{R}$.

4. It is possible to prove that if $u \in L^p_{\mathrm{loc}}(G)$, $1 \leq p \leq \infty$, and $\int u\varphi \, dm = 0$ with all $\varphi \in \mathcal{C}_0^{|\alpha|}(G)$, then $u(x) = 0$ almost every $x \in G$.

6 Sobolev spaces $W^{m,p}(G)$

Definition 6.1. The Sobolev space $W^{m,p}(G)$ consists of all $L^p(G)$-functions u such that functions $\mathrm{D}^\alpha u$, $|\alpha| \le m$, $m \in \mathbb{N}$, exist and are also in $L^p(G)$. That is

$$W^{m,p}(G) = L^p(G) \cap \{u | \, \mathrm{D}^\alpha u \in L^p(G), |\alpha| \le m\},$$

where $p \ge 1$ and $m \in \mathbb{N} = \{0, 1, 2, \dots\}$.

Remark 6.2. The notation $W_p^m(G)$ and $H_p^m(G)$ are used also in literature. Especially, $W^{0,p}(G) = L^p(G)$.

Example 6.3.

1. Let $u : \mathbb{R} \to \mathbb{R}$ be a function such that $u(x) = |x|$. Then, $u \in L^p_{\mathrm{loc}}(\mathbb{R})$ for all $p \in [1, \infty]$.
   It is clear that $u \notin \mathcal{C}^1(\mathbb{R})$. But the function u has the generalized first derivative $\mathrm{D}^\alpha u$, $|\alpha| = 1$. Let us set
   $$v(x) = \begin{cases} +1 & , \quad x \ge 0 \\ -1 & , \quad x < 0. \end{cases}$$
   Then $v \in L^p_{\mathrm{loc}}(\mathbb{R})$. Let $\varphi \in \mathcal{C}_0^\infty(\mathbb{R})$. If we show that
   $$\int_{\mathbb{R}} u\varphi' \, dm = -\int_{\mathbb{R}} v\varphi \, dm,$$
   then $v = \mathrm{D}^\alpha u$, $|\alpha| = 1 \in \mathbb{N}$.
   Let us choose $M > 0$ such that $\operatorname{spt}\varphi \subset [-M, M]$. Then
   $$\begin{aligned} \int_{\mathbb{R}} u\varphi' \, dm &= \int_{-M}^{M} u\varphi' \, dt \\ &= \int_{-M}^{0} u\varphi' \, dt + \int_{0}^{M} u\varphi' \, dt \\ &= \Big|_{-M}^{0} u\varphi - \int_{-M}^{0} u'\varphi \, dt + \Big|_{0}^{M} u\varphi - \int_{0}^{M} u'\varphi \, dt \\ &= -\int_{-M}^{0} u'\varphi \, dt - \int_{0}^{M} u'\varphi \, dt \\ &= -\int_{-M}^{M} v\varphi \, dt \\ &= -\int_{\mathbb{R}} v\varphi \, dm. \end{aligned}$$

2. If $u \in \mathcal{C}^m|G|$, then $u \in \mathcal{C}^l(G)$ whenever $0 \le l \le m$. But a similar result does not hold for generalized derivatives.

Remark 6.4. The generalized derivative $\mathrm{D}^\alpha u \in L^p_{\mathrm{loc}}(G)$ of function $u \in L^p_{\mathrm{loc}}(G)$ is defined modulo the equivalence relation:

$$f \sim g \text{ if and only if } (f - g)(x) = 0 \text{ for almost every } x \in G.$$

Theorem 6.5.

1. *The Sobolev space $W^{m,p}(G)$ is a linear subspace of $L^p(G)$.*
2. *The Sobolev space $W^{m,p}(G)$ is a Banach space with a norm*
   $$\|u\|_{W^{m,p}(G)} = \sum_{|\alpha| \le m} \|\mathrm{D}^\alpha u\|_{L^p(G)}.$$

Theorem 6.6. 1. *The space $W^{m+l,p}(G)$, where $l \in \mathbb{N}$, is a linear subspace of $W^{m,p}(G)$.*

2. *If the Lebesgue measure of G, which we denote by $|G|$, is finite, then $W^{m,q}(G)$ is a linear subspace of $W^{m,p}(G)$, whenever $1 \le p \le q \le \infty$.*

Proof. 1. It follows from the definition.

2. Recall from the theory of L^p-spaces: $L^q(G) \subset L^p(G)$, whenever $1 \leq p \leq q < \infty$.

□

Remark 6.7. The space $W^{m,2}(G)$ is a Hilbert space that is a Banach space with a norm which the inner product

$$(u|v) = \sum_{|\alpha| \leq m} \int_G \mathrm{D}^\alpha u \, \mathrm{D}^\alpha v \, dm.$$

introduces. This norm is equivalent to the classical norm of $W^{m,2}(G)$,

$$\frac{1}{C}\|u\|_{W^{m,2}(G)} \leq (u,v)^{\frac{1}{2}} \leq C\|u\|_{W^{m,2}(G)}, \quad \forall u \in W^{m,2}(G).$$

7 Approximation by smooth functions

Mollifier/Regularizer 7.1. *Let $\varphi \in \mathcal{C}_0^\infty(\mathbb{R}^n)$ be a function with the properties*

1. *$\varphi(x) \geq 0$ for all $x \in \mathbb{R}^n$,*
2. *$\operatorname{spt}\varphi \subset \overline{\mathrm{B}(0,1)}$,*
3. *$\int_{\mathbb{R}^n} \varphi(x)\, dm(x) = 1$.*

A classical example is

$$\varphi(x) = \begin{cases} c\exp(-\frac{1}{1-|x|^2}) & , \quad |x| < 1 \\ 0 & , \quad |x| \geq 1 \end{cases}$$

where the constant $c > 0$ is chosen such that

$$\int_{\mathrm{B}^n(0,1)} \exp\left(-\frac{1}{1-|x|^2}\right) dm(x) = c^{-1} > 0\,.$$

Let $\epsilon > 0$. If we set

$$\varphi_\epsilon(x) = \epsilon^n \varphi(\epsilon x)\,, \tag{3}$$

then

1. *$\varphi_\epsilon \in \mathcal{C}_0^\infty(\mathbb{R}^n)$,*
2. *$\varphi_\epsilon(x) \geq 0$ for all $x \in \mathbb{R}^n$,*
3. *$\operatorname{spt}\varphi_\epsilon \subset \overline{\mathrm{B}(0,\epsilon^{-1})}$,*
4. *$\int_{\mathbb{R}^n} \varphi_\epsilon(x)\, dm(x) = 1$.*

The function φ_ϵ is called a mollifier (or regularizer).

Convolution 7.2. *Let φ_ϵ be a mollifier and let f be a function in $L^1_{\mathrm{loc}}(\mathbb{R}^n)$. Then*

$$(f * \varphi_\epsilon)(x) = \int_{\mathbb{R}^n} f(y)\varphi_\epsilon(x-y)\, dm(y) \tag{4}$$

is the convolution of f and φ_ϵ.

Lemma 7.3. *The convolution of a function $f \in L^1_{\mathrm{loc}}(\mathbb{R}^n)$ and a mollifier φ_ϵ, $f * \varphi_\epsilon$, has the properties*

1. *$f * \varphi_\epsilon : \mathbb{R}^n \to \mathbb{R}$,*
2. *$f * \varphi_\epsilon \in \mathcal{C}^\infty(\mathbb{R}^n)$,*
3. *$\mathrm{D}^\alpha(f * \varphi_\epsilon) = f * \mathrm{D}^\alpha\varphi_\epsilon$ with every multi-index α.*

The following lemma is important.

Lemma 7.4. *Let $f \in L^1_{\mathrm{loc}}(\mathbb{R}^n)$. If $\operatorname{spt} f$ is a compact set in an open set $G \subset \mathbb{R}^n$, then $f * \varphi_j \in \mathcal{C}_0^\infty(G)$ whenever $0 < j^{-1} < \operatorname{dist}(\operatorname{spt} f, \partial G)$.*

Approximation 7.5. *Assume that $f \in L^p(G)$ with some $p \in [1,\infty)$. Then there exists a sequence (f_j) of $\mathcal{C}^\infty(\mathbb{R}^n)$-functions $f_j = f * \varphi_j$ such that $f_j \in L^p(G)$ and $f_j \to f$ in the space $L^p(G)$ when $j \to \infty$; this means in the L^p -sense $||f_j - f||_{L^p} \to 0$, when $j \to \infty$.*

8 Partition of unity

Partition of unity 8.1. *Let A be a subset in $\mathbb{R}^n$. Let $\mathcal{U}$ be an open cover of the set A with open sets U that is $\mathcal{U} = \{U\}$. Then there exists a family of functions $\Psi = \{\psi\}$, with $\psi \in \mathcal{C}_0^\infty(\mathbb{R}^n)$, such that*

1. *$0 \leq \psi(x) \leq 1$ for all $x \in \mathbb{R}^n$.*
2. *If K is a compact subset of A, then $\psi|_K \neq 0$ only with a finite number of functions $\psi \in \Psi$.*
3. *For every function $\psi \in \Psi$ there exists $U \in \mathcal{U}$ such that $\operatorname{spt}\psi \subset U$.*
4. *$\sum_{\psi\in\Psi} \psi(x) = 1$ for all $x \in A$.*

9 The space $H^{m,p}(G)$

There exits a natural characterization for the Sobolev space $W^{m,p}(G)$. Let us assume that G is an open subset of $\mathbb{R}^n$ and $p \in [1,\infty)$ and $m \in \mathbb{N} = \{0,1,2,\dots\}$.

Definition 9.1. Function $u \in L^p(G)$ belongs to the space $H^{m,p}(G)$, if there exists a sequence (φ_i) of $\mathcal{C}^\infty(G)$-functions φ_i such that

1. $\|\varphi_i\|_{W^{m,p}(G)} < \infty$ for all i,
2. $\varphi_i \to u$ in the space $L^p(G)$ when $i \to \infty$,
3. $(\mathrm{D}^\alpha \varphi_i)$ is a Cauchy sequence in the space $L^p(G)$ when $|\alpha| \leq m$.

By this definition $H^{0,p} = L^p(G)$, since every function $u \in L^p(G)$ should have the following property, for an arbitrary function $u \in L^p(G)$ there exists $\varphi_i \in \mathcal{C}^\infty(G)$ such that $\|\varphi_i\|_{L^p(G)} < \infty$ and $\varphi_i \to u$ in the space $L^p(G)$ when $i \to \infty$. On the other hand, $W^{0,p}(G) = L^p(G)$ as well.

The fact $W^{m,p}(G) = H^{m,p}(G)$ when $m = 0,1,\dots$ and $p \in [1,\infty)$ was proved by Meyers and Serrin in 1960's, [MeSe]. The inclusion $H^{m,p}(G) \subset W^{m,p}(G)$ is easy to prove. The inclusion $W^{m,p}(G) \subset H^{m,p}(G)$ requires some work. It is possible to prove the inclusion by using mollifiers and the partition of unity.

Warning 9.2. If $u \in W^{m,p}(G)$ and we set $u(x) = 0$ for $x \in \mathbb{R}^n \backslash G$, then $u \in L^p(\mathbb{R}^n)$. However, note that it does not hold always that $u \in W^{m,p}(\mathbb{R}^n)$.

The space $H_0^{m,p}(G)$ 9.3. A function $u \in L^p(G)$ belongs to the space $H_0^{m,p}(G)$, if there exists a sequence of functions (φ_i) such that

1. $\varphi_i \in \mathcal{C}_0^\infty(G)$ for every i,
2. $\|\varphi_i\|_{W^{m,p}(G)} < \infty$ for every i,
3. $\varphi_i \to u$ in the space $L^p(G)$ whenever $i \to \infty$,
4. $(\mathrm{D}^\alpha \varphi_i)$ is a Cauchy sequence in the space $L^p(G)$ when $|\alpha| \leq m$.

Theorem 9.4.

1. *The space $H_0^{m,p}(G)$ is a linear subspace of the space $W^{m,p}(G)$.*
2. *A function $u \in W^{m,p}(G)$ belongs to the space $H_0^{m,p}(G)$, if and only if there exists a sequence of functions (φ_i), where $\varphi_i \in \mathcal{C}_0^\infty(G)$, such that*

$$\varphi_i \to u \text{ in the space } W^{m,p}(G) \text{ when } i \to \infty.$$

3. *The space $H_0^{m,p}(G)$ is closed in the space $W^{m,p}(G)$.*

Remarks 9.5.

1. We write
$$W_0^{m,p}(G) = H_0^{m,p}(G).$$
2. $W^{0,p}(G) = L^p(G) = W_0^{0,p}(G)$.
3. $W_0^{m,p}(\mathbb{R}^n) = W^{m,p}(\mathbb{R}^n)$.

4. If $m \geq 1$, then $u \in W_0^{m,p}(G)$ means that $u \approx 0$ near the boundary of the set G.

Examples 9.6.

1. If $\varphi \in \mathcal{C}_0^\infty(G)$, then $\varphi \in W_0^{m,p}(G)$.
2. There exist functions u such that $u \in \left(W_0^{1,p}(G) \cap \mathcal{C}^\infty(G)\right)$, but $u \notin \mathcal{C}_0(G)$. Hence, $\operatorname{spt} u \not\subset G$. Let us define $u : (0,\pi) \to \mathbb{R}$ such that

$$u(x) = \sin x.$$

Then $u \in \left(W_0^{1,p}((0,\pi)) \cap \mathcal{C}^\infty((0,\pi))\right)$, when $p \in [1,\infty)$ and $\operatorname{spt} u = \overline{\{x \in (0,\pi) | u \neq 0\}} = [0,\pi]$. But $u \notin \mathcal{C}_0(G)$, since $[0,\pi] \not\subset (0,\pi)$. The reason is that we can approximate functions belonging to $W_0^{m,p}(G)$ with functions with compact support although functions in $W_0^{m,p}(G)$ are not necessarily functions with compact support.

Example 9.7. Let $u(x) = \sin x$ and $G = (0,1)$. We show that $u \in W_0^{1,1}(G)$. We have to construct a sequence (φ_i), where $\varphi \in \mathcal{C}_0^\infty(G)$, such that $\varphi \to u$ in $W^{1,1}(G)$.

Set $G_i = [2^{-i}, \pi - 2^{-i}]$ and let χ_{G_i} be the characteristic function of G_i that is

$$\chi_{G_i}(x) = \begin{cases} 1 & , \text{ when } x \in G_i \\ 0 & , \text{ when } x \notin G_i. \end{cases}$$

Let us define a $\mathcal{C}^\infty$-function, which is very near the characteristic function when $j = j_i = 4 \cdot 2^i$. Let φ_i be a mollifier (we refer to (3)). Then,

$$\psi_i = \chi_{G_i} * \varphi_j \in \mathcal{C}^\infty(\mathbb{R})$$

and

$$\psi_i(x) = \int_{\mathbb{R}} \chi_{G_i}(y)\varphi_j(x-y)\, dm(y)$$

and

$$\begin{aligned} \psi_i'(x) &= (\chi_{G_i} * \varphi_j)'(x) \\ &= \left(\chi_{G_i} * \varphi_j'\right)(x) \\ &= \int_{\mathbb{R}} \chi_{G_i}(y)\varphi_j'(x-y)\, dm(y) \\ &= \int_{\mathbb{R}} \chi_{G_i}(x-z)\varphi_j'(z)\, dm(z). \end{aligned}$$

Let us define a new mollifier as in (3) with $\epsilon = j$

$$\varphi_j(x) = j\varphi(jx)$$

and hence

$$\varphi_j'(x) = j^2\varphi'(jx).$$

Thus,

$$|\varphi_j'(x)| \leq M \cdot j^2 = M \cdot 2^{2i}$$

for all $x \in \mathbb{R}$; here $M = \sup_{x\in\mathbb{R}} |\varphi'(x)|$. Hence,

$$|\psi_i(x)| \leq M_1 \cdot 2^{2i} \cdot \frac{2}{j} = M_2 \cdot 2^i.$$

Further,

$$\psi_i'(x) = 0, \text{ kun } x \in G_{i-1}.$$

Let us define

$$\varphi_i = \psi_i u,$$

so that now $\varphi_i \in \mathcal{C}_0^\infty(G)$. We have obtained

$$\begin{aligned} \|\varphi_i - u\|_{L^1(G)} &= \int_{(0,\pi)} |\varphi_i - u|\, dm \\ &\leq \int_{G\setminus G_{i-1}} 1\, dm \to 0, \end{aligned}$$

when $i \to \infty$, and

$$\begin{aligned}
\|\varphi_i' - u'\|_{L^1(G)} &= \int_{(0,\pi)} |\varphi_i' - u'|\, dm \\
&= \int_{(0,\pi)} |\psi_i' u + \psi_i u' - u'|\, dm \\
&\le \int_{(0,\pi)} |\psi_i' u|\, dm + \int_{(0,\pi)} |\psi_i u' - u'|\, dm \\
&\le \int_{(0,\pi)\setminus G_{i-1}} |\psi_i' u|\, dm + \int_{(0,\pi)} |u'||\psi_i - 1|\, dm \\
&\le 2 \cdot \frac{1}{2^{i-1}} \cdot M_2 \cdot 2^i \cdot \frac{1}{2^{i-1}} + \int_{(0,\pi)} |\psi_i - 1|\, dm \\
&\le \frac{2 \cdot M_2 \cdot 2^i}{2^{2(i-1)}} + \int_{G\setminus G_{i-1}} 1\, dm \to 0,
\end{aligned}$$

when $i \to \infty$.

Thus,

$$\|\varphi_i - u\|_{W^{1,1}(G)} = \|\varphi_i - u\|_{L^1(G)} + \|\varphi_i' - u'\|_{L^1(G)} \to 0\,,$$

when $i \to \infty$.

Remark 9.8. The most important spaces of $W^{m,p}(G)$ and $W_0^{m,p}(G)$ are $W^{1,p}(G)$ and $W_0^{1,p}(G)$. Many proofs of the results considering the space $W^{m,p}(G)$ can be reduced to the results in $W^{1,p}(G)$.

10 One version of the Poincaré inequality

Assume that $u \in \mathcal{C}^1(\mathbb{R}^n)$. Then

$$\begin{aligned}
Du(x_0)h &= \nabla u(x_0) \cdot h \\
&= \partial_1 u(x_0)h_1 + \partial_2 u(x_0)h_2 + \cdots + \partial_n u(x_0)h_n,
\end{aligned}$$

where $\nabla u(x_0) = (\partial_1 u(x_0), \partial_2 u(x_0), \ldots, \partial_n u(x_0))$ is the gradient of u at x_0 and $h = (h_1, \cdots, h_n)$ is a vector in $\mathbb{R}^n$.

Lemma 10.1. *Let $u \in \mathcal{C}^1(\mathbb{R}^n)$ and $h \in \mathbb{R}^n$. Then*

$$u(x+h) - u(x) = \int_0^1 \nabla u(x+th) \cdot h\, dt.$$

Lemma 10.2. *If $u \in \mathcal{C}_0^1(\mathbb{R}^n)$ and $h \in \mathbb{R}^n$, then*

$$\int_{\mathbb{R}^n} |u(x+h) - u(x)|^p\, dm(x) \le n^{\frac{3p}{2}-1}|h|^p \sum_{i=1}^n \int_{\mathbb{R}^n} |\partial_i u(x)|^p\, dm(x).$$

Proof. Let $p > 1$. By Lemma 10.1, Hölder's inequality with $\left(p, \frac{p}{p-1}\right)$, and Fubini's theorem

$$\begin{aligned}
&\int_{\mathbb{R}^n} |u(x+h) - u(x)|^p \, dm(x) \\
&= \int_{\mathbb{R}^n} \left| \int_0^1 \nabla u(x+th) \cdot h \, dt \right|^p \, dm(x) \\
&\leq \int_{\mathbb{R}^n} \left(\int_0^1 |\nabla u(x+th) \cdot h| \, dt \right)^p \, dm(x) \\
&\leq \int_{\mathbb{R}^n} \left(\int_0^1 |\nabla u(x+th)||h| \, dt \right)^p \, dm(x) \\
&\leq |h|^p \int_{\mathbb{R}^n} \left(\int_0^1 |\nabla u(x+th)| \, dt \right)^p \, dm(x) \\
&\leq |h|^p \int_{\mathbb{R}^n} \left(\int_0^1 |\nabla u(x+th)|^p \, dt \right) \, dm(x) \\
&\leq |h|^p \int_0^1 \int_{\mathbb{R}^n} |\nabla u(x+th)|^p \, dm(x) \, dt \\
&= |h|^p \int_{\mathbb{R}^n} |\nabla u|^p \, dm.
\end{aligned}$$

By setting $|\partial_j u(y)|^2 = \max_{i=1,\ldots,n} |\partial_i u(y)|^2$ and by estimating we obtain

$$\begin{aligned}
|\nabla u(y)| &= \left(\sum_{i=1}^n \partial_i u(y)^2 \right)^{1/2} \\
&\leq \left(n \partial_j u(y)^2 \right)^{1/2} \\
&\leq n^{1/2} |\partial_j u(y)| \\
&\leq n^{1/2} \sum_{i=1}^n |\partial_i u(y)| \, .
\end{aligned}$$

Hence,

$$|\nabla u(y)|^p \leq n^{p/2} n^{p-1} \sum_{i=1}^n |\partial_i u(y)|^p.$$

Thus,

$$\int_{\mathbb{R}^n} |u(x+h) - u(x)|^p \, dm(x) \leq n^{\frac{3p}{2}-1} |h|^p \sum_{i=1}^n \int_{\mathbb{R}^n} |\partial_i u(y)|^p \, dm.$$

If $p = 1$, the claim is clear. □

Theorem 10.3. *Assume that $u \in W_0^{1,p}(G)$ and $\operatorname{diam}(G) < \infty$. Then,*

$$\|u\|_{L^p(G)} \leq c(n) \operatorname{diam}(G) \sum_{i=1}^n \|\partial_i u\|_{L^p(G)},$$

where

$$\partial_1 u = D^{(1,0,\ldots,0)} u, \ldots, \partial_n u = D^{(0,0,\ldots,0,1)} u$$

are the generalized first derivatives of u and the constant $c(n) < \infty$ depends only on the dimension n.

Remarks 10.4.

1. If G an open bounded set, it is possible to use in $W_0^{1,p}(G)$ the norm
   (a) $u \mapsto \|u\|_{W^{1,p}(G)} = \|u\|_{L^p(G)} + \sum_{i=1}^n \|\partial_i u\|_{L^p(G)}$
   (b) $u \mapsto \sum_{i=1}^n \|\partial_i u\|_{L^p(G)}$,

   because by Theorem 10.3 these are equivalent.

2. Note that $\sum_{i=1}^n \|\partial_i u\|_{L^p(G)}$ is not a norm in $W^{1,p}(G)$. If u is a non-zero constant then $\partial_i u = 0$ with all $i = 1, \ldots, n$, but $\|u\|_{L^p(G)} \neq 0$. This gives just a seminorm which does not have the condition: if $\|u\| = 0$ then $u = 0$.

11 Sobolev spaces in $\mathbb{R}$

Definition 11.1. Recall that a function $f : [a,b] \to \mathbb{R}$ is absolutely continuous on the interval $[a,b]$, if and only if for every $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $\sum_{i=1}^{k}(b_i - a_i) < \delta$ then

$$\sum_{i=1}^{k} |f(b_i) - f(a_i)| < \varepsilon,$$

where the intervals (a_i, b_i) are disjoint intervals of $[a,b]$.

Remember 11.2.

1. An absolutely continuous function $f : [a,b] \to \mathbb{R}$ is continuous.
2. If the function $h : [a,b] \to \mathbb{R}$ is Lebesque integrable with respect to the Lebesque measure m, then the function
$$f(x) = k + \int_{[a,x]} h(y)\, dm(y)$$
is absolutely continuous, $f(a) = k$, and $f'(x) = h(x)$ for almost every $x \in [a,b]$.

Theorem 11.3. *Let G be an open set in $\mathbb{R}$ and let $1 \le p \le \infty$. If $u \in W^{1,p}(G)$, then there exists a function $g : G \to \mathbb{R}$ such that*

1. *the function g is absolutely continuous on every closed interval $[a,b] \subset G$,*
2. *$(u-g)(x) = 0$ for almost every $x \in G$,*
3. *$\mathrm{D}\, u(x) = g'(x)$ for almost every $x \in G$.*

On the other hand,

Theorem 11.4. *Let G be an open set in $\mathbb{R}$ and let $1 \le p \le \infty$. If a function $g : G \to \mathbb{R}$ is absolutely continuous on every interval $[a,b] \subset G$, and $g \in L^p(G)$, and $g' \in L^p(G)$, then*

$$g \in W^{1,p}(G).$$

There is a local version of Theorem 11.3

Theorem 11.5. *Let $G \subset \mathbb{R}$ be an open set and $u \in L^1_{\mathrm{loc}}(G)$. If the function u has the weak derivative $\mathrm{D}\, u \in L^1_{\mathrm{loc}}(G)$ then there exists a function $g : G \to \mathbb{R}$ such that*

1. *the function g is absolutely continuous on every interval $[a,b] \subset G$.*
2. *$u = g$ almost everywhere in G.*
3. *$\mathrm{D}\, u = g'$ almost everywhere in G.*

Remark 11.6. Let G be an open set in $\mathbb{R}$. Then by Theorem 11.3

$$W^{1,p}(G) = \mathrm{AC}([a,b]),$$

where $[a,b] \subset G$. Thus, functions $u : G \to \mathbb{R}$ which have weak derivatives and functions $v : G \to \mathbb{R}$ which are absolutely continuous on every closed interval in G form the same function class.

12 *ACL*-property

Let G be an open set in $\mathbb{R}^n$. Function $u : G \to \mathbb{R}$ is ACL, absolutely continuous on lines, if for almost every line L which is parallel to a co-ordinate axis the restriction $u|_L$ is absolutely continuous on every closed interval $J \subset (G \cap L)$.

Remark 12.1. Let L be a straight line which is perpendicular to the co-ordinate plane $\mathbb{R}^{n-1}$. Let x_L be the point where the line L meets the plane $\mathbb{R}^{n-1}$. The property P holds on almost every line L, if

$$m_{n-1}\left(\{x_L \in \mathbb{R}^{n-1} |\ P \text{ does not hold on } L\}\right) = 0.$$

Example 12.2. Let $G = \mathbb{R}^2$ and

$$u(x_1, x_2) = \begin{cases} x^2, & \text{if } x \neq 0 \\ 1, & \text{if } x = 0. \end{cases}$$

The function u is an ACL-function in $\mathbb{R}^2$, since on every line which is parallel to x_1-axis or x_2-axis but which does not go through the origin the function u is absolutely continuous on every closed interval, since the function u is continuously differentiable in $\mathbb{R}^2 \setminus \{0\}$.

Remark 12.3. Let $G \subset \mathbb{R}^n$. If the function $u : G \to \mathbb{R}$ is ACL, then the function u has the ordinary partial differential derivatives $\partial_i u$, $i = 1, \ldots, n$, almost everywhere in G.

The following two theorems correspond to Theorem 11.3 and Theorem 11.4 in $\mathbb{R}^n$, $n \geq 2$:

Theorem 12.4. *Let G be an open set in $\mathbb{R}^n$ and let $1 \leq p < \infty$. If $u \in W^{1,p}(G)$, then the exists a function $g : G \to \mathbb{R}$ such that*

1. *g is* ACL.
2. *$u = g$ almost everywhere in G.*
3. *$\partial_i u = \partial_i g$, $i = 1, \ldots, n$, almost everywhere in G, where $\partial_i u$ are the weak derivatives of u and $\partial_i g$ the ordinary partial derivatives.*

Theorem 12.5. *Let $G \subset \mathbb{R}^n$ be an open set and let $1 \leq p < \infty$. If $g : G \to \mathbb{R}$ is an* ACL*-function and $g \in L^p(G)$ and $\partial_i g \in L^p(G)$ for every $i = 1, \cdots, n,$, then*

$$g \in W^{1,p}(G).$$

The local versions of Theorem 12.4 and Theorem 12.5:

Theorem 12.6. *Let $G \subset \mathbb{R}^n$ be an open set and $u \in L^1_{\mathrm{loc}}(G)$. If the function u has weak derivatives $\partial_i u$, $i = 1, \ldots, n$, then there exists a function $g : G \to \mathbb{R}$ such that*

1. *the function g is* ACL.
2. *$u = g$ almost everywhere in G.*
3. *$\partial_i u = \partial_i g$ almost everywhere in G.*

Theorem 12.7. *Let G be an open set in $\mathbb{R}^n$. Let $g : G \to \mathbb{R}$ be an* ACL*-function with $g \in L^1_{\mathrm{loc}}(G)$ and $g' \in L^1_{\mathrm{loc}}(G)$. Then $\partial_i g$, $i = 1, \cdots, n$, are the weak derivatives of the function g.*

13 Corollaries

Lemma 13.1. *If functions $f, g : [a, b] \to \mathbb{R}, a < b$, are absolutely continuous, then the function h,*

$$h(x) = \max\{f(x), g(x)\}$$

is absolutely continuous. Further,

$$|h'(x)| \leq |f'(x)| + |g'(x)|$$

for almost every x in $[a, b]$.

Corollary 13.2. *If $u \in W^{1,p}(G)$ and $v \in W^{1,p}(G)$, then*

$$\max(u, v) \in W^{1,p}(G)$$

and

$$\min(u, v) \in W^{1,p}(G).$$

The following corollary is important,

Corollary 13.3. *If $u \in W^{1,p}(G)$, then*

$$|u| \in W^{1,p}(G).$$

Proof. Since $|u| = \max(u, -u)$, Corollary 13.2 implies $|u| \in W^{1,p}(G)$. □

Warning 13.4. Assume that $u \in W^{m,p}(G)$ and $v \in W^{m,p}(G)$ with $m > 1$. Then it might happen that $\max(u, v) \notin W^{m,p}(G)$ and $|u| \notin W^{m,p}(G)$.

14 Embedding theorems

Generalized Hölder's inequality 14.1. Let $g_i \in L^{p_i}(G)$, $i = 1, \ldots, k$, and $\frac{1}{p_1} + \frac{1}{p_2} + \cdots + \frac{1}{p_k} = 1$. Then,

$$\left| \int_G g_1 \cdots g_k \mathrm{d}m \right| \leq \left(\int_G |g_1|^{p_1} \mathrm{d}m \right)^{1/p_1} \cdots \left(\int_G |g_k|^{p_k} \mathrm{d}m \right)^{1/p_k}.$$

Lemma 14.2 (The inequality between the geometric mean and the arithmetic mean.)**.** *Let $a_i \geq 0$. Then,*

$$\left(\prod_{i=1}^{k} a_i \right)^{\frac{1}{k}} \leq \frac{1}{k} \sum_{i=1}^{k} a_i.$$

Gagliardo-Nirenberg-Sobolev embedding theorem 14.3. *Let D be a domain in $\mathbb{R}^n$. If $u \in W_0^{1,p}(D)$, $1 \leq p < n$, $n \geq 2$, then*

$$u \in L^{\frac{np}{n-p}}(D).$$

Further, there exists a constant $c(n,p)$ such that

$$\|u\|_{L^{\frac{np}{n-p}}(D)} \leq c(n,p) \sum_{i=1}^{n} \|\partial_i u\|_{L^p(D)}.$$

Proof. It is enough to prove the claim for $u \in \mathcal{C}_0^1(\mathbb{R}^n)$, $D = \mathbb{R}^n$. (In the end of the claim's proof we show this, too.) Let us prove the claim when $p = 1$. So we have to show that

$$\|u\|_{L^{\frac{n}{n-1}}(\mathbb{R}^n)} \leq c(n) \sum_{i=1}^{n} \|\partial_i u\|_{L^1(\mathbb{R}^n)}.$$

For every $i = 1, \cdots, n$,

$$|u(x)| \leq \int_{-\infty}^{x_i} |\partial_i u| \mathrm{d}x_i \leq \int_{-\infty}^{\infty} |\partial_i u| \mathrm{d}x_i.$$

Hence,

$$|u(x)|^n \leq \prod_{i=1}^{n} \int_{-\infty}^{\infty} |\partial_i u| \mathrm{d}x_i,$$

and

$$|u(x)|^{\frac{n}{n-1}} \leq \left(\prod_{i=1}^{n} \int_{-\infty}^{\infty} |\partial_i u| \mathrm{d}x_i \right)^{\frac{1}{n-1}}.$$

We integrate with respect to the variable x_1. By the generalized Hölder inequality when $p_2 = \cdots = p_n = n-1$,

$$\begin{aligned}
&\int_{-\infty}^{\infty} |u(x)|^{\frac{n}{n-1}} \mathrm{d}x_1 \\
&\leq \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |\partial_1 u| \mathrm{d}x_1 \right)^{\frac{1}{n-1}} \cdots \left(\int_{-\infty}^{\infty} |\partial_n u| \mathrm{d}x_n \right)^{\frac{1}{n-1}} \mathrm{d}x_1 \\
&= \left(\int_{-\infty}^{\infty} |\partial_1 u| \mathrm{d}x_1 \right)^{\frac{1}{n-1}} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |\partial_2 u| \mathrm{d}x_2 \right)^{\frac{1}{n-1}} \cdots \left(\int_{-\infty}^{\infty} |\partial_n u| \mathrm{d}x_n \right)^{\frac{1}{n-1}} \mathrm{d}x_1 \\
&\leq \left(\int_{-\infty}^{\infty} |\partial_1 u| \mathrm{d}x_1 \right)^{\frac{1}{n-1}} \left(\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |\partial_2 u| \mathrm{d}x_2 \right)^{\frac{n-1}{n-1}} \mathrm{d}x_1 \right)^{\frac{1}{n-1}} \\
&\cdots \left(\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |\partial_n u| \mathrm{d}x_n \right)^{\frac{n-1}{n-1}} \mathrm{d}x_1 \right)^{\frac{1}{n-1}}.
\end{aligned}$$

Next, we integrate with respect to the x_2. By the generalized Hölder inequality, when $p_1 = p_3 = p_4 = \cdots = p_n = n-1$, we obtain

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |u(x)|^{\frac{n}{n-1}} \mathrm{d}x_1\mathrm{d}x_2 \\
&\leq \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} |\partial_1 u|\mathrm{d}x_1\right)^{\frac{1}{n-1}} \left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\partial_2 u|\mathrm{d}x_2\mathrm{d}x_1\right)^{\frac{1}{n-1}} \\
&\cdots \left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\partial_n u|\mathrm{d}x_n\mathrm{d}x_1\right)^{\frac{1}{n-1}} \mathrm{d}x_2 \\
&= \left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\partial_2 u|\mathrm{d}x_2\mathrm{d}x_1\right)^{\frac{1}{n-1}} \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} |\partial_1 u|\mathrm{d}x_1\right)^{\frac{1}{n-1}} \\
&\cdots \left(\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} |\partial_i u|\mathrm{d}x_i\right)\mathrm{d}x_1\right)^{\frac{1}{n-1}} \\
&\cdots \left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\partial_n u|\mathrm{d}x_1\mathrm{d}x_n\right)^{\frac{1}{n-1}} \mathrm{d}x_2 \\
&\leq \left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\partial_2 u|\mathrm{d}x_2\mathrm{d}x_1\right)^{\frac{1}{n-1}} \left(\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} |\partial_1 u|\mathrm{d}x_1\right)^{\frac{n-1}{n-1}} \mathrm{d}x_2\right)^{\frac{1}{n-1}} \\
&\cdots \left(\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\partial_i u|\mathrm{d}x_i\mathrm{d}x_1\right)^{\frac{n-1}{n-1}} \mathrm{d}x_2\right)^{\frac{1}{n-1}} \\
&\cdots \left(\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\partial_n u|\mathrm{d}x_n\mathrm{d}x_1\right)^{\frac{n-1}{n-1}} \mathrm{d}x_2\right)^{\frac{1}{n-1}} .
\end{aligned}$$

We continue in this way for $i = 3, \ldots, n$. Hence,

$$\begin{aligned}
&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} |u(x)|^{\frac{n}{n-1}} \mathrm{d}x_1\cdots\mathrm{d}x_n \\
&\leq \prod_{i=1}^{n}\left(\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} |\partial_i u(x)|\mathrm{d}x_1\cdots\mathrm{d}x_n\right)^{\frac{1}{n-1}} \\
&= \left(\prod_{i=1}^{n}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} |\partial_i u(x)|\mathrm{d}x_1\cdots\mathrm{d}x_n\right)^{\frac{1}{n-1}} .
\end{aligned}$$

Thus,

$$\begin{aligned}
\|u\|_{L^{\frac{n}{n-1}}(\mathbb{R}^n)} &\leq \left(\prod_{i=1}^{n}\int_{\mathbb{R}^n} |\partial_i u|\mathrm{d}m\right)^{\frac{1}{n}} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^n} |\partial_i u|\mathrm{d}m \\
&= \frac{1}{n}\sum_{i=1}^{n}\|\partial_i u\|_{L^1(\mathbb{R}^n)}.
\end{aligned}$$

This is the claim, when $p = 1$.

Let us prove the claim, when $1 < p < n$. Let $\gamma > 1$ and let $u \in \mathcal{C}_0^1(\mathbb{R}^n)$. Then $|u|^\gamma \in \mathcal{C}_0^1(\mathbb{R}^n)$. We have proved already

$$\|u\|_{L^{\frac{n}{n-1}}(\mathbb{R}^n)} \leq \sum_{i=1}^{n}\|\partial_i u\|_{L^1(\mathbb{R}^n)}.$$

We apply this inequality to the function $|u|^\gamma$. By Höder's inequality with $(p, \frac{p}{p-1}), p > 1$, we obtain

$$\begin{aligned}
\||u|^\gamma\|_{L^{\frac{n}{n-1}}(\mathbb{R}^n)} &\leq \sum_{i=1}^n \int_{\mathbb{R}^n} |\partial_i(|u|^\gamma)| \mathrm{d}m \\
&\leq \gamma \sum_{i=1}^n \int_{\mathbb{R}^n} |u|^{\gamma-1} |\partial_i u| \mathrm{d}m \\
&\leq \gamma \||u|^{\gamma-1}\|_{L^{\frac{p}{p-1}}(\mathbb{R}^n)} \sum_{i=1}^n \|\partial_i u\|_{L^p(\mathbb{R}^n)}.
\end{aligned}$$

Let us choose $\gamma = \frac{p(n-1)}{n-p} > 1$. Then $\gamma \frac{n}{n-1} = \frac{np}{n-p}$ and

$$\left(\int |u|^{\gamma \frac{n}{n-1}} \right)^{\frac{n-1}{n}} \leq \gamma \left(\int |u|^{(\gamma-1)\frac{p}{p-1}} \right)^{\frac{p-1}{p}} \sum_{i=1}^n \|\partial_i u\|_{L^p(\mathbb{R}^n)}$$

and

$$\left(\int |u|^{\frac{np}{n-p}} \right)^{\frac{n-1}{n}} \leq \gamma \left(\int |u|^{\frac{np}{n-p}} \right)^{\frac{p-1}{p}} \sum_{i=1}^n \|\partial_i u\|_{L^p(\mathbb{R}^n)}.$$

Hence,

$$\left(\int |u|^{\frac{np}{n-p}} \right)^{\frac{n-1}{n} - \frac{p-1}{p}} \leq \gamma \sum_{i=1}^n \|\partial_i u\|_{L^p(\mathbb{R}^n)}$$

which means

$$\left(\int_{\mathbb{R}^n} |u|^{\frac{np}{n-p}} \right)^{\frac{n-p}{np}} = \|u\|_{L^{\frac{np}{n-p}}(\mathbb{R}^n)} \leq \gamma \sum_{i=1}^n \|\partial_i u\|_{L^p(\mathbb{R}^n)},$$

where $\gamma = \frac{p(n-1)}{n-p} = c(n,p)$.

We have proved the claim using $\mathcal{C}_0^1$-functions u. Now we show that the claim follows from this to Sobolev functions $u \in W_0^{1,p}(D)$. Assume that $u \in W_0^{1,p}(D)$. Then there exists a function sequence (φ_j) of functions $\varphi_j \in C_0^\infty(D)$ such that $\varphi_j \to u$, when $j \to \infty$, in $W^{1,p}(D)$. By taking a subsequence we may assume that $\varphi_j \to u$, when $j \to \infty$, for almost every x in D. By Theorem 14.3 the sequence (φ_j) is a Cauchy sequence in $L^{\frac{np}{n-p}}(D)$. Then, there is a function $v \in L^{\frac{np}{n-p}}(D)$ such that $\varphi_j \to v$ in $L^{\frac{np}{n-p}}(D)$. Further, there exists a subsequence φ_{j_h} such that $\varphi_{j_h} \to v$, when $h \to \infty$, for almost every x in D. Thus, $u(x) = v(x)$ for almost every x in D. Hence,

$$\begin{aligned}
\|u\|_{L^{\frac{np}{n-p}}(D)} &= \lim_{n\to\infty} \|\varphi_j\|_{L^{\frac{np}{n-p}}(D)} \\
&\leq \lim_{n\to\infty} \sum_{i=1}^n \|\partial_i \varphi_j\|_{L^p(D)} = c \sum_{i=1}^n \|\partial_i u\|_{L^p(D)}.
\end{aligned}$$

□

Remark 14.4. If $u \in L^q(D)$, $|D| = \infty$, and $q > p \geq 1$, then it might happen that

$$u \notin L^p(D).$$

Corollary 14.5. *If $u \in W_0^{1,p}(D)$, $1 \leq p < n$, and $|D| < \infty$, then for every q, $1 \leq q \leq \frac{np}{n-p}$,*

$$\|u\|_{L^q(D)} \leq c(n,p,q,|D|) \sum_{i=1}^n \|\partial_i u\|_{L^p(D)}.$$

Proof. Assume that $1 \leq p \leq q \leq \infty$ and let $|D| < \infty$. Then by Hölder's inequality with $\left(\frac{q}{p}, \frac{\frac{q}{p}}{\frac{q}{p}-1}\right)$,

$$\begin{aligned}
\left(\int_D |u|^p \right)^{\frac{1}{p}} &\leq \left(\int_D |u|^{p\frac{q}{p}} \right)^{\frac{1}{p}\cdot\frac{p}{q}} \left(\int_D 1^{\frac{\frac{q}{p}}{\frac{q}{p}-1}} \right)^{\frac{\frac{q}{p}-1}{\frac{q}{p}}} \\
&= |D|^{\frac{q-p}{pq}} \left(\int_D |u|^q \right)^{\frac{1}{q}},
\end{aligned}$$

which is

$$\|u\|_{L^p(D)} \le |D|^{\frac{1}{p}-\frac{1}{q}} \|u\|_{L^q(D)}\,.$$

We apply this inequality and prove the claim. □

Remarks 14.6.

1. If $u \in W^{1,p}(D)$, then the claim in Theorem 14.3 does not hold unless we require extra assumptions.
2. Theorem 14.3 is known as the Gagliardo-Nierenberg-Sobolev embedding theorem. It means that the space $W_0^{1,p}(D)$ can be embedded into $L^{\frac{np}{n-p}}(D)$, when $1 \le p < n$.
3. Theorem 14.3 does not hold when $p = n$.

We need the following lemma when we prove the case $p > n$.

Lemma 14.7. *Assume that the function $u : G \to \mathbb{R}$ is measurable. Then,*

$$\operatorname*{ess\,sup}_G |u| = \|u\|_{L^\infty(G)} = \lim_{p\to\infty} \left(\frac{1}{|G|}\int_G |u|^p \mathrm{d}m\right)^{\frac{1}{p}},$$

whenever $0 < |G| < \infty$.

Proof. For almost every $x \in G$

$$|u(x)| \le \operatorname*{ess\,sup}_{x\in G} |u(x)|.$$

Thus,

$$\left(\frac{1}{|G|}\int_G |u(x)|^p \mathrm{d}m(x)\right)^{\frac{1}{p}} \le \operatorname*{ess\,sup}_{x\in G} |u(x)|,$$

and

$$\overline{\lim_{p\to\infty}} \left(\frac{1}{|G|}\int_G |u|^p \mathrm{d}m\right)^{\frac{1}{p}} \le \operatorname*{ess\,sup}_{x\in G} |u(x)|.$$

If $\operatorname{ess\,sup}_{x\in G} |u(x)| = 0$, then $u(x) = 0$ for almost every x and the claim follows. We may assume that there exists $\lambda \in \mathbb{R}$ such that $0 \le \lambda < \operatorname{ess\,sup}_{x\in G} |u(x)|$. We denote $A = \{x \in G : |u(x)| > \lambda\}$. Now,

$$\left(\frac{1}{|G|}\int_G |u|^p \mathrm{d}m\right)^{\frac{1}{p}} \ge \left(\frac{1}{|G|}\int_A \lambda^p \mathrm{d}m\right)^{\frac{1}{p}} = \lambda\left(\frac{|A|}{|G|}\right)^{\frac{1}{p}} \to \lambda,$$

when $p \to \infty$. Hence,

$$\underline{\lim}_{p\to\infty} \left(\frac{1}{|G|}\int_G |u|^p \mathrm{d}m\right)^{\frac{1}{p}} \ge \lambda \quad \text{for every } \lambda \in [0, \operatorname*{ess\,sup}_G |u|),$$

and

$$\lim_{p\to\infty} \left(\frac{1}{|G|}\int_G |u|^p \mathrm{d}m\right)^{\frac{1}{p}} \ge \operatorname*{ess\,sup}_G |u|.$$

□

Theorem 14.8. *Let D be a bounded domain in $\mathbb{R}^n$. If $u \in W_0^{1,p}(D)$, $p > n$, then there exists a function $v \in C(D)$ such that $(v-u)(x) = 0$ for almost every x in D and*

$$\sup_{x\in D} |v(x)| \le c(n,p)|D|^{\frac{1}{n}-\frac{1}{p}} \sum_{i=1}^n \|\partial u\|_{L^p(D)}.$$

Proof. Let us assume that $u \in \mathcal{C}_0^1(D)$. If $\sum_{i=1}^n \|\partial_i u\|_{L^p(D)} = 0$, then $\partial_i u = 0$ for all $i = 1, \dots, n$ in D. Thus, $u \equiv 0$, since $u \in \mathcal{C}_0^1(D)$. Hence, we may assume that $\sum_{i=1}^n \|\partial_i u\|_{L^p(D)} > 0$ and denote $\alpha = \sum_{i=1}^n \|\partial_i u\|_{L^p(D)}$. Let us define

$$\hat{u} = \frac{|u|}{\alpha}\,.$$

We may assume that $|D| = 1$. Let $\gamma > 1$. By the proof for Theorem 14.3

$$\||u|^\gamma\|_{L^{\frac{n}{n-1}}(D)} \le \gamma \||u|^{\gamma-1}\|_{L^{\frac{p}{p-1}}(D)} \sum_{i=1}^n \|\partial_i u\|_{L^p(D)}.$$

Let us multiply the above inequality by $\frac{1}{\alpha^\gamma}$,

$$\left\| \frac{|u|^\gamma}{\alpha^\gamma} \right\|_{L^{\frac{n}{n-1}}(D)} \leq \gamma \left\| \frac{|u|^{\gamma-1}}{\alpha^{\gamma-1}} \right\|_{L^{\frac{p}{p-1}}(D)}.$$

Hence,

$$\|\hat{u}^\gamma\|_{L^{\frac{n}{n-1}}(D)} \leq \gamma \|\hat{u}^{\gamma-1}\|_{L^{\frac{p}{p-1}}(D)}$$

which is

$$\left(\int |\hat{u}|^{\gamma \frac{n}{n-1}} \right)^{\frac{n-1}{n}} \leq \gamma \left(\int |\hat{u}|^{(\gamma-1)\frac{p}{p-1}} \right)^{\frac{p-1}{p}}.$$

Let us take the power $\frac{1}{\gamma}$ and apply Hölder's inequality with $(\frac{\gamma}{\gamma-1}, \gamma)$,

$$\begin{aligned}
&\left(\int_D |\hat{u}|^{\gamma \frac{n}{n-1}} \right)^{\frac{n-1}{\gamma n}} \\
&\leq \gamma^{\frac{1}{\gamma}} \left(\int_D |\hat{u}|^{(\gamma-1)\frac{p}{p-1}} \right)^{\frac{p-1}{\gamma p}} \\
&\leq \gamma^{\frac{1}{\gamma}} \left[\left(\int_D |\hat{u}|^{(\gamma-1)\frac{p}{p-1}} \right)^{\frac{p-1}{p(\gamma-1)}} \right]^{\left(1-\frac{1}{\gamma}\right)} \\
&\leq \gamma^{\frac{1}{\gamma}} \left[\left(\int_D |\hat{u}|^{\frac{(\gamma-1)p}{(p-1)} \frac{\gamma}{(\gamma-1)}} \right)^{\frac{(p-1)}{p(\gamma-1)} \frac{(\gamma-1)}{\gamma}} \left(\int_D 1^\gamma \right)^{\frac{p-1}{p(\gamma-1)}\gamma} \right]^{\left(1-\frac{1}{\gamma}\right)} \\
&= \gamma^{\frac{1}{\gamma}} |D|^{\frac{p-1}{p}} \left(\|\hat{u}\|_{L^{\frac{p\gamma}{p-1}}} \right)^{\left(1-\frac{1}{\gamma}\right)}.
\end{aligned}$$

Thus,

$$\|\hat{u}\|_{L^{\frac{\gamma n}{n-1}}(D)} \leq \gamma^{\frac{1}{\gamma}} \|\hat{u}\|^{1-\frac{1}{\gamma}}_{L^{\frac{p\gamma}{p-1}}(D)}.$$

Recall $p > n$. Let $\delta = \frac{n}{(n-1)} \frac{(p-1)}{p} > 1$. Set $\gamma = \delta^j$, where $j = 1, 2, \ldots$. Thus,

$$\|\hat{u}\|_{L^{\frac{n}{n-1}\delta^j}(D)} \leq \delta^{j\delta^{-j}} \left(\|\hat{u}\|_{L^{\frac{p}{p-1}\delta^j}(D)} \right)^{1-\delta^{-j}} = \delta^{j\delta^{-j}} \left(\|\hat{u}\|_{L^{\frac{n}{n-1}\delta^{j-1}}(D)} \right)^{1-\delta^{-j}}.$$

Let us assume for awhile that

$$\|\hat{u}\|_{L^{\frac{n}{n-1}\delta^{j-1}}(D)} \geq 1,$$

when $j \geq j_0$ and

$$\|\hat{u}\|_{L^{\frac{n}{n-1}\delta^{j-1}}(D)} \leq 1,$$

when $j < j_0$. Hence,

$$\begin{aligned}
\|\hat{u}\|_{L^{\frac{n}{n-1}\delta^j}(D)} &\leq \delta^{j\delta^{-1}} \left(\|\hat{u}\|_{L^{\frac{n}{n-1}\delta^{j-1}}(D)} \right)^{1-\delta^{-j}} \\
&\leq \delta^{j\delta^{-1}} \|\hat{u}\|_{L^{\frac{n}{n-1}\delta^{j-1}}(D)} \\
&\leq \delta^{j\delta^{-j} + (j-1)\delta^{-(j-1)}} \|\hat{u}\|_{L^{\frac{n}{n-1}\delta^{j-2}}(D)}.
\end{aligned}$$

Let us continue in this way until the number j_0. Thus,

$$\|\hat{u}\|_{L^{\frac{n}{n-1}\delta^j}(D)} \leq \delta \sum_{i=0}^{j} i\delta^{-i} \cdot 1,$$

for all $j = 0, 1, \ldots$. Write $\lambda = \sum_{i=0}^{\infty} i\delta^{-i} \in [i, \infty)$. Thus,

$$\|\hat{u}\|_{L^{\frac{n}{n-1}\delta^j}(D)} \leq \delta^\lambda.$$

By Lemma 14.7,

$$\operatorname*{ess\,sup}_D |\hat{u}| \le \delta^\lambda.$$

Because the function $\hat{u}$ is continuous, we have

$$\sup_D |\hat{u}| \le \delta^\lambda.$$

By the definition of the function $\hat{u}$

$$\sup_D |u| \le \delta^\lambda \alpha = \delta^\lambda \sum_{i=1}^n \|\partial_i u\|_{L^p(D)}.$$

So we have proved the claim of the theorem when $p > n$, $u \in \mathcal{C}_0^1(D)$, and $|D| = 1$.

Let us remove the restriction $|D| = 1$. Assume that $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear mapping such that $Tx = |D|^{\frac{1}{n}} x$. When $Tx = \lambda x$ and A is a measurable set in $\mathbb{R}^n$, then $|T(A)| = \lambda^n |A|$. Write $w(x) = u(T(x))$. Then, $w \in \mathcal{C}_0^1$. Hence,

$$\begin{aligned}
\sup_{D'} |w| &\le \lambda^\delta \sum_{i=1}^n \|\partial_i w\|_{L^p(D')} \\
&\le \lambda^\delta |D|^{\frac{1}{n}} \sum_{i=1}^n \left(\int_{D'} |\partial_i u(T(x))|^p \mathrm{d}m(x) \right)^{\frac{1}{p}} \\
&= \lambda^\delta |D|^{\frac{1}{n}-\frac{1}{p}} \sum_{i=1}^n \left(\int_{D'} |D| |\partial_i u(T(x))|^p \mathrm{d}m(x) \right)^{\frac{1}{p}} \\
&= \lambda^\delta |D|^{\frac{1}{n}-\frac{1}{p}} \sum_{i=1}^n \left(\int_{D} |\partial_i u|^p \mathrm{d}m(x) \right)^{\frac{1}{p}} \\
&= \lambda^\delta |D|^{\frac{1}{n}-\frac{1}{p}} \sum_{i=1}^n \|\partial_i u\|_{L^p(D)}.
\end{aligned}$$

Thus,

$$\sup_D |u| = \sup_{D'} |w| \le \lambda^\delta |D|^{\frac{p-n}{np}} \sum_{i=1}^n \|\partial_i u\|_{L^p(D)}. \tag{5}$$

The claim of the theorem holds for functions $u \in \mathcal{C}_0^1(D)$.

Let us prove that the claim holds for functions $u \in W_0^{1,p}(D)$. Let $u \in W_0^{1,p}(D)$. Then there exists a function sequence (φ_i), $\varphi_i \in \mathcal{C}_0^\infty(D)$, such that $\varphi_i \to u$ in $W^{1,p}(D)$. Especially, $\varphi_i \to u$ in $L^p(D)$. If we take a subsequence we may assume that $\varphi_i \to u$ for almost every x in D. We may assume that

$$\|\varphi_i\|_{W^{1,p}(D)} \le \|u\|_{W^{1,p}(D)} + 1.$$

The sequence (φ_i) converges uniformly in the set D, since if we apply the inequality (5) to the function $\varphi_i - \varphi_j$ we obtain

$$\sup_D |\varphi_i - \varphi_j| \le \lambda^\delta |D|^{\frac{1}{n}-\frac{1}{p}} \sum_{h=1}^n \|\partial_h \varphi_i - \partial_h \varphi_j\|_{L^p(D)} \to 0,$$

when $i, j \to \infty$. By the Cauchy criteria of the uniform convergence the sequence (φ_i) converges uniformly in D. Let us denote

$$v(x) = \lim_{i \to \infty} \varphi_i(x).$$

Since the functions φ_i are continuous and the sequence (φ_i) converges uniformly, the function v is continuous. On the other hand,

$$\lim_{i \to \infty} \varphi_i(x) = u(x)$$

for almost every $x \in D$. Hence, $(u - v)(x) = 0$ for almost every $x \in D$. The claim follows when $p > n$. □

Remark 14.9. Theorem 14.3 does not hold when $p = n$. The following example shows that there exists a function $u \in W_0^{1,n}(D)$ such that $u \notin L^\infty(D)$.

Example 14.10. Let

$$u(x) = \begin{cases} \log\log(1+|x|^{-1}) - \log\log 2 & , \quad 0 < |x| < 1 \\ 0 & , \quad |x| \geq 1. \end{cases}$$

Then,

$$\int_{|x|<1} |\nabla u(y)|^n \mathrm{d}y < \infty\,,$$

but

$$\operatorname{ess\,sup} |u(x)| = \infty.$$

15 Embedding theorems in the space $W_0^{m,p}(D)$

Theorem 15.1. *Let D be a domain in $\mathbb{R}^n$, $n \geq 2$. Let $u \in W_0^{m,p}(D)$ and $mp < n$. Then, $u \in L^{\frac{np}{n-mp}}(D)$ and*

$$\|u\|_{L^{\frac{np}{n-mp}}(D)} \leq c(m,n,p) \sum_{|\alpha|=m} \|\,\mathrm{D}^\alpha\, u\|_{L^p(G)}.$$

Proof. We may assume that $u \in \mathcal{C}_0^\infty(D)$ and prove the claim for $u \in \mathcal{C}_0^\infty(D)$. We apply the Gagliardo-Nirenberg-Sobolev embedding theorem to the inequality

$$\|u\|_{L^{\frac{nq}{n-q}}(\mathbb{R}^n)} \leq c \sum_{i=1}^n \|\partial_i u\|_{L^q(\mathbb{R}^n)},$$

when $\frac{nq}{n-q} = \frac{np}{n-mp}$ that is $q = \frac{pn}{n-(m-1)p}$. By iteration,

$$\begin{aligned} \|u\|_{L^{\frac{np}{n-mp}}(D)} &\leq c_1 \sum_{i=1}^n \|\partial_i u\|_{L^{\frac{pn}{n-(m-1)p}}(D)} \\ &\leq \cdots \\ &\leq c_{m-1} \sum_{|\alpha|=m} \|\,\mathrm{D}^\alpha\, u\|_{L^p(D)}\,. \end{aligned}$$

□

Example 15.2.

1. Let $u \in W_0^{1,1}(\mathbb{R}^3)$. Here, $n = 3$ and $p = m = 1$. By the Gagliardo-Nirenberg-Sobolev embedding theorem $u \in L^{3/2}(\mathbb{R}^3)$.
2. If $u \in W_0^{2,1}(\mathbb{R}^3)$. Here, $n = 3$, $p = 1$ and $m = 2$. Hence, $u \in L^3(\mathbb{R}^3)$.

It seems that the integrability gets better if the degree of differentiability increases.

Theorem 15.3. *Let D be a bounded domain in $\mathbb{R}^n$, $n \geq 2$. Assume that $u \in W_0^{m,p}(D)$ and*

$$0 \leq k < m - \frac{n}{p} = \frac{pm-n}{p}\,.$$

Then there exists a function $v \in \mathcal{C}^k(D)$ such that $(v-u)(x) = 0$ for almost every x in D and

$$\sum_{|\alpha|\leq k} \sup_{x\in D} |\,\mathrm{D}^\alpha\, v(x)| \leq c \sum_{|\alpha|=m} \|\,\mathrm{D}^\alpha\, u\|_{L^p(D)}\,;$$

here $c = c(k,m,n,p,\operatorname{diam}(D))$.

Proof. Let $u \in \mathcal{C}_0^\infty(D)$ and $|\alpha| \leq k$. Then,

$$\begin{aligned} \sup |\,\mathrm{D}^\alpha\, v| &\leq c \sum_{i=1}^n \|\partial_i\, \mathrm{D}^\alpha\, v\|_{L^p(D)} \\ &\leq c \sum_{i=1}^n \sum_{j=1}^n \|\partial_{ij}\, \mathrm{D}^\alpha\, v\|_{L^p(D)} \\ &\leq c \sum_{i_1=1}^n \cdots \sum_{i_k=1}^n \|\partial_{i_1 i_2 \ldots i_k}\, \mathrm{D}^\alpha\, v\|_{L^p(D)} \\ &= c \sum_{|\beta|=m} \|\,\mathrm{D}^\beta\, v\|_{L^p(D)}. \end{aligned}$$

If we take a sum over all α we obtain

$$\sum_{|\alpha|\le k} \sup |\,\mathrm{D}^\alpha v| \le c \sum_{|\alpha|=m} \|\,\mathrm{D}^\alpha v\|_{L^p(D)}.$$

Also, $v \to u$ as in Theorem 14.8. □

Examples 15.4.

1. Let $u \in W_0^{1,10}(\mathbb{R}^2)$. Here, $m = 1$, $p = 10$, and $n = 2$, we have

$$\frac{mp-n}{p} = \frac{10-2}{10} = 1 - \frac{1}{5}$$

which means $k = 0$. Hence, the function u is not continuously differentiable but there exists $v \in \mathcal{C}(\mathbb{R}^2)$ such that $(v-u)(x) = 0$ for almost x in $\mathbb{R}^2$.

2. Let $u \in W_0^{10,1}(\mathbb{R}^2)$. Here, $m = 10$, $p = 1$, and $n = 2$, and

$$\frac{mp-n}{p} = \frac{10\cdot 1-2}{1} = 8\,.$$

Thus, there exists a function $v \in \mathcal{C}^7(\mathbb{R}^2)$ such that $(v-u)(x) = 0$ for almost every x in $\mathbb{R}^2$.

Remark 15.5. The results are for functions $u \in W_0^{1,p}(D)$; for example, if $p < n$, then

$$\|u\|_{L^{\frac{np}{n-p}}} \le c \sum_{i=1}^{n} \|\partial_i u\|_{L^p(D)}.$$

If $u \in W^{1,p}(D)$ the previous inequality does not hold at least when D is a bounded domain and u is a non-zero constant function. So we do not obtain information of the global integrability of u. We are able to obtain local information of the differentiability of the function u.

16 Embedding theorems in the space $W^{m,p}(D)$

Lemma 16.1. *Let A be an open set in $\mathbb{R}^n$ and $K \subset A$ compact. Then there exists a function $\psi \in \mathcal{C}_0^\infty(A)$ such that*

1. *$0 \le \psi \le 1$ and*
2. *$\psi|_K = 1$.*

Proof. Choose j such that $\frac{1}{j} < \frac{1}{2}\operatorname{dist}(F, \complement A)$. Let $i > j$. Define

$$\psi = \chi_{K+\overline{\mathrm{B}}(0,\frac{1}{j})} * \varphi_i.$$

Since $\frac{1}{j} < \frac{1}{2}\operatorname{dist}(K, \complement A)$, then $\frac{1}{i} < \frac{1}{2}\operatorname{dist}(K, \complement A)$. Thus, $\psi(x) = 0$, when $x \notin A$. Hence, $\psi \in \mathcal{C}_0^\infty(A)$. Also $\psi|_K = 1$, since $\frac{1}{i} < \frac{1}{j}$. □

Remember 16.2. If the set $\operatorname{spt} f$ is compact in an open set A in $\mathbb{R}^n$, then

$$f * \varphi_i \in \mathcal{C}_0^\infty(A),$$

whenever $0 < \frac{1}{i} < \operatorname{dist}(\operatorname{spt} f, \partial A)$.

Theorem 16.3. *1. If D is a domain in $\mathbb{R}^n$ and $u \in W^{m,p}(D)$, then, $u \in L_{\mathrm{loc}}^{\frac{np}{n-mp}}(D)$, if $mp < n$.*

2. If D is a bounded domain in $\mathbb{R}^n$ and $u \in W^{m,p}(D)$, then there exists a function $v \in \mathcal{C}^k(D)$ such that $(u-v)(x) = 0$ for almost every x in D, if $0 \le k < \frac{mp-n}{p}$.

Proof. Let $K \subset D$ be compact. Choose a function $\psi \in \mathcal{C}_0^\infty(D)$ such that $0 \le \psi \le 1$ and $\psi|_K = 1$ as in Lemma 16.1. Then, the function $v = \psi u \in W_0^{m,p}(D)$. Theorem 14.3 implies that

$$v \in L^{\frac{np}{n-mp}}(D)$$

and

$$\|u\|_{L^{\frac{np}{n-mp}}(K)} \le \|v\|_{L^{\frac{np}{n-mp}}(D)} < \infty.$$

□

Lemma 16.4. *There exists a constant $K = K(n) \in [1, \infty)$ such that for all $r \in (0, \infty]$ there exists a function $\psi \in \mathcal{C}_0^\infty(\mathbb{R}^n)$ with the following properties*

1. $0 \le \psi \le 1$.
2. $\psi|_{\mathrm{B}(0,r)} = 1$.
3. $|\partial_i \psi| \le K$ *for every* $x \in \mathbb{R}^n$, $i = 1, \ldots, n$.

Proof. Let $\varphi \in \mathcal{C}_0^\infty(\mathbb{R}^n)$ be a mollifier with $\operatorname{spt} \varphi \subset \overline{\mathrm{B}}(0,1)$. Let us define

$$\psi = \chi_{\mathrm{B}(0,r+3)} * \varphi.$$

Then, $\psi \in \mathcal{C}_0^\infty(\mathbb{R}^n)$ and $|\partial_i \psi(x)| \le K$. □

Theorem 16.5. *If $p \in [1, \infty)$, then $W^{1,p}(\mathbb{R}^n) = W_0^{1,p}(\mathbb{R}^n)$.*

Proof. Obviously, $W_0^{1,p}(\mathbb{R}^n) \subset W^{1,p}(\mathbb{R}^n)$.

By Theorem 9.4 it is enough to prove: if $u \in W^{1,p}(\mathbb{R}^n)$, then for every $\varepsilon > 0$ there exists a function $\varphi \in \mathcal{C}_0^\infty(\mathbb{R}^n)$ such that

$$\|u - \varphi\|_{W^{1,p}(\mathbb{R}^n)} < \varepsilon.$$

Since $W^{1,p}(\mathbb{R}^n) = H^{1,p}(\mathbb{R}^n)$, there exists a function sequence (φ_i), $\varphi_i \in \mathcal{C}^\infty(\mathbb{R}^n)$, such that $\|\varphi_i\|_{W^{1,p}(\mathbb{R}^n)} < \infty$, $\varphi_i \to u$ in $L^p(G)$ and $(\mathrm{D}^\alpha \varphi_i)$ is a Cauchy sequence in $L^p(G)$ for every α, $|\alpha| \le 1$. Thus, there is a function $\varphi_1 \in \mathcal{C}^\infty(\mathbb{R}^n)$ such that

$$\|u - \varphi_1\|_{W^{1,p}(\mathbb{R}^n)} < \frac{\varepsilon}{2}.$$

Let us choose $r > 0$ with

$$\|\varphi_1\|_{W^{1,p}(\complement\, \mathrm{B}(0,r))} < \frac{\varepsilon}{4nK}.$$

Define

$$\varphi = \psi\varphi_1 \in \mathcal{C}_0^\infty(\mathbb{R}^n),$$

where $\psi \in \mathcal{C}_0^\infty(\mathbb{R}^n)$ is a function as in Lemma 16.4. Thus,

$$\begin{aligned} & \|u - \varphi\|_{W^{1,p}(\mathbb{R}^n)} \\ \le\ & \|u - \varphi_1\|_{W^{1,p}(\mathbb{R}^n)} + \|\varphi_1 - \varphi\|_{W^{1,p}(\mathbb{R}^n)} \\ \le\ & \frac{\varepsilon}{2} + \|\varphi_1 - \psi\varphi_1\|_{W^{1,p}(\mathrm{B}^n(0,r))} + \|\varphi_1 - \psi\varphi_1\|_{W^{1,p}(\complement\, \mathrm{B}^n(0,r))}, \end{aligned}$$

where

$$\|\varphi_1 - \psi\varphi_1\|_{W^{1,p}(\mathrm{B}^n(0,r))} = 0$$

and

$$\begin{aligned} & \|\varphi_1 - \psi\varphi_1\|_{W^{1,p}(\complement\, \mathrm{B}^n(0,r))} \\ =\ & \||\varphi_1|(1-\psi)\|_{L^p(\complement\, \mathrm{B}^n(0,r))} + \sum_{i=1}^n \|\partial_i\varphi_1 - \psi\partial_i\varphi_1 - \partial_i\psi\varphi_1\|_{L^p(\complement\, \mathrm{B}^n(0,r))} \\ \le\ & \||\varphi_1|(1-\psi)\|_{L^p(\complement\, \mathrm{B}^n(0,r))} \\ & + \|\partial_i\varphi_1(1-\psi)\|_{L^p(\complement\, \mathrm{B}^n(0,r))} + \|\partial_i\psi\varphi_1\|_{L^p(\complement\, \mathrm{B}^n(0,r))} \\ \le\ & \frac{\varepsilon}{4nK} + \frac{\varepsilon}{4nK} n + \frac{\varepsilon}{4nK} nK < \varepsilon\,. \end{aligned}$$

Thus,

$$\|u - \varphi\|_{W^{1,p}(\mathbb{R}^n)} \le \frac{\varepsilon}{2} + \varepsilon = \frac{3}{2}\varepsilon.$$

Hence the claim of the theorem holds. □

Definition 16.6. Let D be a domain in $\mathbb{R}^n$. Let $m \ge 0$, $p \in [1, \infty)$, and $u \in W^{m,p}(D)$. The domain D is an (m,p)-extension domain if there exists a function $u^* \in W^{m,p}(\mathbb{R}^n)$ such that $u^* = u$ in D and

$$\|u^*\|_{W^{m,p}(\mathbb{R}^n)} \le c\|u\|_{W^{m,p}(D)},$$

where $c = c(m, n, p, D)$.

Theorem 16.7. *Assume that D is a bounded (m,p)-extension domain. Then,*

1. *for $u \in W^{m,p}(D)$ and $mp < n$ there exists a constant $c < \infty$ such that*

$$\|u\|_{L^{\frac{np}{n-mp}}(D)} \le c\|u\|_{W^{m,p}(D)}.$$

2. *for $u \in W^{m,p}(D)$, where $0 \le k < \frac{mp-n}{p}$, there exists a function $v \in \mathcal{C}^k(\mathbb{R}^n)$ such that*

   (a) *$(u-v)(x) = 0$ for almost every x in D.*

   (b) *$\sum_{|\alpha|\le k} \sup_{x\in D} |\,\mathrm{D}^\alpha v(x)| \le c\|u\|_{W^{m,p}(D)}$ where the constant c is independent of u.*

Proof. (1) If $u \in W^{m,p}(D)$ and D is an (m,p)-extension domain, then by the definition of the extension domains there is a function $u^* \in W^{m,p}(\mathbb{R}^n)$ such that $u = u^*$ in D. Theorem 16.5 implies that $u^* \in W_0^{m,p}(\mathbb{R}^n)$. By Theorem 15.1

$$\|u^*\|_{L^{\frac{np}{n-mp}}(\mathbb{R}^n)} \le c \sum_{|\alpha|=m} \|\,\mathrm{D}^\alpha u^*\|_{L^p(\mathbb{R}^n)}.$$

The definition of the extension domain yields that

$$\|u\|_{L^{\frac{np}{n-mp}}(D)} = \|u^*\|_{L^{\frac{np}{n-mp}}(D)}.$$

Thus,

$$\begin{aligned}
\|u\|_{L^{\frac{np}{n-mp}}(D)} &= \|u^*\|_{L^{\frac{np}{n-mp}}(D)} \\
&\le \|u^*\|_{L^{\frac{np}{n-mp}}(\mathbb{R}^n)} \\
&\le c \sum_{|\alpha|=m} \|\,\mathrm{D}^\alpha u^*\|_{L^p(\mathbb{R}^n)} \\
&\le c\|u^*\|_{W^{m,p}(\mathbb{R}^n)} \\
&\le c\|u\|_{W^{m,p}(D)}.
\end{aligned}$$

(2) If $u \in W^{m,p}(D)$, then the definition of the extension domain implies that there exists a function $u^* \in W^{m,p}(\mathbb{R}^n)$ such that $u = u^*$ in D. Theorem 15.3 yields that there exists a function $v_1 \in \mathcal{C}^k(\mathbb{R}^n)$ such that $v_1(x) = u^*(x)$ for almost every x in the space $\mathbb{R}^n$. Hence, $v_1(x) = u^*(x) = u(x)$ for almost every x in D. Recall that we assume D is bounded. Let us choose a number $r > 0$ such that $D \subset \mathrm{B}^n(0,r)$. Then we may choose a function $\psi \in \mathcal{C}_0^\infty(\mathrm{B}^n(0,r+3))$ such that $0 \le \psi \le 1$, $\psi|_{\mathrm{B}^n(0,r)} = 1$ and $|\partial_i \psi| \le K(n)$. Let us set $v = \psi v_1$. Then, $v = u$ for almost every x in D, since $v = v_1$ in $\mathrm{B}^n(0,r)$, $D \subset \mathrm{B}^n(0,r)$ and $v_1 = u$ for almost every x in D. Since $v \in \mathcal{C}^k$ and $v \in W_0^{m,p}(\mathrm{B}^n(0,r+3))$, Theorem 15.3 implies

$$\begin{aligned}
\sum_{|\alpha|\le k} \sup |\,\mathrm{D}^\alpha v| &\le c \sum_{|\alpha|=m} \|\,\mathrm{D}^\alpha v\|_{L^p(\mathrm{B}^n(0,r+3))} \\
&\le c \sum_{|\alpha|\le m} \|\,\mathrm{D}^\alpha v\|_{L^p(\mathrm{B}^n(0,r+3))} \\
&\le c \sum_{|\alpha|\le m} \|\,\mathrm{D}^\alpha v\|_{L^p(\mathbb{R}^n)} \\
&= c\|v\|_{W^{m,p}(\mathbb{R}^n)} = \|\psi v_1\|_{W^{m,p}(\mathbb{R}^n)} \\
&= \|\psi v_1\|_{L^p(\mathbb{R}^n)} + \sum_{|\alpha|} \|\,\mathrm{D}^\alpha(\psi v_1)\|_{L^p(\mathbb{R}^n)} \\
&\le \|v_1\|_{L^p(\mathbb{R}^n)} + \|\psi\,\mathrm{D}^\alpha v_1\|_{L^p(\mathbb{R}^n)} + \|\,\mathrm{D}^\alpha \psi v_1\|_{L^p(\mathbb{R}^n)} \\
&\le c\|v_1\|_{W^{m,p}(\mathbb{R}^n)} = c\|u^*\|_{W^{m,p}(\mathbb{R}^n)} \\
&\le c\|u\|_{W^{m,p}(D)}.
\end{aligned}$$

□

Definition 16.8. A domain D in $\mathbb{R}^n$ is a Lipschitz domain if the boundary of the domain can be expressed locally as a graph of a Lipschitz mapping which is defined in an open ball in the space $\mathbb{R}^{n-1}$.

The following theorem and its proof are found in the book [St].

Theorem 16.9 (A. P. Calderon, Elias M. Stein)**.** *Every Lipschitz domain is an extension domain.*

Definition 16.10 (Peter W. Jones)**.** A domain D in $\mathbb{R}^n$ is an (ε,δ)- domain if for every pair of points $x,y\in\mathbb{R}^n$, when $|x-y|<\delta$, there exists a rectifiable curve γ in D joining the points x and y in D such that

$$l(\gamma)\le\varepsilon^{-1}|x-y|$$

and

$$\operatorname{dist}(z,\mathbb{R}^n\setminus D)\ge\varepsilon\frac{|x-z||y-z|}{|x-y|}$$

for all $z\in\gamma$.

The following theorem and its proof are in [Jo].

Theorem 16.11 (Peter W. Jones)**.** *A domain D in $\mathbb{R}^n$ is an extension domain if and only if D is an (ε,δ)-domain with some ε and δ.*

Theorem 16.12. *A domain D in R^n, $n\ge 2$, is a uniform domain if there are constants α and β such that every pair of points x_1, x_2 can be joined by a rectifiable curve γ_{x_1,x_2} in D such that*

$$l(\gamma_{x_1,x_2})\le\alpha|x_1-x_2|$$

and

$$\min_{j=1,2} l(\gamma(x_j,x))\le\beta\operatorname{dist}(x,\partial D)$$

for all $x\in\gamma_{x_1,x_2}$.

Theorem 16.13. *A bounded domain is a bounded uniform domain, if and only if it is an (ε,δ)-domain with some ε and δ.*

Definition 16.14. The boundary of a domain D in $\mathbb{R}^n$ is a $\mathcal{C}^1$ -boundary, if for every $x\in\partial D$ there exists a neighbourhood U_x of the point x and a mapping $\varphi: U_x\to \mathrm{B}^n(0,1)$ such that

1. φ is homeomorphism,
2. $\varphi(U\cap\partial D)=\mathrm{B}^n(0,1)\cap\{x\in\mathbb{R}^n|x_n=0\}$,
3. $\varphi(x_0)=0$ and $\varphi(U\cap D)=\mathrm{B}^n(0,1)\cap\{x\in\mathbb{R}^n|x_n>0\}$.

Theorem 16.15. *A bounded domain with a $\mathcal{C}^1$ -boundary is an extension domain.*

17 Potential inequality

Theorem 17.1. *Assume that D is a bounded domain in $\mathbb{R}^n$, $f\in L^p(D)$, and $1\le p\le\infty$. Let*

$$I_1f(x)=\int_D|x-y|^{1-n}f(y)\mathrm{d}m(y),$$

when $x\in D$. Then

$$\|I_1f\|_{L^p(D)}\le n|B^n(0,1)|^{1-\frac{1}{n}}|D|^{\frac{1}{n}}\|f\|_{L^p(D)}.$$

Proof. Let us fix a point $x\in D$. Let us choose a positive real number R such that

$$|D|=|\mathrm{B}^n(x,R)|=|B^n(0,1)|R^n.$$

Then,

$$\begin{aligned}\int_D|x-y|^{1-n}\mathrm{d}m(y)&\le\int_{\mathrm{B}^n(x,R)}|x-y|^{1-n}\mathrm{d}m(y)\\&=\int_{\mathrm{B}^n(0,R)}|z|^{1-n}\mathrm{d}m(z)\\&=m_{n-1}(S^{n-1}(0,1))\int_0^R r^{1-n}r^{n-1}\mathrm{d}r\\&=m_{n-1}(S^{n-1}(0,1))R\\&=n|B^n(0,1)||D|^{\frac{1}{n}}|B^n(0,1)|^{-\frac{1}{n}}\\&=n|B^n(0,1)|^{1-\frac{1}{n}}|D|^{\frac{1}{n}}.\end{aligned}$$

Since the point $x \in D$ is arbitrarily chosen and then fixed,

$$\int_D |x-y|^{1-n} \mathrm{d}m(y) \leq n|B^n(0,1)|^{1-\frac{1}{n}}|D|^{\frac{1}{n}}$$

for every x in D.

If $p < \infty$, then by the Hölder inequality with $\left(p, \frac{p}{p-1}\right)$

$$\begin{aligned}
&\left(\int_D \left(\int_D |x-y|^{1-n} f(y)\mathrm{d}y\right)^p \mathrm{d}x\right)^{\frac{1}{p}} \\
\leq\ &\left(\int_D \left(\int_D |x-y|^{1-n}|f(y)|^p \mathrm{d}y\right)^{\frac{p}{p}} \left(\int_D |x-y|^{1-n}\mathrm{d}y\right)^{\frac{p(p-1)}{p}} \mathrm{d}x\right)^{\frac{1}{p}} \\
\leq\ &\left(n|B^n(0,1)|^{1-\frac{1}{n}}|D|^{\frac{1}{n}}\right)^{\frac{p-1}{p}} \left(\int_D \left(\int_D |x-y|^{1-n}|f(y)|^p \mathrm{d}y\right) \mathrm{d}x\right)^{\frac{1}{p}} \\
\leq\ &\left(n|B^n(0,1)|^{1-\frac{1}{n}}|D|^{\frac{1}{n}}\right)^{\frac{p-1}{p}} \left(\sup_{y\in D} \int_D |x-y|^{1-n}\mathrm{d}x\right)^{\frac{1}{p}} \left(\int_D |f(y)|^p \mathrm{d}y\right)^{\frac{1}{p}} \\
\leq\ &n|B^n(0,1)|^{1-\frac{1}{n}}|D|^{\frac{1}{n}} \left(\int_D |f(y)|^p \mathrm{d}y\right)^{\frac{1}{p}}.
\end{aligned}$$

If $p = \infty$, then

$$\begin{aligned}
|I_1 f(x)| &\leq \left(\int_D |x-y|^{1-n}\mathrm{d}m(y)\right)\|f\|_{L^\infty(D)} \\
&\leq n|B^n(0,1)|^{1-\frac{1}{n}}|D|^{\frac{1}{n}}\|f\|_{L^\infty(D)}.
\end{aligned}$$

□

Theorem 17.2. *Assume that D is a bounded convex domain in $\mathbb{R}^n$ and $u \in W^{1,p}(D)$. Then*

$$\|u - u_D\|_{L^p(D)} \leq \left(\frac{|B^n(0,1)|}{|D|}\right)^{1-\frac{1}{n}} \mathrm{diam}(D)^n \|\nabla u\|_{L^p(D)},$$

whenever $1 \leq p \leq \infty$.

Proof. We may assume that $u \in \mathcal{C}^1(D)$. Since D is a convex domain, for every x and y in D holds that

$$u(x) - u(y) = -\int_0^{|x-y|} \frac{\partial}{\partial r} u(x + r\theta)\mathrm{d}r,$$

where $\theta = \frac{y-x}{|x-y|}$. Let us denote

$$\varphi(x) = \begin{cases} |\partial_r u(x)| & , \text{ kun } x \in D \\ 0 & , \text{ kun } x \notin D. \end{cases}$$

Integrating over all x in D and using Fubini's theorem we obtain

$$\begin{aligned}
|D||u(x) - u_D| &\leq \int_D \int_0^{|x-y|} \left|\frac{\partial}{\partial r} u(x + r\theta)\right| \mathrm{d}r\mathrm{d}m(y) \\
&\leq \int_{\mathrm{B}^n(x,\mathrm{diam}(D))} \int_0^\infty \varphi(x + r\theta)\mathrm{d}r\mathrm{d}m(y) \\
&= \int_0^\infty \int_{|\theta|=1} \int_0^{\mathrm{diam}(D)} \varphi(x + r\theta)\rho^{n-1}\mathrm{d}\rho\mathrm{d}\theta\mathrm{d}r \\
&= \frac{\mathrm{diam}(D)^n}{n} \int_0^\infty \int_{|\theta|=1} \varphi(x + r\theta)\mathrm{d}\theta\mathrm{d}r \\
&\leq \frac{\mathrm{diam}(D)^n}{n} \int_D |\nabla u(y)||x-y|^{1-n}\mathrm{d}y.
\end{aligned}$$

The claim follows from the potential inequality. □

Remark 17.3. If D is a bounded convex domain in $\mathbb{R}^n$ and u is a function from $W^{1,p}(D)$, then

$$\|u-u_D\|_{L^p(D)} \le \left(\frac{|B^n(0,1)|}{|D|}\right)^{1-\frac{1}{n}} \operatorname{diam}(D)^n \|\nabla u\|_{L^p(D)}\,.$$

Hence,

$$\|u\|_{L^p(D)} \le \left(\frac{|B^n(0,1)|}{|D|}\right)^{1-\frac{1}{n}} \operatorname{diam}(D)^n \|\nabla u\|_{L^p(D)} + |D|^{-1+\frac{1}{p}} \left|\int_D u(x)\mathrm{d}x\right|.$$

Also the other direction is valid as soon as we choose $v = u - u_D$. Note that $v_D = 0$.

Theorem 17.4. *Assume that D in $\mathbb{R}^n$ is a bounded uniform domain, $p \in [1,n)$ and $q \in [p, \frac{np}{n-p}]$. Then there exists a constant c such that for every function $u \in W^{1,p}(D)$*

$$\|u-u_D\|_{L^q(D)} \le c\|\nabla u\|_{L^p(D)}.$$

Proof. Since D is a bounded uniform domain, D is an extension domain. It is possible to prove that a uniform domain satisfies the Poincaré inequality, (6). Hence, for $u \in W^{1,p}(D)$

$$\begin{aligned}\|u-u_D\|_{L^q(D)} &\le c\sum_{|\alpha|\le 1} \|\,\mathrm{D}^\alpha(u-u_D)\|_{L^p(D)}\\ &= c\left(\|u-u_D\|_{L^p(D)} + \|\nabla u\|_{L^p(D)}\right)\\ &\le c\|\nabla u\|_{L^p(D)}.\end{aligned}$$

□

18 On Poincaré domains

This chapter is based on some parts of my thesis, [Hu].

Lemma 18.1. *Assume that D is a bounded domain in $\mathbb{R}^n$. Let A be a subset of D with $|A| > 0$ and $u \in L^p(D)$. Then, for every real number c*

$$\|u-u_A\|_{L^p(D)} \le 2\left(\frac{|D|}{|A|}\right)^{\frac{1}{p}} \|u-c\|_{L^p(D)}.$$

Proof. By the Minkowski inequality

$$\begin{aligned}&\left(\int_D |u(x)-u_A|^p\mathrm{d}x\right)^{\frac{1}{p}}\\ &\le \left(\int_D |u(x)-c|^p\mathrm{d}x\right)^{\frac{1}{p}} + \left(\int_D |c-u_A|^p\mathrm{d}x\right)^{\frac{1}{p}}\\ &= \left(\int_D |u(x)-c|^p\mathrm{d}x\right)^{\frac{1}{p}} + |D|^{\frac{1}{p}}|c-u_A|\\ &= \left(\int_D |u(x)-c|^p\mathrm{d}x\right)^{\frac{1}{p}} + |D|^{\frac{1}{p}}\left|c-\frac{1}{|A|}\int_A u(y)\mathrm{d}y\right|\\ &\le \left(\int_D |u(x)-c|^p\mathrm{d}x\right)^{\frac{1}{p}} + \frac{|D|^{\frac{1}{p}}}{|A|}\int_A |u(y)-c|\mathrm{d}y\,.\end{aligned}$$

If $p = 1$, the claim follows.

By the Hölder inequality with $(p, p/(p-1))$, $p > 1$,

$$\begin{aligned}&\left(\int_D |u(x)-u_A|^p\mathrm{d}x\right)^{\frac{1}{p}}\\ &\le \left(\int_D |u(x)-c|^p\mathrm{d}x\right)^{\frac{1}{p}} + \frac{|D|^{\frac{1}{p}}}{|A|}\left(\int_A |u(y)-c|^p\mathrm{d}y\right)^{\frac{1}{p}} |A|^{\frac{p-1}{p}}\\ &\le \left(\int_D |u(x)-c|^p\mathrm{d}x\right)^{\frac{1}{p}} + \frac{|D|^{\frac{1}{p}}}{|A|^{\frac{1}{p}}}\left(\int_D |u(y)-c|^p\mathrm{d}x\right)^{\frac{1}{p}},\end{aligned}$$

which yields the claim. □

Definition 18.2. Assume that D is a bounded domain in $\mathbb{R}^n$, $n \geq 2$, and $1 \leq p < \infty$. The domain D is called a p-Poincaré domain, if there exists a constant $\kappa = \kappa_p(D)$ such that the inequality

$$\|u - u_D\|_{L^p(D)} \leq \kappa \|\nabla u\|_{L^p(D)} \tag{6}$$

holds for all functions $u \in W^{1,p}(D)$. We write $D \in \mathcal{P}_p$ or $D \in \mathcal{P}(p)$.

The inequality (6) is called the Poincaré inequality and κ is a Poincaré constant of D, $\kappa = \kappa_p(D)$.

Lemma 18.3. *Assume that $D_i \in \mathcal{P}_p$ is a Poincaré domain with a constant $\kappa_p(D_i)$, $i = 1, 2$, respectively, and $|D_1 \cap D_2| > 0$. Then $D_1 \cup D_2$ is a p-Poincaré domain with a constant*

$$\kappa_p(D_1 \cup D_2) \leq \frac{8}{|D_1 \cap D_2|^{\frac{1}{p}}} \left(|D_1|\kappa_p(D_1)^p + |D_2|\kappa_p(D_2)^p\right).$$

Proof. Let us write $D = D_1 \cup D_2$. Applying Lemma 18.1 three times we obtain

$$\begin{aligned}
&\int_D |u(y) - u_D|^p \mathrm{d}y \\
&\leq 2^p \frac{|D|}{|D|} \int_D |u(y) - u_{D_1 \cap D_2}|^p \mathrm{d}y \\
&\leq 2^{2p-1} \left(\int_{D_1} |u(y) - u_{D_1 \cap D_2}|^p \mathrm{d}y + \int_{D_2} |u(y) - u_{D_1 \cap D_2}|^p \mathrm{d}y \right) \\
&\leq 2^{3p-1} \Bigg(\frac{|D_1|}{|D_1 \cap D_2|} \int_{D_1} |u(y) - u_{D_1}|^p \mathrm{d}y \\
&+ \frac{|D_2|}{|D_1 \cap D_2|} \int_{D_2} |u(y) - u_{D_2}|^p \mathrm{d}y \Bigg) \\
&\leq \frac{2^{3p-1}}{|D_1 \cap D_2|} \sum_{i=1}^{2} |D_i| \int_{D_i} |u(y) - u_{D_i}|^p \mathrm{d}y \\
&\leq \frac{2^{3p-1}}{|D_1 \cap D_2|} \sum_{i=1}^{2} |D_i| \kappa(D_i)^p \int_{D_i} |\nabla u(y)|^p \mathrm{d}y,
\end{aligned}$$

whenever $u \in W^{1,p}(D_i)$, $i = 1, 2$. □

Definition 18.4. Domains D_i in $\mathbb{R}^n$, $i = 0, 1, \ldots, k$, form a chain $\mathcal{C}(D_k) = (D_0, D_1, \ldots, D_k)$ whenever $|D_i \cap D_j| \neq 0$, if and only if $|i - j| \leq 1$. The length of the chain $\mathcal{C}(D_k)$ is $l(\mathcal{C}(D_k)) = k$. The notation $\mathcal{C}(D_k)$ means also a collection of the sets in the chain.

Poincaré decomposition 18.5. Let $\mathcal{W}$ be a family of domains $D \in \mathcal{P}(p)$ such that $\kappa_p(D) \leq c_1 < \infty$. Let $N \geq 1$. The family $\mathcal{W}$ is called a (c_1, N, p)-Poincaré decomposition of G, if the following claims hold.

1. $G = \bigcup_{D \in \mathcal{W}} D$.
2. $\sum_{D \in \mathcal{W}} \chi_D(x) \leq N \chi_G(x)$ for all $x \in R^n$.
3. There exists a fixed domain $D_0 \in \mathcal{W}$ such that for every $D \in \mathcal{W}$ there is a chain $\mathcal{C}(D) = (D_0, D_1, \ldots, D)$ of domains from the family $\mathcal{W}$ such that

$$\max\{|D_i|, |D_{i+1}|\} \leq c_1 |D_i \cap D_{i+1}|$$

for all $i = 0, 1, \ldots, l(\mathcal{C}(D_k)) - 1$.

Let us fix for each $D \in \mathcal{W}$ a chain $\mathcal{C}(D)$. This chain is called a Poincaré chain from D_0 to D.

Let us write $A(\mathcal{W}) = \{D \in \mathcal{W} | A \in \mathcal{C}(D)\}$.

Theorem 18.6. *Assume that G is a domain in $\mathbb{R}^n$ and $\mathcal{W}$ is a (c_1, N, p) Poincaré decomposition of G. If there is a constant $c_2 < \infty$ such that for every $A \in \mathcal{W}$*

$$\sum_{D \in A(\mathcal{W})} l(\mathcal{C}(D))^{p-1} |D| \leq c_2 \kappa_p(A)^{-p} |A|,$$

then G is a p-Poincaré domain.

Proof. By Lemma 18.1 it is enough to prove that the inequality

$$\int_G |u(y) - u_{D_0}|^p \mathrm{d}y \le c \int_G |\nabla u(y)|^p \mathrm{d}y$$

holds for every $u \in W^{1,p}(G)$.

We may assume that $u \in \mathcal{C}^1(G)$. Since each D is a p -Poincaré domain with a Poincaré constant $\kappa_p(D)$,

$$\begin{aligned}
&\int_G |u(y) - u_{D_0}|^p \mathrm{d}y = \sum_{D\in\mathcal{W}} \int_D |u(y) - u_{D_0}|^p \mathrm{d}y \\
&\le 2^p \left(\sum_{D\in\mathcal{W}} \int_D |u(y) - u_D|^p \mathrm{d}y + \sum_{D\in\mathcal{W}} \int_D |u_D - u_{D_0}|^p \mathrm{d}y \right) \\
&\le 2^p \left(\sum_{D\in\mathcal{W}} \kappa_p(D)^p \int_D |\nabla u(y)|^p \mathrm{d}y + \sum_{D\in\mathcal{W}} \int_D |u_D - u_{D_0}|^p \mathrm{d}y \right) \\
&\le 2^p \left(c_1^p N \int_G |\nabla u(y)|^p \mathrm{d}y + \sum_{D\in\mathcal{W}} \int_D |u_D - u_{D_0}|^p \mathrm{d}y \right) .
\end{aligned}$$

Let us fix $D \in \mathcal{W}$. We may choose a Poincaré chain $(D_0, D_1, \dots, D_k)$ such that $D_k = D$. Then, by Lemma 18.1

$$\begin{aligned}
&|u_D - u_{D_0}|^p \le \left(\sum_{j=i}^{k} |u_{D_j} - u_{j-1}| \right)^p \\
&\le k^{p-1} \sum_{j=1}^{k} |u_{D_j} - u_{D_j-1}|^p = k^{p-1} \sum_{j=1}^{k} \int_{D_{j-1}\cap D_j} |u_{D_j} - u_{D_j-1}|^p \mathrm{d}y \\
&\le (2k)^{p-1} \sum_{j=1}^{k} \frac{1}{|D_{j-1} \cap D_j|} \Big(\int_{D_{j-1}} |u(y) - u_{D_{j-1}}|^p \mathrm{d}y \\
&+ \int_{D_j} |u(y) - u_{D_j}|^p \mathrm{d}y \Big) \\
&\le (2k)^{p-1} c_1 \sum_{j=1}^{k} \Big(\frac{1}{|D_{j-1}|} \int_{D_{j-1}} |u(y) - u_{D_{j-1}}|^p \mathrm{d}y \\
&+ \frac{1}{|D_j|} \int_{D_j} |u(y) - u_{D_j}|^p \mathrm{d}y \Big) \\
&\le (2k)^{p-1} c_1 \sum_{j=1}^{k} \Big(\frac{\kappa_p(D_{j-1})^p}{|D_{j-1}|} \int_{D_{j-1}} |\nabla u(y)|^p \mathrm{d}y \\
&+ \frac{\kappa_p(D_j)^p}{|D_j|} \int_{D_j} |\nabla u(y)|^p \mathrm{d}y \Big) .
\end{aligned}$$

Thus,

$$|u_D - u_{D_0}|^p \le 2^p k^{p-1} c_1 \sum_{j=0}^{k} \kappa_p(D_j)^p \int_{D_j} |\nabla u(y)|^p \mathrm{d}y.$$

Hence,

$$\begin{aligned}
&\sum_{D\in\mathcal{W}} \int_D |u_D - u_{D_0}|^p \mathrm{d}y \le \sum_{D\in\mathcal{W}} \int_D \left(\sum_{j=1}^{k} |u_{D_j} - u_{D_j-1}| \right)^p \mathrm{d}y \\
&\le 2^p c_1 \sum_{D\in\mathcal{W}} \int_D l(\mathcal{C}(D))^{p-1} |D| \sum_{A\in\mathcal{C}(D)} \kappa_p(A)^p \int_A |\nabla u(y)|^p \mathrm{d}y \\
&= \sum_{A\in\mathcal{W}} \sum_{D\in A(\mathcal{W})} l(\mathcal{C}(D))^{p-1} |D| \kappa_p(A)^p \int_A |\nabla u(y)|^p \mathrm{d}y \\
&\le c_2 \sum_{A\in\mathcal{W}} \int_A |\nabla u(y)|^p \mathrm{d}y \le c_2 N \int_G |\nabla u(y)|^p \mathrm{d}y .
\end{aligned}$$

Namely, in the previous estimate it is possible to change the order of summation:
Let us write $a(D_j) = l(\mathcal{C}(D_j))^{p-1}|D_j|$, $b(D_j) = \kappa_p(D_j)^p \int_{D_j} |\nabla u(y)|^p \mathrm{d}y$ and

$$\chi_{\mathcal{C}(D_i)}(D_j) = \begin{cases} 1 & , \text{ if } D_j \in \mathcal{C}(D_i) \\ 0 & , \text{ otherwise.} \end{cases}$$

Then,

$$\begin{aligned} &\sum_{D\in\mathcal{W}} l(\mathcal{C}(D))^{p-1}|D| \sum_{A\in\mathcal{C}(D)} \kappa_p(A)^p \int_A |\nabla u(y)|^p \mathrm{d}y \\ &= \sum_{i=0}^{\infty} a(D_i) \sum_{j=0}^{\infty} b(D_j)\chi_{\mathcal{C}(D_i)} D_j = \sum_{i=0}^{\infty} b(D_i) \sum_{j=0}^{\infty} a(D_j)\chi_{\mathcal{C}(D_j)}(D_i) \\ &= \sum_{A\in\mathcal{W}} \sum_{D\in A(\mathcal{W})} l(\mathcal{C}(D))^{p-1}|D|\kappa_p(A)^p \int_A |\nabla u(y)|^p \mathrm{d}y\,. \end{aligned}$$

The theorem is proved. □

Example 18.7. Let $0 < h_i \le 1$ and $0 < \delta_{2i} \le 1$ such that $\sum_{i=1}^{\infty} h_i = l < \infty$, $0 < b_1 \le \frac{h_{i+1}}{h_i} \le 1$ and $0 < \delta_{2i} \le h_{2i+1}$. Let $\sum_{i=1}^{k} h_i = d_k$, where $k = 1, 2, \dots$. Let

$$D_{2i-1} = (d_{2i-1} - h_{2i-1}, d_{2i-1}) \times (-\frac{1}{2}h_{2i-1}, \frac{1}{2}h_{2i-1})^{n-1}$$

and

$$P_{2i} = [d_{2i-1}, d_{2i-1} + h_{2i}] \times (-\frac{1}{2}\delta_{2i}, \frac{1}{2}\delta_{2i})^{n-1},$$

where $i = 1, 2, \dots$. Let us define

$$D = \bigcup_{i=1}^{n} (D_{2i-1} \cup P_{2i})\,.$$

Assume that $\delta_{2i} = b_2 h_{2i}^a$, where $a > 1$ and $b_2 > 0$. If $p > (n-1)(a-1)$, then $D \in \mathcal{P}(p)$. Details are provided in [Hu, Chapter 5]. If $p < (n-1)(a-1)$, then $D \notin \mathcal{P}(p)$. Let (u_k), $k = 1, 3, 5, \dots$ be a sequence of piecewise linear continuous functions such that

$$u_k(x) = \begin{cases} h_k^{-\frac{n}{p}} & \text{in } D_k \\ 0 & \text{in } D \setminus \{P_{k-1} \cup D_k \cup P_{k+1}\}. \end{cases}$$

Let us extend u_k into the set $D \cup D_1 \cup (-\frac{h_1}{2}, \frac{h_1}{2})^n$ as odd functions with respect to x_1. Then

$$\int_D |u_{2i-1}(x)|^p \mathrm{d}x \ge \int_{D_{2i-1}} |u_{2i-1}(x)|^p \mathrm{d}x > c_1$$

and

$$\begin{aligned} &\int_D |\nabla u_{2i-1}(x)|^p \mathrm{d}x \\ = \;& 2\int_{P_{2i-2}} |\nabla u_{2i-1}(x)|^p \mathrm{d}x + 2\int_{P_{2i}} |\nabla u_{2i-1}(x)|^p \mathrm{d}x \\ \le \;& c_2 h_{2i-2}^{(n-1)(a-1)-p} + c_3 h_{2i}^{(n-1)(a-1)-p} \to 0, \end{aligned}$$

if $i \to \infty$ and $p < (n-1)(a-1)$.

Also, $D \in \mathcal{P}((n-1)(a-1))$, but for the proof we need a theorem we do present here. We refer to [Hu, 5.9 Remark].

Definition 18.8. Let D be a proper domain in $\mathbb{R}^n$. The quasihyperbolic distance between the points x_1 in D and x_2 in D is defined as

$$k_D(x_1, x_2) = \inf_{\gamma} \int_{\gamma} \frac{\mathrm{d}s}{\mathrm{dist}(x, \partial D)},$$

where the infimum has been taken over all rectifiable paths γ which join the points x_1 and x_2 in D.

Whitney decomposition 18.9. *[St] Let $\mathcal{W}$ be the Whitney decomposition of D. This means that $\mathcal{W}$ is a family of closed dyadic cubes Q such that the interiors of the cubes are pairwise disjoint and*

1. $D = \bigcup_{Q\in\mathcal{W}} Q$
2. $1 \le \frac{\operatorname{dist}(Q,\partial D)}{\operatorname{diam}(Q)} \le 4$
3. $\frac{1}{4} \le \frac{\operatorname{diam}(Q_1)}{\operatorname{diam}(Q_2)} \le 4$ *when* $Q_1 \cap Q_2 \neq \emptyset$.

Also, at most 12^n cubes can touch a fixed cube Q, and for $\sigma \in (1, \frac{5}{4})$ fixed each point of D lies in at most 12^n of the cubes σQ, $Q \in \mathcal{W}$.

Lemma 18.10. *Let $\mathcal{W}$ be a Whitney decomposition of a domain G. Let us fix $Q_0 \in \mathcal{W}$ with $x_0 \in Q_0$. For each $Q \in \mathcal{W}$ there exists a chain $\mathcal{C}(Q)$ which joins the cubes Q and Q_0 such that for every x in Q*

$$l(\mathcal{C}(Q)) \le c(n)k_D(x_0,x) + 1 \le 5c(n)(l(\mathcal{C}(Q)) + 1).$$

If the cubes in the chain $\mathcal{C}(Q)$ are expanded by a factor $\frac{9}{8}$, then the chain $\mathcal{C}(\operatorname{int}\frac{9}{8}Q)$ is a Poincaré chain.

18.1 Conditions for Poincaré domains

Let D be a bounded domain in the Euclidean n- space R^n. Let $\mathcal{W}$ be a Whitney decomposition of D.

(P_1) Let $p \in [1,n)$. Suppose that for some $x_0 \in D$

$$\int_D (k_D(x_0,x)+1)^{p-1} 2^{c(n)(n-p)k_D(x_0,x)}\, dx < \infty$$

for some constant $c(n) < \infty$.

(P_2) Let $p \in [n,\infty)$. Suppose that for some $x_0 \in D$

$$\int_D (k_D(x_0,x))^{p-1}\, dx < \infty\,.$$

(P_3) Let $p \in [1,\infty)$. Suppose that for some constnat $c < \infty$ and $x_0 \in D$ the chains $\mathcal{C}(Q)$ satisfy

$$\sum_{Q\in A(\mathcal{W})} \int_Q (k_D(x_0,x)+1)^{p-1}\, dx \le c \operatorname{diam}(A)^{n-p}$$

for all $A \in \mathcal{W}$.

Theorem 18.11. *If one of the conditions $(P_1), (P_2)$ or (P_3) is satisfied, then D is a p-Poincaré domain.*

References

[AdHe] David R. Adams and Lars Inge Hedberg, *Function Spaces and Potential Theory*. Grundlehren der mathematischen Wissenschaften 314. Springer, Springer-Verlag Berlin Heidelberg, 1996.

[Ad] Robert A. Adams, *Sobolev Spaces*. Academic Press, Inc., Orlando, 1989.

[EdEv] D. E. Edmunds and W. D. Evans, *Hardy Operators, Function Spaces and Embeddings*. Springer, Springer-Verlag Berlin Heidelberg, 2004.

[EvGa] Craig Evans and Ronald F. Gariepy, *Measure Theory and Fine Properties of Functions*. CRC Press, 1992.

[GiTu] David Gilbarg and Neil S. Trudinger, *Elliptic Partial Differential Equations of Second Order*. 2nd Edition. Revised 3rd Printing. Springer-Verlag, Berlin-Heidelberg-New York, 1998.

[Ha] P. Hajlasz, *Sobolev inequalities, truncation method, and John domains*. Papers on analysis, Rep. Univ. Jyväskylä, Dep. Math. Stat., University of Jyväskylä, Jyväskylä (2001), 109–126.

[HK] P. Hajlasz and P. Koskela, *Isoperimetric inequalities and imbedding theorems in irregular domains.* J. London Math. Soc (2) **58**(1998), 425–450.

[Hu] Ritva Hurri, *Poincaré domain in* $\mathbb{R}^n$. Ann. Acad. Sci. Fenn., Series A, I. Math. Dissertationes **71**(1988), 1–42.

[Jo] Peter W. Jones, *Quasiconformal mappings and extendability of functions in Sobolev spaces.* Acta Math., **147**(1981)71 - 88.

[M] O. Martio, *Sobolev avaruudet.* Lecture notes (in Finnish).

[Maz] Vladimir G. Maz'ja, *Sobolev Spaces.* Springer-Verlag Berlin Heidelberg, 1985.

[MazPo] Vladimir G. Maz'ya and Segei V. Poborchi, *Differentiable Functions on Bad Domains.* World Scientific Publishing Co. Pte. Ltd., Singapore, 1997.

[MeSe] Meyers, Norman G. and James Serrin $H = W$. Proc. Nat. Acad. Sci. U. S. A. **51**(1964), 1055–1056.

[St] Elias M. Stein, *Singular Integrals and Differentiability Properties of Functions.* Princeton University Press, Princeton, New Jersey, 1970.

[Zi] William Ziemer, *Weakly Differentiable Functions.* Springer-Verlag, New York Inc., 1989.

Hausdorff measures and dimensions

Maarit Järvenpää
Department of mathematics and statistics, University of Jyväskylä, Finland

Abstract

In fractal geometry various notions of dimensions play an important role. In these lecture notes we discuss the most commonly used concepts of dimension. The emphasis is given to Hausdorff dimensions which are defined in terms of Hausdorff measures.

1 Introduction and notation

The aim of these lecture notes is to consider general features of fractal geometry at a level which is accessible to those who have various mathematical backgrounds. The emphasis is given to different concepts of dimensions, their basic properties and methods for calculating their values. The attempt is to provide mathematical insight into the subject without going into the abstract measure theory. A good overview of the topic can be found from [1], [2], [3], [4] and [5]. This exposition which is based on [3] and [5] contains some crucial definitions, examples and results and it will be used as a basis of my lectures.

In this section we make a rapid survey of some basic definitions and results in measure theory that we will need later. We restrict our consideration to the n-dimensional Euclidean space $\mathbb{R}^n$. We denote by $\operatorname{diam}(E)$ the diameter of a set $E \subset \mathbb{R}^n$, i.e. $\operatorname{diam}(E) = \sup\{|x-y| : x, y \in E\}$ where $|\cdot|$ is the usual Euclidean norm. The closed ball with centre $x \in \mathbb{R}^n$ and with radius $r > 0$ is denoted by $B(x,r)$, i.e. $B(x,r) = \{y \in \mathbb{R}^n : |x-y| \le r\}$. Let $\mathcal{P}(\mathbb{R}^n) = \{A : A \subset \mathbb{R}^n\}$ be the power set of $\mathbb{R}^n$.

Definition 1.1. A function $\mu : \mathcal{P}(\mathbb{R}^n) \to [0,\infty]$ is *a measure* if

(1) $\mu(\emptyset) = 0$,
(2) $\mu(A) \le \mu(B)$ provided that $A \subset B \subset \mathbb{R}^n$,
(3) $\mu(\bigcup_{i=1}^{\infty} A_i) \le \sum_{i=1}^{\infty} \mu(A_i)$ for all $A_i \subset \mathbb{R}^n$.

Contrary to Definition 1.1, in measure theory a measure means usually a non-negative, countably additive set function which is defined in a σ-algebra of $\mathbb{R}^n$. The set function defined in Definition 1.1 is often called an outer measure. There is a close relation between these concepts, see [5].

Definition 1.2. *The support of a measure* μ is the smallest closed set E such that $\mu(\mathbb{R}^n \setminus E) = 0$. It is denoted by $\operatorname{spt}\mu$. We say that μ is a measure on A if $\operatorname{spt}\mu \subset A$.

The following consequence of Fubini's theorem turns out to be be very useful.

Lemma 1.3. *Assume that μ is a measure and $f : \mathbb{R}^n \to [0,\infty)$ is a Borel function. Then*

$$\int f \, d\mu = \int_0^{\infty} \mu(\{x : f(x) \ge t\}) \, dt.$$

Proof. Letting $A = \{(x,t) : f(x) \ge t\}$ and denoting by $\mathcal{L}^1$ the 1-dimensional Lebesgue measure, we have by Fubini's theorem

$$\begin{aligned}\int_0^{\infty} \mu(\{x : f(x) \ge t\}) \, dt &= \int_0^{\infty} \mu(\{x : (x,t) \in A\}) \, dt \\ &= \int \mathcal{L}^1(\{t \in [0,\infty) : (x,t) \in A\}) \, d\mu x \\ &= \int \mathcal{L}^1([0, f(x)]) \, d\mu x = \int f(x) \, d\mu x.\end{aligned}$$

□

2 Hausdorff dimensions and measures

The notion of dimension is crucial in fractal geometry. Among different concepts of dimensions, the Hausdorff dimension is the oldest and the most important one. It measures the metric size of a general set. The Hausdorff dimension has the disadvantage that in many cases it is rather hard to calculate and to estimate. The definition is based on Hausdorff measures.

Definition 2.1. Suppose that $A \subset \mathbb{R}^n$. Let $s \geq 0$ and $\delta > 0$. Define

$$\mathcal{H}^s_\delta(A) = \inf\{\sum_{i=1}^{\infty} \operatorname{diam}(E_i)^s : A \subset \bigcup_{i=1}^{\infty} E_i,\ \operatorname{diam}(E_i) \leq \delta\}$$

and

$$\mathcal{H}^s(A) = \lim_{\delta \to 0} \mathcal{H}^s_\delta(A). \tag{1}$$

The limit in (1) exists since $\mathcal{H}^s_\delta(A) \leq \mathcal{H}^s_\varepsilon(A)$ for all $0 < \varepsilon \leq \delta$. In fact,

$$\mathcal{H}^s(A) = \sup_{\delta > 0} \mathcal{H}^s_\delta(A).$$

It is not difficult to show that equation (1) defines a measure. The measure $\mathcal{H}^s$ is called the *s-dimensional Hausdorff measure*.

For integer values of s the Hausdorff measures generalize the length, area and volume measures. Indeed, it turns out that the n-dimensional Hausdorff measure is a constant multiple of the n-dimensional Lebesgue measure $\mathcal{L}^n$. Moreover, $\mathcal{H}^0(A)$ gives the number of points in A.

One may use more restrictive coverings in the Definition 2.1. Indeed, instead of arbitrary coverings may take closed, open or convex sets [5].

The definition of the Hausdorff dimension is based on the following property of Hausdorff measures.

Proposition 2.2. *Suppose that $A \subset \mathbb{R}^n$. Let $0 \leq s < t$.*

(1) *If $\mathcal{H}^s(A) < \infty$, then $\mathcal{H}^t(A) = 0$.*
(2) *If $\mathcal{H}^t(A) > 0$, then $\mathcal{H}^s(A) = \infty$.*

Proof. We verify Claim 1. Claim (2) is a restatement of Claim 1. Let $\delta > 0$. Assume that $A \subset \cup_{i=1}^{\infty} E_i$ where $\operatorname{diam}(E_i) \leq \delta$ and $\sum_{i=1}^{\infty} \operatorname{diam}(E_i)^s \leq \mathcal{H}^s_\delta(A) + 1$. Then

$$\mathcal{H}^t_\delta(A) \leq \sum_{i=1}^{\infty} \operatorname{diam}(E_i)^t = \sum_{i=1}^{\infty} \operatorname{diam}(E_i)^{t-s} \operatorname{diam}(E_i)^s \leq \delta^{t-s} \sum_{i=1}^{\infty} \operatorname{diam}(E_i)^s \leq \delta^{t-s}(\mathcal{H}^s_\delta(A) + 1).$$

Letting $\delta \to 0$ gives (1). □

By Proposition 2.2 there is a unique value of s at which the s-dimensional Hausdorff measure 'jumps' from infinity to zero. This critical value is called the Hausdorff dimension.

Definition 2.3. *The Hausdorff dimension* of a set $A \subset \mathbb{R}^n$ is defined as follows:

$$\dim_{\mathrm{H}} A = \inf\{s \geq 0 : \mathcal{H}^s(A) = 0\} = \sup\{s \geq 0 : \mathcal{H}^s(A) = \infty\}.$$

For the critical value $s = \dim_{\mathrm{H}} A$, the s-dimensional Hausdorff measure may be zero, infinite or positive and finite. On the other hand, if $0 < \mathcal{H}^s(A) < \infty$, then $s = \dim_{\mathrm{H}} A$.

The Hausdorff dimension satisfies the following properties which follow easily from the definition.
- *Monotonicity.* If $A \subset B$ then $\dim_{\mathrm{H}} A \leq \dim_{\mathrm{H}} B$.
- *Countable stability.* If $A_1, A_2, \cdots \subset \mathbb{R}^n$ then

$$\dim_{\mathrm{H}}(\bigcup_{i=1}^{\infty} A_i) = \sup_i \dim_{\mathrm{H}} A_i.$$

- *Hausdorff dimension does not increase under Lipschitz mappings*, i.e. if $f : A \to \mathbb{R}^m$ is a Lipschitz mapping, that is, there is a constant $0 < L < \infty$ such that

$$|f(x) - f(y)| \leq L|x - y|$$

for all x, $y \in A$, then $\dim_{\mathrm{H}} f(A) \le \dim_{\mathrm{H}} A$.

• *bi-Lipschitz invariance* If $f : A \to \mathbb{R}^m$ is a bi-Lipschitz mapping, i.e. there are constants $0 < L_1 \le L_2 < \infty$ such that

$$L_1|x-y| \le |f(x)-f(y)| \le L_2|x-y|$$

for all x, $y \in A$, then $\dim_{\mathrm{H}} f(A) = \dim_{\mathrm{H}} A$.

• If $A \subset \mathbb{R}^n$ is countable then $\dim_{\mathrm{H}} A = 0$. If $A \subset \mathbb{R}^n$ is open then $\dim_{\mathrm{H}} A = n$.

• For all $A \subset \mathbb{R}^n$ we have $0 \le \dim_{\mathrm{H}} A \le n$.

The Cantor set is one of the best known fractals. It is easy to construct and it illustrates many typical fractal characteristics.

Example 2.4. Let $0 < \lambda < 1/2$. Denote by $I_{0,1}$ the closed unit interval $[0,1]$. At first step of the construction we delete from the middle of $I_{0,1}$ an open interval of length $1-2\lambda$. The remaining two closed intervals of length λ are denoted by $I_{1,1} = [0,\lambda]$ and $I_{1,2} = [1-\lambda, 1]$. We continue this process. Having defined closed intervals $I_{k-1,1}, \dots, I_{k-1,2^{k-1}}$ of length λ^{k-1}, we delete from the middle of each of them an open interval of length $(1-2\lambda)\lambda^{k-1}$. This gives closed intervals $I_{k,1}, \dots, I_{k,2^k}$ of length λ^k. Define

$$C(\lambda) = \bigcap_{k=0}^{\infty} \bigcup_{j=1}^{2^k} I_{k,j}.$$

The Cantor set $C(\lambda)$ is an uncountable compact set without interior points. Moreover, it has Lebesgue measure zero.

We prove that

$$\dim_{\mathrm{H}}(C(\lambda)) = \frac{\log 2}{\log(1/\lambda)}. \tag{2}$$

Since for all k

$$C(\lambda) \subset \bigcup_{j=1}^{2^k} I_{k,j}.$$

we obtain that

$$\mathcal{H}^s_{\lambda^k}(C(\lambda)) \le \sum_{j=1}^{2^k} \operatorname{diam}(I_{k,j}) = (2\lambda^s)^k.$$

Choosing

$$s = \frac{\log 2}{\log(1/\lambda)}$$

gives

$$\mathcal{H}^s(C(\lambda)) = \lim_{k\to\infty} \mathcal{H}^s_{\lambda^k}(C(\lambda)) \le 1,$$

and therefore, $\dim_{\mathrm{H}}(C(\lambda)) \le s$.

For the purpose of proving (2) it suffices to verify that $\mathcal{H}^s(C(\lambda)) \ge 1/4$. Since $C(\lambda)$ is compact the lower bound for the measure follows once we show that

$$\sum_{i=1}^{l} \operatorname{diam}(I_i)^s \ge \frac{1}{4} \tag{3}$$

for open intervals $I_1, I_2, \dots, I_l$ covering the set $C(\lambda)$. Since $C(\lambda)$ has no interior points we may assume that the end points of each I_i are outside $C(\lambda)$. Choose $\delta > 0$ such that the distance from all these end points to $C(\lambda)$ is at least $\delta > 0$. Letting k be so large that $\lambda^k < \delta$, each construction interval $I_{k,j}$ is contained in some interval I_i.

Let I be on open interval and let n be a positive integer. We proceed by showing that

$$\sum_{I_{n,j} \subset I} \operatorname{diam}(I_{n,j})^s \le 4 \operatorname{diam}(I)^s. \tag{4}$$

From (4) we get immediately (3) since

$$4\sum_{i=1}^{l} \operatorname{diam}(I_i) \ge \sum_{i=1}^{l} \sum_{I_{n,j} \subset I_i} \operatorname{diam}(I_{n,j})^s \ge \sum_{j=1}^{2^n} \operatorname{diam}(I_{n,j})^s = 1.$$

Finally, to verify (4) suppose that the open interval I contains construction intervals $I_{n,j}$. Let n_0 be the smallest integer such that I contains a construction interval $I_{n_0,j}$. Clearly, $n_0 \le n$. Let $I_{n_0,j_1}, \dots, I_{n_0,j_p}$ be the construction intervals at step n_0 that intersect I. Then $p \le 4$. Indeed, if this is not the case then I would contain some construction interval $I_{n_o-1,j}$. Hence

$$4\,\mathrm{diam}(I)^s \ge \sum_{m=1}^{p} \mathrm{diam}(I_{n_0,j_m})^s = \sum_{m=1}^{p} \sum_{I_{n,j}\subset I_{n_0,j_m}} \mathrm{diam}(I_{n,j})^s \ge \sum_{I_{n,j}\subset I} \mathrm{diam}(I_{n,j})^s$$

giving (4).

To conclude, we have $1/4 \le \mathcal{H}^s(C(\lambda)) \le 1$ implying that $\dim_{\mathrm{H}}(C(\lambda)) = s$. Similar methods can be used to prove that $\mathcal{H}^s(C(\lambda)) = 1$. For general sets the exact value of the Hausdorff measure in impossible to compute. However, in many cases Hausdorff measures can be estimated by generalizing the above argument.

As an immediate consequence of Example 2.4 we get:

Corollary 2.5. *For all $0 \le s \le 1$ there exists a set $A \subset \mathbb{R}$ with $\dim_{\mathrm{H}} A = s$ and with $0 < \mathcal{H}^s(A) < \infty$.*

3 Box-counting dimensions

Box-counting dimension is one of the most commonly used dimensions. The value of box-counting dimension is relatively easy to compute mathematically and estimate empirically.

Definition 3.1. Let $A \subset \mathbb{R}^n$ be a non-empty bounded set. For all $\delta > 0$ denote by $N(A,\delta)$ the smallest number of sets of diameter at most δ that cover A. *The lower and upper box-counting dimensions* of A are defined as

$$\underline{\dim}_{\mathrm{B}} A = \liminf_{\delta\to 0} \frac{\log N(A,\delta)}{-\log \delta}$$

and

$$\overline{\dim}_{\mathrm{B}} A = \limsup_{\delta\to 0} \frac{\log N(A,\delta)}{-\log \delta}.$$

If these are equal the common value is called *the box-counting dimension* of A and denoted by $\dim_{\mathrm{B}} A$, that is,

$$\dim_{\mathrm{B}} A = \lim_{\delta\to 0} \frac{\log N(A,\delta)}{-\log \delta}.$$

Note that

$$\dim_{\mathrm{H}} A \le \underline{\dim}_{\mathrm{B}} A \le \overline{\dim}_{\mathrm{B}} A. \tag{5}$$

Clearly, $\underline{\dim}_{\mathrm{B}} A \le \overline{\dim}_{\mathrm{B}} A$. Moreover, if $\dim_{\mathrm{H}} A > s$, then

$$1 < \mathcal{H}^s(A) = \lim_{\delta\to 0} \mathcal{H}^s_\delta(A) \le \lim_{\delta\to 0} N(A,\delta)\delta^s$$

giving $\log N(A,\delta) + s\log\delta > 0$ for sufficiently small δ. This implies $s \le \underline{\dim}_{\mathrm{B}} A$. Hence $\dim_{\mathrm{H}} A \le \underline{\dim}_{\mathrm{B}} A$. In general we do not have equalities in (5).

The box-counting dimensions have the following elementary properties which are easily verified:

- *Monotonicity.* If $A \subset B$ then $\underline{\dim}_{\mathrm{B}} A \le \underline{\dim}_{\mathrm{B}} B$ and $\overline{\dim}_{\mathrm{B}} A \le \overline{\dim}_{\mathrm{B}} B$.
- *Finite stability.* The upper box counting dimension is finitely stable, i.e.

$$\overline{\dim}_{\mathrm{B}}\Big(\bigcup_{i=1}^{k} A_i\Big) = \max_i \overline{\dim}_{\mathrm{B}} A_i$$

where $A_1, \dots, A_k \subset \mathbb{R}^n$ are non-empty and bounded sets. The lower box-counting dimension does not satisfy the corresponding property.

- The lower and upper box-counting dimensions do not increase under Lipschitz mappings.
- *bi-Lipschitz invariance.* The lower and upper box-counting dimensions are bi-Lipschitz invariant.
- If $A \subset \mathbb{R}^n$ is open and bounded, then $\dim_{\mathrm{B}} A = n$.
- For all bounded sets $A \subset \mathbb{R}^n$ we have $0 \le \underline{\dim}_{\mathrm{B}} A \le \overline{\dim}_{\mathrm{B}} A \le n$.

There are several equivalent ways to define the box-counting dimension which are stated in Proposition 3.2. For $\delta > 0$ the cubes of the form

$$[k_1\delta, (k_1+1)\delta] \times \cdots \times [k_n\delta, (k_n+1)\delta]$$

where $k_1, \dots, k_n$ are integers are called δ-mesh cubes.

Proposition 3.2. *The lower and upper box-counting dimensions of a non-empty compact set $A \subset \mathbb{R}^n$ are given by*

$$\underline{\dim}_{\mathrm{B}} A = \liminf_{\delta \to 0} \frac{\log M(A,\delta)}{-\log \delta}$$

and

$$\overline{\dim}_{\mathrm{B}} A = \limsup_{\delta \to 0} \frac{\log M(A,\delta)}{-\log \delta}$$

where $M(E,\delta)$ is any of the following:

(1) *the smallest number of sets of diameter at most δ that cover A,*
(2) *the smallest number of closed balls of radius δ that cover A,*
(3) *the smallest number of cubes of side length δ that cover A,*
(4) *the smallest number of δ-mesh cubes that intersect A,*
(5) *the largest number of disjoint balls of radius δ with centers in A.*

Proof. See [3] or [5]. □

The Cantor set is an example of a set for which Hausdorff and box-counting dimensions agree.

Example 3.3. For $0 < \lambda < 1/2$ let $C(\lambda)$ be the Cantor set. Then

$$\dim_{\mathrm{B}} C(\lambda) = \frac{\log 2}{\log(1/\lambda)}.$$

Let $\delta > 0$. Choose a positive integer k such that $\lambda^k < \delta \le \lambda^{k-1}$. Since $C(\lambda)$ can be covered by the 2^k construction intervals $I_{k,j}$ we have $N(C(\lambda),\delta) \le 2^k$. This gives

$$\overline{\dim}_{\mathrm{B}} C(\lambda) = \limsup_{\delta \to 0} \frac{\log N(C(\lambda),\delta)}{-\log \delta} \le \limsup_{k \to \infty} \frac{\log 2^k}{\log \lambda^{k-1}} = \frac{\log 2}{\log(1/\lambda)}.$$

To see that the reverse inequality is valid, choose a positive integer k such that $(1-2\lambda)\lambda^k \le \delta < (1-2\lambda)\lambda^{k-1}$. Then any interval of length at most δ intersects at most one of construction intervals of length λ^k. Since there are 2^k such construction intervals at least 2^k intervals of length at most δ are needed to cover $C(\lambda)$. This gives $N(C(\lambda),\delta) \ge 2^k$ leading to $\underline{\dim}_{\mathrm{B}} C(\lambda) \ge \log 2/\log(1/\lambda)$.

The following proposition indicates a disadvantage of box-counting dimension:

Proposition 3.4. *Let $A \subset \mathbb{R}^n$ be non-empty and bounded. Denoting by $\overline{A}$ the closure of A , i.e. the smallest closed set containing A, we have*

$$\underline{\dim}_{\mathrm{B}} \overline{A} = \underline{\dim}_{\mathrm{B}} A$$

and

$$\overline{\dim}_{\mathrm{B}} \overline{A} = \overline{\dim}_{\mathrm{B}} A.$$

Proof. If closed balls $B_1, \dots, B_k$ of radius δ cover A, i.e. $A \subset \bigcup_{i=1}^k B_i$, then they also cover $\overline{A}$. Denoting by $M(A,\delta)$ the smallest number of closed balls of radius δ that cover A, we get $M(A,\delta) = M(\overline{A},\delta)$. The claim follows. □

An immediate consequence of Proposition 3.4 is that the box-counting dimensions are not countably stable. Indeed, if A is the set of all rational numbers in the closed unit interval $[0,1]$, then $\dim_{\mathrm{B}} E = 1$ even though the box-counting dimension of each singleton is zero. It turns out that similar difficulties remain even if one restricts the attention to closed sets:

Example 3.5. The set $E = \{0, 1, \frac{1}{2}, \frac{1}{3}, \dots\}$ is compact and $\dim_{\mathrm{B}} E = \frac{1}{2}$.

4 Methods for calculating dimensions

In this section we consider some of the basic tools that can be utilized in dimension calculations. The following theorem is often called the Mass distribution principle. A measure μ on a bounded subset of $\mathbb{R}^n$ is called *a mass distribution* if $0 < \mu(\mathbb{R}^n) < \infty$.

Theorem 4.1. *Let $A \subset \mathbb{R}^n$ be bounded and let μ be a measure on A such that $0 < \mu(\mathbb{R}^n) < \infty$. Suppose that for some s there exist $c > 0$ and $\delta_0 > 0$ such that*

$$\mu(E) \leq c \operatorname{diam}(E)^s$$

for all $E \subset \mathbb{R}^n$ with $\operatorname{diam}(E) \leq \delta_0$. Then $\mathcal{H}^s(A) \geq \mu(A)/c$, and moreover,

$$s \leq \dim_{\mathrm{H}} A \leq \underline{\dim}_{\mathrm{B}} A \leq \overline{\dim}_{\mathrm{B}} A.$$

Proof. Let $0 < \delta \leq \delta_0$. Assume that $A \subset \bigcup_{i=1}^{\infty} E_i$ where $\operatorname{diam}(E_i) \leq \delta$ for all i. Then

$$0 < \mu(A) \leq \mu(\bigcup_{i=1}^{\infty} E_i) \leq \sum_{i=1}^{\infty} \mu(E_i) \leq c \sum_{i=1}^{\infty} \operatorname{diam}(E_i)^s.$$

Taking infima gives $\mathcal{H}^s_\delta(A) \geq \mu(A)/c$. Thus $\mathcal{H}^s(A) \geq \mu(A)/c > 0$ implying $\dim_{\mathrm{H}} A \geq s$. □

The Mass distribution principle is an important tool for estimating dimensions from above.

Example 4.2. Consider the $\frac{1}{3}$-Cantor set $C(\frac{1}{3})$. Let μ be the natural mass distribution on $C(\frac{1}{3})$, i.e. at step k each construction interval $I_{k,j}$ of length $\frac{1}{3}$ carries a mass 2^{-k}. Defining for all $A \subset \mathbb{R}$

$$\mu(A) = \inf\{\sum_i \mu(I_{k,i}) : A \subset \bigcup_{k,i} I_{k,i}\}$$

gives a mass distribution on $C(\frac{1}{3})$. We omit the of the proof of this fact.

Let E be a set with $\operatorname{diam}(E) < 1$. Choosing a positive integer k such that $3^{-(k+1)} \leq \operatorname{diam}(E) < 3^{-k}$, the set E intersects at most one of the construction intervals $I_{k,i}$. Since

$$\mu(E) \leq 2^{-k} = (3^{-k})^{\log 2/\log 3} \leq (3 \operatorname{diam}(E))^{\log 2/\log 3}$$

we get from the Mass distribution principle that

$$\mathcal{H}^{\log 2/\log 3}(C(\tfrac{1}{3})) \geq 3^{-\log 2/\log 3} = \tfrac{1}{2}$$

and $\dim_{\mathrm{H}} C(\frac{1}{3}) \geq \log 2/\log 3$.

The following theorem is called Frostman's lemma. In a sense it generalizes the Mass distribution principle and gives a converse of it. For the proof and generalizations of Frostman's lemma see [5].

Theorem 4.3. *Let $A \subset \mathbb{R}^n$ be compact. Then $\mathcal{H}^s(A) > 0$ if and only if there exists a measure μ on A such that $0 < \mu(\mathbb{R}^n) < \infty$ and $\mu(B(x,r)) \leq r^s$ for all $x \in \mathbb{R}^n$ and $r > 0$.*

Potential theoretic methods turn out to be useful when calculating dimensions. Indeed, using Frostman's lemma we may relate Hausdorff dimension to capacities

Definition 4.4. Let $s > 0$. *The s-capacity* of a compact set $A \subset \mathbb{R}^n$ is defined by

$$C_s(A) = \sup\{I_s(\mu)^{-1} : \mu \text{ is a measure on } A \text{ with } \mu(\mathbb{R}^n) = 1\}$$

where

$$I_s(\mu) = \iint |x-y|^{-s} \, d\mu x \, d\mu y$$

is *the s-energy of μ.*

Note that $C_s(A) > 0$ if and only if there is a measure μ on A such that $\mu(\mathbb{R}^n) = 1$ and $I_s(\mu) < \infty$. The Hausdorff dimension is closely related to capacities:

Theorem 4.5. *Suppose that $A \subset \mathbb{R}^n$ is compact. Then*

$$\dim_{\mathrm{H}} A = \sup\{s > 0 : C_s(A) > 0\} = \inf\{s > 0 : C_s(A) = 0\}.$$

Proof. We prove that

$$\dim_{\mathrm{H}} A = \inf\{s > 0 : C_s(A) = 0\}. \tag{6}$$

The remaining equality follows easily from definitions.

Let $s > 0$. For the purpose of verifying (6) we need to show that the following claims hold:

$$C_s(A) = 0 \text{ provided that } \mathcal{H}^s(A) < \infty \tag{7}$$

and

$$\mathcal{H}^t(A) = 0 \text{ for all } t > s \text{ provided that } C_s(A) = 0. \tag{8}$$

We proceed by showing first that (7) holds. Assume to the contrary that $C_s(A) > 0$. Then there is a measure μ on A such that $\mu(\mathbb{R}^n) = 1$ and $I_s(\mu) < \infty$. Thus $\int |x-y|^{-s}\, d\mu y < \infty$ for μ-almost all $x \in \mathbb{R}^n$ giving

$$\lim_{r\to 0} \int_{B(x,r)} |x-y|^{-s}\, d\mu y = 0$$

for μ-almost all $x \in \mathbb{R}^n$. Let $\varepsilon > 0$. Then there are $B \subset A$ and $\delta > 0$ such that $\mu(B) > 1/2$ and

$$\mu(B(x,r)) \le r^s \int_{B(x,r)} |x-y|^{-s}\, d\mu y \le \varepsilon r^s$$

for all $x \in B$ and $0 < r \le \delta$. Let $E_1, E_2, \cdots \subset \mathbb{R}^n$ such that $B \subset \cup_i E_i$, $\operatorname{diam}(E_i) \le \delta$, $E_i \cap B \ne \emptyset$ and

$$\sum_i \operatorname{diam}(E_i)^s \le \mathcal{H}^s_\delta(B) + 1.$$

Choosing $x_i \in E_i \cap B$ and setting $r_i = \operatorname{diam}(E_i)$ gives

$$1/2 < \mu(B) \le \sum_i \mu(B(x_i, r_i)) \le \varepsilon \sum_i r_i^s \le \varepsilon(\mathcal{H}^s(B) + 1) \le \varepsilon(\mathcal{H}^s(A) + 1).$$

Letting $\varepsilon \to 0$ implies that $\mathcal{H}^s(A) = \infty$ proving (7).

For (8), assume that $\mathcal{H}^t(A) > 0$ for some $t > s$. Applying Frostman's lemma we find a measure μ on A such that $0 < \mu(\mathbb{R}^n) < \infty$ and $\mu(B(x,r)) \le r^t$ for all $x \in \mathbb{R}^n$ and $r > 0$. By Lemma 1.3 and a change of variables we get

$$\begin{aligned}
\int |x-y|^{-s}\, d\mu y &= \int_0^\infty \mu(\{y : |x-y|^{-s} \ge u\})\, du \\
&= \int_0^\infty \mu(B(x, u^{-1/s}))\, du \\
&= s \int_0^\infty r^{-s-1} \mu(B(x,r))\, dr \\
&\le s \int_0^1 r^{t-s-1}\, dr + s\mu(\mathbb{R}^n) \int_1^\infty r^{-s-1}\, dr \\
&= \frac{s}{t-s} + \mu(\mathbb{R}^n) < \infty.
\end{aligned}$$

Thus $C_s(A) > 0$ contradicting (8). This completes the proof of (6). □

5 Iterated function systems

Iterated function systems is an important family of fractals for which there is a simple way of calculating dimensions.

Definition 5.1. A mapping $S : \mathbb{R}^n \to \mathbb{R}^n$ is called *a contraction* if there exists $0 < c < 1$ such that

$$|S(x) - S(y)| \le c|x-y|$$

for all x, $y \in \mathbb{R}^n$. Moreover, S is called *a similitude* provided that

$$|S(x) - S(y)| = c|x-y|$$

for all x, $y \in \mathbb{R}^n$. A finite family of contractions $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ is called *an iterated function system.* A non-empty compact set $K \subset \mathbb{R}^n$ is *an attractor* of $\mathcal{S}$ if it is invariant under $\mathcal{S}$, i.e.

$$K = \bigcup_{i=1}^m S_i(K).$$

It turns out that each iterated function system has a unique attractor.

Theorem 5.2. *Suppose that $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ is an iterated function system. There there exists a unique attractor of $\mathcal{S}$.*

Proof. The family of all non-empty compact subsets of $\mathbb{R}^n$ equipped with the Hausdorff metric

$$\rho(E, F) = \max\{\operatorname{dist}(x, F), \operatorname{dist}(y, E) : x \in E, y \in F\}$$

is a complete metric space. Here $\operatorname{dist}(x, F) = \inf\{|x - y| : y \in F\}$ is the distance from x to F. The mapping $F : E \mapsto \cup_{i=1}^{m} S_i(E)$ is a contraction, and therefore, it has a unique fixed point which it the attractor of $\mathcal{S}$. □

Definition 5.3. Let $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ be an iterated function system consisting of similitudes $S_i : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$|S_i(x) - S_i(y)| = c_i|x - y|.$$

for all x, $y \in \mathbb{R}^n$. Then the unique attractor of $\mathcal{S}$ is called *a self-similar set*. We say that $\mathcal{S}$ satisfies *the open set condition* if there is a non-empty bounded open set $O \subset \mathbb{R}^n$ such that

$$\bigcup_{i=1}^{m} S_i(O) \subset O \text{ and } S_i(O) \cap S_j(O) = \emptyset \text{ for } i \neq j.$$

The Cantor set is an example of an iterated function system satisfying the open set condition.

Example 5.4. The middle third Cantor set $C(\frac{1}{3})$ is the attractor of the iterated function system $\{S_1, S_2\}$ where S_1, $S_2 : \mathbb{R}^n \to \mathbb{R}^n$ are given by

$$S_1(x) = \frac{1}{3}x \text{ and } S_2(x) = \frac{1}{3}x + \frac{2}{3}.$$

Hence $C(\frac{1}{3})$ is a self-similar set. Moreover, $\{S_1, S_2\}$ satisfies the open set condition with O as the open unit interval $(0, 1)$.

According to the following theorem, the dimensions of an iterated function system satisfying the open set condition are easy to calculate. For the proof see [3].

Theorem 5.5. *Suppose that $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ is an iterated function system satisfying the open set condition. Let K be the unique attractor of $\mathcal{S}$. Then $\dim_{\mathrm{H}} K = \dim_{\mathrm{B}} K = s$ where s is the unique number satisfying*

$$\sum_{i=1}^{m} c_i^s = 1. \tag{9}$$

Example 5.6. For the middle third Cantor set $C(\frac{1}{3})$ the formula (9) is of the form $2 \cdot 3^s = 1$ giving $\dim_{\mathrm{H}} C(\frac{1}{3}) = \log 2 / \log 3$.

References

[1] G. A. Edgar. Measure, Topology, and Fractal Geometry. *Springer-Verlag*, 1990.

[2] K. J. Falconer. Geometry of fractal sets. *Cambridge University Press*, 1985.

[3] K. J. Falconer. Fractal Geometry. Mathematical Foundations and Applications. Second edition. *Wiley*, 2003.

[4] K. J. Falconer. Techniques in Fractal Geometry. *John Wiley*, 1997.

[5] P. Mattila. Geometry of Sets and Measures in Euclidean Spaces. Fractals and rectifiability. *Cambridge University Press*, 1995.

Continued fractions

Lisa Lorentzen
Norwegian University of Science and Technology

Abstract

This is an introduction to the field of continued fractions, with emphasis on how they are used in function theory. In particular we shall look at these structures as alternatives to series expansions. The question of convergence therefore arises naturally, and so does the question of what they can be used for. We shall touch on both of these aspects. The philosophy of this paper is to present ideas through simple examples. It does not take much imagination to see how these can be extended and generalized.

The analytic theory of continued fractions is a fascinating study with connections to many areas of mathematics. This course is meant to be just a taster of what to expect from these structures. The emphasis will be on convergence questions, both the convergence theory for continued fractions which can be quite involved, and how this often surprisingly good convergence appear in applications.

The paper consists of four different parts. The first part gives an overall picture of what continued fractions really are, and some examples to show why they are of interest. In particular their strong connection to linear fractional transformations is explained. Actually, a continued fraction can be interpreted as a sequence of such transformations. And it is amazing how much one can do with these very simple transformations, and how important they are in so many branches of mathematics.

Continued fractions have many applications, mostly due to their roles as expansions of a given number or function. And as such, their convergence properties are important. The second part goes a little deeper into the convergence theory for linear fractional transformations, although still on an elementary level.

In part 3 we describe some areas where linear fractional transformations; i.e., continued fractions, and their convergence enter the picture. The idea is to show examples of areas where continued fractions are in use. A good thing is that results and ideas from the various fields can be exchanged; i.e., adapted and integrated wherever needed.

A particular topic close to the heart of the author is the equivalence between properties of continued fractions, moment problems and orthogonal polynomials or Laurent polynomials. This is the theme of part 4.

The paper concludes with references to some additional literature and a list of open problems.

1 Continued fractions

1.1 Some basics

1.1.1 Definitions

Let us first agree on what we are talking about. A *continued fraction*

$$b_0 + \mathop{\mathrm{K}}_{n=1}^{\infty} \frac{a_n}{b_n} := b_0 + \mathbf{K}(a_n/b_n) := b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \ddots}}}$$

is an infinite structure which works as an alternative to infinite series. If all $a_n = 1$ and all $b_n \in \mathbb{N}$, we are back to the regular continued fractions in number theory, and if a_n and/or b_n are complex-valued functions, we move into the field of complex function theory. For simplicity we let its *elements* a_n and b_n be complex numbers to start with, and we assume that all $a_n \neq 0$. We write this continued fraction as

$$b_0 + \mathop{\mathrm{K}}_{n=1}^{\infty} \frac{a_n}{b_n} =: b_0 + \frac{a_1}{b_1} + \frac{a_2}{b_2} + \frac{a_3}{b_3} + \cdots .$$

Note where we place the plus signs to indicate the fraction structure, as opposed to series notation.

As for series and infinite products, we need a concept of convergence. The classical way to define convergence of a continued fraction, is to form the *approximants*

$$f_0 := b_0, \quad f_1 := b_0 + \frac{a_1}{b_1}, \quad f_2 := b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2}}, \quad f_3 := b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3}}},$$

and so on. We get f_n by truncating the continued fraction after n fraction terms a_k/b_k. (We shall soon see that the approximants are always well defined in $\widehat{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$.) Then $b_0 + \mathbf{K}(a_n/b_n)$ *converges* to the *value* f of $b_0 + \mathbf{K}(a_n/b_n)$ if and only if $\lim f_n = f$. We also allow convergence to $f = \infty$, and we adopt a tradition from the theory of series and infinite products and write

$$f = b_0 + \mathop{\mathrm{K}}_{n=1}^{\infty} \frac{a_n}{b_n} = b_0 + \mathbf{K}_{n=1}^{\infty}(a_n/b_n) = b_0 + \mathbf{K}(a_n/b_n),$$

both for the continued fraction structure and for its value when it converges.

1.1.2 Connection to linear fractional transformations

We have already claimed that the approximants f_n always exist in $\widehat{\mathbb{C}}$ for a continued fraction $b_0 + \mathbf{K}(a_n/b_n)$ with all $0 \neq a_n \in \mathbb{C}$ and $b_n \in \mathbb{C}$. This can be explained by the following connection between continued fractions and non-singular linear fractional transformations

$$\tau(w) := \frac{aw + b}{cw + d} \quad \text{where } \Delta := ad - bc \neq 0. \tag{1}$$

For the time being, the coefficients a, b, c and d are just complex numbers. We let $\mathcal{M}$ denote the family of these transformations. If we write

$$s_0(w) := b_0 + w \quad \text{and} \quad s_n(w) := \frac{a_n}{b_n + w} \text{ for } n = 1, 2, 3, \ldots,$$

then the condition $a_n \neq 0$ implies that $s_n \in \mathcal{M}$ for all n. Their compositions

$$S_n(w) := s_0 \circ s_1 \circ s_2 \circ \cdots \circ s_n(w) = b_0 + \frac{a_1}{b_1 +} \frac{a_2}{b_2 +} \cdots + \frac{a_n}{b_n + w} \tag{2}$$

therefore also belong to $\mathcal{M}$. Indeed, $\mathcal{M}$ is a group with composition as the group operation. The identity in $\mathcal{M}$ is the identity function $I(w) = w$ ($a = d \neq 0$, $b = c = 0$), and τ_1, $\tau_2 \in \mathcal{M}$ implies that $\tau_1 \circ \tau_2 \in \mathcal{M}$ and τ_1^{-1}, $\tau_2^{-1} \in \mathcal{M}$. Since the mappings from $\mathcal{M}$ are univalent mappings of $\widehat{\mathbb{C}}$ onto $\widehat{\mathbb{C}}$, this means that the approximants $f_n = S_n(0)$ are always well defined in $\widehat{\mathbb{C}}$.

1.1.3 Tail sequences

A classical approximant f_n is obtained by truncating the continued fraction after n fraction terms. The part we cut away,

$$f^{(n)} := \frac{a_{n+1}}{b_{n+1}} + \frac{a_{n+2}}{b_{n+2}} + \frac{a_{n+3}}{b_{n+3}} + \cdots \tag{3}$$

is called the nth *tail* of $b_0 + \mathbf{K}(a_n/b_n)$. This is also a continued fraction, and it converges if and only if $b_0 + \mathbf{K}(a_n/b_n)$ converges. Indeed, (3) converges to $f^{(n)}$ if and only if $b_0 + \mathbf{K}(a_n/b_n)$ converges to

$$f := S_n(f^{(n)}). \tag{4}$$

The sequence $\{f^{(n)}\}$ is then called *the sequence of tail values* for $b_0 + \mathbf{K}(a_n/b_n)$.

Definition 1.1. For every $t \in \widehat{\mathbb{C}}$, the sequence

$$t_n := S_n^{-1}(t) \quad \text{for } n = 0, 1, 2, \ldots \tag{5}$$

is called a tail sequence for $\mathbf{K}(a_n/b_n)$ or for $\{S_n\}$.

Properties:

1. The sequence $\{f^{(n)}\}$ of tail values for a convergent continued fraction $\mathbf{K}(a_n/b_n)$ is evidently an example of a tail sequence.

2. Since $S_n = s_1 \circ s_2 \circ \cdots \circ s_n$, we have

$$t_{n-1} = s_n(t_n) \text{ for } n = 1, 2, 3, \ldots . \tag{6}$$

3. The asymptotic behavior of tail sequences $\{t_n\}$ actually determines the convergence properties of $\mathbf{K}(a_n/b_n)$.

A tail sequence can also be expressed as a kind of „reversed approximants":

Theorem 1.2. *Let $\{t_n\}$ be a tail sequence for $\mathbf{K}(a_n/b_n)$. Then*

$$\begin{aligned} t_n &= S_n^{-1}(t_0) = s_n^{-1} \circ s_{n-1}^{-1} \circ \cdots \circ s_1^{-1}(t_0) \\ &= -\left\{ b_n + \frac{a_n}{b_{n-1}} + \frac{a_{n-1}}{b_{n-2}} + \cdots + \frac{a_2}{b_1} + \frac{a_1}{(-t_0)} \right\} \text{ for } n \geq 1 . \end{aligned}$$

Proof. The result follows since

$$s_k^{-1}(w) = -b_k + \frac{a_k}{w} = -\left\{ b_k + \frac{a_k}{(-w)} \right\} \quad \text{for } k \geq 1 .$$

□

The use of tail sequences to describe the action of a continued fraction was introduced and advocated in a number of papers by Waadeland and Jacobsen.

1.2 Why continued fractions?

Now, what can we do with these continued fractions? Quite a lot, actually! In this section we shall see a few examples of applications in function theory. This means that we allow $\{a_n\}$ and $\{b_n\}$ to depend on a complex variable z. By the way, this is the reason why we always use w to denote the variable in transformations from $\mathcal{M}$.

1.2.1 Computation of functions.

Example 1.3. The principal branch $\mathrm{Ln}(1+z)$ of the natural logarithm $\ln(1+z)$ has the continued fraction expansion

$$\mathop{\mathrm{K}}_{n=1}^{\infty} \frac{a_n z}{1} = \frac{z}{1} + \frac{z/2}{1} + \frac{z/6}{1} + \frac{2z/6}{1} + \frac{2z/10}{1} + \frac{3z/10}{1} + \cdots \tag{7}$$

where

$$a_{2k} := \frac{k}{2(2k-1)} , \qquad a_{2k+1} := \frac{k}{2(2k+1)} .$$

This can be derived from the Taylor series

$$\mathrm{Ln}(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \cdots . \tag{8}$$

What we do is to expand the approximants f_n of $\mathbf{K}(a_n z/1)$ in Taylor series and require that they coincide with (8) as far out as possible. There exist several algorithms for deriving such a continued fraction expansion, depending on the desired form of the continued fraction. The form $\mathbf{K}(a_n z/1)$ with all $a_n > 0$, as used here, is called a Stieltjes fraction.

The continued fraction (7) converges locally uniformly to $\mathrm{Ln}(1+z)$ for z in the cut plane $D_{-1} := \{z \in \mathbb{C};\, |\arg(1+z)| < \pi\}$. It can't get any better than this, since the approximants of $\mathbf{K}(a_n z/1)$ are rational functions, whereas $\ln(1+z)$ has logarithmic branch points at -1 and ∞. Hence we need a branch cut connecting -1 and ∞. The continued fraction has chosen the most "economical one"; i.e., the branch cut along the real half line $(\infty, -1]$.

For comparison we note that the Taylor series diverges outside the closed unit disk $\overline{\mathbb{D}}$. But not only is the domain of convergence for the continued fraction much, much larger than the one for the series – the continued fraction converges faster than the series in their common convergence disk. In the extreme case $z = 1$, their value is

$$\mathrm{Ln}(2) = .69314718, \quad \text{correctly rounded in the 8th place.}$$

The first approximants of the continued fraction are

$$f_1 = 1.000000,\ \ f_2 = .666667,\ \ f_3 = .700000,\ \ f_4 = .692308,$$
$$f_5 = .693333,\ \ f_6 = .693121,\ \ f_7 = .693152$$

In order to get the polynomial approximation with the same accuracy as f_7, we need $n > 100\,000$ terms of the power series. Of course, the convergence of the power series can be improved by for instance using the average value of two consecutive partial sums as an approximation to $\ln(1+z)$, but also the convergence of $\mathbf{K}(a_n z/1)$ can be accelerated by simple means: Since $a_n \to \frac{1}{4}$ and the periodic continued fraction $\mathbf{K}(\frac{1}{4}z/1)$ converges in D_{-1} with value

$$g(z) := \frac{\frac{1}{4}z}{1} \underset{+}{} \frac{\frac{1}{4}z}{1} \underset{+}{} \frac{\frac{1}{4}z}{1} \underset{+}{} \cdots = \frac{\frac{1}{4}z}{1+g(z)}; \quad \text{i.e., } g(z) = -\frac{1}{2} + \frac{1}{2}\sqrt{1+z},\ \ \mathrm{Re}\sqrt{\ldots} \geq 0,$$

we can replace the classical approximants

$$f_n(z) = \frac{a_1 z}{1} \underset{+}{} \frac{a_2 z}{1} \underset{+}{} \cdots \underset{+}{} \frac{a_n z}{1}$$

by

$$S_n(g(z)) = \frac{a_1 z}{1} \underset{+}{} \frac{a_2 z}{1} \underset{+}{} \cdots \underset{+}{} \frac{a_n z}{1} \underset{+}{} \frac{\frac{1}{4}z}{1} \underset{+}{} \frac{\frac{1}{4}z}{1} \underset{+}{} \cdots = \frac{a_1 z}{1} \underset{+}{} \cdots \underset{+}{} \frac{a_n z}{1+g(z)}$$

which converges considerably faster to $\mathrm{Ln}(1+z)$ than $f_n(z)$ for $z \in D_{-1}$. Even faster convergence can be obtained by better approximations to the value of the tail

$$\frac{a_{n+1}z}{1} \underset{+}{} \frac{a_{n+2}z}{1} \underset{+}{} \frac{a_{n+3}z}{1} \underset{+}{} \cdots .$$

(For a number of ideas on how to approximate the value of a convergent tail, see for instance [17] or [22, Section 5.1, p.218 ff].)

1.2.2 Summation method for divergent series.

Example 1.3 already showed that replacing the Taylor series by a continued fraction expansion can work as a summation method, since the continued fraction extended the domain of convergence considerably. Here we shall look at an even more dramatic example. We start with a series which diverges for all z. We convert this series into a continued fraction and look at what we then have got:

Example 1.4. The series

$$L(z) := \sum_{n=0}^{\infty} n!(-z)^{-n} = 1 - \frac{1!}{z} + \frac{2!}{z^2} - \frac{3!}{z^3} + \cdots$$

diverges for all $z \in \mathbb{C}$. Let us convert this series into a continued fraction of the form

$$z\,\mathbf{K}(a_n z^{-1}/1) = \frac{a_1}{1} \underset{+}{} \frac{a_2/z}{1} \underset{+}{} \frac{a_3/z}{1} \underset{+}{} \frac{a_4/z}{1} \underset{+}{} \cdots .$$

By purely formal operations we get

$$\begin{aligned}
L(z) &= 1 - \frac{1!}{z} + \frac{2!}{z^2} - \cdots = \frac{1}{\dfrac{1}{L(z)}} = \frac{1}{1 + \dfrac{1}{z} - \dfrac{1}{z^2} + \dfrac{3}{z^3} - \cdots} \\
&= \frac{1}{1 + \dfrac{1/z}{\dfrac{1}{1 - \dfrac{1}{z} + \dfrac{3}{z^2} - \cdots}}} = \frac{1}{1} \underset{+}{} \frac{1/z}{1 + \dfrac{1}{z} - \dfrac{2}{z^2} + \cdots} \\
&= \frac{1}{1} \underset{+}{} \frac{1/z}{1 + \dfrac{1/z}{\dfrac{1}{1 - \dfrac{2}{z} + \cdots}}}
\end{aligned}$$

which eventually leads to the continued fraction

$$\frac{1}{1} \underset{+}{} \frac{1/z}{1} \underset{+}{} \frac{1/z}{1} \underset{+}{} \frac{2/z}{1} \underset{+}{} \frac{2/z}{1} \underset{+}{} \frac{3/z}{1} \underset{+}{} \frac{3/z}{1} \underset{+}{} \cdots \underset{+}{} \frac{n/z}{1} \underset{+}{} \frac{n/z}{1} \underset{+}{} \cdots.$$

This continued fraction is known to converge locally uniformly in the cut plane $D_0 := \{z \in \mathbb{C};\ |\arg z| < \pi\}$ to the analytic function

$$f(z) = \int_0^\infty \frac{ze^{-t}}{z+t} dt. \tag{9}$$

We have thereby given the divergent series a value in D_0. The question is now: Does this value $f(z)$ have anything to do with the series?

The answer is YES! The series $L(z)$ is an asymptotic expansion of $f(z)$ in angular openings $|\arg z| \le \alpha < \frac{\pi}{2}$; that is, for each fixed $n \in \mathbb{N}$ and $0 < \alpha < \frac{\pi}{2}$,

$$\lim_{z\to\infty,\, |\arg z| \le \alpha} |f(z) - \sigma_n(z)| = 0 \quad \text{where} \quad \sigma_n(z) := \sum_{k=0}^{n} (-1)^k \frac{k!}{z^k}.$$

To see this we observe that

$$\begin{aligned}
f(z) &= \int_0^\infty e^{-t} \cdot \frac{1}{1 - (-t/z)}\, dt \\
&= \int_0^\infty e^{-t} \left(1 - \frac{t}{z} + \frac{t^2}{z^2} - \cdots + \frac{(-t)^n}{z^n} + \frac{(-t)^{n+1}}{z^n(z+t)}\right) dt \\
&= 1 - \frac{1!}{z} + \frac{2!}{z^2} - \frac{3!}{z^3} + \cdots + (-1)^n \frac{n!}{z^n} + \int_0^\infty e^{-t} \frac{(-t)^{n+1}}{z^n(z+t)}\, dt,
\end{aligned}$$

and thus

$$|f(z) - \sigma_n(z)| \le \frac{1}{|z|^n} \int_0^\infty \frac{e^{-t} t^{n+1}}{\operatorname{Re}(z+t)}\, dt \le \frac{(n+1)!}{|z|^{n+1} \cos\alpha} \to 0 \ \text{ as } \ |z| \to \infty.$$

(Remember, the gamma function $\Gamma(z)$ is defined by

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$$

and $\Gamma(n) = (n-1)!$.)

Not every sequence $L(z) = \sum (-1)^n c_n z^{-n}$ can be converted to a continued fraction of the form $z\,\mathbf{K}(a_n z^{-1}/1)$, not to mention a continued fraction of this form with only positive coefficients a_n. But if we are in luck, just as in Example 2, then

- the two sequences $\{f_{2n}(z)\}$ and $\{f_{2n+1}(z)\}$ of approximants for $z\,\mathbf{K}(a_n z^{-1}/1)$ converge locally uniformly in the cut plane $D_0 := \{z \in \mathbb{C};\ |\arg z| < \pi\}$ to some analytic functions $f(z)$ and $\tilde{f}(z)$.
- $L(z)$ is an asymptotic expansion of both $f(z)$ and $\tilde{f}(z)$, and thus of convex combinations of f and $\tilde{f}$; i.e., $\beta f + (1-\beta)\tilde{f}$ for some $0 \le \beta \le 1$.

- $f(z) = \tilde{f}(z)$ if and only if

$$\sum_{n=1}^{\infty} \frac{a_1 a_3 \cdots a_{2n-1}}{a_2 a_4 \cdots a_{2n}} + \sum_{n=1}^{\infty} \frac{a_2 a_4 \cdots a_{2n}}{a_1 a_3 \cdots a_{2n+1}} = \infty. \tag{10}$$

The convergence is actually a consequence of the Parabola Theorem which we partly prove in Section 2.5.

1.2.3 Solving moment problems

Moment problems come in many different forms. As an example at this stage we settle for the following one due to Stieltjes [29]:

Stieltjes Moment Problem For a given sequence $\{c_n\}_{n=0}^{\infty}$ of positive numbers, find a real non-decreasing function $\Psi(t)$ on $\mathbb{R}^+$ with infinitely many points of increase, such that

$$c_n := \int_0^{\infty} t^n d\Psi(t) \text{ for all } n.$$

It is customary to require that Ψ is normalized so that

$$c_0 = \int_0^{\infty} d\Psi(t) = 1.$$

An integral of the form

$$\int_a^b g(t)\, d\Psi(t)$$

is called a *Stieltjes integral*. It is defined just like a Riemann integral, except that the Riemann sum is replaced by

$$S_{\mathcal{P},\mathcal{S}} := \sum_{k=1}^{n} g(t_k^*)(\Psi(t_k) - \Psi(t_{k-1}))$$

where $\mathcal{P}$ is the partition $a = t_0 < t_1 < \cdots < t_n = b$ and $\mathcal{S}$ the selection $t_k^* \in [t_{k-1}, t_k]$.

In more modern language we say that $d\Psi(t)$ is a *positive measure* with *infinite support* on $\mathbb{R}^+$, and if $c_0 = 1$, then $d\Psi(t)$ is a *probability measure* on $\mathbb{R}^+$.

Example 1.5. Let $\Psi(t)$ have a continuous derivative on the finite interval $[a, b]$. Then we say that the measure $d\Psi(t)$ is *absolutely continuous* with finite support. In this case

$$\int_a^b g(t)\, d\Psi(t) = \int_a^b g(t)\Psi'(t)\, dt. \quad \diamondsuit$$

Example 1.6. Let $\Psi(t)$ be the „infinite step function"

$$\Psi(t) := \begin{cases} 0 & \text{for } 0 \le t \le t_1 \\ p_1 & \text{for } t_1 < t \le t_2 \\ p_1 + p_2 & \text{for } t_2 < t \le t_3 \\ \cdots & \\ P & \text{for } t \ge \lim t_n \end{cases}$$

for some positive numbers p_n where $P := \sum_{n=1}^{\infty} p_n < \infty$. Then

$$\int_a^b g(t)\, d\Psi(t) = \sum_{n=1}^{\infty} g(t_n) p_n$$

when this sum converges.

Of course, such a function Ψ does not always exist, but if it exists and can be found, at least approximately, then

$$L(z) := \sum_{n=0}^{\infty} (-1)^n \frac{c_n}{z^n} = \sum_{n=0}^{\infty} \int_0^{\infty} \frac{(-1)^n}{z^n}\, t^n d\Psi(t),$$

so, if we could interchange $\sum$ and $\int$, then

$$L(z) \sim \int_0^{\infty} \sum_{n=0}^{\infty} \left(\frac{-t}{z}\right)^n d\Psi(t) = \int_0^{\infty} \frac{z\, d\Psi(t)}{z+t} =: f(z).$$

So in some sense,

$$f(z) := \int_0^{\infty} \frac{z\, d\Psi(t)}{z+t} \sim L(z) := \sum_{n=0}^{\infty} (-1)^n \frac{c_n}{z^n}\,. \tag{11}$$

Properties:

- The Stieltjes moment problem has a solution if and only if $L(z)$ can be brought to a continued fraction of the form $z\,\mathbf{K}(a_n z^{-1}/1)$ where $a_n > 0$ for all n.
- The solution is unique (up to an additive constant) if and only if this continued fraction converges for some $z > 0$. It turns out that in this case this continued fraction $z\mathbf{K}(a_n z^{-1}/1)$ converges locally uniformly in the cut plane $D_0 := \{z \in \mathbb{C};\, |\arg z| < \pi\}$ to the analytic function $f(z)$ in (11).
- The distribution function $\Psi(t)$ can be derived from $f(z)$.
- If $z\mathbf{K}(a_n z^{-1}/1)$ fails to converge, then $\{f_{2n}(z)\}$ and $\{f_{2n+1}(z)\}$ still converge locally uniformly to analytic functions $f(z)$ and $\tilde{f}(z)$ in D_0. In this case we get two corresponding distribution functions $\Psi(t)$ and $\tilde{\Psi}(t)$, and every convex combination $\beta f(z) + (1-\beta)\tilde{f}(z)$ with $\beta \in [0,1]$ gives a new solution $\beta\Psi(t) + (1-\beta)\tilde{\Psi}(t)$ of the moment problem.

This was proved by Stieltjes in his famous 1894-paper [29]. The particular Stieltjes integral (11) is called the *Stieltjes transform* of the measure $d\Psi(t)$.

Note that this means that every function $f(z)$ of the form (11) has a continued fraction expansion of the form $z\mathbf{K}(a_n z^{-1}/1)$ with all $a_n > 0$. Similarly, every function

$$f(z) := \int_0^{\infty} \frac{d\Psi(t)}{1+tz}$$

has a continued fraction expansion of the form $\frac{1}{z}\mathbf{K}(a_n z/1)$ with all $a_n > 0$. This continued fraction converges locally uniformly in D_0 if and only if it converges at a point $z_0 \in D_0$, which happens if and only if (10) holds. A second criterion is due to Carleman [4], [3]:

Theorem 1.7. *Let $L(z) := \sum_{n=0}^{\infty}(-1)^n c_n z^n$ with all $c_n > 0$ have a continued fraction expansion of the form $\frac{1}{z}\mathbf{K}(a_n z/1)$ with all $a_n > 0$. Then $\frac{1}{z}\mathbf{K}(a_n z/1)$ converges if $\sum c_n^{-1/2n} = \infty$.*

There is also a third method based on properties of the distribution function $\Psi(t)$ in the Stieltjes integral (11). If $\Psi(t)$ has a continuous derivative such that $d\Psi(t) = \Psi'(t)\,dt$, then conditions on $\Psi'(t)$ makes it possible to conclude that $\mathbf{K}(a_n z/1)$ is limit periodic with a certain limit az; i.e., $a_n \to a$. From this we may conclude convergence of $\mathbf{K}(a_n z/1)$ and even accelerate its convergence.

The conditions are placed on the function

$$Q(t) := -\mathrm{Ln}(t\Psi'(t^2)),$$

and they are:

- there exist numbers $M > 0$ and $\varepsilon > 0$ such that $|tQ'(t)| \le M$ for $0 < t < \varepsilon$,
- there exist numbers $k > 0$ and $B > 0$ such that for all $t > k$

$$Q'(t) > 0 \quad \text{and} \quad \Big|\frac{t^2 Q'''(t)}{Q'(t)}\Big| \le B,$$

- $tQ''(t)/Q'(t)$ approaches a limit as $t \to \infty$,
- there exist numbers $\alpha > 0$, $\delta > 0$ and $c > 0$ such that

$$Q'(t) = c\, t^{\alpha-1}(1 + \mathcal{O}(t^{-\delta}) \quad \text{as } t \to \infty.$$

The following was then proved by Jones and Van Assche [14] based on results from [23]:

Theorem 1.8. *Let $f(z)$ be a Stieltjes transform satisfying the conditions described above. Then its corresponding S-fraction expansion $z^{-1}\mathbf{K}(a_n z/1)$ has coefficients*

$$a_n \sim a\, n^{2/\alpha} \ \text{ as } t \to \infty \ \text{ where } a := \frac{1}{4}\Big(\frac{\alpha\sqrt{\pi}\,\Gamma(\frac{\alpha}{2})}{c\,\Gamma(\frac{\alpha+1}{2})}\Big)^{2/\alpha}.$$

For the particular value $\alpha = 2$, the value a reduces to $a = 1/c$ since $\Gamma(1) = 1$ and $\Gamma(\frac{3}{2}) = \frac{1}{2}\Gamma(\frac{1}{2}) = \frac{1}{2}\sqrt{\pi}$.

The connection to Example 1.4 is clear: $f(z)$ in (9) is a Stieljes integral of the measure $d\Psi(t) = e^{-t}dt$, and thus $\Psi(t) = C - e^{-t}$ for some arbitrary real constant C. Hence this $\Psi(t)$ is the unique solution (up to an additive constant) of the Stieltjes moment problem with given $c_n = n!$.

We shall go in more detail on this important application of continued fractions in Part 3 and 4, where we also include the connection to the sequence of orthogonal polynomials with respect to the measure $d\Psi(t)$.

1.2.4 Sequences from $\mathcal{M}$

Let $\{\tau_n\}$ be a sequence from $\mathcal{M}$. We can use continued fractions to determine convergence properties for $\{\tau_n\}$. This follows from the following observations:

1. For a given sequence $\{\tau_n\}$ from $\mathcal{M}$, we can let $\varphi_n := \tau_{n-1}^{-1} \circ \tau_n$ with $\tau_0 := I$ to get

$$\tau_n = \varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_n \quad \text{for all } n. \tag{12}$$

2. Similarly, with $\psi_n := \tau_n \circ \tau_{n-1}^{-1}$, $\psi_0 := I$, we get

$$\tau_n = \psi_n \circ \psi_{n-1} \circ \cdots \circ \psi_1 \quad \text{for all } n. \tag{13}$$

3. Moreover,

$$\begin{aligned} (\varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_n)^{-1} &= \varphi_n^{-1} \circ \varphi_{n-1}^{-1} \circ \cdots \circ \varphi_1^{-1} \\ \text{and} \quad (\psi_n \circ \psi_{n-1} \circ \cdots \circ \psi_1)^{-1} &= \psi_1^{-1} \circ \psi_2^{-1} \circ \cdots \circ \psi_n^{-1}, \end{aligned} \tag{14}$$

   something which also is exploited in tail sequences $\{S_n^{-1}(t_0)\}$ for arbitrary $t_0 \in \widehat{\mathbb{C}}$ in continued fraction theory.

4. If $a_n \neq 0$, then

$$\varphi_n(w) = \frac{a_n w + b_n}{c_n w + d_n} = \tilde{\varphi}_n \circ \varphi_n^*(w) \quad \text{where } \tilde{\varphi}_n(w) := \frac{a_n}{c_n + w},\ \varphi_n^*(w) := \frac{\Delta_n}{a_n w + b_n}$$

   where $\Delta_n = a_n d_n - b_n c_n$ is the determinant for φ_n.

5. If $\{\vartheta_n\}$ and $\{\varphi_n\}$ are two sequences from $\mathcal{M}$, and we define $\tilde{\varphi}_n := \vartheta_{n-1}^{-1} \circ \varphi_n \circ \vartheta_n$ for all n, then

$$\tilde{\varphi}_1 \circ \tilde{\varphi}_2 \circ \cdots \circ \tilde{\varphi}_n \circ = \vartheta_0^{-1} \circ \varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_n \circ \vartheta_n.$$

1.2.5 Continued fractions and beauty

The regular continued fraction expansion for π gives an infinite representation of π, but no-one can really claim that it is beautiful. The same can be said for the decimal representation of π. In a way this is strange and not very satisfactory. A number like π, the ratio between the circumference and the diameter of the most perfect object, the circle, ought to be beautiful! And it is! Here is one way to write π:

$$\pi = \frac{4}{1} + \frac{1^2}{3} + \frac{2^2}{5} + \frac{3^2}{7} + \frac{4^2}{9} + \cdots .$$

The pattern stays all the way through the continued fraction.

While we are at it, we can also mention that Euler's number e, the base for the natural logarithm, which also ought to have a beautiful representation, can be written for instance as

$$e = 2 + \frac{2}{2} + \frac{3}{3} + \frac{4}{4} + \frac{5}{5} + \frac{6}{6} + \cdots .$$

This number even has a beautiful regular continued fraction expansion

$$e = 2 + \frac{1}{1} + \frac{1}{2} + \frac{1}{1} + \frac{1}{1} + \frac{1}{4} + \frac{1}{1} + \frac{1}{1} + \frac{1}{6} + \cdots .$$

Also here the pattern stays all through the continued fraction.

1.3 Some basic theory

Before we go on to study convergence properties of continued fractions, it is useful to know some standard and basic continued fraction theory.

1.3.1 Recurrence relations

With basis in the linear fractional transformations

$$s_0(w) := b_0 + w, \qquad s_n(w) := \frac{a_n}{b_n + w} \text{ for } n = 1, 2, 3, \ldots$$

and

$$S_n(w) := b_0 + \frac{a_1}{b_1} + \frac{a_2}{b_2} + \cdots + \frac{a_n}{b_n + w} = s_0 \circ s_1 \circ s_2 \circ \cdots \circ s_n(w) \tag{15}$$

connected to a continued fraction $\mathbf{K}(a_n/b_n)$, we immediately find that

$$S_n(\infty) = S_{n-1}(0),$$

so S_n has the following representation:

Lemma 1.9. *Let S_n be given by (15). Then*

$$S_n(w) = \frac{A_{n-1}w + A_n}{B_{n-1}w + B_n} \quad \textit{for } n = 1, 2, 3, \ldots$$

where

$$A_n = b_n A_{n-1} + a_n A_{n-2}, \qquad B_n = b_n B_{n-1} + a_n B_{n-2} \tag{16}$$

with initial values $A_{-1} = 1$, $A_0 = b_0$, $B_{-1} = 0$ and $B_0 = 1$.

Proof. It is clear that

$$S_0(w) = b_0 + w = \frac{b_0 + w}{1 + 0w}, \qquad S_1(w) = b_0 + \frac{a_1}{b_1 + w} = \frac{b_0 b_1 + a_1 + b_0 w}{b_1 + w},$$

so (16) holds for $n = 1$. To see that it holds for general $n \in \mathbb{N}$, we observe that

$$S_n(w) = S_{n-1}(s_n(w)) = \frac{A_{n-1} + A_{n-2}\frac{a_n}{b_n + w}}{B_{n-1} + B_{n-2}\frac{a_n}{b_n + w}} = \frac{(b_n + w)A_{n-1} + a_n A_{n-2}}{(b_n + w)B_{n-1} + a_n B_{n-2}}. \tag{17}$$

□

A_n and B_n are called the nth *canonical numerator* and *denominator* of $b_0 + \mathbf{K}(a_n/b_n)$, or just its nth *numerator* and *denominator* for short. These names are quite natural since

$$f_n = S_n(0) = A_n/B_n\,.$$

If a_n depend on a complex variable z, but $a_n(z) \neq 0$ for all n and all z in some set D, then $A_n(z)$ and $B_n(z)$ have no common zeros. This follows since $f_n(z)$ is a well defined element in $\widehat{\mathbb{C}}$ for each $z \in D$.

The additive term b_0 is not so important in convergence investigations of continued fractions, and is often set equal to zero. Then $s_0(w) \equiv w$, and $S_n = s_1 \circ s_2 \circ \cdots \circ s_n$.

1.3.2 A uniqueness question

The representation

$$\tau(w) = \frac{aw+b}{cw+d}$$

of $\tau \in \mathcal{M}$ is not unique. For an arbitrary complex number $r \neq 0$ we have

$$\tau(w) = \frac{aw+b}{cw+d} = \frac{arw+br}{crw+dr},$$

so there is a whole equivalence class of representations for τ. In particular, the determinant Δ for τ varies with r. In function theory one often requires that $\Delta := 1$ to obtain (almost) uniqueness of the representation (one can still use $r = -1$). In classical continued fraction theory we have traditionally gone the other way. By picking out the *canonical representation* in Lemma 1.9, we have identified τ with the corresponding matrix

$$M_\tau := \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

and distinguished between different matrices, and thus between different representations, even though they correspond to the same *function* τ. In this setting we have to be careful when we define compositions of mappings from $\mathcal{M}$. For $\tau_k := (a_k w + b_k)/(c_k w + d_k)$, the composition $\tau_1 \circ \tau_2$ shall mean

$$\tau_1 \circ \tau_2(w) := \frac{a_1(a_2w+b_2)+b_1(c_2w+d_2)}{c_1(a_2w+b_2)+d_1(c_2w+d_2)}$$

as in (17). Then $M_{\tau_1 \circ \tau_2}$ is equal to the matrix product $M_{\tau_1} M_{\tau_2}$.

The reason for this choice is that it gives a close connection between continued fractions, recurrence relations of the form (16) and matrices. This is useful in applications, such as for instance for orthogonal polynomials.

Classical convergence theory for continued fractions was mainly based on manipulation of the recurrence relations (16). In more modern theory the representation of S_n is no longer so important, since we often base the analysis on the mapping properties of $S_n \in \mathcal{M}$. Still, we want to combine classical results with the newer ideas, so we shall at least have the canonical representation for S_n as an option.

2 Convergence of sequences from $\mathcal{M}$

2.1 Introduction.

Let $\{\tau_n\}$ be a sequence from $\mathcal{M}$. What kind of convergence can we expect for such a sequence? It is natural to regard the compact set $\widehat{\mathbb{C}}$ as the Riemann sphere, and use the chordal metric

$$\mathfrak{m}(w_1, w_2) := \frac{2|w_1 - w_2|}{\sqrt{1+|w_1|^2}\sqrt{1+|w_2|^2}}$$

in $\widehat{\mathbb{C}}$. ($\mathfrak{m}(w_1, w_2)$ takes the natural limit values if w_1 and/or w_2 are $= \infty$.) This bounded metric has the useful property that $w_n \to w$ in $\widehat{\mathbb{C}}$ if and only if $\mathfrak{m}(w, w_n) \to 0$, something which is not true in general in the

euclidean metric since $w_n \to \infty$ is not equivalent to $|w_n - \infty| \to 0$. On the Riemann sphere equipped with the chordal metric, the point at ∞ is treated just like any other point on $\widehat{\mathbb{C}}$. In fact,

$$\mathfrak{m}(w_1, w_2) = \mathfrak{m}\left(\frac{1}{w_1}, \frac{1}{w_2}\right). \tag{18}$$

We shall consider *uniform* convergence in $\widehat{\mathbb{C}}$ with respect to the chordal metric. This is what we call $\mathfrak{m}$-*uniform convergence*. With the standard metric

$$\sigma(f_1, f_2) := \sup_{w \in \widehat{\mathbb{C}}} \mathfrak{m}(f_1(w),\, f_2(w)) \tag{19}$$

for functions $f_1,\, f_2 : \widehat{\mathbb{C}} \to \widehat{\mathbb{C}}$, we find that $\{f_n\}$ converges $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}}$ to some function f if and only if $\sigma(f_n, f) \to 0$. Similarly, a sequence $\{f_n\}$ converges $\mathfrak{m}$-uniformly on some set D if and only if $\sigma_D(f_n, f) \to 0$ where

$$\sigma_D(f_n, f) := \sup_{w \in D} \mathfrak{m}(f_n(w), f(w)), \tag{20}$$

and it converges *locally* $\mathfrak{m}$-*uniformly* on D if and only if it converges $\mathfrak{m}$-uniformly on compact subsets of D. If $\{f_n(z)\}$ is uniformly bounded in D, then $\mathfrak{m}$-uniform convergence in D is equivalent to uniform convergence with respect to the euclidean metric. It is also clear that $\{f_n\}$ converges $\mathfrak{m}$-uniformly in D to $f(w)$ if and only if

$$\begin{aligned} &\{f_n\} \text{ converges uniformly in } D_1 := \{w \in D;\, |f(w)| < M\} \text{ and} \\ &\{1/f_n\} \text{ converges uniformly in } D_2 := \{w \in D;\, |f(w)| > 1/M\} \end{aligned} \tag{21}$$

for some constant $M > 1$.

2.2 The metric space $(\mathcal{M}, \sigma)$.

The convergence in this space is very nice indeed:

Theorem 2.1. *Let $\{\tau_n\}$ be a sequence from $\mathcal{M}$. Then $\tau_n \to \tau \in \mathcal{M}$ $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}}$, if and only if there exist three distinct points $w_1, w_2, w_3 \in \widehat{\mathbb{C}}$ such that $\tau_n(w_k) \to \gamma_k$ for $k = 1, 2, 3$ where $\gamma_k \in \widehat{\mathbb{C}}$ are distinct.*

Proof. The only-if-part holds trivially, so let $\tau_n(w_k) \to \gamma_k$ for $k = 1, 2, 3$. The pointwise convergence $\tau_n \to \tau$ follows then from the invariance of the cross ratio,

$$\frac{\tau_n(w) - \tau_n(w_1)}{\tau_n(w) - \tau_n(w_2)} \cdot \frac{\tau_n(w_3) - \tau_n(w_2)}{\tau_n(w_3) - \tau_n(w_1)} = \frac{w - w_1}{w - w_2} \cdot \frac{w_3 - w_2}{w_3 - w_1}, \tag{22}$$

and the $\mathfrak{m}$-uniform convergence from its chordal version

$$\frac{\mathfrak{m}(\tau_n(w),\, \tau_n(w_1))}{\mathfrak{m}(\tau_n(w),\, \tau_n(w_2))} \cdot \frac{\mathfrak{m}(\tau_n(w_3),\, \tau_n(w_2))}{\mathfrak{m}(\tau_n(w_3),\, \tau_n(w_1))} = \frac{\mathfrak{m}(w,\, w_1)}{\mathfrak{m}(w,\, w_2)} \cdot \frac{\mathfrak{m}(w_3,\, w_2)}{\mathfrak{m}(w_3,\, w_1)}. \tag{23}$$

□

Since the metric space $(\widehat{\mathbb{C}}, \mathfrak{m})$ is compact, we thus have:

Corollary 2.2. *Let $\{\tau_n\}$ be a sequence from $\mathcal{M}$. Then there exist three sequences $\{w_{1,n}\}$, $\{w_{2,n}\}$, $\{w_{3,n}\}$ from $\widehat{\mathbb{C}}$ such that*

$$\liminf \mathfrak{m}(w_{j,n}, w_{k,n}) > 0 \quad \textit{and} \quad \liminf \mathfrak{m}(\tau_n(w_{j,n}),\, \tau_n(w_{k,n})) > 0 \quad \textit{for } k \neq j,$$

if and only if every subsequence of $\{\tau_n\}$ has a subsequence converging $\mathfrak{m}$-uniformly to some $\tau \in \mathcal{M}$.

Corollary 2.3. *A sequence $\{\tau_n\}$ from $\mathcal{M}$ converges to a $\tau \in \mathcal{M}$ if and only if there exists a sequence $\{r_n\}$ from $\mathbb{C} \setminus \{0\}$ such that the finite limits*

$$r_n a_n \to a, \quad r_n b_n \to b, \quad r_n c_n \to c \quad \textit{and} \quad r_n d_n \to d, \quad ad - bc \neq 0$$

exists. In this case $\tau(w) = (aw + b)/(cw + d)$.

Proof. The if-part holds trivially, so let $\tau_n(w) := (a_n w + b_n)/(c_n w + d_n)$ converge poinwise in $\widehat{\mathbb{C}}$ to a $\tau \in \mathcal{M}$ (and thus $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}}$). Then

$$\tau_n(\infty) = \frac{a_n}{c_n} \to \tau(\infty) = \frac{a}{c}, \quad \tau_n(0) = \frac{b_n}{d_n} \to \tau(0) = \frac{b}{d},$$
$$\text{and} \quad \tau_n^{-1}(\infty) = -\frac{d_n}{c_n} \to \tau^{-1}(\infty) = \frac{d}{c}.$$

□

Corollary 2.4. *A sequence $\{\tau_n\}$ from $\mathcal{M}$ converges $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}}$ if and only if its limit also belongs to $\mathcal{M}$.*

Proof. In view of the proof of Theorem 2.1 it suffices to prove that if $\{\tau_n\}$ converges $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}}$, then its limit τ belongs to $\mathcal{M}$. Assume that $\tau \notin \mathcal{M}$. Then $\tau(w) \equiv \gamma_0$ for some constant $\gamma_0 \in \widehat{\mathbb{C}}$. Let $\gamma \neq \gamma_0$ and $w_n := \tau_n^{-1}(\gamma)$. Then $\tau_n(w_n) = \gamma \to \gamma$, a contradiction. Hence $\tau \in \mathcal{M}$. □

Corollary 2.5. *Every Cauchy sequence $\{\tau_n\}$ from $(\mathcal{M}, \sigma)$ converges to some $\tau \in \mathcal{M}$.*

In other words, the metric space $(\mathcal{M}, \sigma)$ is both compact and complete. For a sequence $\{\tau_n\}$ from

$$\mathcal{M}_{\overline{D}} := \{\tau \in \mathcal{M};\, \tau(\overline{D}) \subseteq \overline{D}\} \tag{24}$$

where $\overline{D}$ is a closed subset of $\widehat{\mathbb{C}}$ with at least three elements, the situation is different. The sequence may well converge generally to a constant even if the convergence is $\mathfrak{m}$-uniform in $\overline{D}$. The thing is that the exceptional sequence may have all its limit points well outside $\overline{D}$. But if $\{\tau_n\}$ converges to some $\tau \in \mathcal{M}$, then this τ also belongs to $\mathcal{M}_{\overline{D}}$.

Example 2.6. The sequence $\{\tau_n\}$ with $\tau_n(w) := w/n$ maps $\overline{\mathbb{D}}$ into $\overline{\mathbb{D}}$ and converges uniformly in $\overline{\mathbb{D}}$ to 0.

Hence $(\mathcal{M}_{\overline{D}}, \sigma_{\overline{D}})$ may be neither compact nor complete. Similarly, let

$$\mathcal{M}_{\overline{D}}^{(\varepsilon)} := \{\tau \in \mathcal{M};\, \tau(\overline{D}) \subseteq \overline{D}_{z_0}^{(\varepsilon)}\} \text{ where } \overline{D}_{z_0}^{(\varepsilon)} := \overline{D} \setminus B(z_0, \varepsilon) \quad \text{for some given } \varepsilon > 0 \tag{25}$$

where $B(z_0, \varepsilon) := \{w \in \widehat{\mathbb{C}};\, \mathfrak{m}(z_0, w) < \varepsilon)$ and $z_0 \in \overline{D}$ may depend on τ. Then a sequence from this space which converges $\mathfrak{m}$-uniformly in $\overline{D}$ will either converge to a $\tau \in \mathcal{M}_{\overline{D}}^{(\varepsilon)}$ or to a constant, just as above. $(\mathcal{M}_{\overline{D}}^{(\varepsilon)}, \sigma_{\overline{D}_{z_0}^{(\varepsilon)}})$ may therefore also be neither compact nor complete, [20].

2.3 Locally $\mathfrak{m}$-uniform convergence of sequences from $\mathcal{M}$.

In this section we look at convergence which no longer is $\mathfrak{m}$-uniform in $\widehat{\mathbb{C}}$. Then the limit function no longer belongs to $\mathcal{M}$.

Example 2.7. For the sequence $\{\tau_n\}$ from $\mathcal{M}$ given by

$$\tau_n(w) := \frac{a_n w + b_n}{w + 1} \quad \text{where } a_n \to 0,\ b_n \to 0, \tag{26}$$

$\tau_n(w) \to 0$ for every $w \in \widehat{\mathbb{C}} \setminus \{-1\}$, and $\tau_n(-1) = \infty$. Similarly, if

$$\tau_n(w) := \frac{a_n w + b_n}{w + n} \quad \text{where } a_n \to 0,\ b_n \to 0,\ b_n \neq n a_n, \tag{27}$$

then $\tau_n \in \mathcal{M}$, and $\tau_n(w) \to 0$ for every $w \in \widehat{\mathbb{C}}$, but not $\mathfrak{m}$-uniformly, since $\tau_n(-n) = \infty \to \infty$.

Theorem 2.8. *Let $\{\tau_n\}$ be a sequence from $\mathcal{M}$. If there exist three distinct points w_1, w_2, $w_3 \in \widehat{\mathbb{C}}$ such that $\tau_n(w_k) \to \gamma_k$ for $k = 1, 2, 3$ where $\gamma_k \in \widehat{\mathbb{C}}$ and $\gamma_1 = \gamma_2 \neq \gamma_3$, then $\tau_n(w) \to \tau$ locally $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}} \setminus \{w_3\}$ where*

$$\tau(w) := \begin{cases} \gamma_1 & \text{for all } w \neq w_3, \\ \gamma_3 & \text{for } w = w_3. \end{cases} \tag{28}$$

Proof. From (23), interchanging w_2 and w_3, we find that for $w \neq w_1, w_2, w_3$

$$\lim_{n\to\infty} \frac{\mathfrak{m}(\tau_n(w),\, \tau_n(w_1))}{\mathfrak{m}(\tau_n(w),\, \tau_n(w_3))} = \frac{\mathfrak{m}(w,\, w_1)}{\mathfrak{m}(w,\, w_3)} \cdot \frac{\mathfrak{m}(w_2,\, w_3)}{\mathfrak{m}(w_2,\, w_1)} \cdot \lim_{n\to\infty} \frac{\mathfrak{m}(\tau_n(w_2),\, \tau_n(w_1))}{\mathfrak{m}(\tau_n(w_2),\, \tau_n(w_3))} \tag{29}$$

where the right hand side $\to 0$, uniformly with respect to w with $\mathfrak{m}(w, w_3) \geq \varepsilon$ for some $\varepsilon > 0$. □

In other words, transformations τ of the form (28) are possible limits for $\{\tau_n\}$ when we no longer require $\mathfrak{m}$-uniform convergence in $\widehat{\mathbb{C}}$. This is evidently not the whole story. The sequence (27) converges to 0 for all $w \in \widehat{\mathbb{C}}$, but the convergence is not $\mathfrak{m}$-uniform in $\widehat{\mathbb{C}}$, only locally $\mathfrak{m}$-uniform in $\widehat{\mathbb{C}} \setminus \{\infty\}$. The crucial thing is that the limits $\lim \tau_n(w_1)$ and $\lim \tau_n(w_2)$ exist and are equal at two distinct points w_1, w_2, or more generally, that $\lim \tau_n(w_{1,n}) = \lim \tau_n(w_{2,n})$ for two asymptotically distinct sequences $\{w_{1,n}\}$ and $\{w_{2,n}\}$ from $\widehat{\mathbb{C}}$:

Theorem 2.9. *Let $\{\tau_n\}$ be a sequence from $\mathcal{M}$. Then there exist two sequences $\{w_{1,n}\}$ and $\{w_{2,n}\}$ from $\widehat{\mathbb{C}}$ such that*

$$\liminf \mathfrak{m}(w_{1,n}, w_{2,n}) > 0 \quad \text{and} \quad \lim \tau_n(w_{1,n}) = \lim \tau_n(w_{2,n}) = \gamma,$$

if and only if there exists a sequence $\{w_n^\dagger\}$ from $\widehat{\mathbb{C}}$ such that for every $\varepsilon > 0$

$$\lim \sigma_{D_n}(\tau_n, \gamma) = 0 \quad \text{for} \quad D_n := \{w \in \widehat{\mathbb{C}} :\, \mathfrak{m}(w, w_n^\dagger) \geq \varepsilon\}.$$

Proof. Again the if-part holds trivially. Let $q_n := \tau_n^{-1}(\gamma)$ and $w_n^\dagger := \tau_n^{-1}(\gamma^\dagger)$ for some fixed $\gamma^\dagger \neq \gamma$ for all n. Let further

$$w_n := \begin{cases} w_{1,n} & \text{if } \mathfrak{m}(w_{1,n}, w_n^\dagger) \geq \mathfrak{m}(w_{2,n}, w_n^\dagger) \\ w_{2,n} & \text{otherwise.} \end{cases}$$

Then $\tau_n(q_n) \to \gamma$, $\tau_n(w_n) \to \gamma$, $\tau_n(w_n^\dagger) \to \gamma^\dagger$ and $\liminf \mathfrak{m}(w_n, w_n^\dagger) > 0$. The result follows therefore from (29) with w_1 replaced by w_n, w_2 by q_n and w_3 by $w_n^\dagger$. □

Remarks.

1. We say that $\{\tau_n\}$ *converges generally* to γ with *exceptional sequence* $\{w_n^\dagger\}$ in this case. A possible choice for $\{w_n^\dagger\}$ is $w_n^\dagger := \tau_n^{-1}(\gamma^\dagger)$ for a $\gamma^\dagger \neq \gamma$. This concept of convergence was introduced by Jacobsen (now Lorentzen) [10] for continued fractions, and extended to quasi-normal function families in [18]. It has now gained acceptance in the continued fraction community.

2. We notice that the exceptional sequence $\{w_n^\dagger\}$ can not be avoided in Theorem 2.9 since $\mathfrak{m}$-uniform convergence leads to a limit function from $\mathcal{M}$.

3. General convergence can also be formulated as follows: every subsequence of $\{\tau_n\}$ has a subsequence converging locally $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}} \setminus \{w_3\}$ to some τ of the form (28) for some $w_3 \in \widehat{\mathbb{C}}$ depending on the subsequence (possibly with $\gamma_3 = \gamma_1$). (Just let the subsequence be chosen such that $\{w_{n_k}^\dagger\}$ converges in $\widehat{\mathbb{C}}$ and let w_3 be its limit.)

4. The convergence of $\{\tau_n\}$ to γ can not be $\mathfrak{m}$-uniform in $\widehat{\mathbb{C}}$, but it is locally $\mathfrak{m}$-uniform in the following sense:

$$\lim_{n\to\infty} \sup_{w\in\widehat{\mathbb{C}},\, \mathfrak{m}(w,w_n^\dagger)\geq\varepsilon} \mathfrak{m}(\tau_n(w), \gamma) = 0 \quad \text{for every } \varepsilon > 0.$$

Corollary 2.10. *Let $\{\tau_n\}$ be a sequence from $\mathcal{M}$ which converges locally $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}} \setminus \{w^\dagger\}$ to a constant γ. Then $\{\tau_n^{-1}\}$ converges locally $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}} \setminus \{\gamma\}$ to $w^\dagger$.*

Proof. Let $\tilde{\gamma} \in \widehat{\mathbb{C}} \setminus \{\gamma\}$ be arbitrarily chosen, and set $w_n := \tau_n^{-1}(\tilde{\gamma})$ for all n. Then $\tau_n(w_n) = \tilde{\gamma} \not\to \gamma$. That is, every subsequence of $\{w_n\}$ converges to $w^\dagger$. The local $\mathfrak{m}$-uniformity follows since also $\{\tau_n^{-1}\}$ is a sequence from $\mathcal{M}$. □

In the next corollary we consider the asymptotics of the coefficients of

$$\tau_n(w) := \frac{a_n w + b_n}{c_n w + d_n}; \quad \tau_n \in \mathcal{M}. \tag{30}$$

Corollary 2.11. *Let $\{\tau_n\}_{n=1}^{\infty}$ given by (30) converge locally $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}} \setminus \{w^\dagger\}$ to γ. Then there exists a sequence $\{r_n\}$ from $\mathbb{C} \setminus \{0\}$ such that the finite limits*

$$r_n a_n \to a, \quad r_n b_n \to b, \quad r_n c_n \to c, \quad r_n d_n \to d$$

exist, with $ad - bc = 0$ and $|a| + |b| + |c| + |d| \neq 0$.

Proof. Case 1: $w^\dagger \neq \infty$ and $\gamma \neq \infty$. Then $\tau_n(\infty) = a_n/c_n \to \gamma$ and $\tau_n(-d_n/c_n) = \infty$, so $w^\dagger = -\lim d_n/c_n$. Therefore there exists a sequence $\{r_n\}$ from $\widehat{\mathbb{C}} \setminus \{0\}$ such that

$$r_n c_n \to 1, \quad r_n a_n \to \gamma =: a, \quad r_n d_n \to -w^\dagger =: d.$$

Moreover,

$$\tau_n(w) = \frac{a_n w + b_n}{c_n w + d_n} = \frac{r_n a_n w + r_n b_n}{r_n c_n w + r_n d_n},$$

so for $w \neq w^\dagger$,

$$r_n b_n = (r_n c_n w + r_n d_n)\tau_n(w) - r_n a_n w \to (w - w^\dagger)\gamma - \gamma w = -\gamma w^\dagger =: b.$$

That is,

$$\tau_n(w) \to \tau(w) = \frac{aw + b}{cw + d} = \frac{\gamma(w - w^\dagger)}{w - w^\dagger} \quad \text{for } w \neq w^\dagger, \tag{31}$$

where $ad - bc = -\gamma w^\dagger + \gamma w^\dagger = 0$ and $|a| + |b| + |c| + |d| \geq |c| = 1$.

Case 2: $w^\dagger \neq \infty$, $\gamma = \infty$. The sequence $\{\tilde{\tau}_n\}$ given by

$$\tilde{\tau}_n(w) := \frac{1}{\tau_n(w)} = \frac{c_n w + d_n}{a_n w + b_n} \quad \text{for } n = 1, 2, 3, \ldots$$

is also a sequence from $\mathcal{M}$. It converges locally $\mathfrak{m}$-uniformly to $\tilde{\gamma} := 1/\gamma = 0$ in $\widehat{\mathbb{C}} \setminus \{w^\dagger\}$, so the result is a consequence of Case 1. In particular, the limit function τ has the form

$$\tau(w) = \frac{1}{\tilde{\tau}(w)} = \frac{w - w^\dagger}{\tilde{\gamma}(w - w^\dagger)} = \frac{\gamma(w - w^\dagger)}{w - w^\dagger}.$$

Case 3: $w^\dagger = \infty$. This time we consider the transformations

$$\widehat{\tau}_n(w) := \tau_n(1/w) = \frac{a_n + b_n w}{c_n + d_n w}$$

which also belong to $\mathcal{M}$. These transformations converge locally $\mathfrak{m}$-uniformly to γ in $\widehat{\mathbb{C}} \setminus \{1/w^\dagger\} = \widehat{\mathbb{C}} \setminus \{0\}$, so $\{\widehat{\tau}_n\}$ belongs to Case 1 or to Case 2. Hence the result follows. In particular,

$$\tau(w) = \widehat{\tau}(1/w) = \frac{\gamma(1/w - 0)}{1/w - 0} = \frac{\gamma(1 - 0w)}{1 - 0w} \quad \text{for } w \neq \infty. \tag{32}$$

□

Remarks.

1. The possible limit functions in Corollary 2.11 are therefore singular linear fractional transformations

$$\tau(w) = \frac{aw + b}{cw + d}, \quad a, b, c, d \in \mathbb{C} \text{ with } ad - bc = 0, \quad |a| + |b| + |c| + |d| \neq 0. \tag{33}$$

   We let $\mathcal{M}^s$ denote the family of transformations (33).

2. We note that τ is given by (31) if $w^\dagger \neq \infty$ and by (32) if $w^\dagger = \infty$.

3. Throughout this analysis we have chosen to require that also the singular linear fractional transformations shall have finite coefficients, and thus their determinants are $ad - bc = 0$. Another possibility would have been to require that $a_n d_n - b_n c_n = 1$ for all n and thus also in the limit, whereas $|a_n| + |b_n| + |c_n| + |d_n| \to \infty$ when τ_n approaches a singular τ. Our choice seems to give a clearer picture of the situation.

4. Let $\mathcal{M}^* := \mathcal{M} \cup \mathcal{M}^s$. Then every sequence $\{\tau_n\}$ from $\mathcal{M}^*$ which converges in $\widehat{\mathbb{C}} \setminus \{w^\dagger\}$, converges to a transformation from $\mathcal{M}^*$.

5. It seems natural to ask what possibilities can occur if $\{\tau_n\}$ from $\mathcal{M}$ (or $\mathcal{M}^*$) is known to converge at two distinct points, but to distinct values. However, this does not lead to anything substantially new. Indeed, let w_3 be a third point from $\widehat{\mathbb{C}}$. Since $(\widehat{\mathbb{C}}, \mathfrak{m})$ is compact, there exists a subsequence of $\{\tau_n(w_3)\}$ which converges to some $\gamma_3 \in \widehat{\mathbb{C}}$, and we are back to the $\mathfrak{m}$-uniform convergence in Theorem 2.1 or the locally $\mathfrak{m}$-uniform convergence in Theorem 2.9 for this subsequence.

The following behavior proved by Piranian and Thron [26] is now easy to explain:

Theorem 2.12. *Let $\{\tau_n\}$ be a sequence from $\mathcal{M}$ which converges at two distinct points $w_1, w_2 \in \widehat{\mathbb{C}}$. Then, either*

(i) $\{\tau_n\}$ converges $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}}$ to some $\tau \in \mathcal{M}$, or

(ii) $\{\tau_n\}$ converges in $\widehat{\mathbb{C}}$ to some function

$$\tau(w) := \begin{cases} \gamma_1 & \text{for all } w \in \widehat{\mathbb{C}} \setminus \{w^\dagger\} \\ \gamma_2 & \text{for } w = w^\dagger \end{cases} \tag{34}$$

where $\gamma_1 \neq \gamma_2$, and the convergence is locally $\mathfrak{m}$-uniform in $\widehat{\mathbb{C}} \setminus \{w^\dagger\}$, or

(iii) $\{\tau_n\}$ converges only at these two points, or

(iv) $\{\tau_n\}$ converges to a constant function on its set $D \subseteq \widehat{\mathbb{C}}$ of convergence.

Possibility (i) is the convergence in $(\mathcal{M}, \sigma)$, possibility (ii) is described in Theorem 2.8, and possibility (iv) is a consequence of Theorem 2.9 with D being the complement of the set of limit points for $\{w_n^\dagger\}$. Possibility (iii) is either a consequence of Theorem 2.9 or as described in Remark 5 above.

2.4 The value set technique

A sequence $\{V_n\}_{n=0}^\infty$ of sets $V_n \subseteq \widehat{\mathbb{C}}$ is a sequence of *value sets* for $\mathbf{K}(a_n/b_n)$ if

$$s_n(V_n) := \frac{a_n}{b_n + V_n} := \left\{ w = \frac{a_n}{b_n + v};\ v \in V_n \right\} \subseteq V_{n-1} \quad \text{for } n = 1, 2, 3, \ldots . \tag{35}$$

It is normally quite difficult to find value sets for a given continued fraction. So what we do is to start with a sequence $\{V_n\}$ of sets, and the determine for which continued fractions this is a sequence of value sets. The significance of value sets is that

$$S_n(V_n) = S_{n-1}(s_n(V_n)) \subseteq S_{n-1}(V_{n-1}) \subseteq \cdots \subseteq V_0 \quad \text{for all } n.$$

This means that we have better control over the modified approximants $S_n(w_n)$ with $w_n \in V_n$. Indeed, if the closed, nested sets $S_n(\overline{V}_n)$ converge to a one-point set $\{\gamma\}$ and $\liminf \operatorname{diam}_{\mathfrak{m}}(V_n) > 0$, then $\{S_n\}$ converges generally to γ. If all $V_n = V$, we say that V is a *simple value set* for $\mathbf{K}(a_n/b_n)$.

Of course, this idea also works for more general function families.

The nestedness may also be useful in cases where the diameter of $S_n(\overline{V}_n)$ converges to a positive number:

Theorem 2.13. *[19]. Let $\operatorname{rad} \tau_n(\mathbb{D}) \leq r < 1$ for all n and $\operatorname{rad} \mathcal{T}_n(\mathbb{D}) \to R > 0$ for $\mathcal{T}_n := \tau_1 \circ \tau_2 \circ \cdots \circ \tau_n$. If there exists a sequence $\{w_n\}$ of complex numbers such that*

$$\liminf ||w_n| - 1| > 0 \quad \text{and} \quad \liminf ||\tau_n(w_n)| - 1| > 0,$$

then $\{\mathcal{T}_n\}$ converges pointwise in $\mathbb{D}$ to a constant. The convergence is absolute for each $\zeta \in \mathbb{D}$.

Idea of proof: We know that $\mathcal{T}_n(\mathbb{D})$ is a circular disk. Let C_n and R_n be its center and radius. Since $\mathcal{T}_n(\mathbb{D}) \subseteq \mathbb{D}$, $\mathcal{T}_n(w)$ can then be written

$$\mathcal{T}_n(w) = C_n + R_n e^{i\omega_n} \frac{w - Q_n}{1 - \overline{Q}_n w} \quad \text{where } |C_n| + R_n \le 1,\ \omega_n \in \mathbb{R} \text{ and } |Q_n| < 1.$$

Straightforward computation shows that

$$\frac{R_{n+1}}{R_n} \le 1 - \delta_n \quad \text{where} \quad \delta_n := \frac{(1-r)(1-|Q_n|)}{1 - |Q_n| + 2r|Q_n|}.$$

If $\sum(1 - |Q_n|) = \infty$, then $R_n \to 0$, and $\mathcal{T}_n(\zeta) \to C := \lim C_n$ for all $\zeta \in \mathbb{D}$.

If $\sum(1 - |Q_n|) < \infty$, then $\sum |\mathcal{T}_{n+1}(w_{n+1}) - \mathcal{T}_n(w_n)| < \infty$, and thus $\{\mathcal{T}_n(w_n)\}$ converges absolutely, since

$$\begin{aligned} &|\mathcal{T}_{n+1}(w_{n+1}) - \mathcal{T}_n(w_n)| = |\mathcal{T}_n(\tau_{n+1}(w_{n+1})) - \mathcal{T}_n(w_n)| \\ &= R_n \frac{(1 - |Q_n|^2)|\tau_{n+1}(w_{n+1}) - w_n|}{|1 - \overline{Q}_n \tau_{n+1}(w_{n+1})| \cdot |1 - \overline{Q}_n w_n|}. \end{aligned} \tag{36}$$

This equality also proves that $\sum |\mathcal{T}_{n+1}(w_{n+1}) - \mathcal{T}_n(\zeta)| < \infty$ for every $\zeta \in \mathbb{D}$. That is, $\{\mathcal{T}_n(\zeta)\}$ converges absolutely to the same value. □

That $\{\mathcal{T}_n(\zeta)\}$ converges absolutely is the same as $\sum |\mathcal{T}_{n+1}(\zeta) - \mathcal{T}_n(\zeta)| < \infty$. Hence, this theorem implies that $\{\mathcal{T}_n\}$ converges locally uniformly in $\mathbb{D}$ to a constant.

2.5 The Parabola Theorem

The possibly most important convergence criterion for continued fractions $\mathbf{K}(a_n/1)$ is the Parabola Theorem due to Thron [30]:

Theorem 2.14 (The Parabola Theorem)**.** *Let $P_\alpha := \{a \in \mathbb{C};\ |a| - \mathrm{Re}(a\, e^{-2i\alpha}) \le \frac{1}{2}\cos^2\alpha\}$ for a given $\alpha \in \mathbb{R}$ with $|\alpha| < \pi/2$, and let $\mathbf{K}(a_n/1)$ be a continued fraction from P_α; i.e., all $a_n \in P_\alpha$. Then:*

A. $\{f_{2n}\}$ and $\{f_{2n+1}\}$ converge to finite values.

B. $\mathbf{K}(a_n/1)$ converges if and only if

$$\sum_{n=1}^{\infty} |d_n| = \infty \quad \textit{where} \quad d_n := \prod_{k=1}^{n} a_k^{(-1)^{n+k+1}}. \tag{37}$$

C. $$|f - f_n| \le \frac{2|a_1|/\cos\alpha}{\prod_{k=2}^{n}\left(1 + \dfrac{\cos^2\alpha}{4(k-1)|a_k|}\right)}.$$

The essence of this theorem is that we know everything about convergence for continued fractions $\mathbf{K}(a_n/1)$ if all a_n are taken from a parabolic region P_α with focus at the origin and axis along the ray $\{z = r\, e^{2i\alpha};\ r \ge -\frac{1}{4}\cos^2\alpha\}$ through the origin which does not go out towards ∞ along the negative real axis. The half plane

$$V_\alpha := \{w \in \mathbb{C};\ \mathrm{Re}((w + \tfrac{1}{2})e^{-i\alpha}) \ge 0\}$$

is then a simple value set for $\mathbf{K}(a_n/1)$; i.e., $V_n = V_\alpha$ for all n. The condition (37) is equivalent to the condition (10) on page 128. Indeed, the convergence results for the Stieltjes fractions $z\mathbf{K}(a_n z^{-1}/1)$ in Sections 1.2.2 and 1.2.3 are consequences of this Parabola Theorem.

We shall prove the convergence part of this theorem. The purpose is to demonstrate some of the techniques in this area. An important tool is the Lane-Wall Characterization [16] whose proof can be found in [22, p.103].

Theorem 2.15 (The Lane-Wall Characterization)**.** *If*

$$\sum |f_{n+1} - f_{n-1}| < \infty, \tag{38}$$

then $\mathbf{K}(a_n/b_n)$ converges if and only if (37) holds.

Of course, (38) means that $\{f_{2n-1}\}$ and $\{f_{2n}\}$ both converge absolutely to some finite values f and $\tilde{f}$ respectively.

Proof of the convergence part of the parabola theorem. Let $\varphi \in \mathcal{M}$ map the unit disk $\mathbb{D}$ onto the interior V_α° of V_α. Then the compositions $\tau_n := \varphi^{-1} \circ s_n \circ \varphi$ map $\mathbb{D}$ into $\mathbb{D}$ and $\mathcal{T}_n := \tau_1 \circ \tau_2 \circ \cdots \circ \tau_n = \varphi^{-1} \circ s_1 \circ s_2 \circ \cdots \circ s_n \circ \varphi = \varphi^{-1} \circ S_n \circ \varphi$. Since we are interested in convergence properties of $\{S_n(0)\}$, we may just as well look at $\{\mathcal{T}_n(\varphi^{-1}(0)\}$.

Evidently the disks $\mathcal{T}_n(\mathbb{D})$ are nested and $\subseteq \mathbb{D}$ since $\mathcal{T}_{n+1}(\mathbb{D}) = \mathcal{T}_n \circ \tau_{n+1}(\mathbb{D}) \subseteq \mathcal{T}_n(\mathbb{D})$. Hence the radius $R_n := \operatorname{rad} \mathcal{T}_n(\mathbb{D})$ of $\mathcal{T}_n(\mathbb{D})$ is decreasing, and thus it converges to some $R \geq 0$. If $R = 0$, then the convergence of $\mathcal{T}_n(\varphi^{-1}(0))$ is clear.

Let $R > 0$. Since $s_n(\infty) = 0 \in V_\alpha^\circ$, it follows that $\operatorname{rad} \tau_n(\mathbb{D}) \leq r$ for some $r < 1$. It therefore follows from (36) that $\sum |\mathcal{T}_n(\varphi^{-1}(-1)) - \mathcal{T}_n(\varphi^{-1}(0))| < \infty$; i.e., $\sum |S_n(-1) - S_n(0)| = \sum |f_{n-2} - f_n| < \infty$. The Parabola Theorem follows therefore from the Lane-Wall Characterization. □

For the more general Parabola Sequence Theorem [30] where V_n is allowed to vary with n, we refer to [22, p.154].

This type of convergence criteria is in particular useful when a_n is a function of a complex variable z, since then we want uniform convergence with respect to z. (Note that this is another type of uniformity that uniform convergence of $\{\tau_n(w)\}$ with respect to w.)

2.6 Equivalence transformations

The Parabola Theorem is valid for continued fractions of the form $\mathbf{K}(a_n/1)$. However, it does not take much effort to see that if $\{r_n\}_{n=0}^\infty$ is a sequence of non-zero complex numbers with $r_0 := 1$, then $\mathbf{K}(a_n r_n r_{n-1}/r_n)$ has the same sequence of approximants as $\mathbf{K}(a_n/1)$. (Just check the first few approximants and see how it works.) Or, conversely, if $\mathbf{K}(a_n/b_n)$ is a continued fraction with all $b_n \neq 0$, then

$$\mathbf{K}(c_n/1) \quad \text{with } c_1 := \frac{a_1}{b_1} \text{ and } c_n := \frac{a_n}{b_n b_{n-1}} \text{ for } n \geq 2$$

has the same approximants as $\mathbf{K}(a_n/b_n)$, and thus the same convergence properties. This means that any continued fraction with all $b_n \neq 0$ can be brought to the form $\mathbf{K}(c_n/1)$, and any continued fraction $\mathbf{K}(a_n/b_n)$ can be brought to the form $\mathbf{K}(1/b_n d_n)$. Indeed, d_n is given by (37) in this case.

This kind of transformation is called an *equivalence transformation.*

2.7 Iterations

2.7.1 Classification of linear fractional transformations

We classify linear fractional transformations $\tau \in \mathcal{M}$ according to the asymptotic behavior of *iterations* $\tau^{[n]} := \tau \circ \tau \circ \cdots \circ \tau$ as the number n of τs in the composition approaches ∞. A linear fractional transformation $\widehat{\tau}$ is *conjugate* to τ if there exists a transformation φ from $\mathcal{M}$ such that $\widehat{\tau} = \varphi \circ \tau \circ \varphi^{-1}$. Since $\widehat{\tau}^{[n]} = \varphi \circ \tau^{[n]} \circ \varphi^{-1}$, the classification shall be invariant under conjugation.

If $\{\tau^{[n]}\}$ converges generally to some $x \in \widehat{\mathbb{C}}$, then x must be a *fixed point* for τ; i.e. $\tau(x) = x$. Unless τ is the *identity transformation* $I(w) \equiv w$; i.e. $a = d \neq 0, b = c = 0$, it follows that τ has two (possibly coinciding) fixed points x and y.

Case 1: τ has only one fixed point.
Let $\widehat{\tau}(w) := w + q$ for a complex constant $q \neq 0$. Then $\widehat{\tau} \in \mathcal{M}$, and the point at ∞ is the only fixed point for $\widehat{\tau}$. Moreover, $\widehat{\tau}^{[n]}(w) = w + nq \to \infty$ for every $w \in \widehat{\mathbb{C}}$. (The convergence is not locally uniform in $\widehat{\mathbb{C}}$, though, since $\widehat{\tau}^{[n]}(-nq) \to 0$.)

Every $\tau \in \mathcal{M}$ with only one fixed point must be conjugate to $\widehat{\tau}$; i.e., $\tau = \varphi \circ \widehat{\tau} \circ \varphi^{-1}$ for some $\varphi \in \mathcal{M}$. This follows since then $x := \varphi(\infty)$ is the only fixed point for τ. Hence, iterations of τ converge generally to this fixed point with exceptional sequence $\{\varphi(-nq)\}$, or, equivalently, exceptional sequence $\{\varphi(\infty)\}$.

We say that τ is a *parabolic transformation* and that the fixed point of a parabolic transformation is *attracting.*

Case 2: τ has exactly two distinct fixed points $x \neq y$.
Then $\tau = \varphi \circ \widehat{\tau} \circ \varphi^{-1}$ for a $\widehat{\tau}(w) := kw$ for a complex constant $k \neq 1$ with $|k| \leq 1$. The points 0 and ∞ are the fixed points for $\widehat{\tau}$, and $\widehat{\tau}^{[n]}(w) = k^n w$.

Case 2A $|k| < 1$. Then $\widehat{\tau}^{[n]}(w) = k^n w \to 0$ for every $w \neq \infty$. Hence $\tau^{[n]}(w) \to x := \varphi(0)$ for all $w \neq y := \varphi(\infty)$. We say that τ is a *loxodromic transformation* with *attracting fixed point* $x = 0$ and *repelling fixed point* $y = \infty$.

Case 2B $|k| = 1$ with $k \neq 1$. This time $\widehat{\tau}^{[n]}(w) = k^n w$ diverges for every $w \in \widehat{\mathbb{C}}$ except at the two fixed points 0 and ∞. Hence $\tau^{[n]}(w)$ diverges for all $w \in \widehat{\mathbb{C}}$ except at its two fixed points x and y. Indeed, no subsequence of $\{\tau^{[n]}\}$ converges generally to a constant. We say that τ is an *elliptic transformation* with (indifferent) fixed points $x = 0$ and $y = \infty$.

The case $k = 1$ naturally gives $\tau(w) = I(w) \equiv w$, the identity transformation.

Remarks.

1. The classification is also invariant under inversion. Indeed, w is a fixed point for τ if and only if w is a fixed point for τ^{-1}. The roles of x and y must be interchanged, though. If τ is loxodromic with attracting fixed point x and repelling fixed point y, then τ^{-1} is loxodromic with attracting fixed point y and repelling fixed point x.
2. The only $\tau \in \mathcal{M}$ for which $\{\tau^{[n]}\}$ converges to some $\tilde{\tau} \in \mathcal{M}$ is the identity transformation $\tau = I$.
3. The only case of divergence of $\{\tau^{[n]}\}$ is the case where τ is elliptic.

2.7.2 Convergence of periodic compositions

Let $\{\varphi_n\}$ and $\{\psi_n\}$ be periodic sequences from $\mathcal{M}$ with period $p \in \mathbb{N}$; i.e., $\varphi_{n+p} = \varphi_n$ and $\psi_{n+p} = \psi_n$ for all $n \in \mathbb{N}$. Let further

$$\tau_n := \varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_n \quad \text{and} \quad \tilde{\tau}_n := \psi_n \circ \psi_{n-1} \circ \cdots \circ \psi_1 \quad \text{for all } n.$$

It is then clear that

$$\tau_{np+m} = \tau_p^{[n]} \circ \tau_m \quad \text{and} \quad \tilde{\tau}_{np+m} = \tilde{\tau}_m \circ \tilde{\tau}_p^{[n]}$$

for $n \in \mathbb{N}$ and $m \in \{0, 1, \ldots, p-1\}$, where $\tau_0 := \tilde{\tau}_0 := I$. Therefore the classifications of τ_p and $\tilde{\tau}_p$ are imperative. If $p = 1$, then the inner compositions τ_n and the outer compositions $\tilde{\tau}_n$ are just iterations, so their convergence properties are already described.

Let $p > 1$. Then we find:

- If τ_p ($\tilde{\tau}_p$) is elliptic, then $\{\tau_n\}$ ($\{\tilde{\tau}_n\}$) diverges, but every subsequence has a subsequence converging to some $\tau \in \mathcal{M}$.
- If τ_p is parabolic or loxodromic, then $\{\tau_n\}$ converges generally to the attracting fixed point x of τ_p. The exceptional sequence $\{w_n^\dagger\}$ is for instance given by $w_{np+m}^\dagger = w_m^\dagger = \tau_m^{-1}(y)$ for $m = 1, 2, \ldots, p$ where y is the repelling fixed of τ_p. ($y = x$ if τ_p is parabolic.)
- If $\tilde{\tau}_p$ is parabolic or loxodromic, then $\{\tilde{\tau}_p^{[n]}\}_n$ converges generally to the attracting fixed point x of $\tilde{\tau}_p$, with exceptional sequence $\{y\}$, where $y = x$ if $\tilde{\tau}_p$ is parabolic, and y is the repelling fixed point of $\tilde{\tau}_p$ otherwise. But this does not necessarily give convergence of $\{\tilde{\tau}_n\}$. Indeed, a necessary condition for convergence is that x is a fixed point for every ψ_m, and a sufficient condition is the x is an attracting fixed point for every ψ_m.

2.8 Limit periodic compositions

Let $\{\varphi_n\}$ and $\{\psi_n\}$ be sequences from $\mathcal{M}$ which converge to some $\varphi \in \mathcal{M}$ and $\psi \in \mathcal{M}$ respectively. Then

$$\tau_n := \varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_n \quad \text{and} \quad \tilde{\tau}_n := \psi_n \circ \psi_{n-1} \circ \cdots \circ \psi_1$$

for all n have more and more character of iterations of φ and ψ as n increases. So, no wonder, the classification of φ and ψ are again vital.

Case 1: φ and ψ are elliptic.

We would expect $\{\tau_n\}$ and $\{\tilde{\tau}_n\}$ to diverge, but this does not always occur. It depends on *how* $\{\varphi_n\}$ and $\{\psi_n\}$ approach their limit functions.

Example 2.16. Let $\varphi_n(w) := \psi_n(w) := k_n w$ where $k_n \to k \neq 1$ with $|k| = 1$. Then

$$\tau_n(w) = \tilde{\tau}_n(w) = K_n w \quad \text{where} \quad K_n := \prod_{j=1}^{n} k_j$$

which converges generally if $K_n \to 0$ or $K_n \to \infty$ and diverges otherwise.

Evidently the cases $\varphi = I$ and $\psi = I$ are also quite intricate cases.

Case 2: φ and ψ are parabolic.

In this case we would expect $\{\tau_n\}$ and $\{\tilde{\tau}_n\}$ to converge, but the catch is again *how* $\varphi_n \to \varphi$ and $\psi_n \to \psi$.

Example 2.17. Let $\varphi_n(w) = \psi_n(w) = w + q_n$ where $q_n \to q \neq 0, \infty$. Then

$$\tau_n(w) = \tilde{\tau}_n(w) = w + Q_n \quad \text{where} \quad Q_n := \sum_{j=1}^{n} q_j$$

which converges to ∞ if $Q_n \to \infty$, and diverges otherwise.

Case 3: φ and ψ are loxodromic.

This is the show case situation. Here we always get a robust general convergence of both $\{\tau_n\}$ and $\{\tilde{\tau}_n\}$. The reason for this is that if φ is loxodromic with attracting fixed point x and repelling fixed point y, then its derivatives at these points satisfy

$$|\varphi'(x)| < 1 \quad \text{and} \quad |\varphi'(y)| > 1.$$

Hence there is a neighborhood U of x and an $n_0 \in \mathbb{N}$ such that $|\varphi'(w)| < 1$ for all $w \in U$ and $n \geq n_0$, and thus $\{\tau_n\}$ converges generally to some value $\gamma \in \widehat{\mathbb{C}}$ with exceptional sequence $\{y\}$. Similarly, $\{\tilde{\tau}_n\}$ converges generally to x with some exceptional sequence $\{w_n^\dagger\}$.

The extensions to cases where $\{\varphi_n\}$ or $\{\psi_n\}$ are limit p-periodic for some $p > 1$; i.e., the limits

$$\tilde{\varphi}_m := \lim_{n\to\infty} \varphi_{np+m} \in \mathcal{M} \quad \text{and} \quad \tilde{\psi}_m := \lim_{n\to\infty} \psi_{np+m} \in \mathcal{M}$$

exist for $m = 1, 2, \ldots, p$, and $\tilde{\varphi} := \tilde{\varphi}_1 \circ \tilde{\varphi}_2 \circ \cdots \circ \tilde{\varphi}_p$ and $\tilde{\psi} := \tilde{\psi}_1 \circ \tilde{\psi}_2 \circ \cdots \circ \tilde{\psi}_p$ are of elliptic, parabolic, loxodromic or identity type follow similarly.

3 Sources for results on convergence of sequences of linear fractional transformations.

Transformations from $\mathcal{M}$ are extreme in several situations, such as for instance in the Schwarz Lemma and Brouwer's fixed point theorem. They also pop up in various disguises throughout the literature of mathematics. In this part we shall present some situations where some kind of convergence of sequences from $\mathcal{M}$ plays a role.

3.1 Schur's algorithm

Schur's algorithm gives a method to determine whether a given (formal) power series $L(z) := \sum_{n=0}^{\infty} c_n z^n$ belongs to the class $\mathcal{L}_1$ of power series which converge in the unit disk $\mathbb{D}$ to some function $f(z)$ bounded by 1. The idea behind this algorithm is that $L \in \mathcal{L}_1$ implies that $|c_0| \leq 1$, and if $|c_0| < 1$, then $L \in \mathcal{L}_1$ if and only if

$$L_1 := \tau_0(L) \in \mathcal{L}_1 \quad \text{where } \tau_0(w) := \frac{1}{z}\,\frac{w - c_0}{1 - \bar{c}_0 w}.$$

(Evidently $\tau_0(w)$ is a transformation from $\mathcal{M}$ with coefficients depending on $z \neq 0$, and $z\tau_0$ maps $\mathbb{D}$ onto $\mathbb{D}$.) Repeating this idea on L_1, and so on, gives a sequence $\{L_n\}$ of power series, and thus a sequence of numbers $\gamma_n := L_n(0)$, the *Schur parameters*. The algorithm stops if $|\gamma_n| \geq 1$ at some point.

If the algorithm stops with a $|\gamma_n| > 1$, then $L \notin \mathcal{L}_1$.

If the algorithm stops with a $|\gamma_n| = 1$, then L is the expansion of a rational function from $\mathcal{L}_1$. Indeed, as proved by Schur [27], this case occurs if and only if L is the expansion of a finite Blaschke product

$$f(z) = e^{i\alpha} \prod_{k=1}^{n} \frac{z + \omega_k}{1 + \overline{\omega}_k z} \quad \text{for some } \alpha \in \mathbb{R} \text{ and } \omega_k \in \mathbb{D}.$$

If all $|\gamma_n| < 1$, then $L \in \mathcal{L}_1$, and

$$L = \varphi_0 \circ \varphi_1 \circ \cdots \circ \varphi_n(L_{n+1}) \quad \text{for all } n, \quad \text{where } \varphi_k := \tau_k^{-1},$$

and the question of convergence arises. Since

$$\varphi_n(w) = \frac{zw + \gamma_n}{1 + \overline{\gamma}_n z w} = \gamma_n + \frac{(1 - |\gamma_n|^2)z}{\overline{\gamma}_n z + 1/w},$$

the corresponding continued fraction is

$$\gamma_0 + \frac{(1 - |\gamma_0|^2)z}{\overline{\gamma}_0 z} + \frac{1}{\gamma_1} + \frac{(1 - |\gamma_1|^2)z}{\overline{\gamma}_1 z} + \frac{1}{\gamma_2} + \frac{(1 - |\gamma_2|^2)z}{\overline{\gamma}_2 z} + \cdots.$$

Let us for convenience assume here that all $\gamma_n \neq 0$. Then Wall [34] proved that the even approximants $\{f_{2n}\}$ of this continued fraction converge locally uniformly with respect to z in $\mathbb{D}$ to the function $f(z) \sim L(z)$. That is, the Schur algorithm converges to the „right function" $f(z)$. This also happens if some or all $\gamma_n = 0$.

3.2 The linear group of 2×2-matrices

$\mathcal{M}$ can be regarded as an embedding in the general linear group $GL(2, \mathbb{C})$ of 2×2 non-singular matrices

$$M := \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad a, b, c, d \in \mathbb{C}, \quad ad - bc \neq 0. \tag{39}$$

In this setting

$$\tau(w) := \frac{aw + b}{cw + d} \tag{40}$$

has the corresponding matrix $M(\tau)$ given by (39), and $M(\tau_1 \circ \tau_2) = M(\tau_1) \cdot M(\tau_2)$. Moreover, the coefficients of τ are unique.

As already mentioned, another standard way to handle the uniqueness question of the coefficients of τ is to require that

$$\Delta = ad - bc = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = 1.$$

The corresponding subgroup of $GL(2, \mathbb{C})$ is called the *special linear group* $SL(2, \mathbb{C})$. To analyze the asymptotic behavior of $\{\tau_n\}$ one can study the asymptotic behavior of $\{M(\tau_n)\}$ and vice versa. The tools at hand are different in $\mathcal{M}$ and $GL(2, \mathbb{C})$, results on sequences of matrix products

$$M_1, \ M_1 M_2, \ M_1 M_2 M_3, \ldots \quad \text{or} \quad M_1, \ M_2 M_1, \ M_3 M_2 M_1, \ldots$$

in $GL(2, \mathbb{C})$ have their equivalent counterparts for sequences

$$\tau_n := \varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_n \quad \text{or} \quad \tau_n := \psi_n \circ \psi_{n-1} \circ \cdots \circ \psi_1 \tag{41}$$

in $\mathcal{M}$. A particularly well studied case is the case where all M_n are equal and all φ_n are equal. The case $M = I$ is trivial, so let $M \neq I$. For M^n we first look for eigenvalues λ_1, λ_2 and the corresponding eigenvectors $\begin{pmatrix} x \\ 1 \end{pmatrix}$, $\begin{pmatrix} y \\ 1 \end{pmatrix}$ for M by solving the equation

$$(\lambda I - M) \begin{pmatrix} x \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

which gives

$$\lambda_{1,2} = \frac{a + d \pm \sqrt{(a + d)^2 - 4\Delta}}{2}, \quad x, y = \frac{a - d \pm \sqrt{(a + d)^2 - 4\Delta}}{2c}.$$

We choose the notation here such that $\lambda_1 = cx + d$ and $\lambda_2 = cy + d$. For simplicity we assume that x and y are finite with $x \neq y$. Then M has the representation

$$M = \begin{pmatrix} x & y \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} x & y \\ 1 & 1 \end{pmatrix}^{-1}$$

which means that

$$M^n = \begin{pmatrix} x & y \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{pmatrix} \begin{pmatrix} x & y \\ 1 & 1 \end{pmatrix}^{-1} = \lambda_2^n \begin{pmatrix} x & y \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \Re^n & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x & y \\ 1 & 1 \end{pmatrix}^{-1} \quad \text{where } \Re := \frac{\lambda_1}{\lambda_2}.$$

The counterpart for the nth iterate $\varphi^{[n]} := \varphi \circ \varphi \circ \cdots \circ \varphi$ (composition of n copies of φ) takes the form

$$\vartheta \circ \varphi(w) = \frac{\lambda_1}{\lambda_2}\, \vartheta(w) \quad \Leftrightarrow \quad \varphi = \vartheta^{-1} \circ \chi \circ \vartheta,$$
$$\text{where } \vartheta(w) := \frac{w - y}{w - x} \text{ and } \chi(w) := \frac{\lambda_1 w}{\lambda_2}.$$

The points x and y are the two fixed points of φ and $\lambda_1 = cx + d$, $\lambda_2 = cy + d$. Hence

$$\varphi^{[n]}(w) = \vartheta^{-1} \circ \chi^{[n]} \circ \vartheta \quad \text{where} \quad \chi^{[n]}(w) = \left(\frac{\lambda_1}{\lambda_2}\right)^n w = \Re^n w.$$

This means that $\{\varphi^{[n]}\}$ converges generally to $\vartheta^{-1}(0) = y$ if $|\Re| < 1$ and to $\vartheta^{-1}(\infty) = x$ if $|\Re| > 1$. If $|\Re| = 1$ (with $\Re \neq 1$), then $\{\varphi^{[n]}\}$ diverges. We recognize the situation from Section 2.7.2.

3.3 More general group theory

If we start with some subset $\mathcal{B}$ of $\mathcal{M}$ and consider the family $\mathcal{G} \subseteq \mathcal{M}$ which contain $\mathcal{B}$ and the identity transformation $I(w) \equiv w$ and has the properties that

$$\tau \in \mathcal{G} \;\Rightarrow\; \tau^{-1} \in \mathcal{G} \quad \text{and} \quad \tau_1, \tau_2 \in \mathcal{G} \;\Rightarrow\; \tau_1 \circ \tau_2 \in \mathcal{G}, \tag{42}$$

then $\mathcal{G}$ is the group under compositions generated by $\mathcal{B}$. It is a very nice group in the following sense: every sequence from $\mathcal{M}$ has a subsequence which either converges $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}}$ to a τ from $\mathcal{M}$, or locally $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}} \setminus \{w^\dagger\}$ for some $w^\dagger \in \widehat{\mathbb{C}}$ to a constant. This is what is called a *convergence group*. A subgroup $\mathcal{G}$ of $\mathcal{M}$ is called *discrete* if no sequence of distinct elements from $\mathcal{G}$ converges to some $\tau \in \mathcal{M}$; i.e., if every sequence from $\mathcal{M}$ has a generally convergent subsequence. This has lead to a theory of convergence groups more generally.

The group properties (42) are rather restrictive. For instance, $\mathcal{G} := \{\tau \in \mathcal{M};\ \tau(V) = V\}$ for a given set $V \subseteq \widehat{\mathbb{C}}$ is a group, but the transformations $\tau \in \mathcal{M}$ which map V into V does not form a group since $\tau(V) \subseteq V$ does not imply that $\tau^{-1}(V) \subseteq V$ when V contains more than one element. This is unfortunate since $\tau_n(V) \subseteq V$ for all n can for instance be an important factor in a proof for convergence. What we want is a theory for semigroups where the condition $\tau \in \mathcal{G} \;\Rightarrow\; \tau^{-1} \in \mathcal{G}$ is no longer part of the requirements. Actually, it is fair to say that convergence criteria for special sequences of elements from $\mathcal{G}$ has not been a hot topic in this area. The emphasis is more on characterizing properties of the elements in the group as a whole.

3.4 Recurrence relations

Let

$$\tau_n(w) := \frac{A_n w + B_n}{C_n w + D_n}, \quad \varphi_n(w) := \frac{a_n w + b_n}{c_n w + d_n}, \quad \tau_n = \varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_n$$

be transformations from $\mathcal{M}$ regarded as embeddings in $GL(2, \mathbb{C})$; i.e., the coefficients are unique and $M(\tau_n) = M(\tau_{n-1} \circ \varphi_n) = M(\tau_{n-1}) \cdot M(\varphi_n)$. This is equivalent to the fact that the coefficients of τ_n satisfy the recurrence relations

$$\begin{pmatrix} A_n & B_n \\ C_n & D_n \end{pmatrix} = \begin{pmatrix} A_{n-1} & B_{n-1} \\ C_{n-1} & D_{n-1} \end{pmatrix} \begin{pmatrix} a_n & b_n \\ c_n & d_n \end{pmatrix},$$

that is,

$$A_n = a_n A_{n-1} + c_n B_{n-1}, \quad C_n = a_n C_{n-1} + c_n D_{n-1},$$
$$B_n = b_n A_{n-1} + d_n B_{n-1}, \quad D_n = b_n C_{n-1} + d_n D_{n-1}.$$

In the following we assume that $c_n = 1$ and $a_n = 0$ for all n, just as they are for continued fractions. (See also remark 4 on page 130.) Then $A_n = B_{n-1}$ and $C_n = D_{n-1}$, and thus

$$B_n = d_n B_{n-1} + b_n B_{n-2}, \quad D_n = d_n D_{n-1} + b_n D_{n-2}. \tag{43}$$

That is, $\{B_n\}$ and $\{D_n\}$ are solutions of the homogeneous, linear recurrence relation

$$X_n = d_n X_{n-1} + b_n X_{n-2} \quad \text{for } n = 1, 2, 3, \ldots \tag{44}$$

with initial values $B_{-1} := 1$, $B_0 := 0$, $D_{-1} := 0$ and $D_0 := 1$. Since $\varphi_n \in \mathcal{M}$, we know that $b_n \neq 0$ for all n.

Conversely, if $\{B_n\}$ and $\{D_n\}$ are solutions of (44) with these initial values, and all $b_n \neq 0$, then the transformations

$$\varphi_n(w) := \frac{b_n}{w + d_n} \quad \text{give } \tau_n(w) = \varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_n(w) = \frac{A_n w + B_n}{C_n w + D_n} \tag{45}$$

where $A_n = B_{n-1}$ and $C_n = D_{n-1}$ for all n. One of the questions asked in this theory is what kind of asymptotic behavior of the solutions of (44) can we expect? For instance, are two solutions asymptotically equal? Since

$$\tau_n(0) = \frac{B_n}{D_n} = \frac{A_{n+1}}{C_{n+1}} = \tau_{n+1}(\infty),$$

a convergence $(B_n/D_n) \to \gamma \in \widehat{\mathbb{C}}$ can only occur if $\{\tau_n\}$ converges generally to γ. Indeed, the solution space of (44) is a linear vector space of dimension 2 since all $b_n \neq 0$. Since $\{B_n\}$ and $\{D_n\}$ are linearly independent, they form a basis in this vector space; i.e., every solution $\{X_n\}$ can be written as a linear combination of $\{B_n\}$ and $\{D_n\}$:

$$\{X_n\} = \beta\,\{B_n\} + \delta\,\{D_n\} \quad \text{for some constants } \beta,\, \delta \in \mathbb{C}. \tag{46}$$

Hence, if $(B_n/D_n) \to \gamma$, then $(X_n/D_n) \to \beta\gamma + \delta$.

A solution $\{X_n\}$ of (44) is said to be *minimal* if it is non-trivial (i.e. not all X_n are equal to 0) and there exists a second solution $\{Y_n\}$ such that $(X_n/Y_n) \to 0$. We say that $\{Y_n\}$ is a *dominant* solution in this case. Then most of the solutions of (44) are dominant, since also $(\{X_n\}, \{Y_n\})$ is a basis of the solution space. Indeed, the minimal solutions form a subspace of dimension 1, and the rest of the solutions are dominant.

Theorem 3.1. *[25] Let $\{P_n\}$ and $\{Q_n\}$ be two linearly independent solutions of (44). Then $\{P_n/Q_n\}$ converges to some $\gamma \in \widehat{\mathbb{C}}$ if and only if (44) has a minimal solution $\{X_n\}$. In particular $(B_n/D_n) \to -X_0/X_{-1}$.*

Proof. Since $\{P_n\}$ and $\{Q_n\}$ are linearly independent, we know that $\Lambda := P_{-1}Q_0 - P_0Q_{-1} \neq 0$. Moreover, a second pair $(\{\tilde{P}_n\}, \{\tilde{Q}_n\})$ of solutions is also linearly independent if and only if

$$\{\tilde{P}_n\} = \beta_1\{P_n\} + \beta_2\{Q_n\} \quad \text{and} \quad \{\tilde{Q}_n\} = \delta_1\{P_n\} + \delta_2\{Q_n\}$$

for some complex constants β_1, β_2, δ_1 and δ_2 with $\beta_1\delta_2 - \beta_2\delta_1 \neq 0$. Since then

$$\frac{\tilde{P}_n}{\tilde{Q}_n} = \frac{\beta_1 P_n + \beta_2 Q_n}{\delta_1 P_n + \delta_2 Q_n} = \frac{\beta_1 \frac{P_n}{Q_n} + \beta_2}{\delta_1 \frac{P_n}{Q_n} + \delta_2}$$

(with natural limit forms if $Q_n = 0$), we find that $\{\tilde{P}_n/\tilde{Q}_n\}$ converges if and only if $\{P_n/Q_n\}$ converges.

In the convergence case, every number $\gamma \in \widehat{\mathbb{C}}$ can be a limit of $\{\tilde{P}_n/\tilde{Q}_n\}$ by appropriate choice of $\beta_1, \beta_2, \delta_1$ and δ_2. In particular $\gamma = 0$ is possible, which makes $\{\tilde{P}_n\}$ minimal.

Let $\{P_n\}$ be such a minimal solution, and set $\{\tilde{P}_n\} := \{B_n\}$ and $\{\tilde{Q}_n\} := \{D_n\}$. Since $B_{-1} = D_0 = 1$ and $B_0 = D_{-1} = 0$, we must have

$$\beta_1 = \frac{Q_0}{\Lambda}, \quad \beta_2 = -\frac{P_0}{\Lambda}, \quad \delta_1 = -\frac{Q_{-1}}{\Lambda} \quad \text{and} \quad \delta_2 = \frac{P_{-1}}{\Lambda},$$

and thus $(B_n/D_n) = (\tilde{P}_n/\tilde{Q}_n) \to \beta_2/\delta_2 = -P_0/P_{-1}$. □

The recurrence relation (44) can be written on matrix form:

$$\begin{pmatrix} -b_1 & -d_1 & 1 & 0 & 0 & 0 & \cdots \\ 0 & -b_2 & -d_2 & 1 & 0 & 0 & \cdots \\ 0 & 0 & -b_3 & -d_3 & 1 & 0 & \cdots \\ 0 & 0 & 0 & -b_4 & -d_4 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \end{pmatrix} \begin{pmatrix} X_{-1} \\ X_0 \\ X_1 \\ X_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}$$

and by induction, the solution $\{D_n\}$ with $D_{-1} := 0$ and $D_0 := 1$ can be written as

$$\begin{aligned} D_n = d_n D_{n-1} + b_n D_{n-2} &= \begin{vmatrix} D_{n-1} & D_{n-2} \\ -b_n & d_n \end{vmatrix} = \begin{vmatrix} D_{n-2} & D_{n-3} & 0 \\ -b_{n-1} & d_{n-1} & 1 \\ 0 & -b_n & d_n \end{vmatrix} \\ &= \begin{vmatrix} D_{n-3} & D_{n-4} & 0 & 0 \\ -b_{n-2} & d_{n-2} & 1 & 0 \\ 0 & -b_{n-1} & d_{n-1} & 1 \\ 0 & 0 & -b_n & d_n \end{vmatrix} = \cdots \\ &= \begin{vmatrix} d_1 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ -b_2 & d_2 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -b_3 & d_3 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -b_n & d_n \end{vmatrix} \end{aligned} \tag{47}$$

which again can be written as the symmetric tridiagonal determinant

$$D_n = \begin{vmatrix} d_1 & \sqrt{-b_2} & 0 & 0 & 0 & \cdots & 0 & 0 \\ \sqrt{-b_2} & d_2 & \sqrt{-b_3} & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sqrt{-b_3} & d_3 & \sqrt{-b_4} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \sqrt{-b_n} & d_n \end{vmatrix} \tag{48}$$

Since $\{B_n\}$ is a similar solution of the shifted recurrence relation ($B_0 = 0,\ B_1 = b_1$), we also get

$$B_n = b_1 \begin{vmatrix} d_2 & \sqrt{-b_3} & 0 & 0 & \cdots & 0 & 0 \\ \sqrt{-b_3} & d_3 & \sqrt{-b_4} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \sqrt{-b_n} & d_n \end{vmatrix} = b_1 \begin{vmatrix} d_2 & 1 & 0 & 0 & \cdots & 0 & 0 \\ -b_3 & d_3 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -b_n & d_n \end{vmatrix}. \tag{49}$$

Similar connections between recurrence relations and linear fractional transformations can be derived from

$$\tau_n(w) = \frac{A_n w + B_n}{C_n w + D_n}, \quad \psi_n(w) = \frac{a_n w + b_n}{c_n w + d_n}, \quad \tau_n = \psi_n \circ \psi_{n-1} \circ \cdots \circ \psi_1.$$

In particular if all $a_n = 0$ and $b_n = 1$, then

$$\begin{pmatrix} A_n & B_n \\ C_n & D_n \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ c_n & d_n \end{pmatrix} \begin{pmatrix} A_{n-1} & B_{n-1} \\ C_{n-1} & D_{n-1} \end{pmatrix}.$$

That is, $A_n = C_{n-1}$ and $B_n = D_{n-1}$, where $\{C_n\}$ and $\{D_n\}$ are solutions of the recurrence relation

$$X_n = d_n X_{n-1} + c_n X_{n-2}.$$

3.5 Orthogonal polynomials

Let $d\Psi(t)$ be a positive measure with infinite support on the real line, with finite moments

$$c_n := \int_{\mathbb{R}} t^n \, d\Psi(t) \quad \text{for } n = 0, 1, 2, \ldots,$$

and let $\{P_n(x)\}_{n=0}^{\infty}$ be the corresponding sequence of orthogonal monic polynomials; i.e., $P_n(x) = x^n +$ lower degree terms, and

$$\int_{\mathbb{R}} P_n(t)P_m(t)\,d\Psi(t) \text{ is } \begin{cases} = 0 & \text{for } n \neq m, \\ \neq 0 & \text{for } n = m. \end{cases} \tag{50}$$

For convenience we write $\mathfrak{L}(f)$ for integrals of functions f over $\mathbb{R}$ with respect to this measure. That is, the integral above can be written $\mathfrak{L}(P_nP_m)$.

Evidently, $xP_{n-1}(x)$ has the orthogonal expansion

$$xP_{n-1}(x) = P_n + \sum_{k=0}^{n-1} q_kP_k(x) \quad \text{where} \quad q_k = \frac{\mathfrak{L}(xP_{n-1}P_k)}{\mathfrak{L}(P_k^2)} \text{ for } k = 0, 1, \ldots, n-1.$$

Since $\mathfrak{L}(xP_{n-1}P_k) = \mathfrak{L}(P_{n-1}(xP_k))$ where xP_k also has such an orthogonal expansion, it follows by (50) that $q_k = 0$ for $k < n-2$. Hence $xP_{n-1} = P_n + q_{n-1}P_{n-1} + q_{n-2}P_{n-2}$ which means that $\{P_n(x)\}$ is the solution of the recurrence relation

$$P_n = (x - b_n)P_{n-1} - a_n^2P_{n-2} \quad \text{for } n = 1, 2, 3, \ldots \tag{51}$$

with $P_{-1} := 0$, $P_0 := 1$ and

$$b_n = q_{n-1} = \frac{\mathfrak{L}(xP_{n-1}^2)}{\mathfrak{L}(P_{n-1}^2)}, \;\; a_n^2 = q_{n-2} = \frac{\mathfrak{L}(xP_{n-1}P_{n-2})}{\mathfrak{L}(P_{n-2}^2)} = \frac{\mathfrak{L}(P_{n-1}^2)}{\mathfrak{L}(P_{n-2}^2)}. \tag{52}$$

Conversely, by Favard's theorem [6] (first proved by Stieltjes [29]) we know that if $\{P_n\}_{n=0}^{\infty}$ is the solution of (51) where a_n, $b_n \in \mathbb{R}$, $a_n \neq 0$ and $P_{-1}(x) \equiv 0$ and $P_0(x) \equiv 1$, then $\{P_n\}$ is sequence of monic orthogonal polynomials with respect to some positive measure $d\mu(t)$ with infinite support $\subseteq \mathbb{R}$. As in (47)-(48), $\{P_n\}$ is then given by

$$P_n(x) = \begin{vmatrix} x-b_1 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ a_2^2 & x-b_2 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & a_3^2 & x-b_3 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & a_n^2 & x-b_n \end{vmatrix} \tag{53}$$

which also can be written

$$P_n(x) = \begin{vmatrix} x-b_1 & a_2 & 0 & 0 & 0 & \cdots & 0 & 0 \\ a_2 & x-b_2 & a_3 & 0 & 0 & \cdots & 0 & 0 \\ 0 & a_3 & x-b_3 & a_4 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & a_n & x-b_n \end{vmatrix}. \tag{54}$$

Asymptotics for $\{P_n\}$ is important. Ratio asymptotics means limiting behavior of P_{n+1}/P_n which in the notation of the previous section can be written $D_{n+1}/D_n = -\tau_{n+1}^{-1}(\infty)$.

The solution $\{P_n^{(1)}\}$ with $P_0^{(1)} = 0$, $P_1^{(1)} = 1$ is called the sequence of *associated polynomials* related to the measure. Also the asymptotics of $\{P_{n-1}^{(1)}/P_n\}$ is of interest. Clearly

$$\frac{P_{n-1}^{(1)}}{P_n} = \frac{1}{z-b_1} - \frac{a_2^2}{z-b_2} - \frac{a_3^2}{z-b_3} - \cdots - \frac{a_n^2}{z-b_n}$$

which for instance can be written as $\varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_n(0)$ where $\varphi_k(w) := 1/(z - b_k - a_{k+1}^2 w)$. If $f(z) := \lim P_{n-1}^{(1)}(z)/P_n(z)$ exists; i.e., if the corresponding continued fraction converges to $f(z)$, then this can be used to determine the measure $d\mu(t)$. (This will be the topic in Section 4.3.)

Of course, here we restricted the situation to the classical cases where the support of the measure is real, its moments are of orders ≥ 0 and the orthogonal functions $\{P_n(x)\}$ are monic polynomials. This has been extended to more general situations.

3.6 Linear operators

We consider linear self adjoint operators A on an infinitely dimensional Hilbert space, and require that the matrix of A has a symmetric tridiagonal form

$$A = \begin{pmatrix} b_1 & a_1 & 0 & 0 & 0 & \cdots \\ a_1 & b_2 & a_2 & 0 & 0 & \cdots \\ 0 & a_2 & b_3 & a_3 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \qquad \text{with all } a_n \neq 0.$$

Searching for eigenvalues λ and eigenvectors $x = (x_0, x_1, x_2, \cdots)^t$ (the superscript t means the transpose), we require that

$$(\lambda I - A) \cdot x = 0$$

i.e.,

$$\begin{pmatrix} \lambda - b_1 & -a_1 & 0 & 0 & 0 & \cdots \\ -a_1 & \lambda - b_2 & -a_2 & 0 & 0 & \cdots \\ 0 & -a_2 & \lambda - b_3 & -a_3 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix},$$

i.e.,

$$-a_{n-1}x_{n-1} + (\lambda - b_n)x_n - a_n x_{n+1} = 0 \quad \text{for } n = 2, 3, 4, \ldots$$

with $x_{-1} := 0$ and $x_0 := 1$. By Favard's theorem this means that x is an eigenvector for A if and only if $\{x_n(\lambda)\}$ is a sequence of orthogonal polynomials of λ and $\sum |x_n(\lambda)|^2 < \infty$. Indeed, one can prove that $\{x_n(\lambda)\}$ is a sequence of orthonormal polynomials with respect to some measure under certain conditions.

3.7 Moment problems

There exists a whole bunch of moment problems. In Section 1.2.3 we looked at the Stieltjes moment problem. Each one of these problems are connected to a continued fraction of a certain shape. In Section 4.3 we shall for instance consider the Hamburger moment problem which is connected to continued fractions of the form

$$\cfrac{z}{z - b_1} - \cfrac{a_2^2}{z - b_2} - \cfrac{a_3^2}{z - b_3} - \cdots .$$

3.8 Discrete dynamical systems

Let f be a function mapping a set V into itself. For given $p_0 \in V$, the structure

$$p_n = f(p_{n-1}) \quad \text{for } n = 1, 2, 3, \ldots$$

is a dynamical system in its simplest form. The sequence $\{p_n\}$ is called an *orbit* or a *graph trajectory* of the system, and the equation is to be understood as an alternative to differential equations.

In more general cases, the function f may also depend on a time parameter; i.e., $f = f(t, w)$ such that $f(t_k, w) = f_k(w)$ and thus

$$p_n = f_n(p_{n-1}) = f_n \circ f_{n-1}(p_{n-2}) = \cdots = f_n \circ f_{n-1} \circ \cdots \circ f_1(p_0). \tag{55}$$

Of course, $\{f_n\}$ is a sequence from $\mathcal{M}$ only in very special cases, but techniques to prove convergence for sequences from $\mathcal{M}$ may sometimes be adapted to (55) and vice versa.

The "filled-in" Julia set for f (or $\{f_n\}$) is the set of points $p_0 \in V$ for which $\{p_n\}$ does not approach infinity. The true Julia set is the boundary of the filled-in set (the set of "exceptional points"). If the Julia set is connected, it is often called a Fatou set. Otherwise, if it is a Cantor set, it may be called Fatou dust.

3.9 Random iteration

We say that $\mathbf{K}(a_n/b_n)$ is a continued fraction from $\Omega \subseteq (\mathbb{C} \setminus \{0\}) \times \mathbb{C}$ if $(a_n, b_n) \in \Omega$ for all n. In continued fraction theory, a number of sufficient convergence criteria has the following form: *If $(a_n, b_n) \in \Omega$ for all n (for some given Ω), then $\mathbf{K}(a_n/b_n)$ converges.* This is a result on what we might call random iteration. Pick a continued fraction from Ω at random. Does it converge or not? Or more generally: what is the probability for convergence, given a probability distribution on the continued fractions from Ω.

As already mentioned, a sequence $\{\tau_n\}$ from $\mathcal{M}$ can always be regarded as a sequence

$$\tau_n = \varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_n \quad \text{or} \quad \tau_n = \psi_n \circ \psi_{n-1} \circ \cdots \circ \psi_1 \tag{56}$$

of compositions of transformations from $\mathcal{M}$. In this case we have:

Lemma 3.2. *Let $\{\varphi_n\}$ and $\{\psi_n\}$ be sequences from $\mathcal{M}$. If $\{\tau_n\}$ given by (56) converges $\mathfrak{m}$-uniformly in $\widehat{\mathbb{C}}$, then $\varphi_n \to I$ and $\psi_n \to I$.*

Proof. $\varphi_n = \tau_{n-1}^{-1} \circ \tau_n \to \tau^{-1} \circ \tau = I, \quad \psi_n = \tau_n \circ \tau_{n-1}^{-1} \to \tau \circ \tau^{-1} = I.$ □

In combination with the following result by Ambroladze and Wallin [1] on transformations from $\mathcal{M}_{\mathbb{H}} := \{\tau \in \mathcal{M};\, \tau(\mathbb{H}) \subseteq \mathbb{H}\}$ where $\mathbb{H}$ is the upper half plane $\mathbb{H} := \{w \in \mathbb{C};\, \mathrm{Im}(w) > 0\}$, this gets quite interesting:

Theorem 3.3. *Let $d\mu$ be a probability measure on $\mathcal{M}_{\mathbb{H}}$, and assume that the transformations in the support of μ have no common fixed point in $\overline{\mathbb{H}}$ and no common invariant hyperbolic line in $\mathbb{H}$. Let $\{\psi_n\}$ be a random sequence from $\mathcal{M}_{\mathbb{H}}$ with distribution μ. Then, for any point $w \in \mathbb{H}$, $\tau_n(w) := \psi_n \circ \psi_{n-1} \circ \cdots \circ \psi_1(w)$ tends to $\overline{\mathbb{R}}$ almost surely.*

This was actually a consequence of Fürstenberg's more general result [7]:

Theorem 3.4. *Let $d\mu$ be a probability measure on $SL(2,\mathbb{R})$ and let $\mathcal{G}_\mu$ be the smallest closed subgroup of $SL(2,\mathbb{R})$ which contains the support of $d\mu$. Assume that $\mathcal{G}_\mu$ is not compact and no subset L of $\mathbb{R}^2$ is a finite union of one-dimensional subspaces with $ML = L$ for some $M \in \mathcal{G}_\mu$. Finally, let $\{M_n\}$ be a random sequence from $SL(2,\mathbb{R})$ with distribution μ. Then, with probability 1, $||\tilde{M}_n x||$ grows exponentially for $x = (x_1, x_2)^t \in \mathbb{R}^2 \setminus \{0\}$, where $\tilde{M}_n := M_n M_{n-1} \cdots M_1$.*

Here $||M|| = \sqrt{|a|^2 + |b|^2 + |c|^2 + |d|^2}$ for $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, the standard norm in $SL(2\mathbb{R})$. Hence, with obvious interpretation, $||\psi_n|| \to \infty$ implies that no subsequence of $\{\tau_n\}$ converges to some $\tau \in \mathcal{M}$.

4 The trinity of moment problems, orthogonal polynomials and continued fractions

There is a beautiful theory of equivalences between the three fields

- continued fractions
- orthogonal polynomials
- moment problems

We shall look at a particular example where the continued fractions have the form

$$\frac{z}{z-b_1} - \frac{a_2^2}{z-b_2} - \frac{a_3^2}{z-b_3} - \cdots \quad \text{where all } a_n^2 > 0,\ b_n \in \mathbb{R}, \tag{57}$$

the orthogonal polynomials are as in Section 3.5 and the moment problem is *the Hamburger moment problem:* For a given sequence $\{c_n\}_{n=0}^{\infty}$ of real numbers with $c_0 := 1$, find a probability measure $d\Psi(t)$ of infinite support on $\mathbb{R}$ such that

$$c_n = \int_{\mathbb{R}} t^n \, d\Psi(t).$$

The idea is to present some theorems without proofs to illustrate the connections.

4.1 J-fractions

A continued fraction of the form (57) is called a *positive definite J-fraction*. It has some nice properties. For instance, the upper half plane $\mathbb{H}$ is a simple value set for this continued fraction when $z \in \mathbb{H}$. This follows easily since with $z := x + iy$ and $w := u + iv$ where $y > 0$ and $v > 0$ we have

$$s_n(w) = \frac{-a_n^2}{z - b_n + w} = \frac{-a_n^2}{(x + u - b_n) + i(y + v)} = \frac{-a_n^2((x + u - b_n) - i(y + v))}{(x + u - b_n)^2 + (y + v)^2} \in \mathbb{H}.$$

The following convergence property is therefore a consequence of a generalization of the Parabola Theorem in Section 2.5.

Theorem 4.1. *The positive definite J-fraction (57) converges locally uniformly with respect to z in the upper half plane $\mathbb{H}$ to a holomorphic function $f(z)$ in $\mathbb{H}$ if*

$$\sum_{n=1}^{\infty} \frac{1}{|a_n|} = \infty \quad \text{or} \quad \sum_{n=1}^{\infty} \left| \frac{b_{n+1}}{a_n a_{n+1}} \right| = \infty. \tag{58}$$

Of course, by complex conjugation, the same holds true if we replace the upper half plane by the lower half plane.

Another important property is its correspondence to some formal series

$$L(z) := \sum_{n=0}^{\infty} (-1)^n \frac{c_n}{z^n}; \tag{59}$$

that is, the Taylor series expansion of the nth approximant $f_n(z)$ of (57) has the form

$$f_n(z) = \frac{A_n(z)}{B_n(z)} \sim c_0 - \frac{c_1}{z} + \frac{c_2}{z^2} - \cdots + \frac{c_{2n}}{z^{2n}} + \frac{d_{2n+1,n}}{z^{2n+1}} + \cdots \quad \text{for } n = 0, 1, 2, \ldots \tag{60}$$

where $\{d_{k,n}\}_{k>2n}$ are some real numbers depending on n.

Theorem 4.2. *The positive definite J-fraction (57) corresponds to a unique series (59).*

The canonical denominators $\{B_n\}$ of (57) are of course monic polynomials in z of exact degree n, with $B_{-1} :\equiv 0$, $B_0 :\equiv 1$ and

$$B_n(z) = (z - b_n)B_{n-1}(z) - a_n^2 B_{n-2}(z) \quad \text{for } n = 1, 2, 3, \ldots. \tag{61}$$

The *Hankel determinants* for a given sequence $\{c_n\}_{n=0}^{\infty}$ of complex numbers (or series $L(z)$ as in (59)) are given by

$$H_0^{(m)} := 1, \quad H_n^{(m)} := \begin{vmatrix} c_m & c_{m+1} & \cdots & c_{m+n-1} \\ c_{m+1} & c_{m+2} & \cdots & c_{m+n} \\ \vdots & \vdots & & \vdots \\ c_{m+n-1} & c_{m+n} & \cdots & c_{m+2n-2} \end{vmatrix} \quad \text{for } m \geq 0, \ n \geq 1. \tag{62}$$

They are useful in this theory. For instance, it turns out that the canonical denominators of (57) can be written

$$B_n(z) = \frac{1}{H_n^{(1)}} \begin{vmatrix} c_1 & c_2 & \cdots & c_{n+1} \\ c_2 & c_3 & \cdots & c_{n+2} \\ \vdots & \vdots & & \vdots \\ z^n & z^{n-1} & \cdots & 1 \end{vmatrix} \qquad \text{where } H_n^{(1)} > 0. \tag{63}$$

Theorem 4.3. *Let $L(z) := \sum_{n=0}^{\infty} (-1)^n c_n / z^n$ have a continued fraction expansion of the form (57). Then all $a_n^2 > 0$ and $b_n \in \mathbb{R}$ for all n if and only if $H_n^{(1)} > 0$ for all n.*

Another interesting feature is the zeros of the denominators of $S_n(\tau)$ for some given $\tau \in \mathbb{R}$; i.e., for

$$Q_n(z) := B_n(z) + \tau B_{n-1}(z) \quad \text{for } n \in \mathbb{N}. \tag{64}$$

Theorem 4.4. *The zeros $\{z_k^{(n)}\}_{k=1}^n$ of Q_n are all real and simple, and they can be ordered such that*

$$z_1^{(n+1)} < z_k^{(n)} < z_{k+1}^{(n+1)} < z_{k+1}^{(n)} < z_{n+1}^{(n+1)} \quad \text{for } k = 1, 2, \ldots, n-1. \tag{65}$$

Moreover, the approximants $S_n(\tau) = \dfrac{A_n + \tau A_{n-1}}{B_n + \tau B_{n-1}} =: \dfrac{P_n}{Q_n}$ has a partial fraction decomposition

$$\frac{P_n}{Q_n} = \sum_{k=1}^{n} \frac{\lambda_k^{(n)}}{z - z_k^{(n)}} \quad \text{where } \lambda_k^{(n)} > 0 \text{ for all } k. \tag{66}$$

4.2 Orthogonal polynomials – J fractions

Let $d\Psi(t)$ be a positive probability measure with infinite support on $\mathbb{R}$ and finite moments

$$c_n := \int_{\mathbb{R}} t^n \, d\Psi(t) \quad \text{for } n = 0, 1, 2, \ldots . \tag{67}$$

As in Section 3.5 we use the notation

$$\mathfrak{L}(f(t)) := \int_{\mathbb{R}} f(t) \, d\Psi(t) \quad \text{for} \quad f : \mathbb{R} \to \mathbb{R}, \tag{68}$$

so that $c_n = \mathfrak{L}(t^n)$. Then $\mathfrak{L}$ is a linear operator on the space $\mathcal{P}$ of polynomials $P(t)$ of arbitrary degree. By applying Gramm-Schmidt orthogonalization to the sequence

$$1, \, t, \, t^2, \, t^3, \, \ldots$$

of monomials in $\mathcal{P}$, we can obtain the corresponding sequence $\{P_n(t)\}_{n=0}^{\infty}$ of monic orthogonal polynomials with respect to $d\Psi(t)$. We have already seen in Section 3.5 that $\{P_n(z)\}$ also is the sequence of canonical denominators of the continued fraction

$$\frac{z}{z - b_1} - \frac{a_2^2}{z - b_2} - \frac{a_3^2}{z - b_3} - \cdots \quad \text{where } a_n^2 = \frac{\mathfrak{L}(P_{n-1}^2)}{\mathfrak{L}(P_{n-2}^2)} > 0, \quad b_n = \frac{\mathfrak{L}(tP_{n-1}^2)}{\mathfrak{L}(P_{n-1}^2)} \in \mathbb{R}. \tag{69}$$

The coefficients can also be represented by Hankel determinants

$$a_n^2 = \frac{H_n^{(1)}/H_{n-1}^{(1)}}{H_{n-1}^{(1)}/H_{n-2}^{(1)}}, \quad b_n = \frac{G_n}{H_n^{(1)}} - \frac{G_{n-1}}{H_{n-1}^{(1)}} \quad \text{where} \quad G_n := \begin{vmatrix} c_1 & \cdots & c_{n-1} & c_{n+1} \\ c_2 & \cdots & c_n & c_{n+2} \\ \vdots & & \vdots & \vdots \\ c_n & \cdots & c_{2n-2} & c_{2n} \end{vmatrix}.$$

Conversely, let $z\mathbf{K}(-a_n^2/(z - b_n))$ be a given continued fraction with $a_n^2 > 0$ and $b_n \in \mathbb{R}$ for all n. Then its canonical denominators P_n are given by

$$P_n(z) = (z - b_n)P_{n-1}(z) - a_n^2 P_{n-2}(z) \quad \text{for } n = 1, 2, 3, \ldots \tag{70}$$

with $P_{-1}(z) \equiv 0$ and $P_0(z) \equiv 1$. In particular $\{P_n\}$ are monic polynomials of exact degree n. By Favard's Theorem they are the monic orthogonal polynomials for some positive measure $d\Psi(t)$ with infinite support on $\mathbb{R}$. To retrieve this measure (or measures) is exactly what the moment problem is all about.

4.3 Moment problems – J-fractions

Theorem 4.5. *The Hamburger moment problem has a solution if and only if*

$$H_n^{(0)} := \begin{vmatrix} c_0 & c_1 & \cdots & c_{n-1} \\ c_1 & c_2 & \cdots & c_n \\ \vdots & \vdots & & \vdots \\ c_{n-1} & c_n & \cdots & c_{2n-2} \end{vmatrix} > 0 \quad \text{for } n = 1, 2, 3, \ldots .$$

Theorem 4.6. *If $d\Psi(t)$ is a solution of the Hamburger moment problem, then $\sum(-1)^n c_n/z^n$ is an asymptotic expansion of*

$$f(z) := \int_{\mathbb{R}} \frac{z}{z+t}\, d\Psi(t) \tag{71}$$

in the upper half plane $\mathbb{H}$ in the sense that

$$\lim_{z=iy,\, y\to\infty} z^n (f(z) - \sigma_n(z)) = 0 \quad \text{for all } n, \text{ where } \sigma_n(z) := \sum_{k=0}^{n} (-1)^k \frac{c_k}{z^k}. \tag{72}$$

Theorem 4.7. *If all $H_n^{(0)} > 0$, then*

(i) $\sum c_n z^n$ has a corresponding continued fraction $\mathbf{K}(-a_n^2/(z-b_n))$ where $\{a_n\}$ and $\{b_n\}$ are given by (52).

(ii) $d\Psi(t)$ is unique if and only if $\mathbf{K}(-a_n^2/(z-b_n)$ converges. This continued fraction then converges to $\int_{\mathbb{R}} \frac{1}{1+tz}\, d\psi(t)$ from which $d\psi(t)$ can be retrieved.

(iii) If $\mathbf{K}(-a_n^2/(z-b_n))$ diverges, then it has exactly two limit functions, and the possibilities for $d\psi(t)$ are all the linear combinations of the two measures retrieved from each of these two limits.

Example 4.8. Let the sequence $\{c_n\}_{n=0}^{\infty}$ with $c_n := n!$ for all n be given. We are looking for a probability measure $d\Psi(t)$ such that

$$c_n = n! = \int_0^{\infty} t^n\, d\Psi(t) \quad \text{for all } n.$$

We pretend that we have never heard about the Gamma function which naturally gives the answer

$$\int_0^{\infty} t^n e^{-t}\, dt = n!$$

and thus that $d\Psi(t) = e^{-t}\, dt$.

This is actually a Stieltjes moment problem (a special case of the Hamburger moment problem). We therefore try to find a continued fraction expansion of the form $z\mathbf{K}(a_n z^{-1}/1)$ with all $a_n > 0$ for

$$L(z) := \sum_{n=0}^{\infty} (-1)^n \frac{c_n}{z^n} = \sum_{n=0}^{\infty} n!(-z)^{-n}.$$

This was exactly what we did in Example 4.5.1 where we got the continued fraction expansion

$$\frac{1}{1} + \frac{1/z}{1} + \frac{1/z}{1} + \frac{2/z}{1} + \frac{2/z}{1} + \frac{3/z}{1} + \frac{3/z}{1} + \cdots + \frac{n/z}{1} + \frac{n/z}{1} + \cdots$$

which converges locally uniformly in $D_0 := \{z \in \mathbb{C};\ |\arg z| < \pi\}$ to some holomorphic function $f(z)$. (This is a consequence of Corollary 10.) This $f(z)$ is the Stieltjes transform of the measure $d\Psi(t)$ we are looking for, so we have at least attached some value to $L(z)$, and we can in principle find this measure. Since

$$f(z) = \int_0^{\infty} \frac{z}{z+t}\, d\Psi(t)$$

in general is unknown, we can also use an approximant $f_n(z)$ of the continued fraction, and use its inverse Stieltjes transform

$$\Psi_n^*(t) = \frac{1}{\pi} \lim_{\varepsilon \to 0^+} \operatorname{Im} \int_{-t}^{-s} f_n(x - i\varepsilon)\, dx$$

as an approximation to $\Psi(t)$. In order to get a good approximation, we normally need a larger value of n. But for demonstration, let us use

$$f_4(z) = \frac{1}{1} + \frac{1/z}{1} + \frac{1/z}{1} + \frac{2/z}{1} = \frac{1 + \dfrac{3}{z}}{1 + \dfrac{4}{z} + \dfrac{2}{z^2}} = \frac{(3 - 2\sqrt{2})/4}{\dfrac{1}{z} + 1 - \dfrac{1}{\sqrt{2}}} + \frac{(3 + 2\sqrt{2})/4}{\dfrac{1}{z} + 1 + \dfrac{1}{\sqrt{2}}}$$

which we want to be equal to $\displaystyle\int_0^\infty \frac{d\Psi_4(t)}{z + t}$. But that is easy to achieve. According to Example 4.6.2 we can use the step function

$$\Psi_4(t) := \begin{cases} 0 \quad \text{for } 0 \le t \le 1 - 1/\sqrt{2}, \\ (3 - 2\sqrt{2})/4 \quad \text{for } 1 - 1/\sqrt{2} \le t < 1 + 1/\sqrt{2}, \\ \sqrt{2} \quad \text{for } 1 + 1/\sqrt{2} < t. \end{cases} \qquad \diamondsuit$$

4.4 Gauss quadrature

The theme in this section is an application of this theory. We shall see how we can apply it to approximate integrals of the form

$$\int_{\mathbb{R}} f(t)\, d\Psi(t) \tag{73}$$

where f is continuous at the support of the probability measure $d\psi(t)$. We first note that since

$$\lim_{z \to z_k} \frac{F(z)}{z - z_k} = F'(z_k) \neq 0 \quad \text{for } F(z) := \prod_{k=1}^{n} (z - z_k),$$

it follows that

$$p_k(z) := \frac{F(z)}{(z - z_k)F'(z_k)}$$

is a polynomial of degree $n - 1$ with $p_k(z_m) = 0$ for $m \neq k$ and $p_k(z_k) = 1$. This means that if we set

$$G_n(z) := \sum_{k=1}^{n} \lambda_k p_k(z),$$

then G_n is a polynomial of degree $\le n - 1$ with $G_n(z_k) = \lambda_k$ for $k \le n$. G_n is called the *Lagrange interpolating polynomial*. We can use this idea to approximate (73). If we know the sequence $\{P_n(z)\}$ of monic orthogonal polynomials for $d\Psi(t)$, then we know the canonical denominators $B_n(z) = P_n(z)$ of the corresponding positive definite J-fraction.

Theorem 4.9. *Let the infinite support of the measure $d\Psi(t)$ be contained in a bounded interval, and let $\{z_k^{(n)}\}_{k=1}^n$ be the zeros of $B_n(z)$. Then*

$$\int_{\mathbb{R}} f(t)\, d\Psi(t) = \sum_{k=1}^{n} \lambda_k^{(n)} f(z_k^{(n)}) + E_n \quad \textit{where} \quad \lim_{n \to \infty} E_n = 0$$

where

$$\lambda_k^{(n)} = \frac{1}{B_n'(z_k^{(n)})} \int_{\mathbb{R}} \frac{B_n(t)}{t - z_k}\, d\Psi(t) = \frac{A_n(z_k^{(n)})}{B_n'(z_k^{(n)})}$$

and $E_n = 0$ if f is a polynomial of degree $\le 2n - 1$.

5 Additional literature

For those who want to go deeper into the material, we can mention the books [24], [15], [34] and [13]. The exposition in the present paper is however closer to [21], or rather the new and improved edition [22]. For number-theoretic aspects we can also mention the recently published [9].

6 Open problems

There is a lot of open problems in this theory, of course, so the list below is just presented for inspiration. They are neither chosen nor ordered by their importance or grade of difficulty.

1. The question of singular points for functions $\mathbf{K}(a_n z^{\alpha_n}/1)$ with $a_n > 0$ and $\alpha_n \in \mathbb{N}$ has been studied by Thron in a number of papers, for instance with Callas [2] and with D. Singh [28]. In particular it is nice to know when $\partial\mathbb{D}$ is a natural boundary for the function. In [13, Thm 12.8, p.383] it is stated that sufficient conditions are essentially

$$\lim(4|a_n|)^{1/\alpha_n} = 1 \quad \text{and} \quad \liminf \frac{\rho_n}{h_n - \rho_n} = 0$$

   where $\rho_n := \max\{\deg A_n, \deg B_n\}$ and $h_n := \sum_{m=1}^{n} \alpha_m$. Also more references can be found in that book. But what about necessary conditions?

2. A related problem is the following: let $L(z)$ be a power series with convergence radius 1, and assume that $L(z)$ has a continued fraction expansion of the form $c_0 + \mathbf{K}(a_n z 71)$ with all $a_n > 0$. Under what conditions will also $\mathbf{K}(a_n z 71)$ converge in $\mathbb{D}$? And when will it converge in a larger region? Does the location of possible branch points have any impact on this question?

3. In [5] they compute special functions by means of continued fractions. Of course they use the fixed point modification $S_n(g(z))$ from Example 5, but the more refined ideas in [17] or [22, p.218ff] have not been fully exploited.

4. It is easy to prove convergence or divergence of a 2-periodic continued fraction. But it is not so clear what happens if the continued fraction $\mathbf{K}(c_n/1)$ has elements picked randomly from a 2-point set $E := \{a_1, a_2\}$. If E is contained in a parabolic region P_α from the Parabola Theorem (or some other well known convergence set), the case is clear. But what can we say otherwise?

5. We have seen that a function has an S-fraction expansion if and only if it essentially is a Stieltjes transform. Similar results are valid for other moment problems. But Carleman's criterion does not yet have any counterparts for these cases. It is also possible to state other types of sufficient conditions for function to have a continued fraction of specified form with complex coefficients [32], [12], but there exist very few results of this nature.

6. For what kind of series and/or continued fractions does the conversion from series to continued fraction represent a summation method or a convergence acceleration method? It is true in all the interesting cases so far, but surely it can't always be true. What kind of sufficient or necessary criteria can be found? This is of course related to Problem 2 above.

7. The idea of general convergence has been extended to sequences of functions from quasi-normal families, such as for instance to sequences of univalent or p-valent functions in some domain D, [18]. This kind of convergence is different from convergence in measure or capacity. But the possibilities here have not yet been exploited.

8. Value sets $\{V_n\}$ for a family $\mathbb{F}$ of continued fractions $\mathbf{K}(a_n/b_n)$ give a domain V_0 for their values f. So at least we know something about the values f even in cases where they are not known explicitly. In [11] we developed the probability distribution for f in some special cases. The idea due to Waadeland is found in [33]. Such results are also of interest for estimation of tail values for a given continued fraction, in order to accelerate its convergence.

9. If $\mathbf{K}(a_n/1)$ converges generally but not in the classical sense, then probably also $\mathbf{K}(c_n/1)$ shares this property if $\sum |c_n - a_n| < \infty$ or $\sum \mathfrak{m}(a_n, c_n) < \infty$ under proper conditions. Is this true? If so, what are the proper conditions?

10. There is a number of generalizations of continued fractions: vector- or matrix-valued continued fractions $\mathbf{K}(A_n/1)$, branched continued fractions meant for expansions of analytic functions of several complex variables, Schur analysis of functions analytic in the unit disk, compositions of functions other than the s_ns that make up a continued fraction (continued radicals, towers of exponentials, continued fractions building upwards instead of downwards, etc). What can we say about convergence etc for such structures? Some is known, but the topic is by far exhausted.

References

[1] Ambroladze and H. Wallin. Random iteration of Möbius transformations and Furstenberg's theorem. *Ergod. Th. & Dynam. Sys.*, 20:953-962, 2000.

[2] N. P. Callas and W. J. Thron. Singularities of a class of meromorphic functions. *Proc. Amer. Math. Soc.*, 33:445-454, 1972.

[3] T. Carleman. Sur les problème des moments. *C. R. Acad. Sci. Paris*, 174:1680-1682, 1922.

[4] T. Carleman. Sur les équations intégrales singulière à noyau symmétriques. *Report, Uppsala*, 289–220, 1923. Also in: Les fonctions quasi analytiques. *C. R. Acad. Sci. Paris*, 228pp, 1926.

[5] A. Cuyt, W. B. Jones, V. B. Petersen, B. Verdonk and H. Waadeland. Handbook of Continued Fractions for Special Functions. Springer, 2007.

[6] J. Favard. Sur les polynomes de Tchebicheff. *C. R. Acad. Sci. Paris*, 200:2052-2053, 1935.

[7] H. Furstenberg. *Noncommuting random products*, Trans. Amer. Math. Soc. **108** (1963), 377-428.

[8] H. Hamburger. Über eine Erwiterung des Stieltjesschen Momentproblems. Parts I, II, III. *Math. Ann.*, 81:235–319, 1920, 82:120-164, 1921, 82:168-187, 1921.

[9] D. Hensley. Continued fractions. *World Scientific Publ. Co.*, 2006.

[10] L. Jacobsen. General convergence of continued fractions. *Trans. Amer. Math. Soc.* 294(2):477-485,1986.

[11] L. Jacobsen, W. J. Thron and H. Waadeland. Some observations on the distribution of values of continued fractions. *Numer. Math.*, 55:711–733, 1989.

[12] L. Jacobsen and H. Waadeland. When does $f(z)$ have a regular C-fraction expansion or a normal Padé table? *J. Comp. and Appl. Math.*, 28:199–206, 1989.

[13] W. B. Jones and W. J. Thron. Continued Fractions. Analytic Theory and Applications. *Encyclopedia of Mathematics and its Applications,* **11***, Addison-Wesley Publishing Company, Reading, Mass.*, 1980. Now distributed by Cambridge University Press, New York.

[14] W. B. Jones and W. Van Assche. Asymptotic behavior of the continued fraction coefficients of a class of Stieltjes transforms including the Binet function. In *Orthogonal Functions, Moment Theory and Continued Fractions; Theory and Applications*, Lecture Notes in Pure and Appl. Math., (eds.: W. B. Jones and A. Sri Ranga) Marcel Dekker, 257–273, 1998.

[15] A. N. Khovanskii. The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory. *P. Noordhoff, Groningen.*, 1963.

[16] R. E. Lane and H. S. Wall. Continued fractions with absolutely convergent even and odd parts. *Trans. Amer. Math. Soc.*, 67:368-380, 1949.

[17] L. Lorentzen. Computation of limit periodic continued fractions. A survey. *Ann. Numer. Math.*, 10:69-111, 1995.

[18] L. Lorentzen. General convergence in quasi-normal families. *Proc. Edinburgh Math. Soc.*, 46:169-183, 2003.

[19] L. Lorentzen. Möbius transformations mapping the unit disk into itself. *The Ramanujan J. Math.*, 13(1/2/3):253-264, 2007.

[20] L. Lorentzen and St. Ruscheweyh. Simple convergence sets for continued fractions $K(a_n/1)$. *J. Math. Anal. and Appl.*, 179:349-370, 1993.

[21] L. Lorentzen and H. Waadeland. Continued Fractions with Applications. *Studies in Computational Mathematics*, 3, Elsevier, 1992.

[22] L. Lorentzen and H. Waadeland. Continued Fractions. Volume 1: Convergence Theory. *Atlantis Studies in Mathematics for Engineering and Science, Atlantis Press/World Scientific*, 2008.

[23] D. S. Lubinsky, H. N. Mashkar and E. B. Saff. A proof of Freud's conjecture for exponential weights. *Constr. Appr.*, 4:65-83.

[24] O. Perron. Die Lehre von den Kettenbrüchen, Band 2, 3. Auflage. *B. G. Teubner, Stuttgart*, 1957.

[25] S. Pincherle. Sur la génération de systemes récurrents au moyen d'une équation linéaire différentielle. *Acta Mathemaica*, 16:341-363, 1892.

[26] G. Piranian and W. J. Thron. Convergence properties of sequences of linear fractional transformations. *Michigan Math. J.*, 4:129-135, 1957.

[27] I. Schur. Über Potenzreihen, die in Innern des Einheitskreises beschränkt sind. *J. reine und angewandte Math.*, 147:205-232, 1917.

[28] D. Singh and W. J. Thron. On the number of singular points located on the unit circle of certain functions represented by C-fractions. *Pacific J. Math.*, 6:135-143, 1956.

[29] T. J. Stieltjes. Recherches sur les fractions continues. *Ann. Fac. Sci. Toulouse*, 8J:1-122, 1894; 9A:1–47, 1895. Also published in: Oevres 2:402-566.

[30] W. J. Thron. On parabolic convergence regions for continued fractions. *Math. Zeitschr.*, 69:173-182, 1958.

[31] E. B. Van Vleck. On the convergence of algebraic continued fractions whose coefficients have limiting values. *Trans. Amer. Math. Soc.*, 5:253-262, 1904.

[32] H. Waadeland. On general T-fractions corresponding to functions satisfying certain boundedness properties. *J. Approx. Theory*, 26:317-328, 1979.

[33] H. Waadeland. Probability distribution of continued fraction values, an invitation and an example. *Comm. Anal. Theory of Cont. Fract.*, 1:51-55, 1992.

[34] H. S. Wall. Continued fractions and bounded analytic functions. *Bull. Amer. Math. Soc.*, 50:110-119, 1944.

The Yamabe problem with singularities

Farid Madani
Institut Mathématiques de Jussieu, Université Pierre et Marie Curie
Équipe d'Analyse Complexe et Géométrie, 175, rue Chevaleret
75013 Paris, France.

Abstract

Let (M, g) be a compact Riemannian manifold of dimension $n \geq 3$. Under some assumptions, we prove that there exists a positive function φ solution of the following Yamabe type equation

$$\Delta\varphi + h\varphi = \tilde{h}\varphi^{\frac{n+2}{n-2}}$$

where $h \in L^p(M)$, $p > n/2$ and $\tilde{h} \in \mathbb{R}$. We give the regularity of φ with respect to the value of p. Finally, we consider the results in geometry when g is a singular Riemannian metric and $h = \frac{n-2}{4(n-1)}R_g$, where R_g is the scalar curvature of g.

1 Introduction

Let (M, g) be a smooth compact Riemannian manifold of dimension $n \geq 3$. Denote by R_g the scalar curvature of g. The Yamabe problem is the following:

Problem 1.1. Does there exists a constant scalar curvature metric conformal to g?

If $\tilde{g} = \varphi^{4/(n-2)}g$ is a conformal metric to g with φ a smooth positive function, then the scalar curvatures R_g and $R_{\tilde{g}}$ are related by the following equation:

$$\frac{4(n-1)}{n-2}\Delta_g\varphi + R_g\varphi = R_{\tilde{g}}\varphi^{N-1} \tag{1}$$

where $N = \frac{2n}{n-2}$ and Δ_g is the geometric Laplacian of the metric g with nonnegative eigenvalues.
To solve the Yamabe problem, it is equivalent to find a function φ solution of equation above where $R_{\tilde{g}}$ is constant. Equation (1) is called Yamabe equation. Yamabe [11] stated the following functional, defined for any $\psi \in H_1(M) - \{0\}$ by

$$I_g(\psi) = \frac{E(\psi)}{\|\psi\|_N^2} = \frac{\displaystyle\int_M |\nabla\psi|^2 + \frac{n-2}{4(n-1)}R_g\psi^2 \mathrm{d}v}{\|\psi\|_N^2} \tag{2}$$

and he considered the infimum of I_g defined as follow

$$\mu(g) = \inf_{\psi \in H_1(M)-\{0\}} I_g(\psi)$$

He solved the case when $\mu(g)$ is nonpositive. Aubin [1] showed that it was sufficient to solve the following conjecture:

Conjecture 1.2 (Aubin [1])**.** *If (M, g) is not conformal to (S_n, g_{can}) then*

$$\mu(M, g) < \mu(S_n, g_{can}) \tag{3}$$

where $\mu(M, g) = \inf\{I_g(\psi),\ \psi \in H_1(M) - \{0\}\}$

It is known that $\mu(S_n, g_{can}) = K^{-2}(n, 2) = \frac{1}{4}n(n-2)\omega_n^{2/n}$, where ω_n is the volume of the unit sphere S_n and $K(n, 2)$ is defined in theorem 2.5.
In the following, we write $\mu(g)$ instead of $\mu(M, g)$.
Aubin proved that the conjecture is valid for all smooth compact non conformally flat Riemannian manifolds of dimension $n \geq 6$ and conformally flat manifolds with finite non trivial fundamental group. The case of conformally flat manifolds and the dimensions 3,4 and 5 were solved by Schoen [9] using positive mass theorem. Hence the conjecture above holds. By works of Yamabe [11], Aubin [1] and Schoen [9], the Yamabe

problem is completely solved, when the manifold is compact and smooth.
The purpose of this paper is to study the following equation

$$\Delta_g \psi + h\psi = \tilde{h}\psi^{\frac{n+2}{n-2}} \tag{4}$$

where $h \in L^p(M)$, and $\tilde{h} \in \mathbb{R}$. We call this kind of equation "Yamabe type equation". We will give a special consideration for the case $h = \frac{n-2}{4(n-1)} R_g$.

2 Regularity theorems for Yamabe type equations

Theorem 2.1. *Let Ω be an open subset of $\mathbb{R}^n$ and L an uniformely elliptic linear operator of the second degree, defined by*

$$L(u) = \sum_{i,j} a_{ij}\partial_{ij} u + \sum_i b_i \partial_i u + hu \tag{5}$$

where the coefficients a_{ij}, b_i and h are real valued bounded functions of class C^k with $k \in \mathbb{N}$. Let u be a weak solution of the equation $Lu = f$.

(i) If $f \in C^{k,\alpha}(\Omega)$ then $u \in C^{k+2,\alpha}(\Omega)$

(ii) If $f \in H^p_k(\Omega)$ then $u \in H^p_{k+2}(\Omega)$

This theorem is the standard regularity theorem, we can find a proof in the book of Gilbarg and Trudinger [7].
The following two theorems allow us to find the best regularity for the solution of Yamabe type equations. Using the theorem 2.2, Trudinger [10] Showed that the weak solutions of the Yamabe equation (1) are smooth. Yamabe [11] had already used implicitely this theorem.

Theorem 2.2. *On a $n-$dimensional compact Riemannian manifold (M, g), if $u \geq 0$ is a non trivial weak solution in $H_1(M)$ of equation $\Delta_g u + hu = 0$, with $h \in L^p(M)$ and $p > n/2$, then $u \in C^{1-[n/p],\beta}(M)$ and positive.*
$[n/p]$ is the integer part of n/p, $\beta \in (0, 1)$.

Notice that if u satisfies the assumptions of this theorem, then $\Delta u \in L^p(M)$. Regularity theorem 2.1 implies that $u \in H^p_2(M)$ and using Sobolev embedding, we find $u \in C^{1-[n/p],\beta}(M)$
Theorem 2.2 permits to proof the following theorem:

Theorem 2.3. *Let (M, g) be $n-$dimensional compact smooth Riemannian manifold. p and $\tilde{h}$ are two reel numbers, with $p > n/2$. If $\varphi \in H_1(M)$ is a non trivial, nonnegative weak solution of*

$$\Delta_g \psi + h\psi = \tilde{h}\psi^{\frac{n+2}{n-2}} \tag{6}$$

then $\varphi \in H^p_2(M) \subset C^{1-[n/p],\beta}(M)$ and φ is positive.

Proof. It is sufficient to show that there exists $\varepsilon > 0$ such that $\varphi \in L^{(\varepsilon+2n)/(n-2)}(M)$. Indeed, if φ satisfies the assumptions of theorem and belongs to $L^{(\varepsilon+2n)/(n-2)}(M)$, then it is a solution of

$$\Delta_g u + (h - \tilde{h}\varphi^{\frac{4}{n-2}})u = 0$$

with $h - \tilde{h}\varphi^{\frac{4}{n-2}} \in L^r(M)$ and $r = \min(p, \frac{2n+\varepsilon}{4}) > n/2$. Using theorem 2.2, we deduce that φ is positive and continuous. Theorem 2.1 and Sobolev embedding imply that $\varphi \in H^p_2(M)$ with $p > n/2$.
Let l be a positive reel number and H, F are two continuous functions in $\mathbb{R}_+$ defined by:

$$H(t) = \begin{cases} t^\gamma & \text{if } 0 \leq t \leq l \\ l^{q-1}(ql^{q-1}t - (q-1)l^q) & \text{if } t > l \end{cases}$$

$$F(t) = \begin{cases} t^q & \text{if } 0 \leq t \leq l \\ ql^{q-1}t - (q-1)l^q & \text{if } t > l \end{cases}$$

$$\text{where } \gamma = 2q - 1, \text{ and } 1 < q < \frac{n(p-1)}{p(n-2)} \tag{7}$$

φ is positive, belongs to $H_1(M)$. $H\circ\varphi$ and $F\circ\varphi$ belong also to $H_1(M)$. Notice that for any $t\in\mathbb{R}_+-\{l\}$

$$qH(t)=F(t)F'(t),\ (F'(t))^2\leq qH'(t) \text{ and } F^2(t)\geq tH(t) \tag{8}$$

If φ is a weak solution of equation (6), then

$$\forall\psi\in H_1(M)\quad \int_M\nabla\varphi\cdot\nabla\psi\mathrm{d}v+\int_M h\varphi\psi\mathrm{d}v=\tilde{h}\int_M\varphi^{N-1}\psi\mathrm{d}v \tag{9}$$

where $N=2n/(n-2)$.
Let us choose $\psi=\eta^2H\circ\varphi$, where η is C^1-function with support in the ball $B_P(2\delta)$ and radius 2δ sufficiently small, such that $\eta=1$ on $B_P(\delta)$. If we substitute in (9), we obtain

$$\int_M\eta^2H'\circ\varphi|\nabla\varphi|^2\mathrm{d}v+2\int_M\eta H\circ\varphi\nabla\varphi\cdot\nabla\eta\mathrm{d}v=\tilde{h}\int_M\varphi^{N-1}\eta^2H\circ\varphi\mathrm{d}v-\int_M h\varphi\eta^2H\circ\varphi\mathrm{d}v \tag{10}$$

Let $f=F\circ\varphi$ be a function. We estimate the forth integrals above, using function f and relations (8). We have $\nabla f=F'\circ\varphi\nabla\varphi$, the second relation in (8) implies

$$|\nabla f|^2=(F'\circ\varphi)^2|\nabla\varphi|^2\leq qH'\circ\varphi|\nabla\varphi|^2$$

We deduce that the first integral of equality (10) is bounded from below.

$$\frac{1}{q}\|\eta\nabla f\|_2^2\leq\int_M\eta^2H'\circ\varphi|\nabla\varphi|^2\mathrm{d}v$$

The first relation of (8) and Cauchy–Schwarz inequality imply that the second integral of (10) is bounded from below by:

$$2\int_M\eta H\circ\varphi\nabla\varphi\cdot\nabla\eta\mathrm{d}v=\frac{2}{q}\int_M\eta f\nabla f\nabla\eta\mathrm{d}v\geq\frac{-2}{q}\|f\nabla\eta\|_2\|\eta\nabla f\|_2$$

By the last relation in (8), we have $\varphi H\circ\varphi\leq f^2$. The two integrals in the right side in (10) are bounded by:

$$\left|\tilde{h}\int_M\varphi^{N-1}\eta^2H\circ\varphi\mathrm{d}v-\int_M h\varphi\eta^2H\circ\varphi\mathrm{d}v\right|\leq|\tilde{h}|\|\varphi\|_{N,2\delta}^{4/(n-2)}\|\eta f\|_N^2+\|h\|_p\|\eta f\|_{2p/(p-1)}^2$$

where $\|\varphi\|_{N,r}^N=\int_{B_P(r)}\varphi^N\mathrm{d}v$. If we take together these estimates, equality (10) becomes:

$$\|\eta\nabla f\|_2^2-2\|f\nabla\eta\|_2\|\eta\nabla f\|_2\leq q(|\tilde{h}|\|\varphi\|_{N,2\delta}^{4/(n-2)}\|\eta f\|_N^2+\|h\|_p\|\eta f\|_{2p/(p-1)}^2) \tag{11}$$

Notice that for all nonnegative reel numbers $a,\ b,\ c$ and d, if $a^2-2ab\leq c^2+d^2$ then $a\leq c+d+2b$. Using this remark, inequality (11) becomes:

$$\|\eta\nabla f\|_2\leq\sqrt{q|\tilde{h}|}\|\varphi\|_{N,2\delta}^{2/(n-2)}\|\eta f\|_N+\sqrt{q\|h\|_p}\|\eta f\|_{2p/(p-1)}+2\|f\nabla\eta\|_2 \tag{12}$$

By Sobolev embedding, we know that there exists a positive constant c, which depends only on n, such that

$$\|\eta f\|_N\leq c(\|\eta\nabla f\|_2+\|f\nabla\eta\|_2+\|\eta f\|_2)$$

The choice of q $(q<N)$ and inequality (12) permit to write

$$(1-c\sqrt{N|\tilde{h}|}\|\varphi\|_{N,2\delta}^{2/(n-2)})\|\eta f\|_N\leq c\big(\sqrt{N\|h\|_p}\|\eta f\|_{2p/(p-1)}+3\|f\nabla\eta\|_2+\|\eta f\|_2\big) \tag{13}$$

We choose δ sufficiently small such that

$$\|\varphi\|_{N,2\delta}^{2/(n-2)}\leq 1/(2c\sqrt{N|\tilde{h}|})$$

when l goes to $+\infty$, we deduce that there exists a postive constant C, which depends on $n,\ \delta,\|\eta\|_\infty,\ \|\nabla\eta\|_\infty$, $\|h\|_p$ and $|\tilde{h}|$ such that

$$\|\varphi^q\|_{N,2\delta}\leq C(\|\varphi^q\|_2+\|\varphi^q\|_{2p/(p-1)})$$

$\frac{2p}{p-1}q<N$ and φ is bounded in L^N, hence

$$\|\varphi\|_{qN,2\delta}\leq C$$

If $(\eta_i)_{i\in I}$ is a partition of unity subordinate to the covering $\{B_{P_i}(\delta)\}_{i\in J}$ on M

$$\|\varphi\|_{qN}^{qN}=\sum_{i\in I}\|\eta_i\varphi\|_{qN,\delta_i}^{qN}\leq C$$

Hence $\varphi\in L^{qN}$ with $qN>N$. The remark in the begining of the proof implies the theorem. □

Proposition 2.4. *Let (M,g) be $n-$dimensional compact smooth Riemannian manifold. $L := \Delta_g + h$ is a linear operator, with $h \in L^p(M)$ and $p > n/2$. If the smallest eigenvalue λ of L is positive then*

i. L est coercive, in other words there exists $c > 0$ such that

$$\forall \psi \in H_1(M) \quad (L\psi, \psi)_{L^2} \geq c(\|\nabla\psi\|_2^2 + \|\psi\|_2^2)$$

ii. The opertor $L : H_2^p(M) \longrightarrow L^p(M)$ is invertible.

Proof. L admits a smallest eigenvalue because if λ is an eigenvalue associated to the eigenfunction ψ then there exists $C > 0$ such that

$$\lambda\|\psi\|_2^2 = (L\psi,\psi)_{L^2} = \|\nabla\psi\|_2^2 + \int_M h\psi^2 \mathrm{d}v \geq -\|h\|_p\|\psi\|_{2p/(p-1)}^2 \geq -C\|h\|_p\|\psi\|_2^2$$

Hence $\lambda \geq -C\|h\|_p$. If λ is the smallest eigenvalue of L then

$$\lambda = \inf_{\varphi \in H_1(M)-\{0\}} \frac{E(\varphi)}{\|\varphi\|_2^2}$$

where

$$E(\varphi) = (L\varphi,\varphi)_{L^2} = \int_M |\nabla\varphi|^2 + h\varphi^2 \mathrm{d}v$$

So, for any $\varphi \in H_1(M)$

$$E(\varphi) \geq \lambda\|\varphi\|_2^2 \tag{14}$$

Suppose that L is non coercive, then there exists a sequence $(\psi_i)_{i\in\mathbb{N}}$ in $H_1(M)$, which satisfies

$$E(\psi_i) < \frac{1}{i}(\|\nabla\psi_i\|_2^2 + vol(M)^{2/n}) \text{ and } \|\psi_i\|_N = 1$$

It implies

$$(1 - \frac{1}{i})E(\psi_i) < \frac{vol(M)^{2/n}}{i} - \frac{1}{i}\int_M h\psi_i^2 \mathrm{d}v$$

because $|\int_M h\psi_i^2 \mathrm{d}v| \leq \|h\|_{n/2}$, $\lim_{i\to+\infty} E(\psi_i) \leq 0$. On other hands $E(\psi_i) \geq \lambda\|\psi_i\|_2^2$ with $\lambda > 0$. Which is impossible.
If $L\psi = 0$, then, using (14), $\varphi = 0$. So L is injective.
Let $f \in L^p(M)$. Let us prove that the following equation admit a solution $\psi \in H_2^p(M)$

$$\Delta\varphi + h\varphi = f \tag{15}$$

We minimize the functional E defined in the begining of the proof. Let define μ as follow

$$\mu = \inf\{E(\varphi)/\varphi \in H_1(M),\ \int_M f\varphi \mathrm{d}v = 1\} \tag{16}$$

and $(\psi_i)_{i\in\mathbb{N}}$ a sequence in $H_1(M)$ which minimizes E, then

$$\lim_{i\to+\infty} E(\psi_i) = \mu \text{ and } \int_M f\psi_i \mathrm{d}v = 1$$

Without loss of generalities, we suppose that for any nonnegative integer i, $E(\psi_i) \leq \mu + 1$. It implies

$$c(\|\nabla\psi_i\|_2^2 + \|\psi_i\|_2^2) \leq E(\psi_i) \leq \mu + 1$$

because L is coercive. We conclude that $(\psi_i)_{i\in\mathbb{N}}$ is bounded in $H_1(M)$. The Kondrakov theorem and Banach theorem imply that there exists a subsequence $(\psi_j)_{j\in\mathbb{N}}$ such that

* $\psi_j \rightharpoonup \psi$ weakly in $H_1(M)$
* $\psi_j \to \psi$ strongly in $L^s(M)$ for all $1 \leq s < N$
* $\psi_j \to \psi$ almost everywhere.

Then (ψ_j) converge strongly in $L^{2p/(p-1)}(M)$ because $2p/(p-1) < N$. So

$$\int_M f\psi \mathrm{d}v = 1 \text{ and } \int_M h\psi_j^2 \mathrm{d}v \to \int_M h\psi^2 \mathrm{d}v$$

The weak convegence in $H_1(M)$ and the strong convergence $L^2(M)$ imply

$$\lim_{j\to+\infty} \|\nabla\psi_j\|_2 \geq \|\nabla\psi\|_2$$

We conclude that $E(\psi) \leq \mu$, hence $E(\psi) = \mu$. If we write Euler–Lagrange equation for ψ, we find that it is a weak solution in $H_1(M)$ of equation (15). It remains to prove that $\psi \in H_2^p(M)$. Suppose that $\psi \in L^{s_i}(M)$. Then $hu \in L^{\frac{ps_i}{p+s_i}}(M)$, Hence $\Delta u \in L^{\frac{ps_i}{p+s_i}}(M)$. Regularity theorem 2.1 assures that $u \in H_2^{\frac{ps_i}{p+s_i}}(M)$. We know that $H_2^r(M) \subset L^s(M)$ if $r \leq n/2$ with $s = nr/(n-2r)$, and $H_2^r(M) \subset C^{1-[n/r],\beta}(M)$ if $r > n/2$. These inclusions imply the following results

$$\begin{cases} s_0 = N \\ u \in L^{s_{i+1}}(M) \text{ where } s_{i+1} = \frac{nps_i}{np-(p-2n)s_i} & \text{if } s_i \leq \frac{np}{2p-n} \\ u \in H_2^p(M) & \text{if } s_i > \frac{np}{2p-n} \end{cases}$$

If there exists $i \in \mathbb{N}$ such that $s_i > \frac{np}{2p-n}$, which is equivalent to $\frac{ps_i}{p+s_i} > n/2$ then $u \in C^{0,\beta}(M)$, which implies $\Delta u \in L^p(M)$, hence $u \in H_2^p(M)$. If there exists $i \in \mathbb{N}$ such that $s_i = \frac{np}{2p-n}$ then $u \in L^\infty(M)$ and we conclude by regularity theorem that $u \in H_2^p(M)$. Suppose that for any $i \in \mathbb{N}$, $s_i < \frac{np}{2p-n}$, the sequence $(s_i)_{i\in\mathbb{N}}$ is increasing and bounded from above, it converges to $s = 0$ which is impossible. □

Theorem 2.5. *Let (M, g) be a $n-$dimensional smooth compact Riemannian manifold. For all $\varepsilon > 0$, there exists $A(\varepsilon) > 0$ such that*

$$\forall \varphi \in H_1(M) \quad \|\varphi\|_N \leq (K(n,2)+\varepsilon)\|\nabla\varphi\|_2 + A(\varepsilon)\|\varphi\|_2$$

where $N = \frac{2n}{n-2}$ and $K(n,2) = \frac{2}{\sqrt{n(n-2)}}\omega_n^{-1/n}$

The inequality of this theorem is a particular case of a more general one. More further details are given in the Aubin's book [2].

3 Existence theorem

We consider the following equation :

$$\Delta_g\psi + h\psi = \tilde{h}\psi^{\frac{n+2}{n-2}} \tag{17}$$

where $\psi \in H_1(M)$, $h \in L^p(M)$ with $p > n/2$ and $\tilde{h}$ is a reel number. As mentionned in the introduction, this kind of equation are called Yamabe type equation. In the particular case when $h = \frac{n-2}{4(n-1)}R_g$, equation (17) is the Yamabe equation (1). To solve this equation, we use the variational method.
We define the energy E of $\psi \in H_1(M)$ by:

$$E(\psi) = \int_M |\nabla\psi|^2 + h\psi^2 \mathrm{d}v \tag{18}$$

and we consider the fonctional I_g defined for all $\psi \in H_1(M) - \{0\}$ by

$$I_g(\psi) = \frac{E(\psi)}{\|\psi\|_N^2} \tag{19}$$

We denote

$$\mu(g) = \inf_{\psi\in H_1(M)-\{0\},\psi\geq 0} I_g(\psi) = \inf_{\|\psi\|_N=1,\psi\geq 0} E(\psi) \tag{20}$$

with $N = \frac{2n}{n-2}$. the main result of this section is

Theorem 3.1. *If $p > n/2$ and*

$$\mu(g) < K^{-2}(n,2)$$

then equation (17) admits a positive solution $\varphi \in H_2^p(M) \subset C^{1-[n/p],\beta}(M)$, which minimizes I_g (i.e. $E(\varphi) = \mu(g) = \tilde{h}$ and $\|\varphi\|_N = 1$). where $\beta \in (0,1)$.

To proof this theorem, we need the following lemma, proven by Brezis and Lieb[4]

Lemma 3.2. *Let $(f_i)_{i\in\mathbb{N}}$ be a sequence of mesurable functions in (Ω,Σ,μ). If $(f_i)_{i\in\mathbb{N}}$ is uniformely bounded in L^p with $0<p<+\infty$ and $f_i\to f$ almost everywhere, then*

$$\lim_{i\to+\infty}[\|f_i\|_p^p-\|f_i-f\|_p^p]=\|f\|_p^p$$

Proof of theorem 3.1. We check that $\mu(g)$ is finite. In fact, using Hölder inequality, we have

$$E(\psi)\geq-\|h\|_{n/2}\|\psi\|_N^2$$

we deduce that $\mu(g)\geq-\|h\|_{n/2}>-\infty$.
Let $(\varphi_i)_{i\in\mathbb{N}}$ be a minimizing sequence:

$$E(\varphi_i)=\mu(g)+o(1),\ \|\varphi_i\|_N=1 \text{ et } \varphi_i\geq 0 \tag{21}$$

Applying Hölder inequality again for the equation above, we obtain

$$\|\nabla\varphi_i\|_2^2\leq\|h\|_{n/2}+\mu(g)+o(1)$$
$$\|\varphi_i\|_2^2\leq(vol(M))^{2/n}$$

We conclude that $(\varphi_i)_{i\in\mathbb{N}}$ is bounded in $H_1(M)$. Without loss of generalities, we suppose that there exists $\varphi\in H_1(M)$ such that

$*$ $\varphi_i\rightharpoonup\varphi$ weakly in $H_1(M)$

$*$ $\varphi_i\to\varphi$ strongly in $L^s(M)$ for any $s\in[1,N)$

$*$ $\varphi_i\to\varphi$ almost everywhere

We deduce that

$$\int_M|h||\varphi_i-\varphi|^2\mathrm{d}v\leq\|h\|_p\|\varphi_i-\varphi\|_{2p/(p-1)}^2\to 0 \text{ strongly because } 2p/(p-1)<N$$

Let $\psi_i=\varphi_i-\varphi$, then $\psi_i\to 0$ weakly in $H_1(M)$, strongly in $L^q(M)$ for any $q<N$.
We have $\|\nabla\varphi_i\|_2^2=\|\nabla\psi_i\|_2^2+\|\nabla\varphi\|_2^2+2\int_M\nabla\psi_i\cdot\nabla\varphi\mathrm{d}v$. Hence

$$E(\varphi_i)=E(\varphi)+\|\nabla\psi_i\|_2^2+o(1)$$

We know that $E(\varphi)\geq\mu(g)\|\varphi\|_N^2$ by definition of $\mu(g)$, and $E(\varphi_i)=\mu(g)+o(1)$ by definition of $(\varphi_i)_{i\in\mathbb{N}}$. We conclude

$$\mu(g)\|\varphi\|_N^2+\|\nabla\psi_i\|_2^2\leq\mu(g)+o(1) \tag{22}$$

Using lemma 3.2 for $(\varphi_i)_{i\in\mathbb{N}}$, we obtain

$$\|\psi_i\|_N^N+\|\varphi\|_N^N+o(1)=1 \tag{23}$$
$$\|\psi_i\|_N^2+\|\varphi\|_N^2+o(1)\geq 1 \tag{24}$$

Theorem 2.5 gives

$$\|\psi_i\|_N^2\leq(K^2(n,2)+\varepsilon)\|\nabla\psi_i\|_2^2+o(1)$$

Inequality (24) becomes

$$(K^2(n,2)+\varepsilon)\|\nabla\psi_i\|_2^2+\|\varphi\|_N^2+o(1)\geq 1$$

Using the last inequality in (22), we obtain

$$\mu(g)\|\varphi\|_N^2+\|\nabla\psi_i\|_2^2\leq\mu(g)[(K^2(n,2)+\varepsilon)\|\nabla\psi_i\|_2^2+\|\varphi\|_N^2]+o(1)$$

Finally

$$[1-\mu(g)(K^2(n,2)+\varepsilon)]\|\nabla\psi_i\|_2^2\leq o(1)$$

If $\mu(g)<K^{-2}(n,2)$, we can choose ε such that the first factor of this inequality becomes positive. We deduce that $(\psi_i)_{i\in\mathbb{N}}$ converges strongly to zero in $H_1(M)$, $\varphi_i\to\varphi$ strongly in $H_1(M)$ and $L^N(M)$. Hence $I_g(\varphi)=\mu(g)$.
We have just found a non trivial solution of the following Yamabe type equation

$$\Delta\psi+h\psi=\mu(g)\psi^{N-1}$$

which satisfies $\|\varphi\|_N=1$ and $\varphi\geq 0$. Theorem 2.3 implies $\varphi\in H_2^p(M)\subset C^{1-[n/p],\beta}(M)$ and $\varphi>0$. □

4 The choice of the metric

From now until the end of this paper, M is a compact smooth manifold of dimension $n \geq 3$. Denote by T^*M the cotangent space of M.

Assumption (H): *g is a metric in the Sobolev space $H_2^p(M, T^*M \otimes T^*M)$ with $p > n$. There exists a point $P_0 \in M$ and $\delta > 0$ such that g is smooth in the ball $B_{P_0}(\delta)$.*

We can suppose that g is C^2 instead of C^∞ in this ball, but it is not an important point.
Actually our objectif, in this section is to study the Yamabe problem when the metric g admits a finite number of points with singularities and smooth out side these points. The assumption (H) generalizes this conditions and define the notion of "singularities".
By Sobolev embedding, $H_2^p(M, T^*M \otimes T^*M) \subset C^{1,\beta}(M, T^*M \otimes T^*M)$ for some $\beta \in (0,1)$. Hence the metrics which satisfy assumption (H) are $C^{1,\beta}$. The Christoffels belong to $H_1^p \subset C^\beta(M)$. Riemann curvature tensor, Ricci tensor and scalar curvature are in L^p. An example of metric which satisfies assumption (H) is $g = (1 + d(P_0, \cdot)^{2-\alpha})^m g_0$ where g_0 is a smooth metric, $\alpha \in (0,1)$ and $d(P_0, \cdot)$ is the distance function.
We obtain many results which are true for metrics in $H_2^p(M, T^*M \otimes T^*M)$, with $p > n/2$. In the assumption (H), we add the condition that $p > n$ to have a continuous Christoffels for $g \in H_2^p(M, T^*M \otimes T^*M)$. The assumption (H) is sufficient to prove the Aubin's conjecture 1.2 (cf. theorem 8.1), and to construct the Green function of the conformal Laplacian (cf. section 7).
We consider the following problem:

Problem 4.1. Let g be a metric which satisfies the assumption (H). Does there exist a conformal metric $\tilde{g}$ for which the scalar curvature $R_{\tilde{g}}$ is constant ?

It is clear that if the initial metric g is smooth then the problem above is the Yamabe problem 1.1, which is completely solved. We will prove that the answer to this problem is positive. The following proposition tell us that the conformal class of the metrics is well defined when the metrics are in H_2^p.

Proposition 4.2. *Let g be a metric in H_2^p and $\psi \in H_2^p(M)$ a positive function. If $p > n/2$ then the metric $\tilde{g} = \psi^{\frac{4}{n-2}} g$ is well defined, and it is in the same space as g.*

Proof. Using Sobolev embedding, it is easy to check that $H_2^p(M)$ is an algebra for any $p > n/2$. This proposition is a consequence of this fact. □

In their paper [6] about the Yamabe problem, Lee and Parker proved that on every compact Riemanniann manifold (M, g), there exist a normal coordinates system $\{(U_i, x_i)\}_{i \in I}$ and metric g' conformal to g such that $\det g' = 1 + O(|x|^m)$ with m as big as we want. Cao [5] and Günther [8] proved that we can get $\det g' = 1$.

Definition 4.3. $\tilde{g}$ is a Cao–Günther metric if it is conformal to g and there exist a coordinates system such that $\det \tilde{g} = 1$.

Theorem 4.4 (Cao–Günther)**.** *Let M be $C^{a+2,\beta}$ compact manifold of dimension n with $a \in \mathbb{N}$, $\beta \in (0,1)$, g be a $C^{a+1,\beta}-$Riemannian metric , and P be a point in M. Then there exists a $C^{a+1,\beta'}-$positive function φ with $\beta' \in (0, \beta)$ such that $\det(\varphi g) = 1$ in a normal coordinates system with origin P.*

Notice that if the metric $g \in H_2^p(M, T^*M \otimes T^*M)$ with $p > n$ then it belongs to $C^{1,\beta}$. Hence the manifold (M, g) admits a Cao–Günther metric. It is not really useful to suppose that the metric is smooth in a ball, for the existence of this kind of metrics.

5 Conformal Laplacian

Definition 5.1. The conformal Laplacian of Riemannian manifold (M, g) is the operator L_g, defined by :

$$L_g = \Delta_g + \frac{n-2}{4(n-1)} R_g$$

It is known that the conformal Laplacian, when g is smooth, is conformally invariant. Actually it verifies (25) strongly. We prove that we have this property even when the metric is in $H_2^p(M, T^*M \otimes T^*M)$.

Proposition 5.2. *$g \in H_2^p(M, T^*M \otimes T^*M)$ is a Riemannian metric on M with $p > n/2$. If $\tilde{g} = \psi^{\frac{4}{n-2}} g$ is a conformal metric to g with $\psi \in H_2^p(M)$ and $\psi > 0$ then L is weakly conformally invariant, which means that*

$$\forall u \in H_1(M) \qquad \psi^{\frac{n+2}{n-2}} L_{\tilde{g}}(u) = L_g(\psi u) \quad \text{weakly} \tag{25}$$

Moreover if $\mu(g) > 0$ then the conformal Laplacian $L_g = \Delta_g + \frac{n-2}{4(n-1)} R_g$ is invertible and coercive.

Proof. Recall that $\mathrm{d}v_{\tilde{g}} = \psi^{\frac{2n}{n-2}}\mathrm{d}v$ and

$$\forall u, w \in L^2(M) \quad (u,w)_{g,L^2} = \int_M uw\mathrm{d}v_g$$

is the scalar product in $L^2(M)$ with the metric g.
For all $u, w \in H_1(M)$:

$$\begin{aligned}(\psi^{\frac{2n}{n-2}}L_{\tilde{g}}u, w)_{g,L^2} &= (L_{\tilde{g}}u, w)_{\tilde{g},L^2} \\ &= \int_M \tilde{g}(\nabla u, \nabla w) + \frac{n-2}{4(n-1)}R_{\tilde{g}}uw\mathrm{d}v_{\tilde{g}} \\ &= \int_M \psi^2 g(\nabla u, \nabla w) + \frac{n-2}{4(n-1)}R_{\tilde{g}}\psi^{\frac{n+2}{n-2}}(uw\psi)\mathrm{d}v_g\end{aligned}$$

We know that the scalar curvatures R_g and $R_{\tilde{g}}$ are related by Yamabe equation (1), which is equivalent to

$$L_g\psi = \frac{n-2}{4(n-1)}R_{\tilde{g}}\psi^{\frac{n+2}{n-2}} \quad weakly$$

then

$$(L_g\psi, uw\psi)_{g,L^2} = \frac{n-2}{4(n-1)}(R_{\tilde{g}}\psi^{\frac{n+2}{n-2}}, uw\psi)_{g,L^2}$$

Hence

$$\begin{aligned}(\psi^{\frac{2n}{n-2}}L_{\tilde{g}}u, w)_{g,L^2} &= \int_M \psi^2 g(\nabla u, \nabla w) + g(\nabla\psi, \nabla(uw\psi)) + \frac{n-2}{4(n-1)}R_g\psi(uw\psi)\mathrm{d}v_g \\ &= \int_M g(\nabla(\psi u), \nabla(w\psi)) + \frac{n-2}{4(n-1)}R_g(\psi u)(w\psi)\mathrm{d}v_g \\ &= (\psi L_g(\psi u), w)_{g,L^2}\end{aligned} \tag{26}$$

We used the fact that $u\psi$ and $w\psi$ belong to $H_1(M)$, indeed we have the the following Sobolev embedding

$$H_2^p(M) \subset C^{1-[n/p],\beta}(M),\ H_1^p(M) \subset L^{\frac{pn}{n-p}}(M) \text{ and } H_1(M) \subset L^{\frac{2n}{n-2}}(M)$$

Let us prove that L_g is invertible and coercive. Let λ be the smallest eigenvalue of L_g with positive eigenfunction $\varphi \in H_1(M)$, then

$$\lambda\|\varphi\|_2^2 = (L_g\varphi, \varphi)_{g,L^2} = I_g(\varphi)\|\varphi\|_N^2 \geq \mu(g)\|\varphi\|_N^2 > 0$$

hence $\lambda > 0$. We conclude the result, by applying proposition 2.4.

□

6 Yamabe conformal invariant

In the case of smooth metrics, $\mu(g)$ is conformally invariant, which means that if g and $\tilde{g}$ are two smooth conformal metrics then $\mu(g) = \mu(\tilde{g})$. The next proposition shows that we can extend this property to metrics in H_2^p.

Proposition 6.1. *Let M be a smooth compact manifold of dimesion $n \geq 3$. Let g and $\tilde{g} = \psi^{\frac{4}{n-2}}g$ be two metrics in H_2^p, with $\psi \in H_2^p(M)$ positive. if $p > n/2$ then*

$$\mu(g) = \mu(\tilde{g})$$

Proof. Let $u \in H_1(M)$ be test function for the Yamabe functional I_g. Notice that $E(u) = (L_g(u), u)_{g,L^2}$. then

$$I_{\tilde{g}}(u) = (L_{\tilde{g}}(u), u)_{\tilde{g},L^2}\|u\psi\|_N^{-2}$$

Using proposition 5.2, we deduce that

$$I_{\tilde{g}}(u) = (L_g(\psi u), \psi u)_{g,L^2}\|u\psi\|_N^{-2}$$

Finally

$$I_{\tilde{g}}(u) = I_g(\psi u) \tag{27}$$

Which implies that $\mu(g) = \mu(\tilde{g})$. So this invariant depends only on the conformal class $[g]$ and the manifold M. □

7 Green Function

Definition 7.1. Let (M, g) be a compact Riemannian manifold and P be a point in M. We call G_P the Green function on P of the linear operator L, if it satisfies

$$LG_P = \delta_P (\Longleftrightarrow \forall f \in C^\infty(M) \quad \langle G_P, Lf\rangle = f(P))$$

Proposition 7.2 shows the existence of such function for the operator $L = \Delta + h$ with a positive continuous function h. Unfortunately, the method used to construct this function doesn't work when h belongs to $L^p(M)$. This case holds for the conformal Laplacian operator L_g, because $R_g \in L^p(M)$. But, using proposition 7.3, we construct this function and we obtain corollary 7.4.

Proposition 7.2. *Let h be a positive continuous and $P \in M$. g is a metric satisfying assumption (H). There exists a unique Green function G_P for the operator $L = \Delta_g + h$ which satisfies $LG_P = \delta_P$ and*

(i) G_P is smooth in $B_{P_0}(\delta) - \{P\}$

(ii) $G_P \in C^2(M - \{P\})$

(iii) There exists $c > 0$ such that for any $Q \in M - \{P\}$, $|G_P(Q)| \leq cd(P, Q)^{2-n}$

Proof. G_P is unique because L is invertible. In fact, if λ is an eigenvalue of L and φ is a positive eigenfunction associated to λ then

$$\lambda \|\varphi\|_2^2 = (L\varphi, \varphi)_{L^2} = E(\varphi) > 0$$

Hence $\lambda > 0$. To conclude, it is sufficient to apply proposition 2.4. For the existence of such function, we follow Aubin's [2] construction for the Laplacian, in the case of smooth metrics. We choose a decreasing positive smooth radial function $f(r)$, equal to 1 for $r < \delta/2$ and zero for $r \geq \delta(M)$ the injectivity radius of M. We define the following functions

$$H(P, Q) = \frac{f(r)}{(n-2)\omega_{n-1}} r^{2-n} \text{ with } r = d(P, Q)$$
$$\Gamma^1(P, Q) = -L_Q H(P, Q)$$
$$\forall i \in \mathbb{N}^* \qquad \Gamma^{i+1}(P, Q) = \int_M \Gamma^i(P, S)\Gamma^1(S, Q)\mathrm{d}v(S)$$

Then

$$|\Gamma^1(P, Q)| \leq cd(P, Q)^{2-n}$$

We show that

$$\forall i \geq 1 \qquad |\Gamma^i(P, Q)| \leq \begin{cases} cd(P, Q)^{2i-n} & \text{if } 2i < n \\ c(1 + \log d(P, Q)) & \text{if } 2i = n \\ c & \text{if } 2i > n \end{cases}$$

In the last case Γ^i is continuous.
More details are given in Aubin' book [2].
The Green function of L is given by

$$G_P(Q) = H(P, Q) + \sum_{i=1}^{k} \int_M \Gamma^i(P, S)H(S, Q)\mathrm{d}v(S) + F_P(Q) \tag{28}$$

where F_P satisfies

$$LF_P = \Gamma^{k+1}(P, \cdot)$$

We choose $k = [n/2]$, $\Gamma^{k+1}(P, \cdot)$ is continuous. Regularity theorem 2.1 implies that F_P is C^2.
(i) $L_g G_P = 0$ in $B_{P_0}(\delta) - \{P\}$ and the metric is smooth on $B_{P_0}(\delta)$, regularity theorem assure that G_P is smooth on $B_{P_0}(\delta) - \{P\}$, with $P \in M$.
(ii) We have also $LG_P = 0$ in $M - \{P\}$. We conclude that G_P is C^2 in $M - \{P\}$.
(iii) In the expression (28), we notice that the leading term, in the neighborhood of P, is $H(P, Q)$, then for all $P \neq Q$,

$$|G_P(Q)| \leq cd(P, Q)^{2-n}$$

□

Proposition 7.3. *Let g be a metric in $H_2^p(M, T^*M \otimes T^*M)$, $\tilde{g} = \psi^{\frac{4}{n-2}} g$ is conformal to g with $\psi \in H_2^p(M)$ positive and $p > n/2$. We suppose that $L_{\tilde{g}}$ admits a Green function on P, denoted $\tilde{G}_P$, then L_g admits a Green function, denoted G_P and it is given by*

$$\forall Q \in M - \{P\} \qquad G_P(Q) = \psi(P)\psi(Q)\tilde{G}_P(Q)$$

Proof. For any function $\varphi \in C^\infty(M)$:

$$\begin{aligned}
\langle \psi(P)\psi\tilde{G}_P, L_g\varphi\rangle_g &= \psi(P)\int_M \tilde{G}_P\psi L_g[\psi(\frac{\varphi}{\psi})]\mathrm{d}v_g \\
&= \psi(P)\int_M \tilde{G}_P L_{\tilde{g}}\frac{\varphi}{\psi}\mathrm{d}v_{\tilde{g}} \\
&= \psi(P)\langle \tilde{G}_P, L_{\tilde{g}}\frac{\varphi}{\psi}\rangle_{\tilde{g}} \\
&= \varphi(P)
\end{aligned}$$

The second equality above is obtained by the weak conformal invariance of the conformal Laplacian (see proposition 5.2). We know that for any $Q \in M - \{P\}$

$$|\tilde{G}_P(Q)| \leq cd(P,Q)^{2-n}$$

then $G_P \in L^s(M)$, for any $s \in [1, n/(n-2))$ and $L_{\tilde{g}}\frac{\varphi}{\psi} \in L^p(M)$ with $p > n/2$. We choose s such that $\langle \tilde{G}_P, L_{\tilde{g}}\frac{\varphi}{\psi}\rangle_{\tilde{g}}$ is finite. Hence the third equality is well defined. □

Corollary 7.4. *g is a Riemannian metric, satisfying assumption (H). If $\mu(g) > 0$ then the conformal Laplacian L_g admits a Green function G_{P_0} which satisfies $LG_{P_0} = \delta_{P_0}$ and*

(i) G_{P_0} is smooth in $B_{P_0}(\delta) - \{P_0\}$

(ii) $G_{P_0} \in H_2^p(M - B_{P_0}(r))$ for any $r > 0$.

(iii) There exists $c > 0$ such that for any $Q \in B_{P_0}(\delta) - \{P_0\}$, $|G_{P_0}(Q)| \leq cd(P_0, Q)^{2-n}$

Proof. $\mu(g) > 0$, L_g is invertible. We deduce that L_g admits a unique Green function. Using standard variational method (see the proof of poroposition 2.4), we can show that the equation

$$\Delta_g\psi + \frac{n-2}{4(n-1)}R_g\psi = \mu_{q,G}(g)\psi^{q-1} \tag{29}$$

admits a positive solution $\psi \in H_2^p(M)$ when $2 \leq q < N$, with

$$\mu_{q,G}(g) = \inf_{\psi \in H_1(M)-\{0\}} \frac{E(\psi)}{\|\psi\|_q^2}$$

Moreover, g is smooth in $B_{P_0}(\delta)$, regularity theorem shows that ψ is also smooth in the same ball. The metric $\tilde{g} := \psi^{\frac{4}{n-2}} g$ satisfies assumption (H). Using Yamabe equation (1), we deduce that the scalar curvature of $\tilde{g}$ is

$$R_{\tilde{g}} = \frac{4(n-1)}{n-2}\mu_{q,G}(g)\psi^{q-N}$$

Hence $R_{\tilde{g}}$ is positive continuous because $\mu_{q,G}(g) > 0$. Now, we are able to use proposition 7.2, which assure the existence of the Green function $\tilde{G}_{P_0}$ for $L_{\tilde{g}}$ with the metric $\tilde{g}$. using proposition 7.3, we conclude that $G_{P_0} = \psi(P_0)\psi\tilde{G}_{P_0}$ is the Green function of L_g. The metrics g and $\tilde{g}$ are smooth in $B_{P_0}(\delta)$ and $\tilde{G}_{P_0}$ satisfies the properties of proposition 7.2, then the properties announced for G_{P_0} are valid. □

8 Existence theorem

Theorem 8.1. *Let M be a smooth compact manifold of dimension $n \geq 3$, g is a Riemannian metric which satisfies the assumption (H). If (M, g) is not conformal to the sphere (S_n, g_{can}) then $\mu(g) < K^{-2}(n, 2)$.*

This theorem assure that Aubin's conjecture 1.2 still valid for any metric satisfying the assumption (H). To prove this theorem, we use the results of Aubin and Schoen, when the metric g is smooth. The strategy is the following: we construct a test function for the functional I_g, with a support in small geodesic ball. Then the problem is local. We know that the metric g is smooth in $B_{P_0}(\delta)$, so the proof of this theorem is the same as when the metric is smooth everywhere (this is the point where we need the assumption : g is smooth in $B_{P_0}(\delta)$). After, we consider Aubin and Schoen's test functions.
We need also the following result obtained by Aubin [3], for the Green function of L_g:

Theorem 8.2. *If g is a Cao–Günther metric, L_g is invertible and the normalized Green function G_{P_0} have the following expression*

$$G_{P_0}(Q) = r^{2-n} + A + O(r)$$

in a neighborhood of P_0 with $r = d(P_0, Q)$, then $A > 0$, except if (M, g) is conformal to (S_n, g_{can}) for which $A = 0$.

Proof of theorem 8.1. If $\mu(g) \leq 0$ then the inequality is obvious. From now until the end of the proof, we suppose that $\mu(g) > 0$. without loss of generalities, we suppose that g is a Cao–Günther metric given in theorem 4.4. In fact, $\mu(g)$ is conformally invariant (see proposition 5.2).
There are two cases which can happen :
(a) The case (M, g) is not conformally flat in a neighborhood of P_0 and $n \geq 6$. We define $\varphi_\varepsilon = \eta v_\varepsilon$, η is a cut-off function with support in $B_{P_0}(2\varepsilon)$, $\eta = 1$ in $B_{P_0}(\varepsilon)$, $2\varepsilon < \delta$ and

$$v_\varepsilon(Q) = \left(\frac{\varepsilon}{r^2 + \varepsilon^2}\right)^{\frac{n-2}{2}} \qquad r = d(P_0, Q)$$

$supp\varphi \subset B_{P_0}(\delta)$ and the metric g is smooth in this ball, we obtain the following lemma (see Aubin [1]):

Lemma 8.3.

$$\mu(g) \leq I_g(\varphi_\varepsilon) \leq \begin{cases} K^{-2}(n,2) - c|W_g(P_0)|^2\varepsilon^4 + o(\varepsilon^4) \text{ si } n > 6 \\ K^{-2}(n,2) - c|W_g(P_0)|^2\varepsilon^4 \log\frac{1}{\varepsilon} + O(\varepsilon^4) \text{ si } n = 6 \end{cases}$$

where $|W_g(P_0)|$ is the norm of the Weyl tensor on P_0.

(Lee et Parker [6] gave a simple proof of this lemma, using the conformal normal coordinates on P_0). Using this lemma, we conclude that $\mu(g) < K^{-2}(n, 2)$.
(b) The case (M, g) is conformally flat in a neighborhood of P_0 or $n = 3,\ 4$ or 5. In this coordinates system, the Taylor expansion of the Green function is:

$$G_{P_0}(Q) = r^{2-n} + A + O(r)$$

with $r = d(P_0, Q)$ (see Lee and Parker's paper [6] for the proof of this expansion).
If g satisfies assumption (H) and (M, g) is not conformal to (S_n, g_{can}), then theorem 8.2 assure that $A > 0$. Hence we can consider Shoen's test function φ_ε, defined for any $Q \in M$ by:

$$\varphi_\varepsilon(Q) = \begin{cases} v_\varepsilon(Q) \text{ if } Q \in B_{P_0}(\rho_0) \\ \varepsilon_0[G_{P_0} - \eta(G_{P_0} - r^{2-n} - A)](Q) \text{ if } Q \in B_{P_0}(2\rho_0) - B_{P_0}(\rho_0) \\ \varepsilon_0 G_{P_0}(Q) \text{ if } Q \in M - B_{P_0}(2\rho_0) \end{cases}$$

with $2\rho_0 < \delta$, $(\frac{\varepsilon}{\rho_0^2+\varepsilon^2})^{(n-2)/2} = \varepsilon_0(\rho_0^{2-n} + A)$ and η is a smooth nonnegative decreasing function on $\mathbb{R}_+$, with support in $(-2\rho_0, 2\rho_0)$, equal to 1 in $[0, \rho_0]$, the gradient $|\nabla\eta(r)| \leq \rho_0^{-1}$. g is smooth in $B_{P_0}(2\rho_0) \subset B_{P_0}(\delta)$ and $G_{P_0} \in H_2^p(M - B_{P_0}(\rho_0))$ (see corollary 7.4), then we have the estimate of $\mu(g)$, obtained by Schoen[9]:

Lemma 8.4.

$$\mu(g) \leq I_g(\varphi_\varepsilon) \leq K^{-2}(n,2) + c\varepsilon_0^2(c\rho_0 - A)$$

The fact that $A > 0$ allows us to choose ρ_0 sufficiently small ($c\rho_0 < A$) such that $\mu(g) < K^{-2}(n, 2)$.

□

Now, we can state the main theorem which solves the problem 4.1 for any metric which satisfies assumption (H).

Theorem 8.5. *Let M be a smooth compact manifold of dimension $n \geq 3$ and g be a metric satisfying assumption (H). There exists a metric $\tilde{g}$ conformal to g such that the scalar curvature $R_{\tilde{g}}$ is constant everywhere. This metric solves the problem 4.1.*

It means that we can always solve the equation of type Yamabe (17) when $h = \frac{n-2}{4(n-1)} R_g$.

Proof. If (M, g) is conformal to (S_n, g_{can}) then the result is obvious because the scalar curvature of (S_n, g_{can}) is constant. Otherwise (M_n, g) is not conformal to (S_n, g_{can}). In this case, we have the inequality

$$\mu(g) < K^{-2}(n, 2)$$

given by theorem 8.1. Using theorem 3.1, we get a positive solution $\psi \in H_2^p(M)$ of (17), where $h = \frac{n-2}{4(n-1)} R_g$ and $\tilde{h} = \mu(g)$. Using Yamabe equation (1), we deduce that the metric $\tilde{g} = \psi^{\frac{4}{n-2}} g$ has a constant scalar curvature $R_{\tilde{g}} = \frac{4(n-1)}{n-2} \mu(g)$. □

9 Uniqueness of solutions

When the metrics are smooth, if $\mu(g)$ is nonpositive then the solutions of the Yamabe equation (1) are proportional. The following theorem generalizes the uniqueness theorem in the singular case.

Theorem 9.1. *Let g be a metric in $H_2^p(M, T^*M \otimes T^*M)$, with $p > n$. If $\mu(g) \leq 0$ then the solutions of (1) are proportional.*

Proof. Let φ_1 and φ_2 two positive solutions of (1). The metrics $g_i = \varphi_i^{\frac{4}{n-2}} g$ have a constant scalar curvatures R_i, where $i = 1$ or 2. Define $\psi = \frac{\varphi_1}{\varphi_2}$, then $g_1 = \psi^{\frac{4}{n-2}} g_2$. It implies that ψ satisfies

$$\Delta_{g_2} \psi + \frac{n-2}{4(n-1)} R_2 \psi = \frac{n-2}{4(n-1)} R_1 \psi^{\frac{n+2}{n-2}} \tag{30}$$

By regularity theorem 2.1, we deduce that ψ is $C^{2,\beta}$ because the coefficient of the Laplacian are C^0. In fact, in a local coordinates system :

$$\Delta_g \psi = -\nabla_i \nabla^i \psi = -g^{ij} (\partial_{ij} \psi - \Gamma_{ij}^k \partial_k \psi)$$

and the Christoffels are in $H_1^p(M)$ then continuous if $p > n$. In other hands, notice that R_1, R_2 have the same sign. Hence, if $\mu(g) < 0$ then $R_i < 0$ for $i = 1$ and 2. Let $Q_1 \in M$ (resp. $Q_2 \in M$) be a point for which ψ is maximal (resp. minimal). Then $\Delta_{g_2} \psi(Q_1) \geq 0$ and $\Delta_{g_2} \psi(Q_2) \leq 0$. Hence, if we evaluate equation (30) at Q_1 and Q_2, we obtain :

$$\psi^{\frac{4}{n-2}}(Q_1) \leq \frac{R_2}{R_1} \text{ and } \psi^{\frac{4}{n-2}}(Q_2) \geq \frac{R_2}{R_1}$$

We conclude that $\psi = \frac{R_2}{R_1}$, φ_1 and φ_2 are proportional.
If $\mu(g) = 0$ then $R_1 = R_2 = 0$ and (30) becomes $\Delta_{g_2} \psi = 0$, hence ψ is constant. □

References

[1] T. Aubin, *Équations différentielles non linéaires et problème de Yamabe*, J. Math. Pures et appl **55** (1976), 269–296.

[2] T. Aubin, *Some Nonlinear Problems in Riemannian Geometry*, Springer, 1998.

[3] T. Aubin, *Démonstration de la conjecture de la masse positive*, J. Funct. Anal **242** (2007), 78–85.

[4] H. Brezis and E. Lieb, *A relation between pointwise convergence*, Proc. Amer. Math. Soc **88** (1983), 486–490.

[5] J. Cao, *The existence of generalized isothermal coordinates for higher dimensional riemannian manifolds*, Trans. Amer. Math. Soc **324** (1991), 901–920.

[6] J.M. Lee et T. Parker, *The Yamabe problem*, Bull. Amer. Math. Soc **17** (1987), 37–91.

[7] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin Heidelberg New York 1983, 1983.

[8] M. Günther, *Conformal normal coordinates*, Ann. Global Anal. Geom **11** (1993), 173–184.

[9] R. Schoen, *Conformal deformation of a riemannian metric to constant scalar curvature*, J. Differ. Geom **20** (1984), 479–495.

[10] N. Trudinger, *Remarks concerning the conformal deformation of riemannian structures on compact manifolds*, Ann. Scuola Norm. Sup. Pisa **22** (1968), 265–274.

[11] H. Yamabe, *On a deformation of riemannian structures on compact manifolds*, Osaka Math. J **12** (1960), 21–37.

Hausdorff dimensions of Good sets and strict Jarník sets for Fuchsian groups with parabolic elements

S. Munday
University of St. Andrews

Abstract

Certain subsets of limit sets of geometrically finite Fuchsian groups with parabolic elements are considered. It is known that Jarník limit sets determine a "weak multifractal spectrum" of the Patterson measure in this situation. The paper will describe generalisations of these Jarník sets. In particular, we will show that a natural generalisation of these sets, which we call strict Jarník limit sets, gives rise to generalised weak multifractal spectra. We will also give number-theoretical interpretations of these results in terms of continued fractions.

1 Introduction

The first two sections of the paper consist of preliminary material. In Section 1 we introduce the Hausdorff dimension of a set. In Section 2 we give the background in hyperbolic geometry necessary to understand the results of the following sections. In particular we introduce Fuchsian groups - discrete groups of isometries of hyperbolic space with the hyperbolic metric - and their limit sets.

In Section 4 and Section 5 we describe certain subsets of the limit set of a non-elementary geometrically finite Fuchsian group with parabolic elements. The first of these we call Good sets and the second we call strict Jarník sets. The main results given in this paper are the calculation of the Hausdorff dimensions of these sets.

In Section 6 we give the briefest of introductions to the so-called Patterson measure, including the global measure formula, which is used in Section 7 to derive a weak multifractal spectrum for the Patterson measure.

Finally, we will attempt to make clear the sense in which the limit set of a Fuchsian group with parabolic elements can be thought of as a generalisation of continued fractions. This idea allows us to derive the Hausdorff dimension of certain sets of continued fractions as immediate corollaries to the results of Sections 4 and 5.

2 Hausdorff Dimension

Felix Hausdorff (1868-1942) introduced the theory of fractional dimension, now called Hausdorff dimension, in his foundational paper from 1918, “Dimension und äußeres Maß” [11]. In this paper, he adapts a definition of dimension given by Carathéodory in [5] so that it makes sense for non-integer values. (Hausdorff very modestly refers to this ground-breaking work as a “small contribution”.)

Definition 2.1. If U is any non-empty subset of $\mathbb{R}^n$, define the *diameter* of U to be $|U| := sup\{|x-y| : x, y \in U\}$.

Definition 2.2. If $\{U_i\}_{i\geq 1}$ is a collection of sets of diameter at most δ with the property that $F \subseteq \bigcup_{i=1}^{\infty} U_i$, we say that $\{U_i\}$ is a δ-*cover* of F.

Definition 2.3. Suppose that F is a subset of $\mathbb{R}^n$. Then for any $\delta > 0$ we define

$$\mathcal{H}^s_\delta(F) := \inf\{\sum_{i=0}^{\infty} |U_i|^s : \{U_i\} \text{ is a } \delta\text{-cover of } F\}.$$

This infimum increases as δ decreases and so it approaches a limit as $\delta \to 0$. Thus, the following definition makes sense for any subset F of $\mathbb{R}^n$.

Definition 2.4. The *s-dimensional Hausdorff measure* of a set $F \subseteq \mathbb{R}^n$ is given by

$$\mathcal{H}^s(F) := \lim_{\delta \to 0} \mathcal{H}^s_\delta(F).$$

Of course, the limiting value can be 0 or ∞. It is not too difficult to show that $\mathcal{H}^s$ is really a measure. In particular we have that $\mathcal{H}^s(\emptyset) = 0$, if E is contained in F then $\mathcal{H}^s(E) \leq \mathcal{H}^s(F)$, and if $\{F_i\}$ is any countable collection of pairwise disjoint Borel sets, then

$$\mathcal{H}^s\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} \mathcal{H}^s(F_i).$$

Remark 2.5. It is equivalent in the definition of the Hausdorff measure to use coverings by open balls.

For a given set F and a given $\delta < 1$, it is clear that $\mathcal{H}^s_\delta(F)$ is a non-increasing function of s. So, $\mathcal{H}^s(F)$ is also non-increasing. In fact, if $t > s$ and $\{U_i\}_{i=1}^{\infty}$ is a δ-cover of F we have that

$$\sum_{i=1}^{\infty} |U_i|^t = \sum_{i=1}^{\infty} |U_i|^{t-s+s} \leq \delta^{t-s} \sum_{i=1}^{\infty} |U_i|^s$$

so, taking the infimum over all δ-covers on both sides, $\mathcal{H}^t_\delta(F) \leq \delta^{t-s}\mathcal{H}^s_\delta(F)$. If we then let $\delta \to 0$, we see that if $\mathcal{H}^s(F) < \infty$ then $\mathcal{H}^t(F) = 0$ for every $t > s$. So there is a critical value of s where $\mathcal{H}^s(F)$ jumps from ∞ to 0. This critical value is called the Hausdorff dimension of F, written $\dim_H(F)$. Explicitly,

$$\dim_H(F) := \sup\{s : \mathcal{H}^s(F) = \infty\} = \inf\{s : \mathcal{H}^s(F) = 0\}.$$

If $s = \dim_H(F)$, then $\mathcal{H}^s(F)$ may be 0 or ∞, or may satisfy $0 < \mathcal{H}^s(F) < \infty$. A set with this last property is called an s-set.

The following proposition collects some of the basic properties of Hausdorff dimension. For the proofs, the reader is referred to [7].

Proposition 2.6. *Let $F \subset \mathbb{R}^n$.*

1. *$\dim_H(F)$ lies between 0 and n, inclusively.*
2. *If $E \subseteq F$, then $\dim_H(E) \leq \dim_H(F)$.*
3. *The Hausdorff dimension is countably stable, that is, if $F_1, F_2, \ldots$ is a countable sequence of sets, then*

$$\dim_H\left(\bigcup_{i\geq 1} F_i\right) = \sup\{\dim_H(F_i) : i \in \mathbb{N}\}.$$

Although it is possible to calculate the Hausdorff dimension of a set using only the definition, it can often involve pages of complicated estimates. Of course to obtain an upper bound for the dimension of a particular set $F \subset \mathbb{R}^n$ is often (although by no means always) easier than obtaining the corresponding lower bound. For the upper bound it is enough to consider specific coverings of F, while for the lower bound we would have to consider *every* covering of our set F. In particular, some of the covers will consist of both very small sets and sets with relatively large diameters, making obtaining estimates more difficult. A good way around this is to use the following lemma, proved by Frostman in 1935, in his doctoral thesis [9]. A mass distribution on F is a finite measure with support contained in F. The proof is not complicated so we include it here for completeness.

Lemma 2.7. *(Frostman's Lemma.) Let F be a bounded subset of $\mathbb{R}^n$. Let μ be a mass distribution on F and suppose that for some $s > 0$ there exist constants $c > 0$ and $\delta > 0$ with the property that*

$$\mu(U) \leq c|U|^s$$

for all sets U with $|U| \leq \delta$. Then $\mathcal{H}^s(F) \geq \frac{\mu(F)}{c}$ and so

$$s \leq \dim_H(F).$$

Proof. If $\{U_i\}$ is any δ-cover of F, then

$$0 < \mu(F) \leq \sum_{i\geq 1} \mu(U_i) \leq c \sum |U_i|^s.$$

Taking the infimum over all δ-covers of F, we obtain that $\mathcal{H}^s_\delta(F) \geq \frac{\mu(F)}{c}$ for all sufficiently small δ. Hence, $\mathcal{H}^s(F) \geq \frac{\mu(F)}{c}$.

□

3 Hyperbolic Geometry.

In this section, we will briefly introduce two models of the hyperbolic plane and give the basic background in hyperbolic geometry necessary for the results of the following sections. A good reference for the missing details is [2].

3.1 The Poincaré Disc

Let $\mathbb{D}^2$ denote the open unit disc in the complex plane,

$$\mathbb{D}^2 := \{z \in \mathbb{C} : |z| < 1\}$$

and let $\mathbb{S}^1$ denote the boundary of $\mathbb{D}^2$, where

$$\mathbb{S}^1 := \partial\mathbb{D}^2 := \{z \in \mathbb{C} : |z| = 1\}.$$

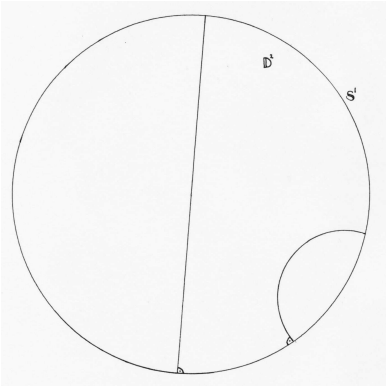


Figure 1: The Poincaré disc.

The Poincaré disc model of hyperbolic space is the metric space $(\mathbb{D}^2, d_h)$ where d_h is the hyperbolic metric, which is defined in the following way. Let $\lambda : \mathbb{D}^2 \to \mathbb{R}$ be given by

$$\lambda(z) := \frac{2}{1 - |z|^2},$$

for all $z \in \mathbb{D}^2$. Then the metric $d_h : \mathbb{D}^2 \times \mathbb{D}^2 \to \mathbb{R}^+$ is given by

$$d_h(u, v) := \inf \left\{ \int_\gamma \lambda(z)|dz| : \gamma \text{ is a smooth curve joining u and v} \right\}.$$

Hyperbolic geodesics in $\mathbb{D}^2$ are given by straight (Euclidean) lines through the origin, or circles orthogonal to $\mathbb{S}^1$.

Recall that a map $g : \mathbb{D}^2 \to \mathbb{D}^2$ is a *conformal automorphism* if and only if it is differentiable and preserves angles (magnitude and orientation) between smooth curves in $\mathbb{D}^2$. The set of all conformal automorphisms of $\mathbb{D}^2$ forms a group under composition of mappings. This group will be denoted by

$$Con(1) := \{g : g \text{ is a conformal automorphism of } \mathbb{D}^2\}.$$

The elements of $Con(1)$ are a certain type of Möbius transformation. It can be shown that if $g \in Con(1)$, there exist complex numbers a and c with the property that $|a|^2 - |c|^2 = 1$ and

$$g(z) = \frac{az + \overline{c}}{\overline{c}z + a}.$$

Lemma 3.1. *For all $g \in Con(1)$ we have that*

1. $|g'(z)| = \frac{1-|g(z)|^2}{1-|z|^2}$ *for all $z \in \mathbb{D}^2$. In particular, $|g'(0)| = 1 - |g(0)|^2$.*
2. $\frac{|g(x)-g(y)|^2}{|x-y|^2} = |g'(x)||g'(y)|$

Proof. By direct calculation, letting $g(z) = \frac{az+\bar{c}}{\bar{c}z+a}$. □

We can now show that the elements of $Con(1)$ are the isometries of the metric space $(\mathbb{D}^2, d_h)$:

Proposition 3.2. *For each $g \in Con(1)$ we have that*

$$d_h(z,w) = d_h(g(z), g(w)) \ \forall z, w \in \mathbb{D}^2.$$

That is, g is an isometry of $(\mathbb{D}^2, d_h)$.

Proof. Let γ be a smooth curve between z and w, also let $g \in Con(1)$. Then, using the substitution $u = g(v)$,

$$\int_{g(\gamma)} \frac{2|du|}{1-|u|^2} = \int_\gamma \frac{2|g'(v)||dv|}{1-|g(v)|^2} = \int_\gamma \frac{2|dv|}{1-|v|^2},$$

by Lemma 3.1. Therefore, since g maps smooth curves to smooth curves, taking the infimum on both sides gives the desired result. □

We will now give one explicit formulation of the hyperbolic distance between points of the unit disc. There are, of course, many other such formulae. For details, refer to [2].

Lemma 3.3. *For all $z \in \mathbb{D}^2$, we have that*

$$d_h(0,z) = \log \frac{1+|z|}{1-|z|}.$$

3.2 The Upper Half-Plane

We will now introduce another model of hyperbolic space, namely the *upper half-plane model*. Let $\mathbb{H}$ denote the upper half of the complex plane $\mathbb{C}$, so

$$\mathbb{H} := \{z = x + iy \in \mathbb{C} : y > 0\}.$$

The boundary of $\mathbb{H}$ is the set $\mathbb{R} \cup \{\infty\}$.

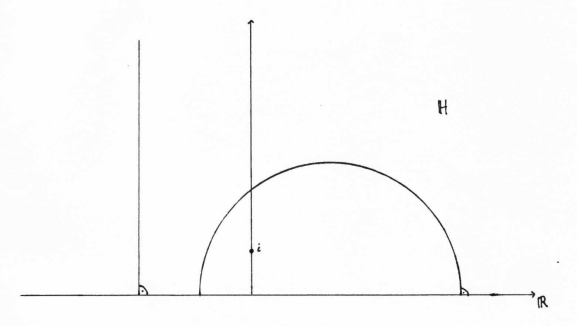


Figure 2: The upper half-plane.

The metric in the upper half-plane is given by the map $d_{\mathbb{H}} : \mathbb{H} \times \mathbb{H} \to \mathbb{R}^+$, which is defined for all z, w in $\mathbb{H}$ by

$$d_{\mathbb{H}}(z,w) := \inf \left\{ \int_\gamma \frac{|dz|}{y} : \gamma \text{ is a smooth curve between } z \text{ and } w \right\}.$$

The geodesics in $\mathbb{H}$ are either vertical Euclidean straight lines (corresponding to geodesics in $\mathbb{D}^2$ with an endpoint at 1) or semicircles orthogonal to the real axis. The following lemma is an easy consequence of this fact.

Lemma 3.4. *For all $z, w \in \mathbb{H}$ with $\mathrm{Re}(z) = \mathrm{Re}(w)$, we have*

$$d_{\mathbb{H}}(z,w) = \left| \log \frac{\mathrm{Im}(z)}{\mathrm{Im}(w)} \right|.$$

The reason for having more than one model of hyperbolic space is purely practical - some results are easier to phrase in terms of one model than another. In order for this to make sense, though, the models must be equivalent in some way. The equivalence we require is *conformal equivalence*, which means that there exists a conformal map from one model to the other. We will now define a conformal map from $\mathbb{H}$ to $\mathbb{D}^2$. Consider the following three maps:

- Let ρ_1 be reflection at the line $\{z = x + iy \in \mathbb{C} : y = 0\}$,

$$\rho_1(z) = \overline{z},$$

  where $\overline{z}$ denotes the complex conjugate of z.

- Let ρ_2 be the reflection at the circle centred at i with radius $\sqrt{2}$,

$$\rho_2(z) = i + \left(\frac{\sqrt{2}}{|z-i|}\right)^2 (z-i).$$

- Let ρ_3 be the map given by clockwise rotation around 0 by $\frac{\pi}{2}$,

$$\rho_3(z) = -iz.$$

Note that each of these three maps is conformal. Now let $\phi := \rho_3 \circ \rho_2 \circ \rho_1$. It is easily verifiable that $\phi(\mathbb{H}) = \mathbb{D}^2$, $\phi(\mathbb{R}) = \mathbb{S}^1 \setminus 1$ and $\phi(\{\infty\}) = 1$. Also, it is easy to check that, for each $z \in \mathbb{H}$,

$$\phi(z) = \frac{z-i}{z+i} \quad \text{and} \quad \phi^{-1}(z) = -i\frac{z+1}{z-1}.$$

Definition 3.5. The map $\phi : \mathbb{H} \to \mathbb{D}^2$ is called the *Cayley transformation*.

It can be directly calculated that $d_{\mathbb{H}}(z, w) := d(\phi(z), \phi(w))$ for each $z, w \in \mathbb{H}$. Also, the group of isometries of $(\mathbb{H}, d_{\mathbb{H}})$ can be obtained by conjugating with $Con(1)$ as follows:

$$Isom(\mathbb{H}) = \phi^{-1} Con(1) \phi.$$

Also, the group of isometries of $(\mathbb{H}, d_{\mathbb{H}})$ is isomorphic to the group $PSL_2(\mathbb{R})$, where

$$PSL_2(\mathbb{R}) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d, \in \mathbb{R} \text{ and } ad - bc = 1 \right\} / \{\pm I\}.$$

The group $PSL_2(\mathbb{R})$ acts on $\mathbb{H}$ via linear fractional transformations:

$$\phi_{\mathbb{H}} : PSL_2(\mathbb{R}) \times \mathbb{H} \to \mathbb{H},$$

where for $g \in PSL_2(\mathbb{R})$, $z \in \mathbb{H}$ and $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, we have

$$\phi_{\mathbb{H}}(g, z) := g(z) = \frac{az+b}{cz+d}.$$

3.3 Classification of Isometries

In this section, we give a classification of hyperbolic isometries in terms of fixed points and geometric actions. For convenience, we will work in the upper half-space model, but it is to be understood that all results here are also valid in the disc model of hyperbolic space (or, for that matter, any other model).

Let $g \in PSL_2(\mathbb{R})$. Then g is of the form $g(z) = \frac{az+b}{cz+d}$ where a, b, c and d are real numbers. It is clear, on setting $g(z) = z$, that the fixed points of g are the roots of a quadratic equation with real coefficients. These will either be two points in $\mathbb{R} \cup \{\infty\}$, one point in $\mathbb{R} \cup \{\infty\}$ or complex conjugate roots, giving one fixed point inside the upper half plane. We make the following definition.

Definition 3.6. Each element g of $PSL(2, \mathbb{R})$ is of exactly one of the following three forms:

1. g is said to be *hyperbolic* if g has exactly two fixed points and these lie on the boundary of hyperbolic space.

2. g is said to be *parabolic* if g has exactly one fixed point that lies on the boundary of hyperbolic space.

3. g is said to be *elliptic* if g has exactly one fixed points that lies in the interior of hyperbolic space.

In the sequel, we will be mostly interested in parabolic points. So, let $g \in PSL_2(\mathbb{R})$ be a parabolic map and let $h \in PSL_2(\mathbb{R})$ be a transformation which maps the fixed point of g to $\{\infty\}$. Then the conjugate hgh^{-1} is called the *standard form* of g. We have the following proposition.

Proposition 3.7. *Let $g \in PSL_2(\mathbb{R})$. Then g is parabolic if and only if the standard form of g maps every horizontal Euclidean straight line in $\mathbb{H}$ into itself. (So the standard form of g is given by a translation $z \mapsto z + b$, for non-zero $b \in \mathbb{R}$.) More generally, if g is parabolic then there exists a Euclidean circle tangent to $\mathbb{R}$ or a horizontal Euclidean straight line in $\mathbb{H}$ left invariant by g.*

These circles and straight lines are called *horoballs*. In the Poincaré disc model of hyperbolic space, the horoballs are Euclidean circles internally tangent to $\mathbb{S}^1$. We can also define horoballs in terms of the *Poisson kernel* $P(z, \xi)$, which is given for $z \in \mathbb{D}^2$ and $\xi \in \mathbb{S}^1$ by

$$P(z, \xi) := \frac{1 - |z|^2}{|z - \xi|^2}.$$

(There is a similar formula in $\mathbb{H}$.) Then we say that the horoball in $\mathbb{D}^2$ at the point ξ of radius k is the set of points $z \in \mathbb{D}^2$ with the property that $P(z, \xi) = \frac{1-k}{k}$. Also (in $\mathbb{D}^2$ or $\mathbb{H}$), the Poisson kernel $P(z, \xi)$ is the "signed distance" between the horoball H_z at ξ through z and the horoball H_0 at ξ through the origin (either 0 or i). If we denote the distance between the horoballs H_0 and H_z by D_z, then the signed distance means that if z is inside H_0 the Poisson kernel $P(z, \xi)$ is equal to e^{D_z} whereas if z is outside H_0 we have that the Poisson kernel $P(z, \xi)$ is equal to e^{-D_z}.

3.4 Fuchsian Groups

We can equip the group $PSL_2(\mathbb{R})$ with a topology inherited from $\mathbb{R}^4$ by identifying $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $(a, b, c, d) \in \mathbb{R}^4$, then defining the norm on $PSL_2(\mathbb{R})$ to be the Euclidean norm on $\mathbb{R}^4$. The norm then induces a metric, which in turn induces the metric topology. Recall that a set E in a topological space (X, τ) is discrete if for each $e \in E$ there exists an open subset $G \in \tau$ such that $E \cap G = \{e\}$. We then define:

Definition 3.8. Let G be a subgroup of $PSL_2(\mathbb{R})$. Then G is said to be a *Fuchsian group* if and only if G is a discrete subset of the topological space $PSL_2(\mathbb{R})$.

Another way to describe a Fuchsian group G is in terms of properly discontinuous group actions. We say that a group G acts properly discontinuously on a metric space X if and only if the orbit $G(x) := \{g(x) : g \in G\}$ is locally finite for all $x \in X$. That is, given an orbit $G(x)$, every compact subset $K \subset X$ contains at most finitely many points of $G(x)$. Note that stating that a group acts properly discontinuously is the same as stating that each orbit of G is a discrete set of points.

Proposition 3.9. *Let G be a subset of $Con(1)$. Then G is Fuchsian group if and only if G acts properly discontinuously on $\mathbb{D}^2$.*

Definition 3.10. Let G be a Fuchsian group. A *fundamental domain* F for G is an open subset of $\mathbb{D}^2$ such that the following conditions are satisfied.

1. $\bigcup_{g \in G} g(\overline{F}) = \mathbb{D}^2$,
2. $g(F) \cap h(F) = \emptyset$, for all $g, h \in G$ with $g \neq h$.

Definition 3.11. A Fuchsian group G is said to be *geometrically finite* if any fundamental region for G has only finitely many edges.

Remark 3.12. That a Fuchsian group G is geometrically finite is equivalent to G being finitely generated. (This is no longer true if we are in higher dimensions).

Recall that a *Riemann surface* is a connected, analytic, complex 1-dimensional manifold. A Riemann surface S is called *simply connected* if every closed curve on S can be continuously deformed into a single point (so the surface of the 2-sphere is simply connected, whereas the torus is not). It is a very deep theorem in the theory of complex functions - the Riemann Mapping Theorem (sometimes called the First Uniformization Theorem) - that every simply connected Riemann surface is conformally equivalent to one of $\mathbb{C}$, $\mathbb{C} \cup \{\infty\}$ or $\mathbb{D}^2$. Further, the Second Uniformization Theorem states that every Riemann surface S is conformally equivalent to a quotient $\tilde{S}/G$ for some simply connected Riemann surface $\tilde{S}$ and for some group

G of conformal automorphisms which acts properly discontinuously on $\tilde{S}$. It follows that if we are in the case where $\tilde{S}$ is conformally equivalent to $\mathbb{D}^2$, every properly discontinuous group G is a Fuchsian group. So, here we always have that a Riemann surface is conformally equivalent to $\mathbb{D}^2/G$ and this is clearly represented by a fundamental domain for the action of G. We can also think of this the other way around - that every Fuchsian group G has an associated Riemann surface, obtained by "gluing" the edges of a fundamental domain F for G, see Figure 3 below.

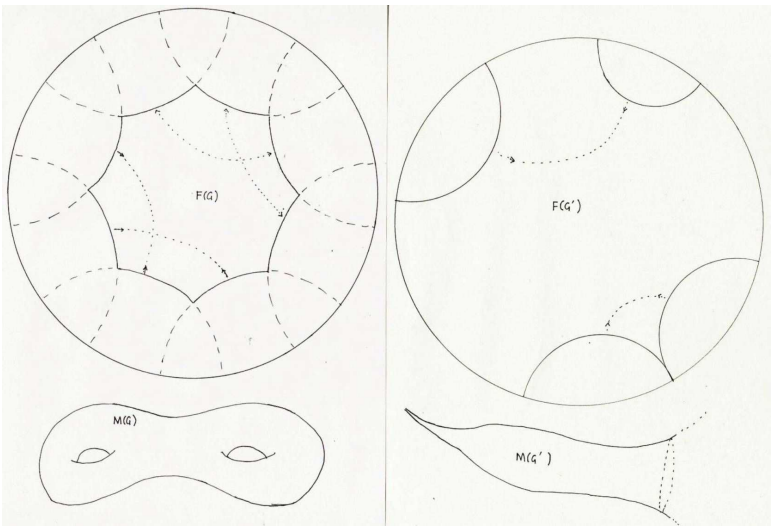


Figure 3: Two Fuchsian groups G, G', their fundamental domains and their associated Riemann surfaces $M(G), M(G')$.

3.5 The Limit Set of a Fuchsian Group

Definition 3.13. Let $w \in \mathbb{D}^2$ (or $\mathbb{H}$) be given. Then the limit set $L(G)$ of the Fuchsian group G is the set

$$L(G) := \{\xi \in \mathbb{C} \cup \{\infty\} : \xi \text{ is an accumulation point of the orbit } G(w)\}.$$

In fact, the limit set is independent of the choice of w in this definition. It is immediately clear that the limit set of a Fuchsian group is always a closed set. It is also clear that the limit set is G-invariant, meaning that $g(L(G)) = L(G)$ for each g in G. It is a consequence of the discontinuous action of a Fuchsian group that $L(G) \subseteq \mathbb{S}^1$.

Theorem 3.14. *If $L(G)$ has more than two points, then $L(G)$ has uncountably many points.*

We say that G is *elementary* if $L(G)$ is either empty (so G generated only by elliptic elements), or consists of only one or two points (so G generated by either a single parabolic element or a single hyperbolic element). Otherwise, G is *non-elementary*. From this point on, unless stated otherwise, assume that G is non-elementary.

We now define certain subsets of the limit set $L(G)$. First we fix some notation. Let s_ξ denote the hyperbolic ray from the origin to the point $\xi \in \mathbb{S}^1$ and, for $t \in \mathbb{R}$, let ξ_t be the point on s_ξ such that $d_h(0, \xi_t) = t$. Also, for a Fuchsian group G and $t > 0$, define $\Delta(\xi_t)$ by setting $\Delta(\xi_t) := d_h(\xi_t, G(0))$. In other words, $\Delta(\xi_t)$ is the smallest hyperbolic distance from the point ξ_t to an orbit point of 0.

Definition 3.15. Let G be a Fuchsian group. A point $\xi \in L(G)$ is said to be a *radial limit point* if there exists a positive constant c such that

$$\liminf_{t\to\infty} \Delta(\xi_t) < c.$$

Denote the set of radial limit points by $L_r(G)$. A point $\eta \in L(G)$ is said to be a *uniformly radial limit point* if there exists a positive constant c such that

$$\limsup_{t\to\infty} \Delta(\xi_t) < c.$$

Denote the set of uniformly radial limit points by $L_{ur}(G)$. Finally, let $L_p(G)$ denote the set of parabolic limit points, where a point p is parabolic if it is the fixed point of some parabolic map in G.

Geometrically, a point $\xi \in L(G)$ is a radial limit point if the ray s_ξ intersects infinitely many balls of radius c around orbit points of 0 and ξ is a uniformly radial limit point if the ray s_ξ is covered by such balls.

Each limit point ξ of G can be represented by the geodesic ray s_ξ from 0 to ξ. So if $L(G)$ is the whole of $\mathbb{S}^1$, every geodesic direction from 0 represents a limit point. If $L(G)$ is a proper subset of $\mathbb{S}^1$, certain directions do not represent limit points. On the surface $M(G)$, the limit points are represented by all those geodesics which do not escape out of a funnel. The parabolic limit points are represented by any geodesic which ends up in a cusp. If ξ is a radial limit point of G, because each of the orbit points 0 projects to the same point $\pi(0)$ on $M(G)$, the ray s_ξ projected to $M(G)$ can be described as "loop approximable", which is illustrated in Figure 4 below.

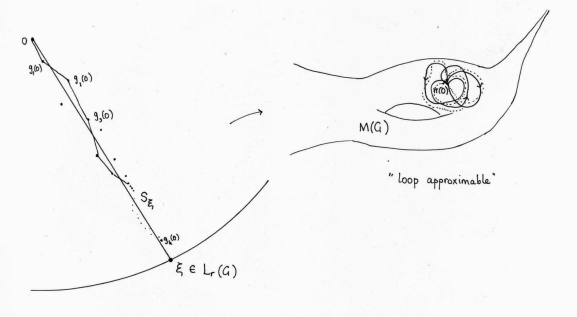


Figure 4: A geodesic on $M(G)$ corresponding to a radial limit point is approximated by loops.

The following result is due to A.F. Beardon and B. Maskit [3].

Theorem 3.16. *Let G be a Fuchsian group. Then G is geometrically finite if and only if*

$$L(G) = L_r(G) \cup L_p.$$

Definition 3.17. For any Fuchsian group G, the Poincaré series is defined for $s \in \mathbb{R}$ and $x, y \in \mathbb{D}^2$ to be

$$\sum_s(x, y) := \sum_{g \in G} e^{-sd_h(x, g(y))}.$$

Definition 3.18. The exponent of convergence $\delta(G)$ is defined to be the infimum of all those s for which the Poincaré series converges. That is

$$\delta(G) := \delta := \inf\{s \in \mathbb{R}^+ : \sum_s(x, y) < \infty\}.$$

More explicitly

$$\sum_s(x, y) = \begin{cases} \infty, & s < \delta; \\ < \infty, & s > \delta. \end{cases}$$

From the triangle inequalities $d_h(x, g(y)) \leq d_h(x, y) + d_h(y, g(y))$ and $d_h(x, g(y)) \geq d_h(y, g(y)) - d_h(x, y)$, we see that

$$e^{-sd_h(x,y)} \sum_s(y, y) \leq \sum_s(x, y) \leq e^{sd_h(x,y)} \sum_s(y, y),$$

so the convergence depends only upon G and not upon x, y and we are justified in writing simply $\delta(G)$.

It was proved by D. Sullivan [22] that if G is a geometrically finite Fuchsian group, then G is of *divergence type*. That is, the Poincaré series $\sum_s(x, y)$ diverges when $s = \delta(G)$. This will be important when we come to define the Patterson measure in Section 6.

It is a result of Bishop and Jones [6] (see also the paper [20]), that for any Fuchsian group G,

$$\dim_H(L_{ur}(G)) = \dim_H(L_r(G)) = \delta(G).$$

This result and Theorem 3.16 above imply that if G is a geometrically finite Fuchsian group, then $\dim_H(L(G)) = \delta(G)$.

Finally, it is a result of Beardon [1] that for G a geometrically finite Fuchsian group with parabolic elements, we have that $\delta(G) > \frac{1}{2}$. Consequently, for such a group G, we obtain the result that $\dim_H(L(G)) > \frac{1}{2}$.

3.6 Standard Horoballs

Throughout the rest of this paper let G be a non-elementary, geometrically finite Fuchsian group with one parabolic element, say γ. Let p be the fixed point of the map γ, so $\gamma(p) = p$. Let F_G be a fundamental domain for G, fixed so that F_G contains the origin of $\mathbb{D}^2$. Denote by $C(L(G))$ the "convex hull" of the limit set $L(G)$. By the "convex hull" we mean the intersection of all convex sets containing the set $L(G) \times L(G) \setminus \{\text{diagonal}\}$ (viewed as a set of geodesics in $\mathbb{D}^2$ with endpoints in $L(G)$).

In the following section we describe a certain set of horoballs associated to the orbit of the parabolic fixed point p under the group G, called a *standard set* of horoballs. This was first introduced, in a more general situation, by B. Stratmann and S. Velani in [21].

Assign a horoball H_γ to the point p, and let H_g be the image of H_γ under the map $g \in G$. Note that if the map g belongs to the stabiliser G_p of p, where $G_p := \{g \in G : g(p) = p\}$, then the horoball H_g is equal to H_γ. It is well known that the set $\{H_g : g \in G/G_p\}$ can be chosen in such a way that it is a pairwise disjoint collection of horoballs. We will call this set a *standard set* of horoballs for G, see Figure 5. Unless stated otherwise, consider $\{H_g : g \in G/G_p\}$ to be a fixed standard set of horoballs for G.

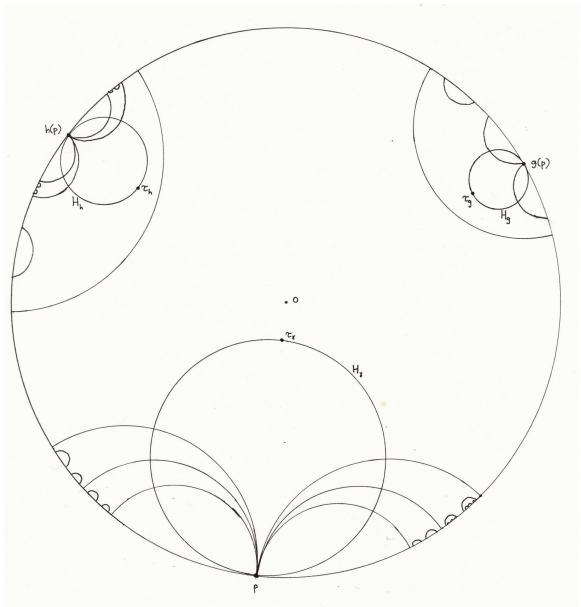


Figure 5: Some standard horoballs for the Fuchsian group G.

Recall that s_ξ denotes the hyperbolic half-ray between the origin and the point ξ on $\mathbb{S}^1$. Define the *top* of the standard horoball H_g to be

$$\tau_g := s_{g(p)} \cap \partial H_g.$$

It was shown in [21] that the point τ_g lies a bounded distance away from the orbit of the origin under G. For completeness, we include the proof here in the Fuchsian groups setting.

Lemma 3.19. *There exists a positive constant ρ, depending only on G, such that for each $g(p) \in G(p)$, there exists $f \in G$ such that*

$$d_h(\tau_g, f(0)) \leq \rho.$$

Proof. First, note that the set $(C(L(G)) \cap \partial H_\gamma)/G_p$ is compact. Then, from this fact and the fact that τ_γ lies in $C(L(G))$, it follows that there exists a compact subset $K(p)$ of $\partial H_\gamma \setminus \{p\}$ containing τ_γ. Denote by d_K the hyperbolic diameter of $K(p)$ and let t_p denote the distance from the origin to τ_g, i.e., $t_p := d_h(0, \tau_\gamma)$. Fix $g \in G$. Then $g(K(p))$ is a compact subset of ∂H_g containing the point $g(\tau_\gamma)$. The point τ_g is contained in $C(L(G))$ and, since $C(L(G))$ is a G-invariant set, the point $g(\tau_\gamma)$ is also contained in $C(L(G))$. Thus, there exists some map h in the stabiliser $G_{g(p)} = gG_pg^{-1}$ such that both τ_g and $h \circ g(\tau_\gamma)$ lie in $h \circ g(K(p))$. Note that the diameter of $h \circ g(K(p))$ is equal to d_K. It follows that

$$\begin{aligned} d_h(\tau_g, h \circ g(0)) &\leq d_h(\tau_g, h \circ g(\tau_\gamma)) + d_h(h \circ g(\tau_\gamma), h \circ g(0)) \\ &= d_h(\tau_g, h \circ g(\tau_\gamma)) + d_h(\tau_\gamma, 0) \\ &\leq d_K + t_p. \end{aligned}$$

Setting $f := h \circ g$ and $\rho := d_K + t_p$, the proof is finished.

□

Examining the proof of Lemma 3.19, we see that it is possible to choose the map g in such a way that $g(K(p))$ and the intersection of ∂H_g with the image of F_G containing τ_g are one and the same thing. We can now choose a set $\mathfrak{T}$ of coset representatives of G/G_p in a geometric way, namely, let g be in $\mathfrak{T}$ if the orbit point $g(0)$ lies in a ρ-neighbourhood of τ_g, the top of the horoball H_g, where ρ comes from Lemma 3.19. That is

$$g \in \mathfrak{T} \Rightarrow d_h(\tau_g, g(0)) \leq \rho.$$

From here on, we will write $\{H_g : g \in \mathfrak{T}\}$ for a fixed standard set of horoballs for G with top representation.

Definition 3.20. The *shadow map* $\Pi : \mathbb{D}^2 \to \mathbb{S}^1$ is defined by

$$\Pi(A) := \{\xi \in \mathbb{S}^1 : s_\xi \cap A \neq \emptyset\}.$$

Using basic hyperbolic geometry, we obtain the following estimate of the size of the shadow of the standard horoball H_g. Recall that x is said to be comparable to y, denoted $x \asymp y$, if there exists a constant $c \geq 1$ such that $c^{-1}y \leq x \leq cy$.

Proposition 3.21. *For every standard horoball with top representation from $\{H_g : g \in \mathfrak{T}\}$ we have that*

$$|\Pi(H_g)| \asymp e^{-d_h(0,\tau_g)}.$$

4 Good Sets

We consider geodesic movements with infinitely many cusp excursions and which spend at most a bounded hyperbolic time between two consecutive cusp excursions. More precisely, for $\xi \in L(G)$, assume that the ray s_ξ intersects infinitely many standard horoballs $H_{g_1}(\xi), H_{g_2}(\xi), \ldots$, which we always assume to be ordered according to their appearance when travelling from 0 to ξ. Then let $d_n(\xi) := \max\{d_h(\eta, \partial H_{g_n}(\xi)) : \eta \in s_\xi \cap H_{g_n}(\xi)\}$ denote the depth of the n-th cusp excursion and let $\tilde{t}_n(\xi) := d_h(0, H_{g_n}(\xi))$. For $\kappa > 0$, we then define

$$\begin{aligned}\mathcal{B}_\kappa(G) := \{\xi \in L(G) : s_\xi \text{ makes infinitely many cusp excursions,} \\ \text{and } d(H_{g_n}(\xi), H_{g_{n+1}}(\xi)) < \kappa, \text{ for all } n \in \mathbb{N}\},\end{aligned}$$

and let

$$\mathcal{B}(G) := \bigcup_{\kappa > 0} \mathcal{B}_\kappa(G).$$

Our first result is to give an estimate for the Hausdorff dimension of the τ-Good set

$$\mathcal{C}_\tau(G) := \{\xi \in \mathcal{B}(G) : d_n(\xi) > \tau, \text{ for all } n \in \mathbb{N}\},$$

for $\tau > 0$ sufficiently large. For these sets we derive the following result, where $\Delta_p := 1/2$ (a topological invariant).

Theorem 4.1.

$$\lim_{\tau \to \infty} \dim_H (\mathcal{C}_\tau(G)) = \Delta_p.$$

Before we give a sketch of the proof, we remark that although the definition of the τ-Good set may seem a little unusual, the description on the surface $M(G)$ associated to G is quite natural. A point ξ is in the set $\mathcal{C}_\tau(G)$ if ξ corresponds to a geodesic on $M(G)$ which makes infinitely many cusp excursions, each time going at least a distance τ into the cusp, and in between these cusp excursions, the geodesic can only spend a bounded time in the "compact part" of M (that is, out of the cusp).

In order to prove Theorem 4.1, we have to establish both an upper bound and a lower bound for the Hausdorff dimension. First fix $\kappa > 0$. For the upper bound, note that the shadows of the standard horoballs $\{H_g : g \in \mathfrak{T}\}$ cover the set $\mathcal{C}_{\tau,\kappa}(G) := \{\xi \in \mathcal{B}_\kappa(G) : d_n(\xi) > \tau, \text{ for all } n \in \mathbb{N}\}$. Fix some $\xi \in \mathcal{C}_{\tau,\kappa}(G)$. Then s_ξ intersects the sequence of horoballs $H_{g_1}(\xi), H_{g_2}(\xi), \ldots$. Suppose we have reached the top of the horoball $H_{g_{n+1}}(\xi)$. In order to reach this point, we must have a sequence of natural numbers $a_1(\xi), a_2(\xi), \ldots$ such that $2\log a_i(\xi) \leq d_i(\xi) \leq 2\log(a_i(\xi) + 1)$ for each $1 \leq i \leq n$.

Then, by Proposition 3.21, we know that

$$|\Pi(H_{g_{n+1}}(\xi))| \leq ce^{-\tilde{t}_n(\xi)} \leq c\frac{1}{(a_1(\xi) \ldots a_n(\xi))^2}.$$

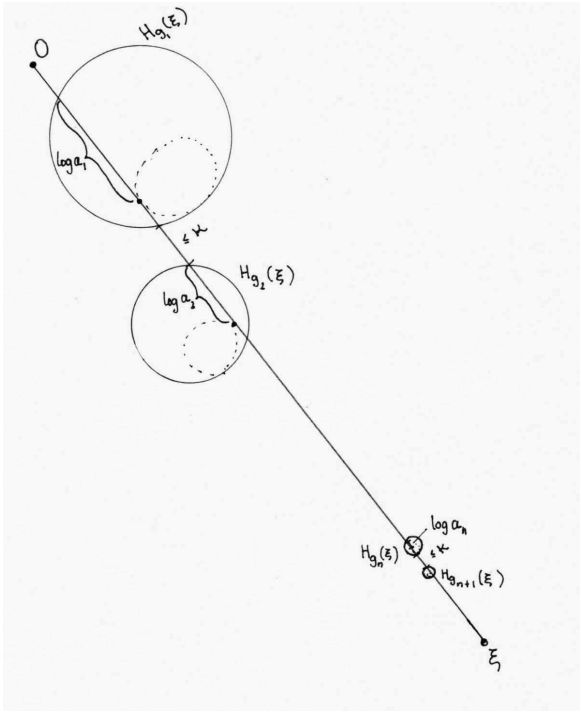


Figure 6: The sequence $a_1(\xi), a_2(\xi), \ldots$ associated to the point $\xi \in L(G)$.

We can clearly choose n large enough that the family $\mathcal{F}$ of $(n+1)$-st level standard horoballs is a γ-cover of $\mathcal{C}_{\tau,\kappa}(G)$ for any positive γ. So, letting $s := \frac{1}{2}(1+\epsilon(\tau))$, we have that

$$\mathcal{H}^s_\gamma(\mathcal{C}_{\tau,\kappa}(G)) \leq \sum_{\mathcal{F}} |\Pi(H_{g_{n+1}}(\xi))|^s = \sum_{\mathcal{F}} c\frac{1}{(a_1(\xi)\ldots a_n(\xi))^{2s}} \leq \ldots \leq c(\tau),$$

where $c(\tau)$ is a positive constant depending only on τ. Hence, $\mathcal{H}^s(\mathcal{C}_{\tau,\kappa}(G)) \leq \infty$ and so

$$\dim_H(\mathcal{C}_{\tau,\kappa}(G)) \leq s = \frac{1}{2}(1+\epsilon(\tau)).$$

Finally, choosing $\epsilon(\tau) \to 0$ as $\tau \to \infty$, we have that

$$\lim_{\tau\to\infty} \dim_H(\mathcal{C}_{\tau,\kappa}(G)) \leq \Delta_p.$$

For the lower bound, define the Cantor-like set $\mathcal{C}'_{\tau,\kappa}(G)$ inside the set $\mathcal{C}_{\tau,\kappa}(G)$ in the following way:

$$\mathcal{C}'_{\tau,\kappa}(G) := \{\xi \in \mathcal{B}_\kappa(G) : \tau < d_n(\xi) < \tau', \text{ for all } n \in \mathbb{N}\}.$$

The upper bound τ' in the definition allows us to control the sizes of the covering horoballs. It also means that there are only finitely many horoballs at any given layer of the construction. We then use a Frostman argument, with the mass distribution $\nu(\Pi(H_{g_n}(\xi))) = \frac{1}{S^{n-1}a_1\ldots a_{n-1}}$, where S is a constant depending on τ', to obtain that $\dim_H(\mathcal{C}'_{\tau,\kappa}(G)) \geq \frac{1}{2}$ for every τ and κ, and by the monotonicity of the Hausdorff dimension (Proposition 2.6), we are done.

5 Strict Jarník Sets

Let $t_n(\xi) := \tilde{t}_n(\xi) + d_n(\xi)$. We consider the *strict (θ,κ)-Jarník limit set* $\mathcal{J}^*_{\theta,\kappa}(G)$ which is given for $\kappa > 0$ and $\theta \in [0,1]$ by

$$\mathcal{J}^*_{\theta,\kappa}(G) := \left\{\xi \in \mathcal{B}_\kappa(G) : \limsup_{n\to\infty} \frac{d_n(\xi)}{t_n(\xi)} = \theta\right\}.$$

The *strict θ-Jarník limit set* $\mathcal{J}^*_\theta(G)$ is then defined by

$$\mathcal{J}^*_\theta(G) := \bigcup_{\kappa>0} \mathcal{J}^*_{\theta,\kappa}(G).$$

We have the following result.

Theorem 5.1. *For each $\theta \in [0,1]$, we have that*

$$\dim_H\left(\mathcal{J}_\theta^*(G)\right) = (1-\theta)\Delta_p.$$

Suppose that $\theta > 0$. If $\xi \in \mathcal{J}_\theta^*(G)$, the intuition is that the hyperbolic time spent making the n-th cusp excursion is longer than all the time taken travelling to the top of the n-th standard horoball $H_{g_n}(\xi)$. So, on the surface $M(G)$ the geodesic movements representing points in $\mathcal{J}_\theta^*(G)$ are those that make successively deeper and deeper cusp excursions but still only visit the compact part of $M(G)$ for a bounded amount of time. Note that Theorem 5.1 implies that the Hausdorff dimension of the strict θ-Jarník limit set is *at most* $\frac{1}{2}$. If $\theta > 0$, then the Hausdorff dimension of $\mathcal{J}_\theta^*(G)$ is strictly less than $\frac{1}{2}$. Recall that in Section 3.5 we stated that the limit set $L(G)$ of a geometrically finite Fuchsian group with parabolic elements has Hausdorff dimension strictly larger than $\frac{1}{2}$. We infer that in order to define a subset of the limit set $L(G)$ with Hausdorff dimension in the range $(\frac{1}{2}, \delta(G))$, we would have to relax the restriction that $\xi \in \mathcal{B}_\kappa$, that is, we would have to increase the time spent outside the cusp.

The proof is a little more involved than that of Theorem 4.1, so we give just a brief sketch. We first observe that the set $F_{\theta,\kappa}$ defined for a fixed integer $N > 2$ by

$$F_{\theta,\kappa} := \left\{ \xi \in \mathcal{B}_\kappa : \log s_n \leq d_n(\xi) \leq \log N s_n, \limsup_{n\to\infty} \frac{\log s_{n+1}}{2\log(s_1 \dots s_n)} = \frac{\theta}{1-\theta} \right\},$$

has the same Hausdorff dimension as the set $\mathcal{J}_\theta^*(G)$. (That $F_{\theta,\kappa} \subset \mathcal{J}_\theta^*(G)$, so that $\dim_H(F_{\theta,\kappa}) \leq \dim_H(\mathcal{J}_\theta^*(G))$, is easy to see. The reverse inequality is somewhat more difficult.) To establish the dimension of the set $F_{\theta,\kappa}$, we proceed in a similar manner to the proof of Theorem 4.1 but instead of coverings by shadows of standard horoballs we now cover our set with shadows of "shrunken" horoballs, which are defined as follows. Let $\widetilde{H}_{g_n}(\xi)$ be the horoball with base point $g_n(p)$ and top $\tilde{\tau}_{g_n}$ given by

$$d_h(0, \tilde{\tau}_{g_n}) = d_h(0, \tau_{g_n}) + \log s_n.$$

Define $s_0 = \frac{1}{2+2\left(\frac{\theta}{1-\theta}\right)}$ and let $s > s_0$ be given. Let $\mathcal{F}$ be a γ-cover of $F_{\theta,\kappa}$ consisting of n-th level shrunken horoballs. So

$$\mathcal{H}^s_\gamma(F_{\theta,\kappa}) \leq \sum_{\mathcal{F}} |\Pi(\widetilde{H}_{g_n}(\xi))|^s \leq \prod_{i=1}^{n-1} (N-1)s_i \left(\frac{1}{s_n (s_1 \dots s_{n-1})^2} \right)^s,$$

where $\prod_{i=1}^{n-1}(N-1)s_i$ is the number of n-th level shrunken horoballs and $\frac{1}{s_n(s_1 \dots s_{n-1})^2}$ is the largest size the shadow of such a horoball could have. By the definition of s, this is bounded for each $\gamma > 0$, so we have $\dim_H(F_{\theta,\kappa}) \leq s$ for every $s > s_0$ and thus $\dim_H(F_{\theta,\kappa}) \leq s_0 = \frac{1}{1+\left(\frac{\theta}{1-\theta}\right)}\Delta_p = (1-\theta)\Delta_p$.
To obtain the lower bound, we again use Frostman's Lemma, but we omit the details here.

6 The Patterson Measure

The Patterson measure was constructed by S.J. Patterson in 1976 [16]. His work was motivated by a number theoretical problem in the theory of Diophantine approximation. The Patterson measure is a very effective tool for examining the limit set of a Fuchsian group. For further details about the Patterson measure, refer to [15].

In the construction of the measure it is important to distinguish between those groups whose Poincaré series converge at $\delta(G)$ and those whose Poincaré series diverge at $\delta(G)$. In the former case we say that the group G is of *convergence type* and in the latter case we say that the group G is of *divergence type*. It was proved by D. Sullivan in [22] that if G is geometrically finite with parabolic elements, then G is of divergence type.

For $x \in \mathbb{D}^2$ and $s > \delta$, the basic idea of the construction is to place a Dirac point mass of weight $\frac{e^{-sd_h(x,g(0))}}{\sum_s(0,0)}$ at each point $g(0)$ in the orbit of 0. Then we invoke Helley's Theorem to obtain a measure in the limit as $s \to \delta$. If the group G is of divergence type, this limit measure will be supported on the limit set $L(G)$. If G is of convergence type, we will simply get another measure supported on the disc $\mathbb{D}^2$ with point masses on the orbit of 0.

In order to get around this problem, Patterson introduced an ingenious multiplicative factor $h(e^{d_h(x,g(0))})$ which does not alter the exponent of convergence, but ensures that the Poincaré series at $\delta(G)$ diverges. However, as we are only interested here in geometrically finite groups and recalling from Section 3.5 that geometrically finite groups are of divergence type, in all that follows the factor h will be set equal to 1.

Let $s > \delta$. We will start by forming the measure

$$\mu_{x,s}(A) := \frac{\sum\limits_{g\in G} e^{-sd(x,g(0))}\delta_{g(0)}(A)}{\sum\limits_{g\in G} e^{-sd(0,g(0))}}.$$

Here, $\delta_{g(0)}$ is a Dirac point-mass at the point $g(0)$, that is

$$\delta_{g(0)}(A) = \begin{cases} 1, & \text{if } g(0) \in A; \\ 0, & \text{otherwise.} \end{cases}$$

Notice that the family of measures $\{\mu_{x,s} : s > \delta\}$ is bounded. We deduce that on a sequence of values monotonically decreasing to δ, the measures μ_{x,s_j} converge weakly to a measure μ_x. Some more work is required to show that such a measure is unique, indeed, in some cases it is not. In the situation where G is a geometrically finite Fuchsian group, it was proved by Sullivan that the Patterson measure is unique. So we make the following definition.

Definition 6.1. Let G be a geometrically finite Fuchsian group with exponent of convergence δ and let (s_j) be a sequence of values monotonically decreasing to δ. Then the Patterson measure with base point x is defined to be

$$\mu_x(A) := \lim_{s_j\to\delta}\left(\frac{\sum\limits_{g(0)\in A} e^{-s_j d(x,g(0))}}{\sum\limits_{g\in G} e^{-s_j d(0,g(0))}}\right) = \lim_{s_j\to\delta}\mu_{0,s_j}(A).$$

We have the following proposition:

Proposition 6.2. *Let G be a geometrically finite Fuchsian group with exponent of convergence δ. Then the Patterson measure μ_x is supported on the limit set of G.*

It was shown by Patterson in [16] that the Patterson measure is a δ-conformal measure. That is to say, for every Borel set $E \subset \mathbb{S}^1$ and every $g \in G$,

$$\mu_0(g^{-1}(E)) = \int_E P(g^{-1}(0),\xi)^\delta \, d\mu_0(\xi).$$

Here, $P(g^{-1}(0),\xi)$ is the Poisson kernel, introduced in Section 3.3. It follows from Lemma 3.1 that it is equivalent to write:

$$\mu_0(g^{-1}(E)) = \int_E |g'(\xi)|^\delta \, d\mu_0(\xi).$$

Sullivan was the first to give a geometric interpretation of the Patterson measure (for this reason, it is sometimes referred to as the Patterson-Sullivan measure). An example of this geometric insight is the interpretation of δ-conformality in the so-called Sullivan Shadow Lemma ([22],[23], see also [15]):

Lemma 6.3. *(Sullivan Shadow Lemma). For all Fuchsian groups G, Δ chosen large enough and for every $g \in G$,*

$$\mu_0(\Pi(B(g(0),\Delta))) \asymp e^{\delta d_h(0,g(0))}$$

The Shadow Lemma gives us a way of estimating the measure of shadows of balls around orbit points of zero. It can also be phrased in terms of $\Delta(\xi_t)$, the distance of the point ξ_t from the orbit of zero. For $\xi \in L(G)$, and positive t, let $b(\xi_t, e^{-t})$ denote the shadow of the geodesic which intersects the ray s_ξ orthogonally at the point ξ_t. One immediately verifies that $b(\xi_t, e^{-t})$ is an arc of $\mathbb{S}^1$ centred at the point ξ with radius comparable to e^{-t}. As long as $\Delta(\xi_t)$ is bounded, which is to say that as long as we are travelling towards a radial limit point, we can use the Shadow Lemma to estimate the measure of $b(\xi_t, e^{-t})$.

The following estimate, called the *Global Measure Formula* by B. Stratmann and S. Velani [21], gives a uniform estimate for the measure of balls in $\mathbb{S}^1$ around any limit point of G. (Note that this estimate was first given by Sullivan in [23].) In order to state the formula, we require the following notation. Define $k(\xi_t)$ to be equal to 1 if ξ_t is inside some standard horoball H_g and let $k(\xi_t)$ be equal to δ otherwise.

Lemma 6.4. *(Global Measure Formula). Let G be a non-elementary geometrically finite Fuchsian group with parabolic elements. If $\xi \in L(G)$ and t is positive, then*

$$\mu_0(b(\xi_t, e^{-t})) \asymp e^{-t\delta} e^{-(\delta - k(\xi_t))\Delta(\xi_t)}.$$

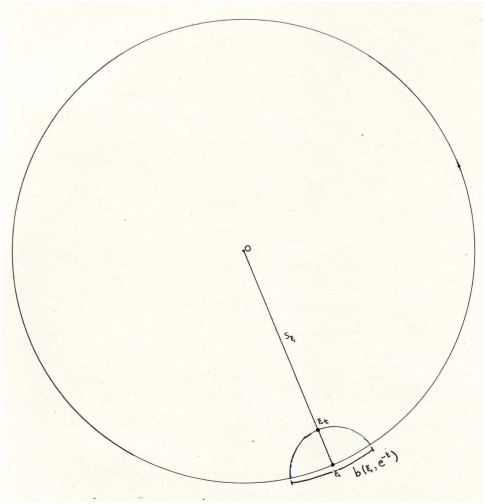


Figure 7: The situation for the global measure formula.

7 Weak Multifractal Spectra for the Patterson Measure

In this section we again assume that G is a non-elementary geometrically finite Fuchsian group with parabolic elements. Recall the definitions of $\mathcal{B}(G)$ and $t_n(\xi)$ given in Sections 4 and 5 respectively.

Let the α-strict-Jarník level sets for the Patterson measure μ_0 be defined by:

$$\mathcal{F}_\alpha^* := \left\{ \xi \in L(G) : \limsup_{n\to\infty} \frac{\log \mu_0(b(\xi, e^{-t_n(\xi)}))}{-t_n(\xi)} = \alpha \right\}.$$

We have the following theorem.

Theorem 7.1. *For each $\alpha \in [2\delta - 1, \delta]$, we have that*

$$\dim_H (\mathcal{F}_\alpha^* \cap \mathcal{B}(G)) = \Delta_p \cdot f_p(\alpha),$$

where $f_p(\alpha) := (\alpha - (2\delta - 1))/(1 - \delta)$.

Proof. The global measure formula for μ_0 gives the existence of a constant $c > 0$ (depending only on G), such that for each $\xi \in L(G)$ and every $t > 0$ we have that

$$\delta + (\delta - 1)\frac{\Delta(\xi_t)}{t} - \frac{c}{t} \le \frac{\log \mu_0(b(\xi, e^{-t}))}{\log e^{-t}} \le \delta + (\delta - 1)\frac{\Delta(\xi_t)}{t} + \frac{c}{t}.$$

(Here we are interested in the case that $t = t_n(\xi)$, $\Delta(\xi_t) = d_n(\xi)$ and $k(\xi_t) = 1$.) From this we immediately deduce that $\xi \in \mathcal{J}_\theta^*(G)$ if and only if $\xi \in \mathcal{B}(G)$ and

$$\limsup_{n\to\infty} \frac{\log \mu_0(b(\xi, e^{-t_n(\xi)}))}{-t_n(\xi)} = \delta - (1 - \delta)\theta.$$

Consequently, if $\alpha := \delta - (1 - \delta)\theta$, the result follows by an application of Theorem 5. □

Remark 7.2. Note that in [18] a "weak multifractal analysis" of the Patterson measure was given. The analysis there was based on investigations of the Hausdorff dimension of the associated *θ-Jarník limit set*

$$\mathcal{J}_\theta(G) := \left\{ \xi \in L(G) : \limsup_{t\to\infty} \frac{\Delta(\xi_t)}{t} \ge \theta \right\}.$$

In [18] (see also [19] and [12]) the result was obtained that

$$\dim_H(\mathcal{J}_\theta(G)) = (1 - \theta)\delta, \text{ for each } \theta \in [0, 1].$$

In [19] it was then shown how to use this result in order to derive the following "weak multifractal spectrum" of the Patterson measure:

$$\dim_H (\mathcal{F}_\alpha) = \begin{cases} 0 & \text{for} \quad 0 < \alpha \le 2\delta - 1 \\ \delta \cdot f_p(\alpha) & \text{for} \quad 2\delta - 1 < \alpha \le \delta \\ \delta & \text{for} \quad \alpha > \delta. \end{cases},$$

where f_p is given as before by $f_p(\alpha) = ((\alpha - (2\delta - 1))/(1 - \delta)$ and where $\mathcal{F}_\alpha(G)$ is defined by

$$\mathcal{F}_\alpha(G) := \left\{ \xi \in L(G) : \liminf_{n\to\infty} \frac{\log \mu_0(b(\xi, e^{-t_n(\xi)}))}{-t_n(\xi)} \leq \alpha \right\}.$$

The outcome here should be compared with the result in Theorem 7.1. We get the following picture:

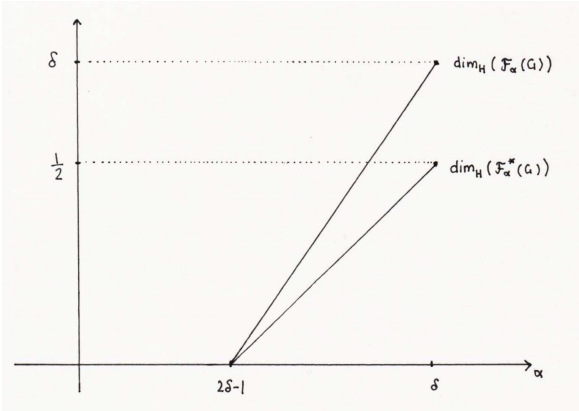


Figure 8: Weak multifractal spectra for the Patterson measure.

8 Application to Continued Fractions

An expression of the form $a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \ldots}}}$ is called a *regular* or *simple continued fraction*. We will use the notation $[a_0; a_1, a_2, a_3, \ldots]$. Here, each of $a_0, a_1, a_2, \ldots$ are positive integers, called the *elements* or *partial quotients* of the continued fraction. The number of elements may be finite or infinite. A finite continued fraction is the result of a finite number of rational operations, so it represents a rational number. Every infinite continued fraction represents an irrational number. The converse of these statements is also true, that is, every real number admits a continued fraction representation.

We now want to make the connection between continued fractions and Fuchsian groups. Consider the group $PSL_2(\mathbb{Z})$, defined by

$$PSL_2(\mathbb{Z}) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d, \in \mathbb{Z} \text{ and } ad - bc = 1 \right\} / \{\pm I\}.$$

This is clearly a discrete subgroup of $PSL_2(\mathbb{R})$ and is thus a Fuchsian group. $PSL_2(\mathbb{Z})$ is often called the *modular group*. It is generated by the maps $S : z \mapsto -\frac{1}{z}$ and $T : z \mapsto z + 1$. The map T is a parabolic transformation with fixed point $\{\infty\}$. In Figure 9, the usual fundamental domain for $PSL_2(\mathbb{Z})$ is shown. Also shown, in Figure 10 below, is the so-called *modular surface*, the Riemann surface associated to $PSL_2(\mathbb{Z})$. The limit set of $PSL_2(\mathbb{Z})$ is the whole of $\mathbb{R} \cup \{\infty\}$.

In [17], C. Series explained how continued fractions can be viewed as geodesic movements on the modular surface, M. Let $\mathcal{L}_G$ be the set of all oriented geodesics in $\mathbb{H}$ with endpoints satisfying $0 < |l_-| < 1$, $|l_+| > 1$ and $l_- \cdot l_+ = -1$, where l_- is the starting point and l_+ is the ending point (for simplicity we also assume that each endpoint is an irrational number). Then $\mathcal{L}_G$ is the set of all geodesics in $\mathbb{H}$ which start in the interval $(-1, 0)$ and end in the interval $(1, \infty)$ or start in the interval $(0, 1)$ and end in the interval $(-\infty, -1)$. The idea is to tessellate the upper half-plane via the *Farey tessellation* (where the vertices of the triangles are exactly the set $\mathbb{Q} \cup \{\infty\}$, see Figure 11), and associate to each geodesic in $\mathcal{L}_G$ a *cutting sequence*, where the geodesic is coded with either an L or an R as it travels through the triangles of the Farey tessellation depending upon whether we see one vertex of a triangle on the left or on the right respectively. The assumption that all endpoints are irrational means that each geodesic l in $\mathcal{L}_G$ has an infinite cutting sequence associated to it.

We say that a geodesic l changes type at each point where we find the code given to two neighbouring triangles to be of different type (for example at points x_0, x_1, x_2 in Figure 11 above). Let the type change point on the imaginary axis be denoted by y_l (note that every geodesic $l \in \mathcal{L}_G$ must have a type change at the imaginary axis). We can code each $l \in \mathcal{L}_G$ by its type changes as follows.

1. $\ldots L^{n_{-2}} R^{n_{-1}} y_l L^{n_1} R^{n_2} \ldots$, if $l_+ > 1$,

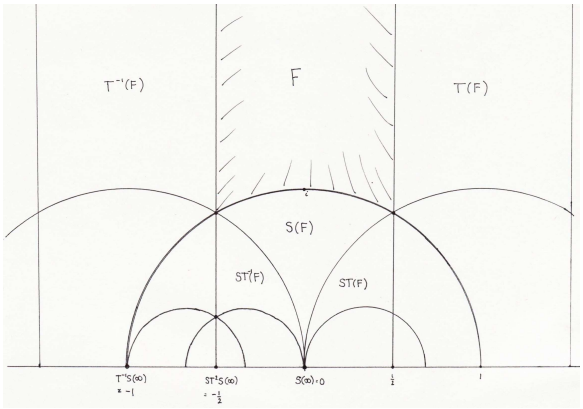


Figure 9: The usual fundamental domain for the modular group, $PSL_2(\mathbb{Z})$.

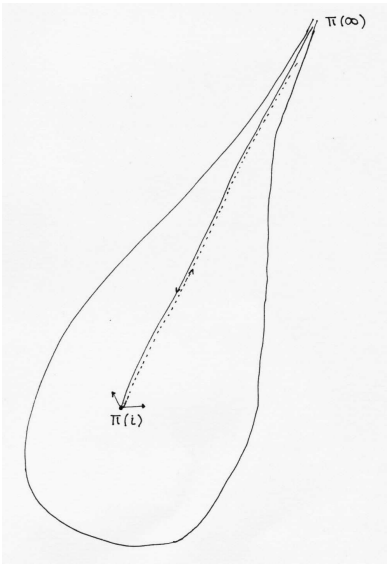


Figure 10: The modular surface, M.

2. $\ldots R^{n_{-2}} L^{n_{-1}} y_l R^{n_1} L^{n_2} \ldots$, if $l_+ < -1$.

The lines of the Farey tessellation project to the singular line S on M (see Figure 10 above). Every geodesic $l \in \mathcal{L}_G$ projects to a geodesic $\hat{l}$ on M and we see that all of the type change points are projected to the line S. Denote this projection map by π. Now consider the subset X of the unit tangent space $UT(M)$ consisting of all those unit tangent vectors with base point $x \in S$ which point along geodesics whose cutting sequences change type at x. Clearly if $l \in \mathcal{L}_G$, then the unit tangent vector $\vec{u}_{y_l}$ to l at the point y_l projects to an element in X. This identification of $\mathcal{L}_G$ with X is almost a homeomorphism.

Theorem 8.1. *[17] The map $i : \mathcal{L}_G \to X$ given by $i(l) = \pi(\vec{u}_{y_l})$ is surjective, continuous and open. It is injective, except that the oppositely oriented geodesics joining $+1$ to -1 have the same image. Moreover, if $\vec{u}_x \in X$ gives rise to a geodesic with cutting sequence $\ldots L^{n_{-2}} R^{n_{-1}} y_l L^{n_1} R^{n_2} \ldots$, then $l = i^{-1}(\vec{u}_x)$ has endpoints given by*

$$l_+ = [n_1; n_2, n_3, \ldots] \text{ and } -\frac{1}{l_-} = [n_{-1}; n_{-2}, n_{-3}, \ldots].$$

Alternatively, if the cutting sequence is $\ldots R^{n_{-2}} L^{n_{-1}} y_l R^{n_1} L^{n_2} \ldots$, then the endpoints are given by

$$l_+ = -[n_1; n_2, n_3, \ldots] \text{ and } \frac{1}{l_-} = [n_{-1}; n_{-2}, n_{-3}, \ldots].$$

So the idea is that a large continued fraction entry corresponds to a deep cusp excursion, or, in other words, to spending a long time inside a standard horoball. If we have a finite continued fraction, i.e. a

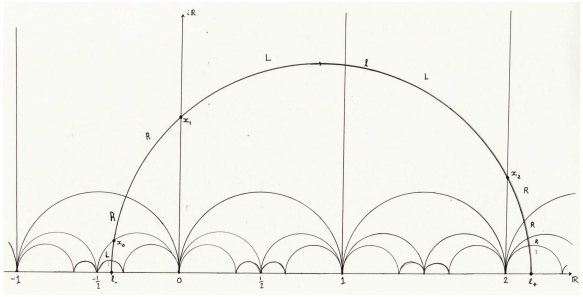


Figure 11: The Farey tessellation.

rational number, this corresponds to a geodesic starting or ending in the cusp, in other words, entering a standard horoball and never leaving it again.

We can then interpret the results given in the sections above in terms of continued fractions, by letting the group G be the modular group $PSL_2(\mathbb{Z})$. In Section 4, we described the τ-Good set $\mathcal{C}_\tau(G)$ by saying that ξ belongs to $\mathcal{C}_\tau(G)$ if s_ξ makes infinitely many cusp excursions and each one is at least of a depth τ. If $G := PSL_2(\mathbb{Z})$, we obtain the following corollary.

Corollary 8.2. *Let the set F_N be defined by*

$$F_N := \{\xi = [a_1(\xi), a_2(\xi), \ldots] : a_i(\xi) > N \text{ for all } i \in \mathbb{N}\}.$$

Then

$$\lim_{N\to\infty} \dim_H(F_N) = \frac{1}{2}.$$

This result can be found in the 1941 paper [10] of I.J. Good (a student of Besicovitch). The proof in that paper does not use Frostman's Lemma, for although it was known at that time it does not seem to have been very well known then. Good derives upper and lower bounds for each set F_N, for $N \geq 20$. Determining the exact Hausdorff dimension of any of the sets F_N is still an open problem.

We can also derive a continued fractions result from Theorem 5.1. Let $G := PSL_2(\mathbb{Z})$ again and define the set K_σ to be

$$K_\sigma := \{\xi = [a_1(\xi), a_2(\xi), \ldots] : \limsup_{n\to\infty} \frac{\log a_n(\xi)}{\log(a_1(\xi)\ldots a_{n-1}(\xi))} = 2\sigma\}.$$

It immediately follows that $\dim_H(K_\sigma) = \frac{1}{2+2\sigma}$.

References

[1] A. F. Beardon. Inequalities for certain Fuchsian groups. *Acta Math.*, 127:221–258, 1971.

[2] A. F. Beardon. *The Geometry of Discrete Groups.*, Springer-Verlag, 1983.

[3] A. F. Beardon and B. Maskit. Limit points of Kleinian groups and finite sided polyhedra. *Acta Math.*, 132:1–12, 1974.

[4] A. S. Besicovitch. Sets of fractional dimension(IV): On rational approximation to real numbers. *Jour. London Math. Soc.*, 9:126–131, 1934.

[5] A. S. Carathéodory. Über das lineare Maß von Punktmengen - eine Verallgemeinerung des Längenbegriffs. *Nachrichten K. Gesell. Wissensch.*, 404–426, 1914.

[6] C. J. Bishop, P.W. Jones. Hausdorff dimension and Kleinian groups. *Acta Math.*, 179:1–39, 1995.

[7] K. Falconer. *Fractal Geometry.*, Wiley, New York, 1990.

[8] A.-H. Fan, L.-M. Liao,B.-W. Wang and J. Wu. On Khintchine exponents and Lyapunov exponents of continued fractions. *Ergod. Th. Dynam. Sys.*, 29:73–109, 2009.

[9] O. Frostman. Potential d'équilibre et capacité des ensembles avec quelque applications à la théorie des fonctions. *Meddel. Lunds Univ. Math. Sem.*, 3:1–118, 1935.

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

[11] F. Hausdorff. Dimension und äusßeres Maß. *Math. Annalen*, 79:157–179, 1918.

[12] R. Hill, S. L. Velani. The Jarník-Besicovitch theorem for geometrically finite Kleinian groups. *Proc. London Math. Soc.*, 77:524–550, 1998.

[13] V. Jarník. Diophantische approximationen and Hausdorff mass. *Mathematicheskii Sbornik*, 36:371–382, 1929.

[14] I.Ya. Khintchine. *Continued Fractions*, Univ. Chicago Press, 1964.

[15] P. J. Nicholls. *The Ergodic Theory of Discrete Groups.*, Springer-Verlag, 1983.

[16] S. J. Patterson. The limit set of a Fuchsian group. *Acta Math.*, 136:241–273, 1976.

[17] C. Series. The modular surface and continued fractions. *J. London Math. Soc.*, 31:69–80, 1985.

[18] B. O. Stratmann. Fractal dimensions for Jarník limit sets; the semi-classical approach. *Ark. för Mat.*, 33:385–403, 1995.

[19] B. O. Stratmann. Weak singularity spectra of the Patterson measure for geometrically finite Kleinian groups with parabolic elements. *Michigan Math. J.*, 46:573–587, 1999.

[20] B. O. Stratmann. The exponent of convergence of Kleinian groups; on a theorem of Bishop and Jones. *Progr. Probab.*, 57:93–107, 2004.

[21] B. O. Stratmann, S. Velani. The Patterson measure for geometrically finite groups with parabolic elements, new and old. *Proc. London Math. Soc.*(2), 71:197–220, 1995.

[22] D. Sullivan. The density at infinity of a discrete group. *I.H.E.S. Publ. Math.*, 50:171–202, 1979.

[23] D. Sullivan. Entropy, Hausdorff measures old and new, and limit sets of geometrically finite Kleinian groups. *Acta Mathematica*, 153:259–277, 1984.

Methods for Symmetric Key Cryptography and Cryptanalysis

Kaisa Nyberg
Helsinki University of Technology and Nokia Research Center, Finland

Abstract

In this lecture, a brief introduction to statistical methods in cryptography will be given. When designing a cryptographic primitive, such as a block cipher, stream cipher or hash function, the goal is to make it resemble a random function as closely as possible. Statistical methods are used in testing test how well this goal has been achieved. Observed statistical deviations can be used not only in distinguishing the primitive from a truly random function, but also in recovering part of the secret key.

One of the most important statistical methods in cryptography is the linear cryptanalysis introduced by Mitsuru Matsui in [3] for the DES block cipher. Similar linear modelling techniques had been previously used in the analysis of stream ciphers and called as correlation attacks. In this lecture, the basic linear cryptanalysis methods on block ciphers and certain types of stream ciphers will be given.

The linear cryptanalysis method is based on derivation of a linear relationship between the plaintext, ciphertext and the key which holds with a probability different from $1/2$. The method can be extended to using multiple such relationships on the same data. Multidimensional extensions of linear cryptanalysis on block ciphers allow a variety of different statistical approaches, which will be discussed based on a recent paper by Miia Hermelin, Joo Yeon Cho and the author [2]. Finally, it will shown how a key recovery attack using linear cryptanalysis on a stream cipher, for example the attack in [1], or a block cipher can also be formulated as a decoding problem of a general linear code.

Draft Table of Contents

1. Cryptographic primitives
2. Linear approximation of block ciphers
3. Linear approximation of filter generators
4. Linear distinguishing attacks
5. Key recovery attacks using linear cryptanalysis
6. Key recovery attack as a decoding problem

References

[1] C. Berbain, H. Gilbert, and A. Maximov. Cryptanalysis of Grain. In M. Robshaw, editor, *Fast Software Encryption*, volume 4047 of *Lecture Notes in Computer Science*, pages 15–29, Berlin/Heidelberg, 2006. Springer.

[2] M. Hermelin, J. Y. Cho, and K. Nyberg. Multidimensional extension of Matsui's Algorithm 2. In *Fast Software Encryption*, Lecture Notes in Computer Science. Springer, 2009. To appear.

[3] M. Matsui. Linear Cryptanalysis Method for DES Cipher. In T. Helleseth, editor, *Advances in Cryptology – EUROCRYPT '93*, volume 765 of *Lecture Notes in Computer Science*, pages 386–397, Berlin/Heidelberg, 1994. Springer.

A glance at Hyperbolic Function Theory in the Context of Geometric Algebras: Hypergenic Operators

*Sirkka-Liisa Eriksson and Heikki Orelma**
Tampere University of Technology

Abstract

In this contribution we consider hyperbolic function theory in the context of geometric algebras. In the first part we give basic definitions, function theoretic fundamentals and integral representations. In the second part of the contribution we define so called hypergenic operators and consider their properties.

Mathematics Subject Classification (2000). Primary 30G35; Secondary 30A05

Keywords. Hypergenic function, modified Dirac operator, hypergenic operator.

1 On Geometric Algebras

In this section we recall some elementary concepts of geometric algebras and geometric algebra valued functions. All results are classical. Let $\{e_0, e_1, ..., e_n\}$ be a basis of $\mathbb{R}^{n+1}$. An arbitrary vector $x \in \mathbb{R}^{n+1}$ may be written as a linear combination of the basis vectors as

$$x = x_0 e_0 + x_1 e_1 + \cdots + x_n e_n$$

where $x_i \in \mathbb{R}$ for each $i = 0, 1, ..., n$. The scalar product $x \cdot y \in \mathbb{R}$ of $x, y \in \mathbb{R}^{n+1}$ is defined as

$$x \cdot y = x_0 y_0 + x_1 y_1 + \cdots + x_n y_n.$$

The norm for $x \in \mathbb{R}^{n+1}$ is defined in terms of the Euclidean scalar product by

$$|x| = \sqrt{x \cdot x}.$$

We shall assume that our basis is orthonormal, i.e., $e_i \cdot e_j = \delta_{ij}$, where δ_{ij} is the Kronecker symbol. The *geometric product* (or the Clifford product) of vectors $x, y, z \in \mathbb{R}^{n+1}$ is defined by three basic axioms:

(A1) The associativity rule

$$(xy)z = x(yz).$$

(A2) The left and right distributivity rules

$$x(y + z) = xy + xz,$$
$$(x + y)z = xz + yz.$$

(A3) The contraction rule

$$x^2 = |x|^2.$$

The vector space $\mathbb{R}^{n+1}$ is not closed under multiplication as the contraction rule shows. However, the addition and the multiplication with the contraction rule of vectors in $\mathbb{R}^{n+1}$ generate a free associative algebra with unity. This algebra is denoted by $\mathcal{C}\ell_{n+1}$ and it is called the *geometric algebra*. Assume $i \neq j$. Using contraction rule for a vector $x = x_i e_i + x_j e_j$ we obtain $e_i^2 = |e_i|^2 = 1$ and

$$e_i e_j + e_j e_i = 0$$

for unequal i and j. Hence we infer the multiplication rule

$$e_i e_j + e_j e_i = 2\delta_{ij}$$

*This paper is worked out in the department of Mathematical Analysis at Ghent University during the spring 2009. Hence the second author wishes to thank all the colleagues in the Galglaan for the nice and inspirational atmosphere.

for each $i, j = 0, 1, ..., n$. The geometric algebra $\mathcal{C}\ell_{n+1}$ has dimension 2^n and the canonical basis is given by $e_A = e_{a_1} \cdots e_{a_k}$ where $A = \{a_1, ..., a_k\} \subset N = \{1, ..., n\}$ and $a_1 < \cdots < a_k$. Especially $e_\emptyset = 1$ and $e_{\{j\}} = e_j$. The space of k-vectors is defined by $\mathcal{C}\ell^k_{n+1} = \text{Span}\{e_A : |A| = k\}$ and thus any $a \in \mathcal{C}\ell_{n+1}$ has the following multivector decomposition:

$$a = [a]_0 + [a]_1 + \cdots + [a]_n$$

with $[a]_k \in \mathcal{C}\ell^k_{n+1}$. The space of 0-vectors and 1-vectors $\mathcal{C}\ell^0_{n+1}$ and $\mathcal{C}\ell^1_{n+1}$ is usually identified with $\mathbb{R}$ and $\mathbb{R}^{n+1}$ respectively.

The *main involution* is an algebra automorphism $' : \mathcal{C}\ell_{n+1} \to \mathcal{C}\ell_{n+1}$ satisfying $e'_i = -e_i$ and $(ab)' = a'b'$ for each $a, b \in \mathcal{C}\ell_{n+1}$.

If $x \in \mathbb{R}^{n+1}$ is a vector such that $x \neq 0$ we define its inverse by $x^{-1} = \frac{x}{|x|^2}$. Let us consider a mapping $f : \Omega \to \mathcal{C}\ell_{n+1}$ where Ω is an open subset of $\mathbb{R}^{n+1}$. In the canonical basis it has the representation

$$f = \sum_A e_A f_A.$$

A function f is called differentiable in Ω if and only if f_A is differentiable in Ω for each A.

The geometric product for any vectors x and y may be decomposed into symmetric and antisymmetric parts defined by

$$x \cdot y = \frac{1}{2}(xy + yx)$$

and

$$x \wedge y = \frac{1}{2}(xy - yx),$$

where $x, y \in \mathbb{R}^{n+1}$, i.e.,

$$xy = x \cdot y + x \wedge y.$$

The antisymmetric part $x \wedge y$ is called the outer product and the symmetric part $x \cdot y$ is called the inner product. It is easy to prove that the inner product $x \cdot y$ is scalar valued and precisely the Euclidean scalar product in above.

2 Hypergenic Functions

If f is differentiable the *left Dirac operator* is defined by

$$D_\ell f = \sum_{k=0}^{n} e_k \frac{\partial f}{\partial x_k}$$

and the *right Dirac operator* is defined by

$$D_r f = \sum_{k=0}^{n} \frac{\partial f}{\partial x_k} e_k$$

where partial derivatives operate componentwise. Denoting $\mathcal{C}\ell_n$ the Clifford algebra generated by $\{e_1, ..., e_n\}$. We may represent the Clifford algebra $\mathcal{C}\ell_n$ as the direct sum

$$\mathcal{C}\ell_{n+1} = \mathcal{C}\ell_n \oplus e_0 \mathcal{C}\ell_n.$$

Let π_1 and π_2 be the corresponding projections, i.e., $\pi_1(a + e_0 b) = a$ and $\pi_2(a + e_0 b) = e_0 b$ and let $\mu : \mathcal{C}\ell_{n+1} \to \mathcal{C}\ell_{n+1}$ be the involution $\mu(a) = e_0 a$. Using the previous mappings we define

$$P_0 := \pi_1 \quad \text{and} \quad Q_0 := \mu \circ \pi_2.$$

Let Ω be an open subset of $\mathbb{R}^{n+1}$ contained in the upper half-space $\mathbb{R}^{n+1}_+ := \mathbb{R}^{n+1} \cap \{x_0 > 0\}$. We define the *left- and right-modified Dirac operator* on the open set Ω using the previous mappings by

$$H^\ell_k f = D_\ell f - \frac{k}{x_0} Q_0 f,$$
$$H^r_k f = D_r f - \frac{k}{x_0} Q'_0 f$$

where k is an arbitrary real number. We shall also use abbreviated notations $H^{\ell} := H^{\ell}_{(n-1)}$ and $H^{r} := H^{r}_{(n-1)}$ for index $k = n-1$. Null solutions of the previous operators are called left- and right- hypergenic functions.

As a technical tool we will need P_0- and Q_0-parts of the operators H^{ℓ}_k and H^r_k, represented in the next lemma.

Lemma 2.1 ([3]). *Let $f : \Omega \to \mathcal{C}\ell_{n+1}$ be a smooth function. Then*

(1) $P_0(H^{\ell}_k f) = D_0 P_0 f + \frac{\partial Q_0 f}{\partial x_0} - \frac{k}{x_0} Q_0 f,$

(2) $Q_0(H^{\ell}_k f) = \frac{\partial P_0 f}{\partial x_0} - D_0 Q_0 f,$

(3) $P_0\big((H^{\ell}_k)^2 f\big) = \Delta P_0 f - \frac{k}{x_0}\frac{\partial P_0 f}{\partial x_0},$

(4) $Q_0\big((H^{\ell}_k)^2 f\big) = \Delta Q_0 f - \frac{k}{x_0}\frac{\partial Q_0 f}{\partial x_0} + \frac{k}{x_0^2} Q_0 f,$

where $D_1 = e_1\frac{\partial}{\partial x_1} + \cdots + e_n\frac{\partial}{\partial x_n}$.

The hypergenic functions are generalizations of the complex analytic functions in the following sense:

Theorem 2.2 ([1]). *If $\Omega \subset \mathbb{R}^{n+1}$ is an open set and $f : \Omega \to \mathcal{C}\ell_{n+1,0}$ is a smooth function the equation $H^{\ell}_k f = 0$ is equivalent to the system*

$$\frac{\partial f_0}{\partial x_0} + \frac{\partial f_1}{\partial x_1} + \cdots + \frac{\partial f_n}{\partial x_n} - \frac{k}{x_0} f_0 = 0,$$
$$\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j} = 0,$$

for $i, j = 0, ..., n$.

We generalize the above theorem to Geometric algebra valued functions:

Theorem 2.3 ([1]). *Let $\Omega \subset \mathbb{R}^{n+1}$ be an open set and $f : \Omega \to \mathcal{C}\ell_{n+1,0}$ be a smooth function. The equation $H^{\ell}_k f = 0$ is equivalent with the following system of equations*

$$\sum_{i=1}^{n} e_i \frac{\partial P_0 f}{\partial x_i} + \frac{\partial Q_0 f}{\partial x_0} - \frac{k}{x_0} Q_0 f = 0,$$
$$\frac{\partial P_0 f}{\partial x_0} - \sum_{i=1}^{n} e_i \frac{\partial Q_0 f}{\partial x_i} = 0.$$

Assume that Ω is an open subset of $\mathbb{R}^{n+1}_+$. The operator

$$\Delta_{LB} g = x_0^2\Big(\Delta g - \frac{k}{x_0}\frac{\partial g}{\partial x_0}\Big)$$

is *the Laplace-Beltrami operator* for $g \in C^2(\Omega, \mathcal{C}\ell_{n+1})$ with respect to the Riemannian metric

$$ds^2 = \frac{dx_0^2 + \cdots + dx_n^2}{x_0^{\frac{2k}{n-2}}}.$$

Theorem 2.4 ([1]). *Let $\Omega \subset \mathbb{R}^{n+1}$ be an open set and $f : \Omega \to \mathcal{C}\ell_{n+1}$ be a k-hypergenic. The function $P_0 f$ is a solution of the Laplace-Beltrami equation i.e. $\Delta_{LB} P_0 f = 0$ and $Q_0 f$ is a solution of the eigenvalue problem $\Delta_{LB} Q_0 f = -k Q_0 f$.*

Hypergenic functions and solutions of the Laplace-Beltrami operator are related as follows.

Theorem 2.5 ([3]). *Let $\Omega \subset \mathbb{R}^{n+1}_+$ be an open set. A smooth function $f : \Omega \to \mathcal{C}\ell_{n+1}$ is hypergenic if and only if for any $a \in \Omega$ there exists $r > 0$ satisfying $B_r(a) \subset \Omega$ and a map $g : B_r(a) \to \mathcal{C}\ell_{n,0}$ satisfying*

$$f = Dg$$

and

$$\Delta_{LB} g = 0.$$

Partial derivatives with respect to x_0 is not generally hypergenic, but the following result holds.

Lemma 2.6 ([1]). *Let $\Omega \subset \mathbb{R}^{n+1}_+$ be a domain and $f \in C^2(\Omega, \mathcal{C}\ell_{n+1})$ be k-hypergenic, then*

$$\frac{\partial f}{\partial x_i}$$

is hypergenic for each $i = 1, ..., n$.

Proposition 2.7 ([1]). *Let $f \in C^1(\Omega, \mathcal{C}\ell_{n+1})$. The following are equivalent:*

(1) $f = 0$,

(2) f and fe_0 are k-hypergenic.

The *Euler operator* is defined by

$$E = \sum_{i=0}^{n} x_i \frac{\partial}{\partial x_i}.$$

We see that it is a scalar operator and measures the degree of homogenicity of a function.

Theorem 2.8 ([3]). *Let $f : \Omega \to \mathcal{C}\ell_{n+1}$ be a smooth function. If the function f is k-hypergenic then Ef is k-hypergenic.*

If f is a hypergenic function the previous theorem gives us a method how to construct more hypergenic functions:

The above theorem implies that for every k-hypergenic function $f : \Omega \to \mathcal{C}\ell_{n+1}$ there exists the sequence of hypergenic functions defined by

$$E^m f := \underbrace{EE \cdots E}_{m \text{ copies}} f$$

for each $m \in \mathbb{N}$ and $E^0 f = f$. This sequence $\{E^m f : m = 0, 1, ...\}$ is called the *homogenized sequence* of f.

As an example we consider monomials

$$X_i = \sum_{j=0}^{n} (-1)^{\delta_{ij}} x_j e_j$$

where $i = 1, ..., n$. Since $H^\ell X_i = 0$ the monomials X_i are hypergenic. Moreover

$$EX_i = X_i$$

for each $i = 1, ..., n$ and hence the homogenized sequence of X_i is just $\{X_i\}$.

3 Integral Formulas for Hypergenic Functions

We recall briefly some preliminaries from integration theory. Let M be a k-dimensional manifold-with-boundary in $\mathbb{R}^{n+1}_+$, see, e.g., [6]. The boundary of M is denoted by ∂M. If moreover

$$\Lambda^* M = \bigoplus_{p=0}^{n+1} \Lambda^p M$$

is the exterior algebra over $\mathbb{R}^{n+1}$ with basis $\{dx_0, dx_1, ..., dx_n\}$ we then construct the bundle $\mathcal{C}\ell_{n+1} \otimes_{\mathbb{R}} \Lambda^k M$. If $\omega(x)$ is a section of the previous bundle over $x \in M$ it is of the form

$$\omega(x) = \sum_{A,B} \omega_{A,B}(x) e_A dx_B$$

where $B = \{b_1, ..., b_k\} \subset N = \{1, ..., n\}$, $dx_B = dx_{b_1} \wedge \cdots \wedge dx_{b_k}$, and $\omega_{A,B}$ are real functions. The meaning of the symbol $e_A dx_B$ is clear. Let furthermore M be an oriented k-dimensional manifold-with-boundary in $\mathbb{R}^{n+1}$, then we define

$$\int_M \omega(x) = \sum_{A,B} e_A \int_M \omega_{A,B}(x) dx_B.$$

In this paper M will be $(n+1)$ dimensional or n-dimensional i.e. the boundary of M. The $(n+1)$-form

$$dV = dx_0 \wedge dx_1 \wedge \cdots \wedge dx_n$$

on M is called the (Riemannian) volume element. In surface integrals we shall often use the n-form

$$d\sigma = \sum_{i=0}^{n} (-1)^i e_i d\check{x}_i$$

where

$$d\check{x}_i = dx_0 \wedge dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_n$$

for each $i = 0, 1, ..., n$. The exterior derivative d for Clifford algebra valued differential forms is defined componentwise, i.e., if

$$\omega = \sum_{A,B} \omega_{A,B} e_A dx_B$$

is a k-form then

$$d\omega = \sum_{A,B} e_A d\omega_{A,B} \wedge dx_B.$$

Applying the classical Stokes theorem (see [6]) it is easy to prove that:

Theorem 3.1 (Stokes)**.** *Let ω be a $\mathcal{C}\ell_{n+1}$-valued k-form in the oriented k-dimensional manifold-with-boundary M. Then*

$$\int_M d\omega = \int_{\partial M} \omega.$$

Let us denote the interior of the subset $U \subset \mathbb{R}^n$ by U°. Next we recall Cauchy's formula for hypergenic functions:

Theorem 3.2 ([1])**.** *Assume that Ω is an open subset of $\mathbb{R}^{n+1}_+$ and $f : \Omega \to \mathcal{C}\ell_{n+1}$ is hypergenic. Let $M \subset \Omega$ be an oriented $(n+1)$-dimensional manifold-with-boundary. Then*

$$f(y) = \frac{2^{n-1} y_0^{n-1}}{\omega_{n+1}} \int_{\partial M} \frac{(x-y)^{-1} d\sigma(x) f(x) - (\widehat{x}-y)^{-1} \widehat{d\sigma}(x) \widehat{f}(x)}{|x-y|^{n-1} |x-\widehat{y}|^{n-1}}$$

for each $y \in M^\circ$.

Cauchy's formula can also be represented with one kernel:

Theorem 3.3 ([3])**.** *Assume that Ω is an open subset of $\mathbb{R}^{n+1}_+$ and $f : \Omega \to \mathcal{C}\ell_{n+1}$ is hypergenic. Let $M \subset \Omega$ be an oriented $(n+1)$-dimensional manifold-with-boundary. Then*

$$f(y) = \frac{2^{n-1} y_0^{n-1}}{\omega_{n+1}} \int_{\partial M} \Lambda(x,y) \Big(P_0\big(d\sigma(x) f(x)\big) + \frac{y - Px}{x_0} Q_0\big(d\sigma(x) f(x)\big) \Big).$$

for each $y \in M^\circ$ where

$$\Lambda(x,y) = \frac{(x-y)^{-1} - (\widehat{x}-y)^{-1}}{|x-y|^{n-1} |x-\widehat{y}|^{n-1}}.$$

For general geometric algebra-valued functions we have:

Theorem 3.4 (Borel-Pompeiu Formula, [3])**.** *Assume that Ω is an open set of $\mathbb{R}^{n+1}_+$ and $f : \Omega \to \mathcal{C}\ell_{n+1}$ is differentiable. Let $M \subset \Omega$ be an oriented $(n+1)$-dimensional manifold-with-boundary. Then*

$$\begin{aligned} f(y) = {} & \frac{2^{n-1} y_0^{n-1}}{\Omega_{n+1}} \int_{\partial M} \Big\{ \frac{1}{x_0^{n-1}} P_0\big(p(x,y) d\sigma(x) f(x)\big) + e_0 Q_0\big(q(x,y) d\sigma(x) f(x)\big) \Big\} \\ & + \frac{2^{n-1} y_0^{n-1}}{\Omega_{n+1}} \int_M \frac{1}{x_0^{n-1}} \Big\{ P_0\big(p(x,y) H^\ell f(x)\big) + e_0 Q_0\big(q(x,y) H^\ell f(x)\big) \Big\} dV. \end{aligned}$$

for each $y \in M^\circ$ where $p(x,y) = x_0^{n-1} \frac{(x-y)^{-1} - (x-\widehat{y})^{-1}}{|x-y|^{n-1} |x-\widehat{y}|^{n-1}}$ and $q(x,y) = \frac{(x-y)^{-1} + (x-\widehat{y})^{-1}}{|x-y|^{n-1} |x-\widehat{y}|^{n-1}}$ are hypergenic with respect to x.

4 On Hypergenic Operators

We denote the class of hypergenic functions in Ω by $\mathcal{H}(\Omega)$ and $H := H^{\ell}_{(n-1)}$. In the previous section of the paper we proved that the Euler operator and derivatives with respect to the variables $x_1, ..., x_n$ preserve hypergenity. In this section we consider more generally that type of operators. Any operator $T : \mathcal{H}(\Omega) \to \mathcal{H}(\Omega)$ is called *a hypergenic operator*, that is, Tf is a hypergenic for any hypergenic function f. We proceed algebraically. Let as consider the free associative algebra over reals with unity generated by operators

$$L_i : f \mapsto e_i f, \qquad R_i : f \mapsto f e_i,$$
$$D_i : f \mapsto \partial_{x_i} f, \qquad P_i : f \mapsto x_i f,$$

where $i = 0, 1, ..., n$. The composition operation is defined by

$$AB \cdots Cf := A(B(\cdots(Cf))).$$

Obviously the identity element of the algebra is the mapping $f \mapsto f$. The previous algebra contain all geometric algebra valued differential operators with polynomial coefficients and denoted by $\mathcal{P}(n)$. Obviously there is a subset of $\mathcal{P}(n)$ which contains all geometric algebra valued hypergenic operators. For the associativity we may still enrich the algebra $\mathcal{P}(n)$ pursuantly. Let us define, in $\mathcal{P}(n)$, the usual commutator relation $[\cdot, \cdot] : \mathcal{P}(n) \times \mathcal{P}(n) \to \mathcal{P}(n)$, i.e., if $A, B \in \mathcal{P}(n)$ then

$$[A, B] = AB - BA.$$

Hence the triplet $(\mathcal{P}(n), +, [\cdot, \cdot])$ is obviously a Lie algebra. Any Lie subalgebra of $\mathcal{P}(n)$ generated by any subset of hypergenic operators is called a hypergenic algebra. Hence we can ask usual questions related to Lie algebras concerning subalgebras and ideals etc.

Since in the class of hypergenic functions an operator T is hypergenic if and only if $HTf = THf = 0$ for each hypergenic f, we obtain:

Theorem 4.1. *An operator T is hypergenic if and only if $[H, T] = 0$ in the class of hypergenic functions.*

Before some practical examples of hypergenic operators, we need to compute:

Proposition 4.2. *For previous operators:*

$$[L_i, P_j] = [D_i, R_j] = [P_i, R_j] = [L_i, D_j] = 0.$$

Also immediately we obtain

Proposition 4.3. *Operators R_i and D_i are hypergenic for $i = 1, ..., n$.*

Proof. In Lemma 2.6 we saw that D_i is hypergenic for any $i = 1, ..., n$. Since

$$Q_0(fe_i) = Q_0(P_0 f + e_0 Q_0 f e_i) = (Q_0 f)e_i$$

and $D(fe_i) = (Df)e_i$ we obtain that

$$HR_i f = D(fe_i) - \frac{n-1}{x_0} Q_0(fe_i) = R_i H f,$$

which completes the proof. □

Hence we see that operators R_i and D_i, for $i = 1, ..., n$, generate a hypergenic algebra.

We cannot give complete presentation of hypergenic operators. We shall consider only some interesting operators. Similar methods works also for more complicated operators.

Proposition 4.4. *Let $\lambda = (\lambda_0, \lambda_1, ..., \lambda_n) \in \mathbb{R}^{n+1}$ and*

$$T_\lambda : f \mapsto \sum_{i=0}^{n} \lambda_i f x_i e_i.$$

Then

$$[H, T_\lambda] f = \sum_{i=0}^{n} \lambda_i e_i f e_i - (n-1)\lambda_0 f'$$

for any smooth $f : \Omega \to \mathcal{C}\ell_n$.

Proof. Let $f : \Omega \to \mathcal{C}\ell_n$ be a smooth function. We need the properties

$$Q_0(fe_i) = \begin{cases} P'f, & \text{for } i = 0, \\ (Q_0 f)e_i, & \text{for } i \neq 0, \end{cases}$$

Hence

$$\begin{aligned} Q_0 T_\lambda f &= \lambda_0 x_0 P_0' f + \sum_{i=1}^{n} \lambda_i x_i (Q_0 f) e_i \\ &= \lambda_0 x_0 (P_0' f - Q_0 f e_0) + T_\lambda Q_0 f \\ &= \lambda_0 x_0 f' + T_\lambda Q_0 f \end{aligned}$$

Since

$$D(x_i f e_i) = e_i f e_i + x_i (Df) e_i$$

we get

$$DT_\lambda f = \sum_{i=0}^{n} \lambda_i e_i f e_i + T_\lambda Df.$$

Thus

$$\begin{aligned} HT_\lambda f &= DT_\lambda f - \frac{n-1}{x_0} Q_0 T_\lambda f \\ &= \sum_{i=0}^{n} \lambda_i e_i f e_i + T_\lambda Df - \frac{n-1}{x_0} \lambda_0 x_0 f' - \frac{n-1}{x_0} T_\lambda Q_0 f \\ &= \sum_{i=0}^{n} \lambda_i e_i f e_i - (n-1)\lambda_0 f' + T_\lambda Hf, \end{aligned}$$

which completes the proof. □

Corollary 4.5. *Let $\lambda = (\lambda_0, \lambda_1, ..., \lambda_n) \in \mathbb{R}^{n+1}$. The operator*

$$T_\lambda : f \mapsto \sum_{i=0}^{n} \lambda_i f x_i e_i$$

is hypergenic for each λ satisfying

$$\sum_{i=0}^{n} \lambda_i e_i f e_i - (n-1)\lambda_0 f' = 0$$

for each hypergenic $f : \Omega \to \mathcal{C}\ell_n$.

One possible and useful way to consider hypergenic operators $T : \mathcal{H}(\Omega) \to \mathcal{H}(\Omega)$ is to restrict the domain set. Hence if $\mathcal{A}(\Omega) \subset \mathcal{H}(\Omega)$ is a nonempty subset we may consider hypergenic operators $T : \mathcal{A}(\Omega) \to \mathcal{H}(\Omega)$. In the above case T is called to the restricted operator to the set $\mathcal{A}(\Omega)$.

Restricting T_λ-operator to the class of real functions we may govern all possible hypergenic monomials of the form $\lambda_0 x_0 e_0 + \lambda_1 x_1 e_1 + \cdots + \lambda_n x_n e_n$.

Proposition 4.6. *For each $(t_1, ..., t_n) \in \mathbb{R}^n$ the function*

$$T_{(\frac{|t|}{n-2}, t_1, ..., t_n)} 1$$

is hypergenic, where $|t| = \sum_{i=1}^{n} t_i$.

Corollary 4.7. *Assume $\mu_k = (1, ..., 1, -1, 1, ..., 1)$ is an $(n+1)$-tuple and (-1) is at the k'th place in the tuple. Then*

$$X_k = T_{\mu_k} 1.$$

Proof of Proposition 4.6. . Consider the equation

$$\sum_{i=0}^{n} \lambda_i e_i f e_i - (n-1)\lambda_0 f' = 0.$$

For $f = 1$ we obtain

$$\sum_{i=0}^{n} \lambda_i - (n-1)\lambda_0 = 0$$

which completes the proof. □

Next we shall restrict the operator T_λ to the set of hypergenic monomials

$$X_k = \sum_{j=0}^{n} (-1)^{\delta_{kj}} x_j e_j$$

where $k = 1, ..., n$.

Proposition 4.8. *In the set of hypergenic monomials X_k the operator T_λ is hypergenic only if $\lambda = 0$.*

Proof. Consider the equation

$$[H, T_\lambda]X_k = \sum_{i=0}^{n} \lambda_i e_i X_k e_i - (n-1)\lambda_0 X_k' = 0.$$

Since

$$e_i e_j e_i = \begin{cases} e_i, & \text{for } i = j, \\ -e_j & \text{for } i \neq j, \end{cases}$$

we obtain

$$\begin{aligned} e_i X_k e_i &= \sum_{j=0}^{n} (-1)^{\delta_{kj}} x_j e_i e_j e_i \\ &= (-1)^{\delta_{ki}} x_i e_i - \sum_{\substack{j=0 \\ j \neq i}}^{n} (-1)^{\delta_{kj}} x_j e_j \\ &= 2(-1)^{\delta_{ki}} x_i e_i - X_k. \end{aligned}$$

Since $X_k' = -X_k$ we infer

$$\begin{aligned} [H, T_\lambda]X_k &= \sum_{i=0}^{n} \lambda_i (2(-1)^{\delta_{ki}} x_i e_i - X_k) + (n-1)\lambda_0 X_k \\ &= \sum_{i=0}^{n} \lambda_i 2(-1)^{\delta_{ki}} x_i e_i - \sum_{i=0}^{n} \lambda_i X_k + (n-1)\lambda_0 X_k \end{aligned}$$

Denoting $|\lambda| = \sum_{i=0}^{n} \lambda_i$ we have

$$\begin{aligned} [H, T_\lambda]X_k &= \sum_{i=0}^{n} \lambda_i 2(-1)^{\delta_{ki}} x_i e_i - |\lambda| X_k + (n-1)\lambda_0 X_k \\ &= \sum_{i=0}^{n} (-1)^{\delta_{ki}} (2\lambda_i - |\lambda| + (n-1)\lambda_0) x_i e_i. \end{aligned}$$

The above expression vanish for each $x \in \Omega$ if and only if

$$2\lambda_i - |\lambda| + (n-1)\lambda_0 = 0$$

for each $i = 0, 1, ..., n$. The corresponding coefficient matrix is

$$A = \begin{pmatrix} n & -\mathbb{I}^T \\ (n-2)\mathbb{I} & B \end{pmatrix}$$

where $\mathbb{I}^T = (1, 1, ..., 1)$ and $B = (b_{ij})$ is the matrix with coefficients $b_{ij} = 2\delta_{ij} - 1$. Hence

$$\det A = \begin{vmatrix} n & -\mathbb{I}^T \\ \mathbf{0} & C \end{vmatrix} = -n \det C$$

where the first equality follows adding the first row to others multiplied by $-(n-2)/n$. Thus $C = (c_{ij})$ is a symmetric matrix with coefficients $c_{ij} = 2\delta_{ij} - 1 + \frac{n-2}{n} = 2\delta_{ij} - \frac{2}{n}$. Using induction argument with respect to the dimension of an n-determinant it is easy to prove that if

$$a_{ij} = \begin{cases} a, & \text{for } i = j, \\ 1 & \text{for } i \neq j, \end{cases}$$

then $\det(a_{ij}) = (a + n - 1)(a - 1)^{n-1}$. Using that information one can show that the matrix A is non-singular. □

Corollary 4.9. *The operator T_λ is hypergenic only for $\lambda = 0$.*

Hence we see that the restriction is essential when we consider hypergenic operators.

Also we introduce one step more complicated operator.

Proposition 4.10. *Let $\lambda = (\lambda_{ij})$ be a real $(n+1) \times (n+1)$-matrix and*

$$S_\lambda : f \mapsto \sum_{i,j=0}^{n} \lambda_{ij} f x_i e_j.$$

Then

$$[H, S_\lambda]f = \sum_{i,j=0}^{n} \lambda_{ij} e_i f e_j - \frac{n-1}{x_0} \sum_{i=0}^{n} \lambda_{i0} x_i f' = 0$$

for any smooth $f : \Omega \to \mathcal{C}\ell_n$.

Proof. We need the following properties

$$Q_0(f e_j) = \begin{cases} P'f, & \text{for } j = 0, \\ (Q_0 f)e_j & \text{for } j \neq 0. \end{cases}$$

The properties implies that

$$\begin{aligned} Q_0 S_\lambda f &= \sum_{i=0}^{n} \sum_{j=1}^{n} \lambda_{ij} x_i (Q_0 f) e_j + \sum_{i=0}^{n} \lambda_{i0} x_i P_0' f \\ &= \sum_{i,j=0}^{n} \lambda_{ij} x_i (Q_0 f) e_j + \sum_{i=0}^{n} \lambda_{i0} x_i (P_0' f - Q_o f e_0). \end{aligned}$$

Hence

$$[Q_0, S_\lambda]f = \sum_{i=0}^{n} \lambda_{i0} x_i f'.$$

Since

$$D(x_i f e_j) = e_i f e_j + x_i (Df) e_j$$

we obtain

$$[D, S_\lambda]f = \sum_{i,j=0}^{n} \lambda_{ij} e_i f e_j.$$

Hence

$$[H, S_\lambda]f = \sum_{i,j=0}^{n} \lambda_{ij} e_i f e_j - \frac{n-1}{x_0} \sum_{i=0}^{n} \lambda_{i0} x_i f',$$

which completes the proof. □

Corollary 4.11. *Let $\lambda = (\lambda_{ij})$ be a real $(n+1)\times(n+1)$-matrix. The operator*

$$S_\lambda : f \mapsto \sum_{i,j=0}^{n} \lambda_{ij} f x_i e_j.$$

is hypergenic for each λ satisfying

$$\sum_{i,j=0}^{n} \lambda_{ij} e_i f e_j - \frac{n-1}{x_0} \sum_{i=0}^{n} \lambda_{i0} x_i f' = 0$$

for any hypergenic $f : \Omega \to \mathcal{C}\ell_n$.

We leave the (possible) deeper study of the operator S_λ to forthcoming papers. Finally we study a generalization of the Euler operator.

Proposition 4.12. *Let $\lambda = (\lambda_0, \lambda_1, ..., \lambda_n) \in \mathbb{R}^{n+1}$. The operator*

$$E_\lambda : f \mapsto \sum_{i=0}^{n} \lambda_i x_i \partial_{x_i} f$$

is hypergenic only for $\lambda = \alpha(1, ..., 1)$, i.e., only

$$E_{(\alpha,...,\alpha)} = \alpha E$$

is hypergenic.

Proof. Since

$$\begin{aligned} DE_\lambda f &= \sum_{i=0}^{n} \lambda_i e_i \partial_{x_i} f + \sum_{i=0}^{n} \lambda_i x_i \partial_{x_i} Df \\ &= \sum_{i=0}^{n} \lambda_i e_i \partial_{x_i} f + E_\lambda Df \end{aligned}$$

we have

$$[D, E_\lambda] f = \sum_{i=0}^{n} \lambda_i e_i \partial_{x_i} f.$$

Computing

$$E_\lambda\Big(\frac{n-1}{x_0} Q_0 f\Big) = -\lambda_0 \frac{n-1}{x_0} Q_0 f + \frac{n-1}{x_0} E_\lambda Q_0 f$$

and $E_\lambda Q_0 = Q_0 E_\lambda$ we infer that

$$\big[\frac{n-1}{x_0} Q_0, E_\lambda\big] f = \lambda_0 \frac{n-1}{x_0} Q_0 f.$$

Hence E_λ is hypergenic if

$$[H, E_\lambda] f = \sum_{i=0}^{n} \lambda_i e_i \partial_{x_i} f - \lambda_0 \frac{n-1}{x_0} Q_0 f = 0.$$

Since f is hypergenic it satisfy the equation

$$\sum_{i=0}^{n} e_i \partial_{x_i} f - \frac{n-1}{x_0} Q_0 f = 0.$$

We obtain that $\lambda_i = \alpha \in \mathbb{R}$ for each $i = 0, 1, ..., n$ and the proof is complete. □

Some remarks and conclusions. In the classical generalized function theory (see, e.g., [4] or [5]) the formula $H(fx) = (Hf)x$ has significant role. But as we saw in above that in our case the operator $f \mapsto fx$ is not hypergenic. Hence the preceding formula is not available in our theory in the similar form. Thus in forthcoming studies hypergenic operators will be in essential role. We should find the restriction $\mathcal{A}(\Omega) \subset \mathcal{H}(\Omega)$ such that there exists a good class of (multiplicative) hypergenic operators $T : \mathcal{A}(\Omega) \to \mathcal{A}(\Omega)$.

References

[1] S.-L. Eriksson, H. Orelma, Hyperbolic Function Theory in the Clifford Algebra $\mathcal{C}\ell_{n+1,0}$, *Adv. Appl. Clifford Algebr.*, 2009

[2] S.-L. Eriksson, H. Orelma, On Modified Dirac Operators in Geometric Algebras: Integration of Multivector Functions, *Proceedings of Graduate Student Workshop on Clifford Algebras and Inverse Problems, 2008* (to appear)

[3] S.-L. Eriksson, H. Orelma, Topics on Hyperbolic Function Theory in $\mathcal{C}\ell_{n+1,0}$ *Comput. Methods Funct. Theory*, (Submitted).

[4] H. Leutwiler, Generalized Function Theory. *Clifford Analysis and Applications, Proceedings of the Summer School held in Tampere University of Technology, August, 2.6, 2004*

[5] H. Leutwiler, Introduction to generalized function theory. *Clifford algebras and potential theory, 65–84, Univ. Joensuu Dept. Math. Rep. Ser., 7, Univ. Joensuu, Joensuu, 2004*

[6] M. Spivak, Calculus on Manifolds, *Benjamin*, New York, 1965.

Interaction of two charges in a uniform magnetic field: symmetries, reduction and non-integrability of the planar problem

D. Pinheiro[*] *and R. S. MacKay*[†]

Abstract

We review some recent results concerning the Hamiltonian system that describes the interaction of two charges moving in a plane under the action of a uniform magnetic field. This is an interesting example of the use of symmetries to reduce the phase space dimension in order to enable a clear analysis of the corresponding Hamiltonian dynamical system[1].

1 Introduction

The interaction of two charges moving in $\mathbb{R}^2$ in a magnetic field $\boldsymbol{B}$ can be formulated as a Hamiltonian system with 4 degrees of freedom. Assuming that the magnetic field is uniform and the interaction potential has rotational symmetry this Hamiltonian system can be reduced to one with 2 degrees of freedom; for certain values of the conserved quantities and choices of parameters, this system is integrable. Furthermore, when the interaction potential is of Coulomb type, for suitable regime of parameters, there are invariant subsets on which this system contains a suspension of a subshift of finite type. This implies non-integrability for this system with a Coulomb type interaction. A detailed study of this problem can be found in [10]. See also [11] for the analogous problem in three dimensional space.

In this paper we briefly review the results in [10]. In section 2 we formulate our problem as a Hamiltonian system with a non–canonical symplectic form, making it easier to identify the system symmetries. We identify translational and rotational symmetries of the system and the corresponding conserved quantities, as weel as an exceptional conserved quantity when the two particles have the same gyrofrequency. In section 3 we state a result regarding the reduction of the Hamiltonian system introduced in the previous section and briefly point out two techniques that can be used to prove the result. In sections 4 and 5, we specialize our analysis of the problem by choosing a specific interaction potential and build on the results of the previous sections to study the dynamics of the Hamiltonian system previously introduced. The natural choice for the potential V is the Coulomb potential

$$V(r) = \frac{e_1 e_2}{4\pi\epsilon_0}\frac{1}{r}\,, \tag{1}$$

where r denotes the distance between the two particles, e_1 and e_2 denote the values of the charges and ϵ_0 denotes the permittivity of the vacuum. Depending on the problem other potentials would be plausible as, for example, in [5] a logarithmic potential is chosen for the interaction of two vortices. In fact, our results are valid for a class of potential functions that includes both the Coulomb potential and the screened Coulomb potential. In section 5 we state two results concerning the existence of periodic and chaotic trajectories shadowing sequences of collision orbits. Those results are based on a method introduced in [3] for a proof of the existence of chaotic orbits of the second species for the circular restricted 3-body problem.

2 Problem Formulation

2.1 One charged particle in a magnetic field

For pedagogical reasons we start by considering the well understood case of one particle moving in a uniform magnetic field $\boldsymbol{B}$ of norm $B \neq 0$, orthogonal to the plane of the motion and pointing upwards. A particle of mass $m > 0$ and charge e moving in $\mathbb{R}^2$ under the action of such a field is subject to a Lorentz force of the

[*]CEMAPRE, ISEG, Universidade Técnica de Lisboa, Lisboa, Portugal, dpinheiro@iseg.utl.pt.

[†]Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK, R.S.MacKay@warwick.ac.uk.

[1]This work was supported by FCT - Fundação para a Ciência e Tecnologia grants with reference SFRH / BD / 9239 / 2002 and SFRH / BPD / 27151 / 2006, CMUP - Centro de Matemática da Universidade do Porto and CEMAPRE - Centro de Matemática Aplicada à Previsão e Decisão Económica.

form $\boldsymbol{F_L} = eB\mathbf{J}\boldsymbol{v}$ where $\boldsymbol{v} = (v_x, v_y) \in \mathbb{R}^2$ is the particle velocity and $\mathbf{J}$ is the standard symplectic matrix in $\mathbb{R}^2$, given by

$$\mathbf{J} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} . \tag{2}$$

This system is known to be Hamiltonian with Hamiltonian function and (non-canonical) symplectic form, given by

$$\begin{aligned} H(\boldsymbol{x}, \boldsymbol{v}) &= \frac{1}{2} m |\boldsymbol{v}|^2 \\ \omega &= m\, \mathrm{d}x \wedge \mathrm{d}v_x + m\, \mathrm{d}y \wedge \mathrm{d}v_y - eB\, \mathrm{d}x \wedge \mathrm{d}y . \end{aligned} \tag{3}$$

where $\boldsymbol{x} = (x, y) \in \mathbb{R}^2$ denotes the particle position (see [6]). To put the Hamiltonian system given by (3) into canonical form it is common to introduce the canonical coordinates $\boldsymbol{q} = (q_x, q_y) \in \mathbb{R}^2$ and $\boldsymbol{p} = (p_x, p_y) \in \mathbb{R}^2$, given by

$$\boldsymbol{q} = \boldsymbol{x} , \qquad \boldsymbol{p} = m\boldsymbol{v} + e\boldsymbol{A}(\boldsymbol{x}) , \tag{4}$$

where $\boldsymbol{A}(\boldsymbol{x}) = (A_x(\boldsymbol{x}), A_y(\boldsymbol{x})) \in \mathbb{R}^2$ is a vector potential for $\boldsymbol{B}$. The new Hamiltonian system is then given by

$$\begin{aligned} H(\boldsymbol{q}, \boldsymbol{p}) &= \frac{1}{2m} |\boldsymbol{p} - e\boldsymbol{A}(\boldsymbol{q})|^2 \\ \omega &= \mathrm{d}q_x \wedge \mathrm{d}p_x + \mathrm{d}q_y \wedge \mathrm{d}p_y - e\left(\frac{\partial A_x}{\partial y} - \frac{\partial A_y}{\partial x} + B\right) \mathrm{d}q_x \wedge \mathrm{d}q_y . \end{aligned}$$

Hence, for the system to be canonical the vector field $\boldsymbol{A}(\boldsymbol{x})$ must be chosen to verify the equation

$$\frac{\partial A_x}{\partial y} - \frac{\partial A_y}{\partial x} + B = 0 ,$$

which is indeed the condition for $\boldsymbol{A}(\boldsymbol{x})$ to be a vector potential for $\boldsymbol{B}$. If needed, we make the choice $\boldsymbol{A}(\boldsymbol{x}) = -\frac{B}{2}\mathbf{J}\boldsymbol{x}$. We consider it better, however, to use the formulation (3) because translation symmetry is more transparent, so instead of the change of variables (4) we just make the change of variables given by

$$\boldsymbol{q} = \boldsymbol{x} , \qquad \boldsymbol{p} = m\boldsymbol{v} \tag{5}$$

obtaining the Hamiltonian system

$$\begin{aligned} H(\boldsymbol{q}, \boldsymbol{p}) &= \frac{1}{2m} |\boldsymbol{p}|^2 \\ \omega &= \mathrm{d}q_x \wedge \mathrm{d}p_x + \mathrm{d}q_y \wedge \mathrm{d}p_y + k\, \mathrm{d}q_x \wedge \mathrm{d}q_y , \end{aligned} \tag{6}$$

where $k = -eB$. The symplectic form in (6) defines a Poisson bracket $\{.,.\} : C^\infty\left(\mathbb{R}^4\right) \times C^\infty\left(\mathbb{R}^4\right) \to C^\infty\left(\mathbb{R}^4\right)$ given by

$$\{F, G\} = \frac{\partial F}{\partial q_x}\frac{\partial G}{\partial p_x} - \frac{\partial G}{\partial q_x}\frac{\partial F}{\partial p_x} + \frac{\partial F}{\partial q_y}\frac{\partial G}{\partial p_y} - \frac{\partial G}{\partial q_y}\frac{\partial F}{\partial p_y} - k\left(\frac{\partial F}{\partial p_x}\frac{\partial G}{\partial p_y} - \frac{\partial G}{\partial p_x}\frac{\partial F}{\partial p_y}\right) .$$

In the formulation (6) the Lorentz force effect can not be seen in the Hamiltonian function but it is present in the $k\, \mathrm{d}q_x \wedge \mathrm{d}q_y$ term of the symplectic form and equivalent term in the Poisson bracket.

2.2 Two charged particles in a magnetic field

We now consider two particles with masses m_1 and m_2 (positive) and non–zero charges e_1 and e_2, respectively, n the same magnetic field as described in section 2.1 (uniform of norm $B \neq 0$, orthogonal to the plane of the motion and pointing upwards). Each one of the particles moving under the action of such a field is subject to a Lorentz force of the form $\boldsymbol{F_L} = e_i B\mathbf{J}\boldsymbol{v}_i$ where $\boldsymbol{v}_i = (v_{x_i}, v_{y_i}) \in \mathbb{R}^2$ is the i-th particle velocity ($i \in \{1, 2\}$) and $\mathbf{J}$ is given by (2). Furthermore, we assume that the interaction of the two particles is determined by a potential $V(r)$ depending on the distance r between the two particles.

The phase space M for this problem is $\mathbb{R}^8$ with the singular points of the interaction potential removed (six-dimensional planes if V is the Coulomb potential (1)). Let $\boldsymbol{q}_i = (q_{x_i}, q_{y_i}) \in \mathbb{R}^2$ denote the vector position of the i-th particle and $\boldsymbol{p}_i = (p_{x_i}, p_{y_i}) \in \mathbb{R}^2$ denote its (non–conjugate) momentum

$$\boldsymbol{p}_i = m\boldsymbol{v}_i , \qquad i \in \{1, 2\} .$$

The motion of the two particles can be described by a Hamiltonian system, with Hamiltonian function $H : M \longrightarrow \mathbb{R}$ and non–canonical symplectic form ω, given by

$$\begin{aligned} H(\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{p}_1, \boldsymbol{p}_2) &= \frac{1}{2m_1} |\boldsymbol{p}_1|^2 + \frac{1}{2m_2} |\boldsymbol{p}_2|^2 + V(|\boldsymbol{q}_1 - \boldsymbol{q}_2|) \\ \omega &= \sum_{i=1,2} \mathrm{d}q_{x_i} \wedge \mathrm{d}p_{x_i} + \mathrm{d}q_{y_i} \wedge \mathrm{d}p_{y_i} + k_i \, \mathrm{d}q_{x_i} \wedge \mathrm{d}q_{y_i} \ , \end{aligned} \tag{7}$$

where, for simplicity of notation, we introduce the constants $k_i = -e_i B \ , i \in \{1, 2\}$. The Poisson bracket associated with this symplectic form, $\{.,.\} : C^\infty(M) \times C^\infty(M) \to C^\infty(M)$, is given by

$$\{F, G\} = \sum_{i=1,2} \frac{\partial F}{\partial q_{x_i}} \frac{\partial G}{\partial p_{x_i}} - \frac{\partial G}{\partial q_{x_i}} \frac{\partial F}{\partial p_{x_i}} + \frac{\partial F}{\partial q_{y_i}} \frac{\partial G}{\partial p_{y_i}} - \frac{\partial G}{\partial q_{y_i}} \frac{\partial F}{\partial p_{y_i}} - k_i \left(\frac{\partial F}{\partial p_{x_i}} \frac{\partial G}{\partial p_{y_i}} - \frac{\partial G}{\partial p_{x_i}} \frac{\partial F}{\partial p_{y_i}} \right) \ . \tag{8}$$

The Hamiltonian system defined by (7) is invariant under the group generated by the following families of symmetries

$$\begin{aligned} \phi_{\boldsymbol{v}}(\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{p}_1, \boldsymbol{p}_2) &= (\boldsymbol{q}_1 + \boldsymbol{v}, \boldsymbol{q}_2 + \boldsymbol{v}, \boldsymbol{p}_1, \boldsymbol{p}_2) \\ \phi_\theta(\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{p}_1, \boldsymbol{p}_2) &= (R_\theta \boldsymbol{q}_1, R_\theta \boldsymbol{q}_2, R_\theta \boldsymbol{p}_1, R_\theta \boldsymbol{p}_2) \ , \end{aligned} \tag{9}$$

where $\boldsymbol{v} = (v_x, v_y) \in \mathbb{R}^2$ is a translation vector and R_θ is the rotation matrix in $\mathbb{R}^2$, given by

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \ .$$

We define the (signed) gyrofrequency Ω_i of each particle as

$$\Omega_i = \frac{k_i}{m_i} \ , \qquad i \in \{1, 2\} \ .$$

Proposition 2.1. *The Hamiltonian System* (7) *has the following conserved quantities:*

- *Linear momentum* $\boldsymbol{P} = (P_x, P_y) = \boldsymbol{p}_1 + \boldsymbol{p}_2 + \mathbf{J}(k_1 \boldsymbol{q}_1 + k_2 \boldsymbol{q}_2)$.
- *Angular momentum* $L = \sum_{i=1,2} \boldsymbol{q}_i . \mathbf{J} \boldsymbol{p}_i - \frac{k_i}{2} |\boldsymbol{q}_i|^2$.

Furthermore, if the particles have equal gyrofrequencies $\Omega_1 = \Omega_2$*, there exists another conserved quantity* W*, given by*

$$W = |\boldsymbol{p}_1 + \boldsymbol{p}_2|^2 \ .$$

The following commutation relations between the conserved quantities given above hold:

$$\begin{array}{lllllllll} \{P_x, P_y\} &=& k_1 + k_2 \ , & \{L, P_x\} &=& P_y \ , & \{L, P_y\} &=& -P_x \ , \\ \{W, L\} &=& 0 \ , & \{W, P_x\} &=& 0 \ , & \{W, P_y\} &=& 0 \ . \end{array}$$

We note that:

i) the conserved quantities $\boldsymbol{P}$ and L are, respectively, the usual linear and angular momenta for the two-body problem with extra terms representing the presence of the magnetic field and hence the effect of the Lorentz force on the particles.

ii) combining P_x and P_y into the conserved quantity

$$P = |\boldsymbol{P}|^2 = {P_x}^2 + {P_y}^2 \tag{10}$$

we obtain the following commutation relations

$$\{L, P\} = 0 \ , \qquad \{L, W\} = 0 \ , \qquad \{P, W\} = 0 \ , \tag{11}$$

which show L, P and W to be in involution.

iii) corresponding to W there is a "hidden" symmetry in the case of equal gyrofrequencies $\Omega_1 = \Omega_2$, given by

$$\begin{aligned}
\boldsymbol{q}_1 &\rightarrow \boldsymbol{q}_1 + \frac{1}{k_1+k_2}\left[R_{2(k_1+k_2)\phi} - \mathbf{Id}_{2\times 2}\right]\mathbf{J}(\boldsymbol{p}_1+\boldsymbol{p}_2)\\
\boldsymbol{q}_2 &\rightarrow \boldsymbol{q}_2 + \frac{1}{k_1+k_2}\left[R_{2(k_1+k_2)\phi} - \mathbf{Id}_{2\times 2}\right]\mathbf{J}(\boldsymbol{p}_1+\boldsymbol{p}_2)\\
\boldsymbol{p}_1 &\rightarrow \boldsymbol{p}_1 + \frac{k_1}{k_1+k_2}\left[R_{2(k_1+k_2)\phi} - \mathbf{Id}_{2\times 2}\right](\boldsymbol{p}_1+\boldsymbol{p}_2)\\
\boldsymbol{p}_2 &\rightarrow \boldsymbol{p}_2 + \frac{k_2}{k_1+k_2}\left[R_{2(k_1+k_2)\phi} - \mathbf{Id}_{2\times 2}\right](\boldsymbol{p}_1+\boldsymbol{p}_2)\ ,
\end{aligned}$$

where $\phi \in \mathbb{R}$.

3 Reduction

In [10] we prove that the Hamiltonian system (7) can be reduced to one with 2 degrees of freedom. Furthermore, when the two particles have the same gyrofrequency we use the exceptional conserved quantity to prove integrability of this Hamiltonian system in this case. We also prove that if the sum of the two charges is zero the dynamics in the zero sets of the linear momenta are also integrable. We do this by constructing a set of coordinates on which the system exhibits a reduction to two degrees freedom, and integrability when it applies. We should remark that a similar reduction is obtained in [5] for the problem of two interacting vortices with mass moving in a plane - in that paper is also given the analogy between that problem and the one we treat here. However, one key point of [10] is that the total change of coordinates that exhibits the reduction is computed. This change of coordinates is just the $SE(2)$ lift that, given the base dynamics of the reduced Hamiltonian systems, enables us to describe the full eight-dimensional dynamics.

The following theorem provides a summary of the results in [10] concerning the reduction of the Hamiltonian system (7). See [10] for more details on the reduced Hamiltonian systems.

Theorem 3.1. *The Hamiltonian system given by* (7) *always reduces to one with at most two degrees of freedom and it is integrable in the following special cases:*

- $\Omega_1 = \Omega_2$.
- $k_1 + k_2 = 0$ *and* ${P_x}^2 + {P_y}^2 = 0$.

In [9] a detailed study is done for the symplectic reduction of the Hamiltonian system (7) by its symmetry group SE(2): the symplectic reduction is mostly regular, making it a standard non-trivial illustration of the theory of symplectic reduction (see [1, 2, 4, 7]). There are, however, level sets of the conserved quantities where the reduction is singular: the level sets of the form $2(k_1+k_2)L + |\boldsymbol{P}|^2 = 0$ in the case $k_1 + k_2 \neq 0$ and $L = 0$ in the case $k_1 + k_2 = 0$ have conical singularities that must be removed for the reduced space to be a smooth manifold. The symplectic reduction for the spatial version of that problem is analogous.

4 Reconstructed Dynamics for a Coulomb potential

The reduced Hamiltonian systems and the corresponding reconstruction maps obtained in [10] can be used to provide a qualitative description of the possible types of dynamics in the full eight-dimensional phase space in terms of the properties of the dynamics of the reduced systems. In this section we consider the interaction potential to be Coulomb

$$V(r) = \frac{e_1 e_2}{4\pi\epsilon_0}\frac{1}{r}\ ,$$

where r is the distance between the particles and ϵ_0 is the permittivity of the vacuum. We should remark, however, that the description given below still holds for a class of Coulomb-type potentials of the form

$$W(r) = \frac{e_1 e_2}{4\pi\epsilon_0}\frac{f(r)}{r}\ ,$$

where $f(r)$ is a positive bounded smooth function. A physically interesting particular case is the screened Coulomb potential where $f(r) = \mathrm{e}^{-r/r_D}$ and r_D is the Debye length.

The reduced Hamiltonian systems exhibit a rich dynamical behaviour:

- In the integrable regimes the energy levels are foliated by periodic orbits.
- Close to the integrable regimes most of the periodic orbits cease to exist but almost all orbits in the energy levels are quasiperiodic and hence the dynamics still look regular.
- As we will state in the next section, for opposite signs of charge (except for the case $\Omega_1 + \Omega_2 = 0$) there is chaotic dynamics which, as said in the Introduction, implies non-integrability for this system.

Using the reconstruction maps we obtain that

- If $k_1 + k_2 \neq 0$ periodic and quasiperiodic base dynamics lift to quasiperiodic dynamics under the reconstruction map. In this case the dynamics are, generically, quasiperiodic with 3 rationally independent frequencies. The particles rotate with these 3 frequencies about a fixed centre determined by the linear momenta.
- If $k_1 + k_2 = 0$ periodic and quasiperiodic base dynamics lift to possibly unbounded motion corresponding to a combination of a drift and quasiperiodic dynamics. The quasiperiodic dynamics have, generically, 2 rationally independent frequencies.
- Chaotic dynamics lift to chaotic dynamics under the reconstruction maps. The motion is always bounded if $k_1 + k_2 \neq 0$ and typically unbounded otherwise.

5 Nonintegrability with a Coulomb-type potential and opposite signs of charge

In this section we state two results that imply that the Hamiltonian system (7) is not integrable for the special case of a Coulomb-type interaction potential and opposite signs of charges with $\Omega_1 + \Omega_2 \neq 0$: there exist regimes of parameters and energy for which there is an invariant subset where the system contains a suspension of a subshift of finite type and has positive entropy. Roughly, this corresponds to the existence of a horseshoe in the dynamics and hence, from a result in [8] we obtain that, for the two degree of freedom reduced Hamiltonian systems in [10], there is no other analytic conserved quantity independent of the Hamiltonian function.

Since the integrable case $\Omega_1 = \Omega_2$ does not have any saddle point in its reduced phase space, there are no possibilities for a simple use of Melnikov method to obtain chaos for nearby $\Omega_1 \neq \Omega_2$.

The condition of opposite signs for the charges is needed to guarantee arbitrarily close approaches on the level sets of the conserved quantities of (7). The construction of a large set of collision orbits forms an important part in the proof of existence of chaotic orbits.

Theorem 5.1. *Let e_1 and e_2 be non-zero and have opposite signs. Furthermore, assume that $e_1 + e_2$ is non-zero and fix values $\ell \in \mathbb{R}$ of L and $h > 0$ of H such that*

$$\xi = \frac{(k_1 + k_2)l}{h} \in (0, m_1 + m_2) \ . \tag{12}$$

Then,

- *if Ω_1 and Ω_2 are rationally independent then for every $\xi \in (0, m_1 + m_2)$ there are infinitely many non-degenerate collision trajectories of energy h and for any finite set K of them there exists $\delta_0 > 0$ such that for every chain $(\gamma_{k_i})_{i\in\mathbb{Z}}$, $k_i \in K$, and $\delta \in (0, \delta_0)$ there is a unique trajectory of energy h near the collision chain and converging to the chain as $\delta \to 0$.*
- *If $|\Omega_1/\Omega_2|$ is rational, say N_1/N_2 in lowest terms, then*
  - *(i) if $\min\{m_1, m_2\} \geq m'$ and $N_1 > 2$ (resp. $N_2 > 2$) there is a subinterval (m_1, m^*) (resp. (m_2, m^*)) of $(0, m_1 + m_2)$ such that for all $\xi \in (m_1, m^*)$ (resp. $\xi \in (m_2, m^*)$) there are at least 4 non-degenerate collision trajectories of energy h, and the set of chains formed from them has positive entropy. Furthermore, if $N_2 - 2 < N_1$ or $N_1 - 2 < N_2$ there is a subinterval (m'', m') of $(0, m_1 + m_2)$ such that for all $\xi \in (m'', m')$ there are at least 4 non-degenerate collision trajectories of energy h, and the set of chains formed from them has positive entropy.*
  - *(ii) if $m_2 < m' < m_1$ (resp. $m_1 < m' < m_2$) and $N_1 > 2$ (resp. $N_2 > 2$) there is a subinterval (m_1, m^*) (resp. (m_2, m^*)) of $(0, m_1 + m_2)$ such that for all $\xi \in (m_1, m^*)$ (resp. $\xi \in (m_2, m^*)$) there are at least 4 non-degenerate collision trajectories of energy h, and the set of chains formed from them has positive entropy.*

(iii) if $m' < \min\{m_1, m_2\}$ there is a subinterval $(m', \min\{m_1, m_2\})$ of $(0, m_1+m_2)$ with $2(N_1+N_2-1)$ non-degenerate collision trajectories of energy h.

Given a finite set K of non-degenerate collision trajectories, there exists $\delta_0 > 0$ such that for every chain $(\gamma_{k_i})_{i\in\mathbb{Z}}$, $k_i \in K$, and $\delta \in (0, \delta_0)$ there is a unique trajectory of energy h near the collision chain and converging to the chain as $\delta \to 0$.

A similar result holds for the case $e_1 + e_2 = 0$.

Theorem 5.2. *Let e_1 and e_2 be non-zero and assume that $e_1 + e_2 = 0$. Fix the values $\boldsymbol{p} \in \mathbb{R}^2$ of $\boldsymbol{P}$ and $h > 0$ of H such that*

$$\xi = \frac{|\boldsymbol{p}|^2}{2h} \in (0, m_1 + m_2) \ . \tag{13}$$

Then,

- *if Ω_1 and Ω_2 are rationally independent then for every $\xi \in (0, m_1 + m_2)$ there are infinitely many non-degenerate collision trajectories of energy h, and for any finite set K of them there exists $\delta_0 > 0$ such that for every chain $(\gamma_{k_i})_{i\in\mathbb{Z}}$, $k_i \in K$, and $\delta \in (0, \delta_0)$ there is a unique trajectory of energy h near the collision chain and converging to the chain as $\delta \to 0$.*
- *If $|\Omega_1/\Omega_2|$ is rational and not equal to 1, say N_1/N_2 in lowest terms, for all $\xi \in (0, m_1 + m_2)$ there is at least one chain and for $\xi \in (0, \min\{m_1, m_2\})$ there is a set of chains with entropy at least $\log(N_1 + N_2 - 1)$. For each finite set K of non-degenerate collision trajectories there exists $\delta_0 > 0$ such that for every chain $(\gamma_{k_i})_{i\in\mathbb{Z}}$, $k_i \in K$, and $\delta \in (0, \delta_0)$ there is a unique trajectory of energy h near the collision chain and converging to the chain as $\delta \to 0$.*

6 Conclusions

This paper provides a short survey of the main results in [10]. Namely, the Hamiltonian system (7) can always be reduced to one with two degrees of freedom. Moreover, for interaction between the two charged particles determined by a Coulomb potential, with opposite sign charges (except for the case $\Omega_1 + \Omega_2 = 0$), this system can not be reduced further, because it contains a suspension of a nontrivial subshift of finite type. On the other hand it is integrable for the special case of same sign charges when the particles have equal gyrofrequencies (equal ratio of charge to mass) and on some special submanifolds.

References

[1] R. Abraham and J. Marsden. *Foundations of Mechanics.* Benjamin/Cummings, Reading, Massachusetts, 2nd edition, 1978.

[2] V. I. Arnold. *Mathematical Methods of Classical Mechanics.* Springer, New York, 1989.

[3] S. V. Bolotin and R. S. MacKay. Periodic and chaotic trajectories of the second species for the n–center problem. *Cel Mech Dyn Astron*, 77:49–75, 2000.

[4] R. Cushman and L. Bates. *Global Aspects of Classical Integrable Systems.* Birkhäuser Verlag, Basel, Switzerland, 1997.

[5] C. Grotta Raggazo, J. Koiller, and W. M. Oliva. On the Motion of Two-Dimensional Vortices with Mass. *J. Nonlinear Sci.*, 4:375–418, 1994.

[6] R. G. Littlejohn. A guiding center hamiltonian: A new approach. *J Math Phys B*, 20:2445–2458, 1979.

[7] J. Marsden and T. Ratiu. *Introduction to Mechanics and Symmetry.* Springer, New York, 1999.

[8] J. Moser. *Stable and Random Motions in Dynamical Systems.* Princeton University Press, Princeton, New Jersey, 1973.

[9] D. Pinheiro. *Interaction of two charges in a uniform magnetic field.* PhD thesis, University of Warwick, 2006.

[10] D. Pinheiro and R. S. MacKay. Interaction of two charges in a uniform magnetic field: I. Planar problem. *Nonlinearity*, 19:1713–1745, 2006.

[11] D. Pinheiro and R. S. MacKay. Interaction of two charges in a uniform magnetic field: II. Spatial problem. *Journal of Nonlinear Science*, 18:615–666, 2008.

On an algorithm for factoring natural numbers

Ramilya Rubtsova
Kazan State University, Russia

Abstract

Let a natural N and a set of prime numbers $F = \{p_0, p_1, \ldots, p_z\}$ called *factor base* be given. We describe a method for a searching pairs (A, C) of naturals satisfying to relation $C = A^2 \mod N$, where C can be factored into a product of elements of F or their powers. Such pairs are later used to represent N as (a unique) product of prime numbers, that is to factor N. The well known method RSA of encryption with a public key is based on a computational hardness of this problem.

1 Introduction

Let factor base $F = \{p_0, p_1, \ldots, p_z\}$ be the set of the first z+1 primes. A natural N is called smooth over F, if it factored into a product $N = p_0^{e_0} \cdot ..., \cdot p_z^{e_z}$ of elements of F. Smooth numbers are in a one-one correspondence with vectors $(e_0, e_1, \ldots, e_z)$. A smooth number N is a square of another number iff all powers e_i in the corresponding vector are even. Simultaneosly with smooth numbers we work with *semismooth* numbers. An integer C is called *semismooth* over F, if $C = A^2 \cdot B$ for an integer A and an F-smooth integer B.

Almost all modern methods of factorization are based on the idea of searching pairs (C, D) with D being F-smooth satisfying to condition

$$C^2 \equiv D \mod N \qquad (1)$$

When a sufficiently large set M of such pairs is accumulated, a system of linear equations is formed to find a subset $M \subseteq H$ satisfying

$$A^2 = \prod_{(C,D)\in M} C^2 \equiv \prod_{(C,D)\in M} D \equiv B^2 \mod N \qquad (2)$$

for a natural B. Then a divisor of N can be found as a greatest common divisor of N and $|A \pm B|$.

Such pairs are searching by various methods and algorithms. We describe in the next section a new strategy for it.

2 Definitions and Algorithms

Let natural N be given such that there exist (unknown) primes p and q such that $N = p \cdot q$. Our main task is to find divisors p and q. We fix a factor base $F = \{p_0, p_1, \ldots, p_z\}$ consisting of $z + 1$ primes. Elements of F are much less than p and q.

A pair of integers (A, C) satisfying (1) with F-smooth C is called *equation*.

A pair (P, Q) consisting of semismooth P and any integer Q is called *relation* (or incomplete equation) if

$$P \equiv Q \mod N \qquad (3)$$

The mail task is to convert *relations* into *equations*. Below we describe a conversion algorithm.

First we fix a constant InitRel denoting the number of initial relations. We need to obtain a system of at least $z + 2$ equations. Since each relation can produce a number of equations, constant InitRel is chosen several times less than z. The work of the algorithm can be divided into two stages, stage of initialization and stage of implementation. At the first stage we form a table of initial relations.

Stage I. Initialization.

1. For each x from 1 to InitRel calculate numbers $R(x) = [\sqrt{N}] + x$ and $Q(x) = R(x)^2 - N$. For each pair $S = (R(x), Q(x))$ (denoting below merely $S = (R, Q)$), carry out the following actions:

2. Factor R and Q into products $R = C_1 \cdot B_1$, $Q = C_2 \cdot B_2$, of F-smooth factors B_1, B_2 and C_1, C_2 do not containing divisors from F. Numbers B_1, B_2 are stored in computer memory as $z + 1$-dimension vectors

with coordinates equal to powers in factoring of B_1, B_2 by F. Factors C_1, C_2 are stored as arrays of 16-bits numbers. Due to the choice of (R, Q) the next relation holds:

$$N + C_2 \cdot B_2 = C_1^2 \cdot B_1^2 \qquad (4)$$

or, replacing B_1^2 by B_1'

$$C_1^2 \cdot B_1' \equiv C_2 \cdot B_2 \mod N \qquad (5)$$

Number C_2 here we call *normable* divisor. The idea of our method is a step by step transformation of a considered relation into an equation by diminishing of the normable divisor.

Consider the pair $S = (C_1' \cdot B_1'; C_2 \cdot B_2)$. If coefficient C_2 is equal to 1 then S is an equation and is added to the table of equations. Otherwise, it is a relation and is added to the table of initial relations.

If coefficient C_2 is equal to 1 then pair (P, Q) is an equation and is added to the table of equations. Otherwise, we add (P, Q) to the table of relations.

The stage of initialization is completed.

Stage II. Implementation.

1. Run over the table of relations to choose a pair (P, Q), $P = C_1^2 \cdot B_1$, $Q = C_2 \cdot B_2$ with a least coefficient C_2. It satisfies to

$$C_1^2 \cdot B_1 \equiv C_2 \cdot B_2 \mod N \qquad (6)$$

2. Calculate a real number $t = N/C_2$. The main idea is to search for a semismooth number D close to t. If such D exists then multiplying both parts of (6) by D and taking the rest of factor $D * C_2$ by module N we obtain in the right part a smooth number B_2' and a semismooth number E in left part. Such a pair can be transformed into an equation. But even if such factor D does not exist, in many cases we can find a suitable D' to obtain a new relation with less factor C_2.

3. Search for semismooth numbers D in interval $[k \cdot t - L;\ k \cdot t + L]$, $k \in \{1, 2, \ldots Bnd\}$ for a small constant L and a bound Bnd. Since it is preferable to find D more close to $k \cdot t$ the internal cycle is needed to take on k. The searching procedure for factors D is the most problem point of the construction.

4. When a suitable $D = C^2 \cdot B$ is found (call it normalizing factor), multiply both parts of (6) by it:

$$D \cdot C_1^2 \cdot B_1 \equiv D \cdot C_2 \cdot B_2 \mod N \qquad (7)$$

5. Replace factor $E = D \cdot C_2$ in the right part of (7) by $C_2' = E \mod N$. Due to the choice of D it should be a small integer (or at least less than C_2). Separate smooth and nonsmooth factors in the both parts of the relation to obtain a new relation:

$$(C_1')^2 \cdot B_1' \equiv C_2' \cdot B_2' \mod N \qquad (8)$$

Call this operation *normalization* of a relation. After this step common divisors of B_1' and B_2' differing from 1 can appear, so we reduce them substracting from presentation vectors of B_1' and B_2' nonzero numbers $r_i = \min\{B_1(i),\ B_2(i)\}$ for i, $0 \le i \le z$.

If after normalization factor C_2' is not equal to 1, we obtain a new relation such that C_2' is less C_2 with a high probability. Otherwise, the aim is reached and a new equation can be obtained. In order to make left part of (8) equal to a square, add 1 to both $B_1'(i)$ and $B_2'(i)$ for each i, $0 \le i \le z$, such that B_1' is odd.

6. Continue the cycle of item 3.

7. When the cycle is completed remove the considered pair (P, Q) from the table of relations.

8. Go to 1 and repeat items 1-7 anew. Finish when the number of found equation will exceed $z + 2$.

Example. Let $N = 1396231$. Factor base B contains 10 first primes

$F = \{2, 3, 5, 7, 11, 13, 17, 19, 23, 29\}$

We need to find $z + 2 = 12$ equations. Set parameters InitRel, L, and Bnd equal to 8, 5, and 10 respectively (these parameters are to be varied). When InitRel is increased, values of L and Bnd can be diminished.

Initialization stage.

$[\sqrt{N}] = 1181$. Pairs (P, Q) form a set {(1182, 893), (1183, 3258), (1184, 5625), (1185, 7994), (1186, 10365), (1187, 12738), (1188, 15113), (1189, 17490) }.

The third pair is an equation since $5625 = 3^2 \cdot 5^4$ and we immediately add it to the table of equations.
Let us consider a pair (1183, 3258). It forms a relation like (6)

$$7^2 \cdot 13^4 \equiv 181 \cdot 2 \cdot 3^2 \mod N \qquad (9)$$

Corresponding $C_2 = 181$, and $t = N/C_2 \approx 7713,98$. A cycle of step 3 of the construction gives us three non-trivial semismooth factors (in fact, smooth) $D_1 = 7714 = 2 \cdot 7 \cdot 19 \cdot 29$, $D_2 = 61710 = 2 \cdot 3 \cdot 5 \cdot 11^2 \cdot 17$, $D_3 = 54000 = 2^4 \cdot 3 \cdot 5$.

Multiplying (9) by these D and taking into account that $181 * 7714 \mod 1396231 = 3$, $181 * 61710 \mod 1396231 = -338 = -2 * 13^2$, $181 * 54000 \mod 1396231 = 383$, we obtain

$$2 \cdot 7^3 \cdot 13^4 \cdot 19 \cdot 29 \equiv 2 \cdot 3^3 \mod 1396231, \qquad 2 \cdot 3 \cdot 5 \cdot 7^2 \cdot 11^2 \cdot 13^4 \cdot 17 \equiv -2^2 \cdot 3^2 \cdot 13^2 \mod 1396231,$$

$$2^4 \cdot 3 \cdot 5 \cdot 7^2 \cdot 13^4 \equiv 383 \cdot 2 \cdot 3^2 \mod 1396231.$$

The equal factors in the left and right parts can be reduced. These formulas will give us 2 new equation and a relation:

$$7^4 \cdot 13^4 \cdot 19^2 \cdot 29^2 \equiv 3^3 \cdot 7 \cdot 19 \cdot 29 \mod 1396231 \rightarrow Eq = (7^2 \cdot 13^2 \cdot 19 \cdot 29;\ 3^3 \cdot 7 \cdot 19 \cdot 29)$$

$$5^2 \cdot 7^2 \cdot 11^2 \cdot 13^2 \cdot 17^2 \equiv -2 \cdot 3 \cdot 5 \cdot 17 \mod 1396231 \rightarrow Eq = (5 \cdot 7 \cdot 11 \cdot 13 \cdot 17;\ -2 \cdot 3 \cdot 5 \cdot 17)$$

$$Rel = (2^3 \cdot 5 \cdot 7^2 \cdot 13^4;\ 383 \cdot 3)$$

Further we carry out the same calculations again for each pair in the table of relation.

After the 14-th cycle the common number of found equations became 12 that satisfied the required condition. Solving system using Gaussian elimination we obtain a pair of numbers (A, B) satisfying to (1): A=1184, B=75. Using Euclid's Algorithm we can find a greatest common divisor $d = gcd(N, A + B) = gcd(1396231, 1259) = 1259$. Dividing N by 1259 we find the second divisor of N equal to 1109. So $N = 1259 \cdot 1109$.

3 Conclusions

The current version of the algorithm does not ensure an essential advantage to existing methods (especially relatively to leaders, the Quadratic Sieve and the General Number Field Sieve) but if the procedure of a searching of normalizing factors will be improved, this method can give in many cases a better performance. An advantage of the method is that it allows for a single relation to find several equations and this gives a hope to make it faster.

References

[1] T.Cormen, C.Leiserson, R.Rivest. Introduction to algorithms. *MIT Press, McGraw Hill*, 1990.

[2] C. Pomerance. A tale of two sieves. *Notices of AMS*, 1473- 1485, 1996

[3] H. Boender. The number of Relations in the Quadratic Sieve Algorithm *NM-R9622, The Netherlands, 1996*

Girls and boys and equity in mathematics: Teachers' beliefs

Riitta Soro
PhD in Education, University of Turku, Loimaa Secondary School, Finland
riitta.soro@loimaa.fi

Abstract

In Finland there are only minor differences between girls' and boys' mathematics achievements in the evaluations of comprehensive school or in the matriculation examinations arranged in upper secondary schools. Girls tend to underestimate their math ability in school. Females do not participate in advanced mathematics courses or in mathematics-related careers at the same level as males do. The focus of my survey study was to examine on one hand, teachers' beliefs about differences between boys and girls as learners of mathematics, and on the other hand, teachers' beliefs about gender equity in mathematics and the means they used to promote equity. Even though many of the teachers did not express very stereotyped beliefs, a great majority held different beliefs about girls and boys and those differences favoured boys. A great majority of teachers did not believe that they had a responsibility to address gender equity and they did not pay any attention to the issue. Gender equity was considered self-evident and mathematics gender-neutral. Many teachers hold very different beliefs and expectations about girls and boys, and at the same time, they believed that they treated a student as an individual and not as a girl or a boy.

Key Words: gender, equity, mathematics, teacher, beliefs.

1 Introduction

A basic belief underlying my presentation is that females' social learning and beliefs about themselves with regard to mathematics are different from those of males. There are significant sex differences in participation rates in mathematics and science education studies, and in related careers. In the field of mathematics a female is still a peculiarity. Mathematics has been and continues to be a critical filter to careers and occupations, which are interesting, challenging, have high status, and are usually well-paid. There are also other interests than personal economic ones or the interests of economic life needing high technology employees.

Council of Europe has defined: "Gender equality means an equal visibility, empowerment and participation of both sexes in all spheres of public and private life. Gender equality is the opposite of gender inequality, not of gender difference." Educational gender equity in mathematics has not been reached. Females have not elected to participate in advanced mathematics courses or in mathematics-related careers at the same level as males have. Girls tend to underestimate their math ability in school, even though their actual performance is just as good as or better than that of the boys. Teachers' values, expectations and beliefs have an influence on girls' self-confidence in mathematics.

The entire field of mathematics might be enriched if more young females were given the opportunity to grow into mathematical scholars and give their unique contribution. Elizabeth Fennema (1990) wrote: "Mathematics is a unique product of human culture. Permitting females to understand this culture is important both for their own appreciation of the beauty of mathematics and the transmission of this culture to future generations."

2 Gender equity

In this presentation the word equity is used instead of equality. In some aspects "equality" is not synonymous with "equity" and, thus, rather than striving for equality in the meaning of 'sameness' amongst girls and boys, teachers should promote equity which reflects the needs and strengths of both groups. Judgements on educational equity have been based on three different definitions (1) equal opportunity, (2) equal treatment, and (3) equal outcome.

(1) Equal opportunity. Many teachers believe that equity has been reached since there are no formal borders and the co-educational school system provides equal opportunity to elect mathematics. However there are far more boys than girls in advanced math classrooms.

(2) Equal treatment, the second definition of equity, is also problematic. Teachers may believe that they treat boys and girls the same way. Classroom observations show that this does not prevail. Males interact more frequently with their teachers. Teachers have different achievement expectations and they vary their explanations for success and failures depending on the sex of the student. Even, if the teacher strives to equal treatment of both sexes, the personal experiences are not similar; girls and boys may perceive it differently.

(3) Gender equity as equal outcome. If equity in mathematics is defined as equal educational outcome, there should not be gender differences in achievement or participation or in how males and females feel about themselves and mathematics. This third definition is consistent with the definition of equality Council of Europe has given i. e. to require equal visibility, empowerment and participation of both sexes.

3 Results of the study on "Girls and boys and equity in mathematics

3.1 Beliefs about boys and girls as learners of mathematics

Most teachers, 86 %, believed in gender differences in mathematics. The most prominent difference concerned working. Careful hard work is addressed to girls and boys are lazy. Secondly the use of cognitive skills seemed different, girls tend to routines and boys use their power of reason. The third difference was found in attitudes, boys are willing to take risks but girls lack self-confidence. The great majority of teachers mentioned different factors for girls' and boys' high achievement. The characteristics of high achieving boys were more varied and many-sided than those of girls.

Even though many of the teachers did not express very stereotyped beliefs, a great majority held different beliefs about girls and boys and those differences favoured boys. The most emergent was the belief in girls employing inferior cognitive skills. No differences were found between the beliefs of female and male teachers.

3.2 Beliefs about gender equity

Only one third of the teachers regarded the equity issue necessary to be brought up. Some of the teachers refused to answer questions concerning equity and wrote: "Mathematics and teaching mathematics is gender free." The teachers were categorized under following labels according to their answers:

(1) Students have no gender. Approximately 41

(2) Equal treatment. 38

(3) Girls' and boys' needs. 21

(4) Favour the weaker. The teachers of this study did not accept this principle of compensation in math teaching and the alternative of "favouring the weaker one" was rejected by all teachers. The Finnish law on gender equity says that it is possible to deviate from equal treatment especially in the favour of females, if it strives to realize the aims of the law for equality. This compensation is not regarded as discrimination.

4 Conclusions

A great majority of the teachers of this study held different beliefs about girls and boys and those differences favoured boys. Teachers' beliefs and (unconscious) expectations discovered a tendency to attribute boys' success to talent and girls' success to hard work. Some of the teachers were concerned about boys, who were underachieving or might fall aside, but girls were supposed to manage thanks to their consciousness. Boys attained most of teacher attention. But this situation was not seen to violate equity. Gender equity was considered self-evident, so it's no need to make any fuss about it and furthermore mathematics is gender-free.

Valerie Walkerdine published 1989 'Counting Girls Out', a book that changed perceptions about the gender problem. The mainstream analyses of the problem had located one or other 'lack' in girls and women as the root of the problem. In the new edition in 1998 Walkerdine writes in the afterword: "Considerable concern is now being expressed about the relatively poor school performance of boys related to girls....Girls' attainment in school is not celebrated as an index of cleverness, brains or intellectuality. Rather those very

factors that [year 1976] were considered a problem in relation to Mathematics, namely rule-following, rote-learning, neatness, good behaviour and so forth, are presented as the keys to female success, downgrading that success, while suggesting that classrooms are too feminine and that masculinity is downgraded and discouraged. The ideal child it seems is still a boy, a boy indeed with potential, whose success is being thwarted by women and girls, indeed by the very notion of female success."

References

[1] Fennema, E. 1990. Teachers' beliefs and gender differences in mathematics. In E. Fennema & G.C. Leder (eds.) Mathematics and Gender. New York: Teachers College Press, 169-187.

[2] Soro, R. 2002. Opettajien uskomukset tytöistä, pojista ja tasa-arvosta matematiikassa. English abstract. Annales Universitatis Turkuensis ser.C, 191. 243 p. [Teachers' Beliefs about Girls and Boys and Equity in Mathematics]. Turku, Finland: Painosalama.

[3] Soro, R. 2002. Teachers' Beliefs about Gender Differences in Mathematics: A Measurement on a Scale with a New Response Format. In A. Cockburn & E. Nardi (eds.): Proceedings of the 26th Conference of the International Group for the Psychology of Mathematics Education. University of East Anglia. UK.Vol 4, 225-232

[4] Walkerdine, V. 1998. Counting girls out: Girls and Mathematics (New Edition). London: Virago.

On entire solutions of some inhomogeneous linear differential equations in a Banach space

Sergey Gefter and Tetyana Stulova†*

Abstract

Let A be a closed linear operator on a Banach space having a bounded inverse operator and f be an entire function of zero exponential type. The problem on the existence and uniqueness of some entire solutions of the differential equation $w' = Aw + f(z)$ is considered in the paper. Moreover the explicit formula for zero exponential type entire solution is founded.

1 Introduction

The present paper studies special holomorphic solutions of inhomogeneous linear differential equation

$$w' = Aw + f(z) \tag{1}$$

in a Banach space. Here A is a closed linear operator on a Banach space E with the domain of definition $D(A)$ ($D(A)$ is not necessarily dense in E), having a bounded inverse operator. These operators appear under studying some boundary-value problems for parabolic type equations (see [1] - [4]). For example this case is appeared in the problem on the heat conduction on the finite segment $[0, l]$ with zero boundary condition. In this situation we can consider $E = C[0, l]$, and operator $A = \frac{d^2}{dz^2}$ with the domain of definition $D(A) = \left\{u \in C^2[0, l] : u(0) = u(l) = 0\right\}$. Studying Equation (1) we suppose that $f(z)$ is an E-valued function, which is holomorphic in a neighborhood of zero and under a solution of the equation we understand a holomorphic in the neighborhood of zero E-valued function $w(z)$, such that $w(z) \in D(A)$ and Equation (1) is fulfilled in the same neighborhood. It is well-known that if the operator A is bounded then under any initial condition $w_0 \in E$ the Cauchy problem for Equation (1)

$$\begin{cases} w' = Aw + f(z) \\ w(0) = w_0 \end{cases} \tag{2}$$

has always a unique holomorphic solution

$$w(z) = e^{zA} w_0 + \int_0^z e^{(z-\zeta)A} f(\zeta)\, d\zeta \tag{3}$$

(see [1], [5], and [6]). The properties of holomorphic and entire solutions of the equation $w'(z) = Aw(z) + f(z)$ for the case when the operator A is unbounded were studied in numerous works (see, for example, [1], [7] - [9]).

The main result of the paper is the existence proof and the uniqueness one of an entire solution of zero exponential type in a case when $f(z)$ is an entire function of zero exponential type (see Theorem 2.3). Let us recall that $f(z)$ is of zero exponential type if for $f(z)$ the following condition is fulfilled: $\forall \varepsilon > 0\ \exists C_\varepsilon > 0\ \forall z \in \mathbb{C} : \|f(z)\| \le C_\varepsilon e^{\varepsilon |z|}$. This case is even interesting for a bounded operator A, because it is clear that the solution (3) of Cauchy problem (2) can not be of zero exponential type. Besides that for this case we obtain an explicit formula for solution of Equation (1). Note, that we get the solution of Equation (1) of zero exponential type under a single assumption on invertibility of A, and we do not impose another conditions on its resolvent. The proof of the main theorem is based on studying the implicit differential equation $Tw' + g(z) = w$, where $T = A^{-1}$ and $g(z) = -A^{-1} f(z)$. The holomorphic solutions behavior of the implicit equation mentioned above were studied with another technique in [10].

*Kharkov National University; Department of Mechanics and Mathematics, 4 sq. Svoboda, Kharkov, 61077 Ukraine
†National AeroSpace University (KhAI), 17 st. Chkalov, Kharkov, 61070 Ukraine

2 Main results

Let E be a Banach space and $g : \mathbb{C} \longrightarrow E$ be an entire function.

Lemma 2.1. *The function $g(z)$ is of zero exponential type if and only if $g'(z)$ is of zero exponential type. Moreover the following condition is fulfilled*
$\forall \varepsilon > 0\ \exists\ M > 0\ \forall n \in \mathbb{N} \cup \{0\} :\ \left\|g^{(n)}(z)\right\| \leq M\ \varepsilon^n e^{\varepsilon|z|},\ \ z \in \mathbb{C}.$

Proof. Let $g(z) = \sum\limits_{m=0}^{\infty} \alpha_m z^m$ have zero exponential type and $\varepsilon > 0$. As $\sqrt[m]{m!\,\|\alpha_m\|} \to 0$ then
$\exists\ C > 0\ \forall m \in N :\ m!\,\|\alpha_m\| \leq C \cdot \varepsilon^m$, i.e. $\|\alpha_m\| \leq C\frac{\varepsilon^m}{m!}$. Then $\left\|g^{(n)}(z)\right\| =$

$$= \left\| \sum_{m=n}^{\infty} \alpha_m m\,(m-1)\ldots(m-(n-1))\, z^{m-n} \right\| = \left\| \sum_{m=0}^{\infty} \frac{(m+n)!}{m!} \alpha_{m+n} z^m \right\| \leq$$

$$\leq \sum_{m=0}^{\infty} \frac{(m+n)!}{m!} \|\alpha_{m+n}\| \cdot |z^m| \leq \sum_{m=0}^{\infty} \frac{C}{m!} \cdot \varepsilon^{m+n} \cdot |z|^m = C \cdot \varepsilon^n \sum_{m=0}^{\infty} \frac{(\varepsilon\,|z|)^m}{m!} = C\varepsilon^n e^{\varepsilon|z|}$$

.

Conversely, let $g'(z) = \sum\limits_{k=0}^{\infty} \beta_k z^k$ have zero exponential type. Then
$g(z) = g(0) + \int\limits_0^z g'(\xi)\, d\xi = g(0) + z \sum\limits_{k=0}^{\infty} \frac{\beta_k}{k+1} z^k$ and $\sqrt[k]{k!\left\|\frac{\beta_k}{k+1}\right\|} = \frac{\sqrt[k]{k!\|\beta_k\|}}{\sqrt[k]{k+1}} \to 0$, that is the function $g(z)$ has zero exponential type too. □

Let $T : E \longrightarrow E$ be an arbitrary bounded linear operator. At first consider the inhomogeneous implicit differential equation of the form

$$Tw' + g(z) = w, \tag{4}$$

Theorem 2.2. *Let g be of zero exponential type. Then Equation (4) has a unique entire solution of zero exponential type $w(z) = \sum\limits_{n=0}^{\infty} T^n g^{(n)}(z)$.*

Proof. Let $g(z) = \sum\limits_{m=0}^{\infty} \alpha_m z^m$ and $0 < \varepsilon < \frac{1}{\|T\|}$. Then by Lemma 2.1
$\exists C > 0\ \forall n \in N : \left\|g^{(n)}(z)\right\| \leq C\varepsilon^n e^{\varepsilon|z|}$, $z \in \mathbb{C}$. Now show that the series $\sum\limits_{n=0}^{\infty} T^n g^{(n)}(z)$ converges uniformly in any disk and the sum is an entire function of zero exponential type. Let $|z| \leq R$. Then
$\left\|T^n g^{(n)}(z)\right\| \leq C \cdot \|T\|^n \cdot \varepsilon^n e^{\varepsilon|z|} \leq C\,\|T\|^n\,\varepsilon^n e^{\varepsilon R} = Ce^{\varepsilon R}(\varepsilon \cdot \|T\|)^n$ and $\sum\limits_{n=0}^{\infty} (\varepsilon\,\|T\|)^n < +\infty$.
Therefore the series $\sum\limits_{n=0}^{\infty} T^n g^{(n)}(z)$ converges uniformly in the disk $|z| \leq R$. So the function
$w(z) = \sum\limits_{n=0}^{\infty} T^n g^{(n)}(z)$ is entire and it is easy to check that $w(z)$ is a solution of Equation (4). Besides that
$\|w(z)\| \leq \sum\limits_{n=0}^{\infty} \left\|T^n g^{(n)}(z)\right\| \leq Ce^{\varepsilon|z|} \sum\limits_{n=0}^{\infty} (\varepsilon\,\|T\|)^n = \frac{C}{1-\varepsilon\|T\|} \cdot e^{\varepsilon|z|}$, $z \in \mathbb{C}$. Hence $w(z)$ has of zero exponential type. Prove the uniqueness of the entire solution of zero exponential type. Let $w(z) = \sum\limits_{n=0}^{\infty} c_n z^n$ be an entire solution of zero exponential type for the homogeneous equation $Tw' = w$. Then one can easy show that $c_0 = n!T^n c_n$ (see Lemma 2.1 [10]). Therefore $\sqrt[n]{\|c_0\|} \leq \sqrt[n]{n!\,\|c_n\|} \cdot \sqrt[n]{\|T^n\|}$. As $\sqrt[n]{n!\,\|c_n\|} \to 0$ and $\sqrt[n]{\|T^n\|}$ converges to the spectral radius of T, then $\sqrt[n]{\|c_0\|} \to 0$ that is $c_0 = 0$. Note that the function $w^{(k)}(z)$ satisfies the homogeneous equation $Tw' = w$ and it is an entire function of zero exponential type (see Lemma 2.1). Therefore $c_k = 0$, $k \in \mathbb{N}$, that is $w = 0$. Theorem is completely proved. □

Theorem 2.3. *Let A be a closed linear operator on a Banach space (domain of definition $D(A)$ of A is not necessarily dense). Consider the following differential equation*

$$w' = Aw + f(z). \tag{5}$$

If the operator A has a bounded inverse one and $f(z)$ is an entire function of zero exponential type (that is $\forall \varepsilon > 0\ \exists C_\varepsilon > 0\ \forall z \in \mathbb{C} : \|f(z)\| \leq C_\varepsilon e^{\varepsilon |z|}$), then Equation (5) has a unique entire solution of zero exponential type $w(z) = -\sum\limits_{n=0}^{\infty} A^{-(n+1)} f^{(n)}(z)$. Moreover the Cauchy problem

$$\begin{cases} w' = Aw + f(z) \\ w(0) = w_0 \end{cases}$$

has an entire solution of zero exponential type if and only if $w_0 + \sum\limits_{n=0}^{\infty} A^{-(n+1)} f^{(n)}(0) = 0$.

Proof. Let $T = A^{-1}$ and $g(z) = -A^{-1} f(z)$. Then $D(T) = E$, T is bounded, $g(z)$ is an entire function of zero exponential type and Equation (4) is equivalent to Equation (5). According to Theorem 2.2 Equation (5) has the unique entire solution of zero exponential type
$w(z) = \sum\limits_{n=0}^{\infty} T^n g^{(n)}(z) = -\sum\limits_{n=0}^{\infty} A^{-(n+1)} f^{(n)}(z)$ and $w(0) = -\sum\limits_{n=0}^{\infty} A^{-(n+1)} f^{(n)}(0)$. Theorem is proved. □

Note that in the case when in Equation (5) the function $f(z)$ is not entire, and it is only holomorphic in the neighborhood of zero, then Equation (4) can not have a holomorphic in a neighborhood of zero solution at all.

Example 2.4. Let $e_1, e_2, ...$ be an orthonormalized basis in space. We define an operator A on the basis vectors as $Ae_2 = e_1, Ae_3 = 2e_2, Ae_4 = 3e_3, ..., Ae_n = (n-1)e_{n-1}, n \geq 2$. Now we extend the operator A the following way $A\left(\sum\limits_{n=2}^{\infty} u_n e_n\right) \stackrel{def}{=} \sum\limits_{n=2}^{\infty} (n-1) u_n e_{n-1}$ in its natural domain of definition
$D(A) = \left\{ u = \sum\limits_{n=2}^{\infty} u_n e_n : \sum\limits_{n=2}^{\infty} (n-1)^2 |u_n|^2 < +\infty \right\}$. It is easy to check that A is a closed invertible operator. If $f(z) = -\frac{e_1}{1-z}$ and $w(z) = \sum\limits_{n=2}^{\infty} w_n(z) e_n$, then Equation (5) has the form as the infinity system of differential equations

$$\begin{cases} 0 = w_2(z) - \frac{1}{1-z} \\ w_2'(z) = 2w_3(z) \\ w_3'(z) = 3w_4(z) \\ w_4'(z) = 4w_5(z) \\ \dots \\ w_n'(z) = nw_{n+1}(z) \\ \dots \end{cases}$$

Hence, we get the *formal solution* of this system

$$\begin{cases} w_2(z) = \frac{1}{1-z} \\ w_3(z) = \frac{1}{2(1-z)^2} \\ w_4(z) = \frac{1}{3(1-z)^3} \\ \dots \\ w_n(z) = \frac{1}{n-1} \frac{1}{(1-z)^{n-1}},\ n \geq 2 \\ \dots \end{cases}$$

But $w(z)$ is not contained in the domain of definition of the operator A, because for $z = 0$

$$\sum_{n=2}^{\infty} (n-1)^2 \cdot \frac{1}{(n-1)^2} \cdot \frac{1}{|1-z|^{2n-2}} = +\infty.$$

The proof technique of Theorem 2.2 make possible to deduce results on an existence and uniqueness of solution of Equation (4) for other classes of entire functions. As an example we consider the result relating to classes of all entire functions. Here it is naturally to appear the strong restrictions to the operator T.

Theorem 2.5. *Let $T : E \longrightarrow E$ be a bounded quasinilpotent linear operator (i.e. the spectrum $\sigma(T)$ of T reduces to the only point $\lambda = 0$), Fredholm resolvent $F_T(z) = \sum\limits_{n=0}^{\infty} T^n z^n$ of T be of exponential type, and $g(z) = \sum\limits_{m=0}^{\infty} \alpha_m z^m$ be an arbitrary entire function. Then Equation (4) has a unique entire solution $w(z) = \sum\limits_{n=0}^{\infty} T^n g^{(n)}(z)$.*

Proof. As $F_T(z)$ is of exponential type, then $\overline{\lim\limits_{n\to\infty}} \sqrt[n]{n!\,\|T^n\|} < \infty$, that is
$\exists C_1 > 0,\ M > 0\ \forall n \in \mathbb{N}:\ \|T^n\| \le C_1 \cdot \frac{M^n}{n!}$. Now let $R > 0$, and $0 < \varepsilon < \frac{1}{M+R}$. As $g(z)$ is entire, then $\lim\limits_{m\longrightarrow 0} \sqrt[m]{\|\alpha_m\|} = 0$. Therefore $\exists C_2 > 0\ \forall m \in \mathbb{N}:\ \|\alpha_m\| \le C_2 \cdot \varepsilon^m$. Then
$\left\|g^{(n)}(z)\right\| = \left\|\sum\limits_{m=n}^{\infty} \alpha_m m(m-1)\dots(m-n+1) z^{m+n}\right\| = \left\|\sum\limits_{m=0}^{\infty} \frac{(m+n)!}{m!}\alpha_{m+n} z^m\right\| \le$
$\le \sum\limits_{m=0}^{\infty} \frac{(m+n)!}{m!}\|\alpha_{m+n}\| \cdot |z^m| \le \sum\limits_{m=0}^{\infty} \frac{(m+n)!}{m!} C_2 \varepsilon^{m+n} |z|^m = \varepsilon^n C_2 \sum\limits_{m=0}^{\infty} \frac{(m+n)!}{m!} (\varepsilon|z|)^m = \frac{n! C_2 \varepsilon^n}{(1-\varepsilon|z|)^{n+1}}$. From here for $|z| \le R$ we obtain
$\left\|T^n g^{(n)}(z)\right\| \le \|T^n\| \left\|g^{(n)}(z)\right\| \le C_1 \frac{M^n}{n!} \cdot \frac{n! C_2 \varepsilon^n}{(1-\varepsilon|z|)^{n+1}} = \frac{C_1 C_2}{1-\varepsilon|z|} \cdot \left(\frac{M\varepsilon}{1-\varepsilon|z|}\right)^n$.
Since $\frac{M\varepsilon}{1-\varepsilon|z|} < 1$ then the series $\sum\limits_{n=0}^{\infty} T^n g^{(n)}(z)$ converges uniformly in the disk $|z| \le R$. So the function $w(z) = \sum\limits_{n=0}^{\infty} T^n g^{(n)}(z)$ is entire.

Prove the uniqueness of the entire solution. Let $w(z) = \sum\limits_{n=0}^{\infty} c_n z^n$ be an entire solution for the homogeneous equation $Tw' = w$. Then one can easy show that $c_0 = n! T^n c_n$ (see Lemma 2.1 [10]). Therefore $\sqrt[n]{\|c_0\|} \le \sqrt[n]{\|c_n\|}\sqrt[n]{n!\,\|T^n\|}$. As $\sqrt[n]{\|c_n\|} \longrightarrow 0$, and $\sqrt[n]{n!\,\|T^n\|}$ is bounded, then $\sqrt[n]{\|c_0\|} \to 0$, that is $c_0 = 0$. Note that the function $w^{(k)}(z)$ satisfies the homogeneous equation $Tw' = w$ and it is an entire function. Therefore $c_k = 0$, $k \in \mathbb{N}$, that is $w = 0$. Theorem is completely proved.

□

References

[1] S.Kreĭn. Linear differential equations in Banach space. *Translations of Mathematical Monographs, Amer. Math. Soc., Providence, R.I.*, 29, 1971.

[2] Yu.T.Sil'chenko, P.E.Sobolevskii. Solvability of the Cauchy problem for an evolution equation in a Banach space with a non-densely given operator coefficient which generates a semigroup with a singularity (Russian). *Siberian. Math. J.*, 27:4:544–553, 1986.

[3] Yu.T.Sil'chenko. Differential equations with non-densely defined operator coefficients, generating semigroups with singularities. *Nonlinear Anal., Ser. A: Theory Methods*, 36:3:345–352, 1999.

[4] G.Da Prato, E.Sinestrati. Differential operators with non dense domain. *Annali della scuola normale superiore. Di Pisa.*, 14:285–344, 1987.

[5] Ju.Dalec'kii and M.Kreĭn. Stability of differential equations in Banach space. *Amer. Math. Soc., Providence, R.I.*, 1974.

[6] E.Hille. Ordinary differential equations in the complex domain. *A Wiley-Interscience publication, New York, London*, 1976.

[7] M.Gorbachuk. An operator approach to the Cauchy-Kovalevskay theorem. *J. Math. Sci.*, 5:1527-1532, 2000.

[8] M.Gorbachuk. On analytic solutions of operator-differential equations. *Ukrainian Math. J.*, 52:5:680-693, 2000.

[9] M.Gorbachuk, and V.Gorbachuk. On the well-posed solvability in some classes of entire functions of the Cauchy problem for differential equations in a Banach space. *Methods Funct. Anal. Topology.*, 11:2:113-125, 2005.

[10] S.Gefter, and T.Stulova. On Holomorphic Solutions of Some Implicit Linear Differential Equation in a Banach Space. *Operator Theory: Advances and Applications, Birkhäuser Verlag, Basel, Switzerland.*, 191:323–332, 2009.

Collective animal behaviour: coming together

David J. T. Sumpter
Mathematics Department, Uppsala University

These notes are taken from chapter 2 of the book, Collective Animal Behaviour, forthcoming from Princeton University Press.

Animal groups vary in size from two magpies sitting on a branch to plagues of millions of locusts crossing the desert. Not only do the sizes of groups vary between species, but they can change dramatically within species. In some cases, a change in group size depends on changes in the environment. For example, locust outbreaks are thought to originate where resources are patchily distributed, causing locusts to move towards these limited resources (Collett et al., 1998; Despland et al., 2004). In other cases, individuals in similar environments are found in very different-sized groups. Fishermen are used to such intrinsic variation in fish school size. Some days a net contains three fish, while the next day it contains tens of thousands (Bonabeau & Dagorn, 1995). Human settlements also show similar variety in size, from tiny villages to massive cities, with differences in size arising without large differences in the environments in which they were originally founded (Reed, 2001).

Can we then make general predictions about animal group sizes? In this chapter I approach the group size question from the two directions of functional and mechanistic explanation. The functional approach looks at how the costs and benefits of group membership can be used to calculate the optimal group size, at which individuals maximise their fitness, and the stable group size, at which no individual can improve its fitness by moving to another group. The mechanistic approach attempts to explain the large variation in group sizes observed empirically. By describing the mechanisms by which individuals join and leave groups a distribution of group sizes is predicted.

2.1 Optimal group size

There are many ways an individual can benefit from being a member of a group. The movement of a water skater as a predator approaches both confuses the predator and alerts other skaters of its presence (Treherne & Foster, 1981); the starling in a flock can invest less time scanning for potential danger and more time probing the ground for food (Fernandez-Juricic et al., 2004); the homing pigeon released with members of its roost can shorten its route home (Biro et al., 2006); the fish at the front of a school is less likely to be attacked than a straggler outside the group (Parrish, 1989); and the pelican at the back of a v-formation saves energy in the wake of those in front (Weimerskirch et al., 2001). These and many other experimental observations explain why individuals form and join groups. There are also always costs associated with group membership. While some less obvious costs, such as increased parasite burden (Brown & Brown, 1986), have been demonstrated, they have not been studied empirically to the same degree as benefits (Krause & Ruxton, 2002). In part this is because it is reasonable to assume that as group size increases, eventually so too does competition for local resources. For an overview and categorisation of the different costs and benefits of group living see Krause & Ruxton (2002). For any single species, Brown & Brown's (1996) study of cliff swallows is probably the most comprehensive investigation of the costs and benefits of group living.

The functional approach to grouping considers how natural selection will act to shape group size. Individuals which live in groups where benefits outweigh costs will have a higher fitness, i.e. relative probability of survival and reproduction, than those in groups where the costs outweigh benefits. Thus a starting point for making predictions about how group sizes will evolve is to identify a group size fitness function. This fitness function can be calculated as the benefit minus the cost for individuals in groups of different size (figure 2.1).

The main practical consideration in determining the group size fitness function is finding a common currency or units, such as energy intake or time budgets, in which to measure costs and benefits (Krebs & Davis, 1993). For example, Caraco used a theoretical model of the percentage of time yellow-eyed junkos feed, fight with each other and scan for predators to make and test predictions about how behaviour changes with group size (Caraco, 1979b; Caraco, 1979a) and how group size changes with food supply and predation risk (Caraco et al., 1980). Whatever the currency chosen, a general observation about the group size function

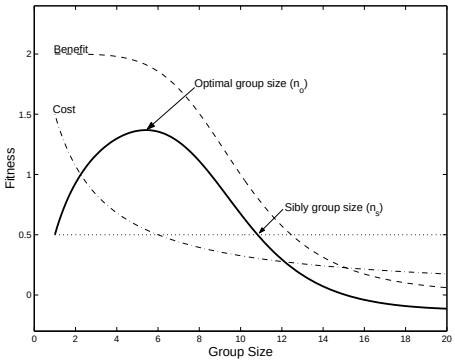


Figure 2.1. Derivation of a theoretical group size fitness function, f(n), where n is group size. The thick dark line is the group size fitness function. This is derived by subtracting the cost function (-.) from the benefit function (-). The cost function in this case is $3.75/(n + 1.5)$ which is a typical predation dilution curve: the rate at which an individual is attacked decreases inversely proportionally to the number of group members. We suppose the benefit function relates to rate of food intake, and is where $K = 10$ is the group size at which individuals forage with exactly half the efficiency they forage with when alone. The optimal group size no is the value of n which gives the maximum difference between costs and benefits. The Sibly group size n_s is the maximum value of n for which $f(n) > f(1)$.

is that as groups become very large the costs will always exceed the benefits. Eventually local competition for resources outweighs any other benefits. The result of this observation is that the fitness function will have at least one maximum. Figure 2.2 gives three examples of theoretical group size fitness functions. In general, even if the group size has more than one local maximum, there is only one global maximum. This maximum is known as the optimal group size.

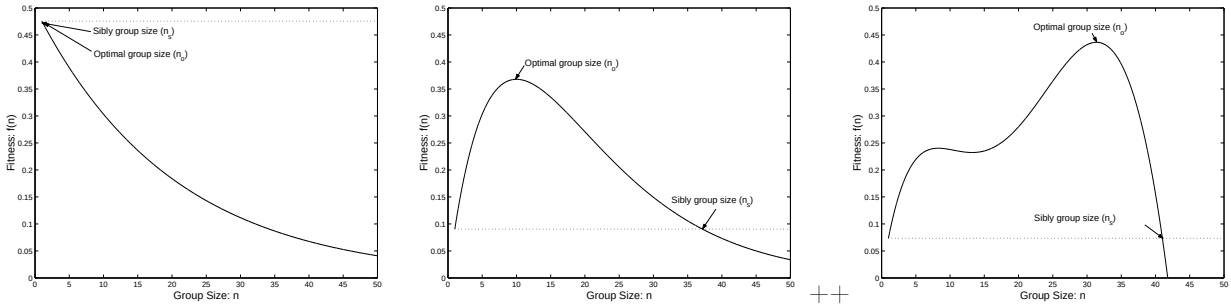


Figure 2.2 Theoretical group size fitness functions. (a) when there is a single maximum at a group size of one it is never advantageous for an individual to join a group; (b) a single maximum gives the optimal group size no, while the group size which has the same fitness as an individual on its own gives the Sibly group size, ns; (c) the fitness function has two local maxima but the global maximum is the optimal group size.

Determining the group size fitness function directly from the energy or time budget of individual animals for different group sizes is difficult in practice, not least because some unknown factor is easily omitted. There are however numerous empirical studies that have been able to relate group size to a particular variable that is likely to contribute to fitness. Pride (2005) found that stress, measured by levels of cortisol concentration, was higher for individuals in smaller and larger groups of Lemur. Individuals in intermediate sized groups showed lower stress (figure 2.3a). Brown and Brown (1996) found that during years where overall survival of young was low, cliff swallows in colonies of between 30 and 80 nests produced more surviving young than smaller or larger colonies. Due to difficulties in measuring survival of these offspring, they were unable to give a clear estimate of survival to adulthood. Without this estimate it is difficult to measure lifetime reproductive success, which accounts for the total number of individuals passed from one adult to the next generation of adults and is thus the preferable measure of fitness. Female lifetime reproductive success has been measured in social spiders (figure 2.3b) and individuals in intermediate sized groups of 23 to 107 had highest fitness (Aviles & Tufino, 1998).

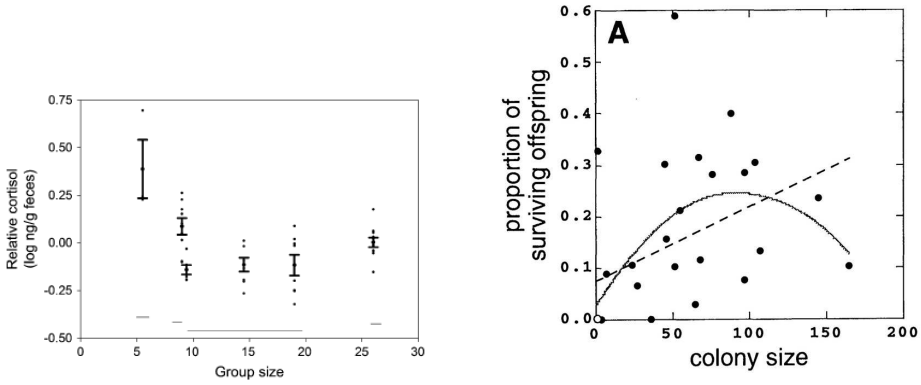


Figure 2.3 Empirical group size fitness functions. (a) Female cortisol levels for female ring-tailed lemurs averaged throughout the year per individual (Pride, 2005); (b) Proportion of surviving offspring per female in the colony for the social spider Amelosimus eximius (Aviles & Tufino, 1998).

2.2 Stable group size

While a particular group size may be optimal, this does not imply that it is stable. One theoretical prediction is that stable group sizes will usually be larger than the optimal group size. The argument for this prediction, first proposed by Sibly (1983), is that there is a benefit for individuals on their own or in smaller groups to join a group of optimal size, thus increasing the group size. More rigorously, this argument is made by considering a series of individuals arriving sequentially and choosing between a number of available resource sites. We assume their choices will be made on the basis of the fitness function, f(n), in figure 2.2b. Further assuming there is no intrinsic difference between sites, the first arriving individual will choose a site at random. The second arriving individual will then choose the same site as the first, since it has higher fitness there than on its own. Further individuals will continue to make the same decision provided the fitness gained from joining the group is larger than that of being on their own, $f(n+1) > f(1)$. The important observation here is the advantage to the arriving individual to join a group even after that group has exceeded the optimal group size, no. If n_s is the largest group size for which $f(n_s) \geq f(1)$ then, under this process, all groups will become of size n_s. For most realistic fitness functions $n_s > n_0$ and the resulting group size will be larger than the optimal size (although see Giraldeau & Gillis (1985) for an exception to this rule where $n_s = n_0$).

The above argument has led some researchers to refer to ns as the stable or equilibrium group size (Beauchamp & Fernandez-Juricic, 2005; Giraldeau, 2000; Clark & Mangel, 1986). This interpretation suggests a paradox whereby groups reach a stable size for which membership confers no benefit over being alone, thus calling into question how grouping can evolve under free entry (Giraldeau, 2000; Giraldeau, 1988). The paradox arises however under three very strict, and in most cases biologically unrealistic, assumptions about how groups are formed: individuals (a) arrive sequentially starting with empty sites; (b) are unable to leave once they have chosen a site; and (c) are naïve to the order in which they arrive.

What happens if we relax assumptions (a) and (b) and individuals are free to move between sites? Box 2.A describes a simulation model, also based on an argument first given by Sibly (1983), in which individuals are free to leave their current resource sites and join a site with higher fitness, with fitness being determined by the same function as in figure 2.2b. Figures 2.4a and 2.4 b show the outcome of this model, given an initially random distribution of individuals between sites. Despite the highly variable starting distribution, the groups quickly converge to a stable size distribution with a mean slightly larger than the optimal group size. This stable group size distribution is not unique. Figures 2.4c and 2.4d show that if individuals are initially distributed with sizes close to n_0 then the mean group size remains close to n_0.

In fact, most distributions of group sizes where all individuals are in groups of size greater than that which is optimal quickly become stable without greatly increasing in size. So unless the initial group size distribution has mean n_s there is no reason that it should be favoured as the mean stable group size over any other mean group size greater than no. Indeed, the simulations suggest that for a wide range of initial group size distributions, stable group sizes will be only slightly larger than optimal.

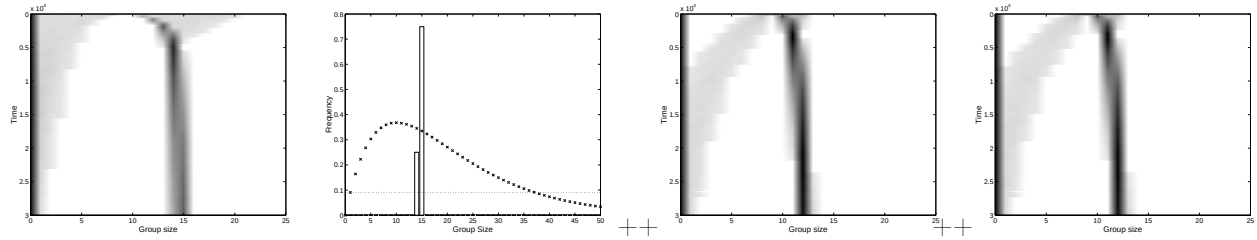


Figure 2.4 Outcome of Sibly's stable group size model for $r = 10$ (a,b) and $r = 2$ (c,d). (a) and (c) show the time evolution of the group size distribution for 30,000 time steps. Shading indicates proportion of sites occupied by a particular number of individuals on a particular time step. (b) and (d) show the stable distribution of site occupation when no further moves are possible for the simulation. Crosses show group size fitness function and thin dotted line gives the Sibly group size, n_s.

Box 2.A Sibly's stable group size model

Consider an environment with $s = 2000$ available sites. Assume initially that at half the sites, $i = 1$ to 1000, the number of individuals at the site, $n_i(0)$, is drawn from a uniform distribution with minimum $10 - r$ and maximum $10 + r$. Thus the average number of individuals at these occupied sites is 10 individuals, equal to the optimal group size in figure 2b. The other half of the sites, $i = 1001$ to 2000, are unoccupied, i.e. $n_i(0) = 0$. The unoccupied sites ensure that grouping in the model does not result simply from a limitation of available sites. The rules of the model are as follows. On each time step t a random individual is picked. It then calculates the fitness function for all of the sites were it to move to that site, i.e. $f(n_j(t) + 1)$ for all sites apart from the site i that is already at. In this case we use the group size fitness function shown in figure 2.2b, which is $f(n) = nexp(n/10)$ If $f(n_j(t) + 1) > f(n_i(t))$ for some j then the individual moves to the site which has the maximum value of $f(n_j(t) + 1)$. If more than one site has the same value of $f(n_j(t) + 1)$ then one of these sites is picked at random. This process is continued until no further moves are possible. An example outcome of this process is shown in figure 2.4. For both wide ($r = 10$) and narrow ($r = 2$) initial distributions of group sizes, small groups quickly reduce in size as members join larger groups. The optimal group size is unstable in both cases and is smaller than the stable group size. The stable group size differs with r, with larger stable group size for larger initial variation in distribution amongst sites. In no case is the stable group size as large as the Sibly group size, n_s.

There are of course many realistic situations in which individuals do arrive sequentially at a resource site and are unable to leave without incurring a cost. A typical example is birds arriving at a nesting site. However, before we predict stable group sizes close to n_s for sequential arrival we must consider what occurs if we remove assumption (c) and allow individuals to know how many individuals will arrive after them. In this case, it is best for early arrivals to occupy empty resource sites, secure in the knowledge that it will be best for later arrivals to join them. Given full knowledge of the sequence of arrivals it is conceivable that the stable strategy will result in group sizes very close to the optimal. A simple example of this can be constructed by considering four birds arriving with a group fitness function: $f(1) = 1$; $f(2) = 3$; $f(3) = 2$; and $f(4) = 1$. To optimize its fitness the third arrival must choose a site on its own (if the second has not already done so) thus ensuring that the fourth joins it. Turning the so-called group size paradox on its head, we see that even if some of the early arrivals are not joined, they will still have a fitness equal to that obtained if they ended up in a group of size n_s. Thus even with a high degree of error group sizes will in general be less than n_s. Although a complete knowledge of arrival sequence is not particularly realistic, changes in strategy dependent on arrival position are observed in birds (Brown & Brown, 1996 chapter 13). The above discussion highlights some of the difficulties in making general predictions about stable group sizes using the evolutionarily stable strategy models first proposed by Sibly. I would agree with the careful conclusion of Sibly (1983): "flocks of optimal size are unstable and will tend to increase in size". However, group sizes only slightly above optimal are stable and only under very limited set of assumptions is there a group size paradox. I thus follow the wording of Krause & Ruxton (2002) and call n_s the Sibly group size. The stable group size lies somewhere between the optimal, n_0, and the Sibly group size, n_s. Group size is likely to be highly dependent on the mechanisms through which groups form and the information available

to potential group members about whether further individuals will join a group.

2.3 Natural group size distributions

How do the actual sizes of animal groups compare to theoretical predictions about optimal and stable group sizes? Data to answer this question is lacking in many of the cases where group size fitness functions have been calculated, and where it is available it is often ambiguous (Krause & Ruxton, 2002). One notable exception is Aviles & Tufino's (1998) study of social spiders. Figure 2.5a shows the distribution of group sizes of spider colonies under natural conditions. Compared to the predicted optimal group size of 50 (figure 2.3b) the mean group size is 425.6. Moreover, of the approximately 18,500 individual spiders surveyed, only 300 were in the optimal group size category of between 50 and 100 spiders. There is little evidence that the spiders usually obtain the optimal group size.

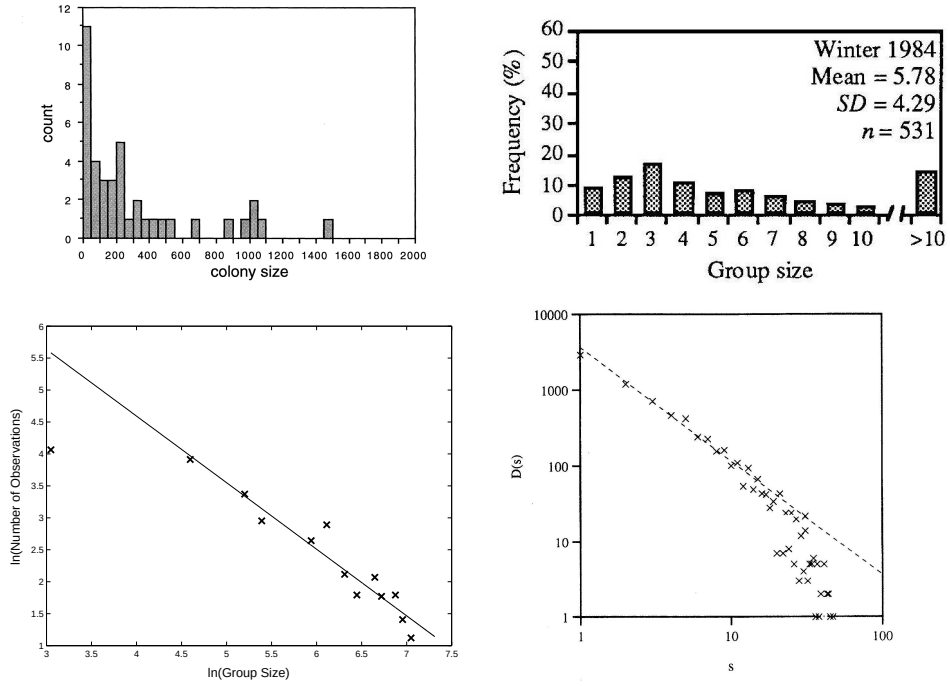


Figure 2.5. Group size distributions for (a) social spiders (Aviles & Tufino, 1998); (b) roe deer in open cultivated planes with population density of 16-18 deer per ha (Gerard et al., 2002) (c) American buffalo (Sinclair, 1977). Here the data is plotted on a log-log scale. The solid line is the best linear regression to all points excluding the first point and gives $\log(\text{frequency}) = 8.76 - 1.04 \cdot \log(\text{group size})$; (d) Frequency of catches in terms of tonnes of Tuna fish caught in a net with a 2km perimeter (Bonabeau & Dagorn, 1995). The fitted line has slope $-3/2$, giving a power law exponent of $\alpha = -3/2$.

While these observations do provide support for the hypothesis that stable groups are larger than optimal, the most striking feature of the spider colony size distributions is that they are highly skewed. There are lots of small groups and a few exceptionally large groups. Similar group size distributions are seen throughout the animal kingdom. In addition to social spiders, figure 2.5b-d shows group size distributions for two mammalian herbivores and tuna fish schools. All these distributions have long tails corresponding to groups that are often several scales of magnitude larger than the modal group size.

Long-tailed group size distributions are clearly not expected from stable group size theory, which predicts a very narrow group size distribution (figure 2.4). This discrepancy between theory and data led Gerard et al. (2002) to question the validity of the stable group size approach to predicting group size. They suggested that although natural selection may play some yet to be established role in determining group size, the dynamics of fission and fusion in mobile mammalian and fish groups means that the sizes of the groups individuals find themselves in will vary widely, are seldom optimal and certainly not stable. Aviles & Tufino (1998) are also sceptical about stable group size theory even for immobile spider aggregations. They cite population growth and dispersal costs as reasons for a wide range of group sizes. Although I would be less inclined than Gerard et al. to dismiss the optimal and stable group size approach entirely, it is clear from these empirical studies that a theory is needed which explains not only why groups of particular sizes arise,

but also why there is such a variation in the size distribution of these groups.

2.4 Power law distributions

Long tailed distributions can often be described as a power law. A simple test of whether group size data might be power law distributed is to plot the logarithm of the group size against the logarithm of the frequency. If the data in this log-log plot is fitted by a straight line then it suggests that the data is power law distributed. Specifically, if the slope $-\alpha$ fits the data then a group size n occurs with frequency $p(n) \propto n^{-\alpha}$. α is referred to as the exponent of the power law.

Figure 2.5c shows such a log-log plot for American buffalo group sizes. This data fits a power law with exponent $\alpha = 1.04$. Figure 2.5d shows that the frequency of sizes of tuna fish catches is also fitted by a power law, with exponent $\alpha = 1.5$, over several orders of magnitude. Once group sizes become very large, frequency distributions usually tail off exponentially. The tuna fish data thus fits a truncated power law: a power law over several orders of magnitude but then tailing off exponentially for very large groups. Since there is usually a limit to how large a real animal group can get we expect most power laws to be truncated at some point. Truncated power laws give a reasonably good fit to many data of animal group sizes from spiders (Aviles & Tufino, 1998), fish (Bonabeau et al., 1999; Niwa, 1998; Niwa, 2003), seals (Sjoberg et al., 2000), and mammalian herbivores (Sinclair, 1977; Gerard et al., 2002).

Long tails in frequency distributions cause great excitement in the minds of theoretical physicists. These distributions are thought to characterise systems with highly non-linear dynamics or amplification of stochastic fluctuations (Sornette, 2004). How are we meant to make biological sense of these ideas? We can start by investigating the assumptions underlying mathematical models that generate power laws. Do models which generate power laws have properties we can relate to the way individual animals interact? It turns out that there are a number of models, each based on reasonable biological assumptions that can generate power laws with slopes that match the data (Newman, 2005; Sornette, 2004). The problem is determining which is most realistic and could actually account for the observations.

2.5 Merge and split models

Power law distributions can be generated from very minimal assumptions about animal behaviour. Bonabeau & Dagorn (1995) proposed a model for animal grouping based on a single assumption: that when groups meet they always merge to form a larger group. The model has s sites, each containing a group of size $n_i(t)$ at time step t. On each time step of the model each group picks a new site to visit at random. If two or more groups choose the same site then they merge, e.g. if the group at site i and the group at site j both move to site k, then $n_k(t+1) = n_i(t) + n_j(t)$. The resulting model is identical to a model of particles that stick together (Takayasu et al., 1988). When these particle (or animal) groups are equally likely to pick any of the available sites, and particles are added to the system at a constant rate, then the probability that a group is of size n is proportional to $n^{-3/2}$ (Takayasu, 1989). This was very close to the power law exponent observed in the catch sizes of tuna (figure 2.5d).

Despite the claim that the above model might provide a universal law for fish school distributions, species other than tropical tuna do not have exponents of $-3/2$. Using computer simulations and further analytical results (Takayasu, 1989; Takayasu et al., 1991), Bonabeau (1999) argued that exponents of between $-4/3$ and $-3/2$ could be accounted for by a reduction in the spatial dimension of the fishes' habitat. For example, attraction to specific resource sites. However, empirically measured exponents have a much wider range of between -0.7 and -1.8 (Bonabeau et al., 1999; Niwa, 1998). A further limitation of the model is that it requires that individuals are continuously added, so that although the scaling rule continues to hold the population increases to infinity with time. If this assumption is removed then the theoretical exponent is -2 and, even less realistically, local populations at sites can become negative (Takayasu et al., 1991). There may be ways to overcome this technical limitation and recover an appropriate range of exponents, but these have not been fully investigated. In summary, while Bonabeau and Dagorn's work was useful in showing power laws in group size distributions, the theoretical model they used is not particularly biologically realistic nor a robust explanation of the available data.

Despite the limitations of early models, there does appear to be a universal scaling law for fish school sizes. Niwa (2003) took all available data on fish school sizes and re-plotted group sizes (N_i) versus frequency (W_i), this time dividing the group sizes by the expected group size experienced by an individual. This expected group size is given by

$$< N >_p = \frac{\sum_{i=1}^{g} N_1^2 W_i}{\sum_{i=1}^{g} N_i W_i}$$

where g is the number of group size classes. $< N >_p$ is not the same as the observed mean group size, which is rather $\sum_{i=1}^{g} N_i W_i$. Rather, $< N >_p$ is the expected group size of an individual picked at random. $< N >_p$ is always equal to or larger than the expected group size, since we are more likely to pick an individual in a larger group. Niwa found that by normalizing the data in this way, distributions for six different fish species all fall on the same curve (figure 2.6a). All these distributions had exponents close to -1 until normalized group size reaches one, at which point they tailed off exponentially.

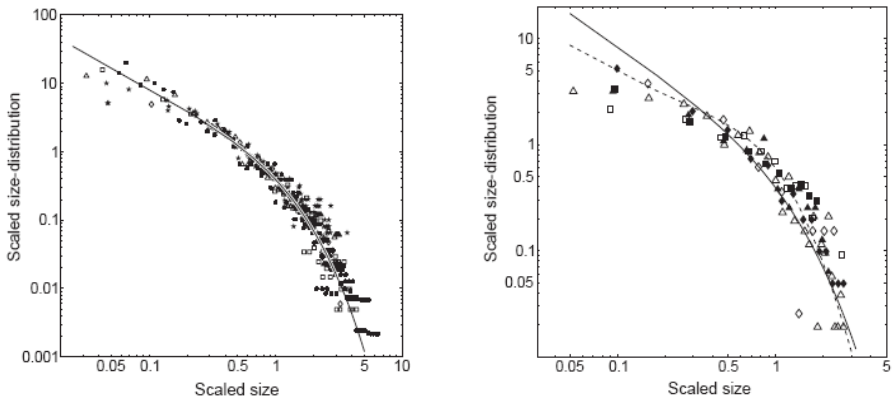


Figure 2.6. Niwa's scaling of fish and mammal group size distributions (Niwa 2003). (a) Empirical distribution of pelagic fish school sizes, six different species represented by different symbols, scaled by the average group size experienced by an individual. The solid line is equation 2.1. (b) Empirical distribution of mammalian herbivore group sizes, six different species represented by different symbols. The solid line is equation 2.1 and the dotted line is a modified version of equation 2.1 with one extra parameter (see Niwa 2003 for details).

The data is well fitted by the predictions of a simple model of group aggregation and breakup. The model's assumptions about aggregation were the same as Bonabeau and Dagorn's - groups move on each time step and when they meet they always merge - but Niwa further assumed that on each time step there is a fixed probability that groups break apart, splitting in to two groups the size of which is uniformly distributed (see Box 2.B for details of the model). The central prediction of this model is that the probability that a site contains a group of size N is

$$W(N) \propto N^{-1} exp\left[-\frac{N}{< N >_p}\left(1 - \frac{e^{-N/<N>_p}}{2}\right)\right] \quad (2.1)$$

This equation captures the qualitative observation that group size distribution at first decreases inversely with N, but once the group size reaches $< N >_p$ it starts to decrease exponentially. It fits both simulations of the above model (figure 2.7b) and the available fish data (figure 2.6a).

Box 2.B Niwa's merge and split model
Assume that space is divided in to s sites on which a total of m individuals are initially randomly distributed. The n_i individuals on site i are said to constitute a group. On each time step there are two stages to the model: move and split. . First all groups move to a new site chosen uniformly at random. If two groups of size ni and nj meet at site k then they form a new group $n_k = n_i + n_j$, thus groups always merge when they meet. The same rule applies if three or more groups meet. After moving each group with a size greater than or equal to 2 will split into a pair of groups with probability p. When a group splits the size of the two components is chosen uniformly at random, so that all group sizes are equally likely. On the next time step the two split groups move separately to new randomly chosen sites, as do all unsplit groups, and the process continues. Figure 2.7a shows a time series of the number of individuals occupying a randomly chosen site for a simulation of this model for parameters $s = m = 2000$ and $p = 0.3$. Figure 2.7b shows the distribution generated by this simulation over $100,000$ time steps.
Niwa derived equation 2.1 by expressing the above simulation model in terms of a stochastic differential equation. He then used simulations to determine a form for the variance in these models and applied results from Richmond (2001) to integrate the model and obtain $W(N) \propto N^{-1} exp\left[-\frac{N}{<N>_p}\left(1 - \frac{e^{-N/<N>_p}}{2}\right)\right]$. Niwa further showed that for a variety of individual based models of schooling

$$< N >_p \approx \frac{3.08\lambda\rho}{p\phi}$$

where λ is the probability per time step that a school merges; p is the probability per time step that a school splits; ρ is the population density, i.e. $\rho = \sum_{i=1}^{g} N_i W_i / s$; and ϕ is the proportion of the s sites occupied by a school (see chapter 5 Box 5.A for details of spatially explicit models).

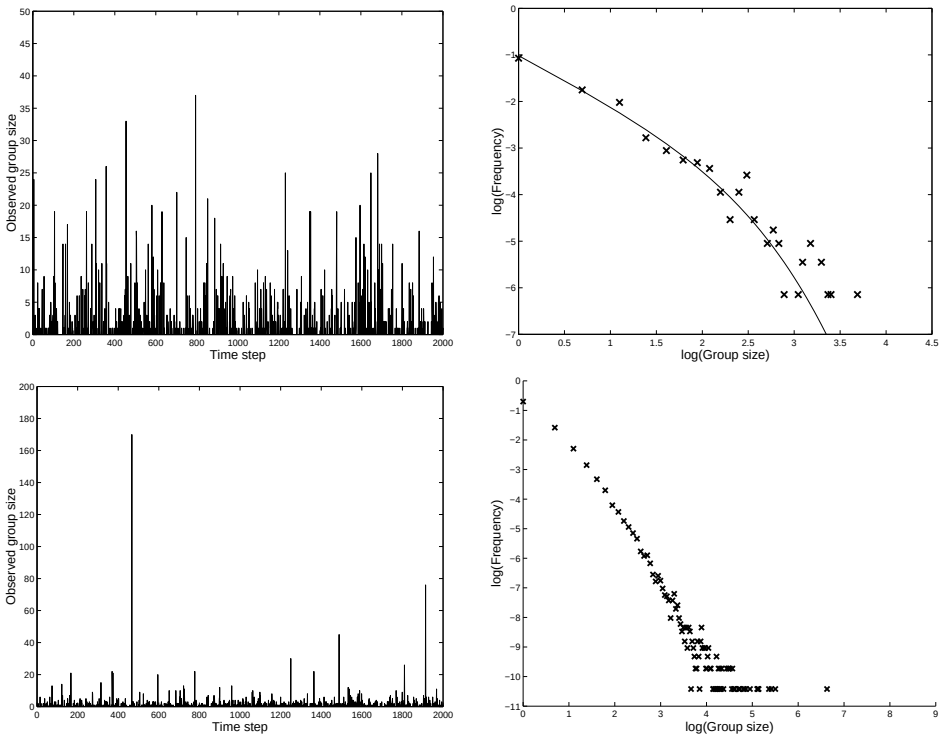


Figure 2.7. Niwa's merge and split model (a and b) and a Preferential attachment model (c and d): (a) shows a time series of the number of individuals occupying a randomly chosen site for a simulation of Niwa's model (see Box 2.B for details) for parameters $s = m = 2000$ and $p = 0.3$; (b) shows the group size distribution generated by this simulation over $100,000$ time steps; (c) shows a time series of the number of individuals occupying a randomly chosen site for a simulation of a preferential attachment model (see Box 2.C for details) for parameters $s = m = 2000$ and $c = 1$; (d) shows the group size distribution generated by this simulation over $100,000$ time steps.

Niwa's work is remarkable in its generality. Bonabeau and Dagorn's model of truncated power laws has 4 parameters, all of which needed to be tuned for particular species. Niwa's model has one parameter which is naturally measured from the data and fits all available fish size data. In theory, measuring the average group

size experienced by an individual allows the entire group size distribution to be predicted. Since equation 2.1 does not contain any model parameters, it is entirely independent of the rates at which groups merge and split. This may seem strange at first, but it should be borne in mind that $< N >_p$ is determined by these rates. Indeed, Niwa (2004) showed that $< N >_p \propto 1/p$ not only for his simple model, but also for a range of spatially explicit simulation models (see also chapter 5). Niwa has established a universal rule for fish schooling that does not depend on specific types of interactions and environmental structure. Provided fish schools merge when they meet and tend to split uniformly at random, we expect Niwa's predictions to hold.

The result does come with a couple of words of warning. When normalised so that they share at least one point in common and stretched out on a log-log plot, very different distributions can begin to appear very similar. A similar method of data fitting has led to misleading conclusions about invariance in life history traits (Nee et al., 2005). Niwa's approach does not suffer from the same deficiencies, because group size and frequency are independent variables. The second warning is that the data used was based on fish catches and observations at fish aggregation devices. Such data is subject to sampling errors, with catches of certain sizes being preferred by the fishermen. With this in mind, it would be reassuring to see a confirmation of these results for more fish species, with data collected using other measuring devices. I make these comments not because I doubt Niwa's findings but because, if the results were confirmed through independent field observations, his work would stand as one of the most fundamental laws of group behaviour.

While Niwa's model might provide a universal rule for fish schooling, it does not appear generalise to mammalian herbivores. Figure 2.6b shows herd size distributions for six different species. Although all six species lie on a similarly shaped curve, the data are not the same as given in equation 2.1. Niwa suggests that mammals might not break up according to the uniform splitting rule given in his model. Another possibility is that groups merge and split as a function of their size, and that the resulting group size distribution is a reflection of this behaviour.

2.6 Preferential attachment

With mammalian herbivores in mind, Gueron & Levin (1995) proposed a general framework for models where the probabilities of fission and fusion are a function of group size. They studied particular examples of this model in which the probability of two groups of size x and y merging could be written as $\psi(x, y) = \alpha a(x)a(y)$, while the probability of a group of size x splitting as $p(x) = \beta x a(x)$. They considered three cases: $a(x) = 1$; $a(x) = x$; and $a(x) = 1/x$. The use of $a(x)$ in both the splitting and joining probabilities produced a mathematical symmetry which allowed them to determine a function for group size frequency (Gueron, 1998). Like Niwa's model they predict that the frequency of larger groups decreases exponentially with group size. This prediction was also made in Okubo's (1986) classic review of animal grouping, where he also argued that the available data on mammalian groups fitted an exponential model. However, closer examination of the data in figures 2.5b & 6c reveals that mammalian data has a longer tail than predicted by these models (see also Bonabeau et al. 1999). Although the framework of Gueron & Levin (1995) may well, for particular functional forms for $\psi(x, y)$ and $p(x)$, produce group size distributions similar to those seen in mammalian herbivores, the details of these forms have yet to be established. One candidate for appropriate joining functions is preferential attachment. Preferential attachment is where the probability of an individual joining a group increases with group size. Box 2.C gives an example of a model where the probability of an individual joining a group is a linearly increasing function of its size, while the probability of a group splitting is independent of group size. This particular model generates a power law distribution of group sizes with an exponent of approximately 2.5. In this model I assumed that the population size remains constant. This assumption is consistent with the dynamics of mobile animal groups where total population size usually changes more slowly than the rate at which individuals leave or join groups. While the mathematical properties of constant population models have not been extensively investigated, analysis of preferential attachment models in which the population continues to grow suggests that, depending on the details of the rules for attachment, power laws with exponents of greater than or equal to two can be generated (again see Box 2.C).

Box 2.C Preferential attachment model

Price (1976) proposed a preferential attachment model for scientific citations. In the model, papers are written one after another with no overlap or delay in publication time and each paper cites b previous papers. When each new paper is published, the probability that a currently existing paper i is cited by this new paper is proportional to the number of times, n_i, that the existing has already been cited. In particular, the probability it is cited is

$$\frac{(n_i + c)}{\sum_{j=1}^{m}(n_j + c)} \quad (2.C.1)$$

where m is the total number of papers and c is a constant. This model is known as preferential attachment since the probability of attachment increases with the number of previous attachments (i.e. citations).

We let p_n be the probability that under this system a paper is cited n times. Newman(2005), following a method first developed by Simon (1955), shows that the tail of the distribution of citations is according to a power law, i.e. for large $k_{p_k} \sim k^{\alpha}$, with

$$\alpha = 2 + \frac{b}{c} \quad (2.C.2)$$

Empirically, we see that paper citations are distributed with a power law with slope 3.04 (figure 2.8b). Were the model to fit the data we would thus predict that $c \approx b$. In general, appropriate choice of c and b can produce a power law with any slope greater than 2.

The above model applies in cases where the population continues to grow, as it clearly does with scientific papers. In modelling animal populations that do not change in total population, as in Niwa's model, that space is divided into s sites on which a total of m individuals are initially randomly distributed. The n_i individuals on site i are said to constitute a group. In the spirit of preferential attachment, on each time step we choose a site i at random and remove all individuals, modeling perhaps a disturbance by a predator. We then redistribute them between the sites according to equation 2.C.1. Figure 2.7c shows a time series of the size of the group at the randomly chosen site and figure 2.7d shows the distribution of these group sizes on a log-log plot. This simulation also appears to give a power law distribution.

Many of the distributions associated with human behaviour exhibit power laws with exponents greater than or equal to two. A now classic example is the growth and connection of websites on the World Wide Web. The frequency of the number of connections to websites follows a power law with a slope 2.1 over four orders of magnitude (Barabasi et al., 1999; Barabasi & Albert, 1999). Figure 2.8 shows a large number of examples of distributions that have been claimed to follow power laws. Newman (2005), who produced this figure, emphasises that it is difficult to confirm that these data really do follow a single power law rather than multiply overlaid power law or non-power law distributions. Furthemore, since power laws often only hold in the tail of a distribution, a somewhat arbitrary cut-off point has to be selected above which the exponent a is estimated. These technical limitations do not substantially detract from the ubiquity of power laws (Ball, 2004; Buchanan, 2000). Across many different types of systems, not only those associated with humans but also in the physical and biological world, power laws provide a good fit to the distribution of events occurring in these systems.

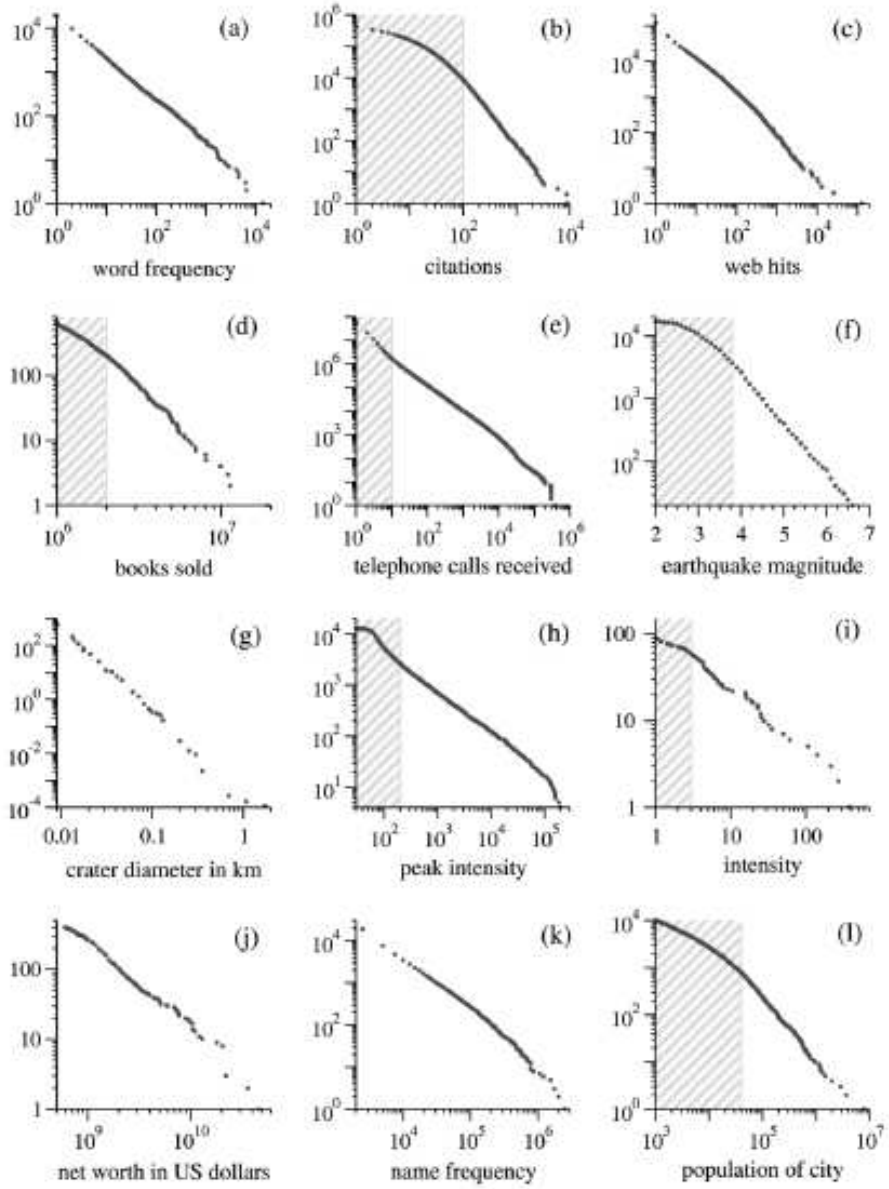


Figure 2.8. Distributions or 'rank/frequency plots' of twelve quantities reputed to follow power laws (reproduced from Newman 2005, figure 3.3). Data in the shaded regions were excluded from the calculation of the estimated power law exponents, α. (a) Numbers of occurrences of words in the novel Moby Dick by Hermann Melville. $\alpha = 2.20$. (b) Numbers of citations to scientific papers published in 1981, from time of publication until June 1997. $\alpha = 3.04$. (c) Numbers of hits on web sites by 60000 users of the America Online Internet service for the day of 1December 1997. $\alpha = 2.40$. (d) Numbers of copies of bestselling books sold in the US between 1895 and 1965. $\alpha = 3.51$. (e) Number of calls received by AT&T telephone customers in the US for a single day. $\alpha = 2.22$. (f) Magnitude of earthquakes in California between January 1910 and May 1992. Magnitude is proportional to the logarithm of the maximum amplitude of the earthquake, and hence the distribution obeys a power law even though the horizontal axis is linear. $\alpha = 3.04$. (g) Diameter of craters on the moon. Vertical axis is measured per square kilometre. $\alpha = 3.14$ (h) Peak gamma-ray intensity of solar flares in counts per second, measured from Earth orbit between February 1980 and November 1989. $\alpha = 1.83$ (i) Intensity of wars from 1816 to 1980, measured as battle deaths per 10000 of the population of the participating countries. $\alpha = 1.80$ (j) Aggregate net worth in dollars of the richest individuals in the US in October 2003. $\alpha = 2.09$ (k) Frequency of occurrence of family names in the US in the year 1990. $\alpha = 1.94$ (l) Populations of US cities in the year 2000. $\alpha = 2.30$

Is the ubiquity of power laws a consequence of preferential attachment mechanisms? The key question is whether the individuals that contribute to systems with power law distributions behave in a way consistent with preferential attachment. We can see how this could be the case with, for example, the growth of the internet or citations of scientific papers, where the probability that we link to a web site or cite a particular paper increases with the number of previous links or citations.

Let's consider preferential attachment applied to scientific citations. Assuming that the constant c in equation 2.C.1 is equal to the mean number of citations per paper then using equation 2.C.2 we recover the power law with exponent approximately 3 seen in the science citation data (figure 2.8). In the model, initial citations are chosen almost entirely at random then further citations are made according to how many previous citations have been made, rather than on the basis of anything written in the papers. We are led to the rather disturbing conclusion that citations may be due entirely to amplifications of initially random decisions on the part of scientists and are independent of the supposed quality of the papers.

While not ruling out the above model of scientific citations it should be pointed out that it is by no means a unique explanation. For example, although the famous bell-shaped or Normal curve is an accurate description of the empirical distribution of IQ near to the mean of the distribution, the tails of this distribution are much wider than predicted by the Normal distribution (Burt, 1963). In general, large deviations in distributions are often better characterised by power laws than the Normal approximation (Sornette, 2004). Let's assume that the quality of papers is proportional to author IQ and scientists working in academia come from the upper tail of the IQ distribution (I do realise the limitations of this assumption). If papers are cited in proportion to their quality then the distribution of citations will simply reflect the power law distribution of the IQ of their authors. Likewise, the links to webpages might be proportional to the intelligence of their designer or the funds possessed by their owner (Figure 2.8j).

Neither preferential attachment nor extreme IQs provide entirely satisfactory mechanistic explanations of the power laws arising in figure 2.8. Indeed, there are at least a dozen distinct mathematical models-from self-organised criticality (Bak, 1996) to highly optimised tolerance (Carlson & Doyle, 2002; Doyle & Carlson, 2000)-in which power laws can be derived (Sornette, 2004; Newman, 2005). Even the causes of power law distributions in physical systems such as meteorite sizes and earthquakes have no generally accepted explanations. In themselves, power laws provide a very weak predictor of the mechanisms which generate them. We should not however be overly discouraged by these observations. Each of the mechanisms for generating power laws has its own set of assumptions, which are experimentally testable. Further experiments can be performed to test the various models against each other.

Bearing in mind our general caution about power laws, we can now begin to think about how to apply our models to explaining the group size distribution of, for example, mammalian herbivores. The example model I give in Box 2.C probably does not encompass the behavioural rules whereby buffalo groups join and split. Furthermore, the model also gives a significantly larger exponent than $\alpha \approx 1.04$ estimated from the data in Figure 2.5c. However, the preferential attachment model incorporates realistic behavioural rules into grouping models: individuals prefer to join larger groups which are then split by random disturbance. With further refinement, this model may begin to capture empirically measured behaviour of real animals. The possibility of including behavioural rules whereby individuals attempt to maximise some variable, in this case group size, brings me back to the functional models with which I began this chapter. Power law distributed group sizes and the instability of the optimal group size become complementary ideas. Preferential attachment is the mechanism by which individuals are more likely to attach themselves to larger groups. The functional reason for this strategy follows from the advantage of being in a larger group, even if that group is larger than the optimal size.

2.7 Group size and population density

Niwa's and Sibly's models give different predictions about how group size changes with population density. While keeping the same basic rules for merging and splitting, Niwa (2004) showed that, for a variety of individual based models of schooling, the mean group size experienced by an individual, $\langle N \rangle_p$, was proportional to the population density. Thus Niwa's model predicts that mean group size will strictly increase with population density. Sibly's stable group size model (Box 2.A) predicts that, provided the total population is larger than the Sibly group size, group size will remain constant as population density increases. Under this model, increases in population density will lead to further groups being created, of stable group size somewhere between the optimal and Sibly group size. Under Niwa's model we also expect increases in group number with population density, but this would be less pronounced than under Sibly's model.

Laboratory experiments on killifish support the predictions of Niwa's model (Hensor et al., 2005). Both group size and group number increased with population density. There was no indication of the modal group size leveling off at a particular 'stable' number and it appears that the distribution of group sizes had a large variance. Hensor et al. (2005) developed an individual based model, based on mechanistic principles of local individual attraction, which gave a very good match to the experimental data.

Field experiments on killifish gave qualitatively similar results to the laboratory experiments (Hensor et al., 2005). Both group number and group size increased with population density. Quantitatively, however, results from laboratory and field were very different. The number of groups was much smaller and group

sizes were much larger in the field than in the laboratory, and were no longer consistent with Hensor et al.'s model. The differences between laboratory experiments and fieldwork may be accounted for in terms of environmental heterogeneity. The fish may be attracted to a certain feature of their environment which simply is not present in homogeneous laboratory conditions. Furthermore, Hensor et al. found that fish body size has an important role in determining group size distribution. The failure of models to accurately predict the outcome of field experiments brings me to a final word of warning about the assumptions that underlie the models discussed in this chapter.

2.8 Alternative explanations for grouping

Most of the models discussed in this chapter assume that groups consist of genetically unrelated individuals that have similarly shaped group size fitness functions and live in relatively homogeneous environments. One species for which these assumptions have been explicitly tested are cliff swallows, which are not genetically related, exhibit no relationship between site availability and group size, but do have between individual differences in group size fitness functions (Brown & Brown, 1996). Indeed, Brown & Brown attribute these last differences to much of the between group size variation observed in cliff swallows. In general we can't hope that these assumptions hold exactly for all the species we are interested in but we can expect them to be a reasonable approximation of reality.

Particular care should be taken with the assumption of environmental homogeneity. Figure 2.8 shows that many features of the physical world have distributions similar to those seen in animal groups. The sizes of animal groups could then simply be attraction to particular physical features, rather than aggregation in response to other animals. Another possibility is that the distribution of a predator species is simply a reflection of the distribution of prey. For example, a predatory fish might gain greatest fitness foraging alone but due to the clustered distribution of its prey it is found in group size distributions similar to that of its prey.

Giraldeau & Caraco (2000) refer to this type of attraction to resources as a 'dispersion economy' (group size fitness function as in figure 2.2a) while attraction to conspecifics is referred to as an 'aggregation economy' (group size fitness function as in figure 2.2b,c). It is usually straightforward to discern if animals are part of a dispersion economy by testing whether individuals in homogeneous environments are attracted or repelled by conspecifics. More difficult is separating effects of attraction to aspects of the environment from those to other individuals in aggregation economies. If an animal is weakly attracted to a particular environmental feature then this weak attraction can be amplified as others copy the choices made by others. One experimental approach are binary choice tests where individuals are presented with two identical environments (Ame et al., 2004; Goss et al., 1989). I will discuss such tests in more detail in the next chapter.

2.9 Linking mechanistic and functional approaches

There is less contradiction between mechanistic models discussed in the second half of this chapter and the functional models than is sometimes supposed. All mechanistic models make implicit assumptions about the group size preferred by individuals in groups. The rules of the models mean that individuals experience a typical group size. For example, in Niwa's model the group size experienced by an individual is $< N >_p \propto 1/p$ and this can be controlled by the individuals by adjusting the rate at which they split, i.e. changing p. What is not investigated in these models is how an individual can adjust its probability of leaving a group in order to increase its own fitness. Indeed, it is usually the probability of a group joining another group which is used in these models, rather than the probability of an individual leaving or joining a group.

Surprisingly, no-one has investigated fission and fusion models within the context of optimising group size. This is unfortunate since basic fission and fusion or joining and leaving rates can be empirically measured and these models could be used to make predictions about what animals are trying to optimise. The approach of Gueron & Levin (1995) would be a good starting point, but the symmetrical fusion and fission used in their model is counter-intuitive. For example, in their model large groups are simultaneously more likely to split and to join other groups. These assumptions are particularly strange in the light of optimal group size theory, where we might expect merging to increase below optimal group size and splitting to increase above optimal group size, and visa versa.

An interesting question is the circumstances under which individuals following a simple set of leaving and joining rules will reach a group size distribution with a mean or mode close to the optimal group size. Beauchamp & Fernandez-Juricic (2005) have made a start on this question. They assumed that individuals decide to leave resource sites on the basis of an estimate of their food intake at that site (Bernstein et al., 1988). The food intake is then a function that first increases but later decreases with group size (e.g figure

2.2b). Using this model they showed that groups formed with a modal size near to that of the optimal group size and much lower than the Sibly group size. Furthermore the distribution of group sizes had a large variance consistent with empirical data.

More work is needed in understanding how and why groups of unrelated individuals form. Indeed, it is quite surprising how little this basic problem of collective behaviour has been studied either theoretically or experimentally. In comparison to aspects of how individuals act once established in groups, the process by which they have formed has received less attention. This disparity may be due to the fact that without understanding aspects such as information transfer, decision-making and synchronisation we cannot discern the benefits and costs of grouping. The models presented in this chapter, and particularly the work of Niwa, should however encourage us that it is possible to make predictions about individuals coming together without knowing the details of what animals do once the group has formed.

References

[1] Ame, J. M., Rivault, C. & Deneubourg, J. L. 2004. Cockroach aggregation based on strain odour recognition. Animal Behaviour, 68, 793-801.

[2] Aviles, L. & Tufino, P. 1998. Colony size and individual fitness in the social spider Anelosimus eximius. American Naturalist, 152, 403-418.

[3] Bak, P. 1996. How Nature Works: The Science of Self-Organized Criticality. Copernicus Books. Ball, P. 2004. Critical Mass: How one thing leads to another.

[4] Heinemann/Farrar. Barabasi, A. L. & Albert, R. 1999. Emergence of scaling in random networks. Science, 286, 509-512.

[5] Barabasi, A. L., Albert, R. & Jeong, H. 1999. Mean-field theory for scale-free random networks. Physica A, 272, 173-187.

[6] Beauchamp, G. & Fernandez-Juricic, E. 2005. The group-size paradox: effects of learning and patch departure rules. Behavioral Ecology, 16, 352-357.

[7] Bernstein, C., Kacelnik, A. & Krebs, J. R. 1988. Individual Decisions And The Distribution Of Predators In A Patchy Environment. Journal Of Animal Ecology, 57, 1007-1026.

[8] Biro, D., Sumpter, D. J. T., Meade, J. & Guilford, T. 2006. From compromise to leadership in pigeon homing. Current Biology, 16, 2123-2128.

[9] Bonabeau, E. & Dagorn, L. 1995. Possible Universality In The Size Distribution Of Fish Schools. Physical Review E, 51, R5220-R5223.

[10] Bonabeau, E., Dagorn, L. & Freon, P. 1999. Scaling in animal group-size distributions. Proceedings Of The National Academy Of Sciences Of The United States Of America, 96, 4472-4477.

[11] Brown, C. R. & Brown, M. B. 1986. Ectoparasitism as a cost of coloniality in cliff swallows (Hirundo pyrrhonota). Ecology, 67, 1206-1218.

[12] Brown, C. R. & Brown, M. B. 1996. Coloniality in the Cliff Swallow. Chicago: The University of Chicago Press.

[13] Buchanan, M. 2000. Ubiquity: The New Science That Is Changing the World. London: Phoenix.

[14] Burt, C. 1963. Is intelligence distributed normally? British Journal of Mathematical & Statistical Psychology, 170-190.

[15] Caraco, T. 1979a. Time Budgeting And Group-Size - Test Of Theory. Ecology, 60, 618-627.

[16] Caraco, T. 1979b. Time Budgeting And Group-Size - Theory. Ecology, 60, 611-617.

[17] Caraco, T., Martindale, S. & Pulliam, H. R. 1980. Avian Flocking In The Presence Of A Predator. Nature, 285, 400-401.

[18] Carlson, J. M. & Doyle, J. 2002. Complexity and robustness. Proceedings of the National Academy of Sciences of the United States of America, 99, 2538-2545.

[19] Clark, C. W. & Mangel, M. 1986. The evolutionary advantages of group foraging. Theoretical Population Biology, 45-75.

[20] Collett, M., Despland, E., Simpson, S. J. & Krakauer, D. C. 1998. Spatial scales of desert locust gregarization. Proceedings Of The National Academy Of Sciences Of The United States Of America, 95, 13052-13055.

[21] Despland, E., Rosenberg, J. & Simpson, S. J. 2004. Landscape structure and locust swarming: a satellite's eye view. Ecography, 27, 381-391.

[22] Doyle, J. & Carlson, J. M. 2000. Power laws, highly optimized tolerance, and generalized source coding. Physical Review Letters, 84, 5656-5659.

[23] Fernandez-Juricic, E., Siller, S. & Kacelnik, A. 2004. Flock density, social foraging, and scanning: an experiment with starlings. Behavioural ecology, 15, 371-379.

[24] Gerard, J. F., Bideau, E., Maublanc, M. L., Loisel, P. & Marchal, C. 2002. Herd size in large herbivores: Encoded in the individual or emergent? Biological Bulletin, 202, 275-282.

[25] Giraldeau, L. A. 1988. The stable group and determinants of foraging group size In: The ecology of social beahviour (Ed. by Slobodchikoff, C. N.). New York: Academeic press.

[26] Giraldeau, L. A. 2000. Social foraging theory. Princeton, New Jersey: Princeton University Press.

[27] Giraldeau, L. A. & Gillis, D. 1985. Optimal group size can be stable: A reply to Sibly. Animal Behaviour, 666-667.

[28] Goss, S., Aron, S., Deneubourg, J. L. & Pasteels, J. M. 1989. Self-Organized Shortcuts in the Argentine Ant. Naturwissenschaften, 76, 579-581.

[29] Gueron, S. 1998. The steady-state distributions of coagulation-fragmentation processes. Journal Of Mathematical Biology, 37, 1-27.

[30] Gueron, S. & Levin, S. A. 1995. The Dynamics Of Group Formation. Mathematical Biosciences, 128, 243-264.

[31] Hensor, E., Couzin, I. D., James, R. & Krause, J. 2005. Modelling density-dependent fish shoal distributions in the laboratory and field. Oikos, 110, 344-352.

[32] Krause, J. & Ruxton, G. D. 2002. Living in groups. Oxford ; New York: Oxford University Press.

[33] Krebs, J. R. & Davis, N. B. 1993. An introduction to behavioural ecology. Oxford: Blackwell Science.

[34] Nee, S., Colegrave, N., West, S. A. & Grafen, A. 2005. The illusion of invariant quantities in life histories. Science, 309, 1236-1239.

[35] Newman, M. E. J. 2005. Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46, 323-351.

[36] Niwa, H. S. 1998. School size statistics of fish. Journal of Theoretical Biology, 195, 351-361.

[37] Niwa, H. S. 2003. Power-law versus exponential distributions of animal group sizes. Journal of Theoretical Biology, 224, 451-457.

[38] Niwa, H. S. 2004. Space-irrelevant scaling law for fish school sizes. Journal of Theoretical Biology, 228, 347-357.

[39] Okubo, A. 1986. Dynamical aspects of animal grouping. . Advances in Biophysics, 22, 1-94.

[40] Parrish, J. K. 1989. Reexamining The Selfish Herd - Are Central Fish Safer. Animal Behaviour, 38, 1048-1053.

[41] Price, D. J. D. 1976. General Theory of Bibliometric and Other Cumulative Advantage Processes. Journal of the American Society for Information Science, 27, 292-306.

[42] Pride, E. 2005. Optimal group size and seasonal stress in ring-tailed lemurs (Lemur catta). Behavioral Ecology, 16, 550-560.

[43] Reed, W. J. 2001. The Pareto, Zipf and other power laws. Economics Letters, 74, 15-19.

[44] Richmond, P. 2001. Power law distributions and dynamic behaviour of stock markets. European Physical Journal B, 20, 523-526.

[45] Sibly, R. M. 1983. Optimal group size is unstable. Animal Behaviour, 31, 947-948.

[46] Simon, H. A. 1955. On a Class of Skew Distribution Functions. Biometrika, 42, 425-440.

[47] Sinclair, A. R. E. 1977. The American Buffalo. Chicago: University of Chicago Press.

[48] Sjoberg, M., Albrectsen, B. & Hjalten, J. 2000. Truncated power laws: a tool for understanding aggregation patterns in animals? Ecology Letters, 3, 90-94.

[49] Sornette, D. 2004. Critical Phenomena in Natural Sciences. Springer Verlag.

[50] Takayasu, H. 1989. Steady-State Distribution Of Generalized Aggregation System With Injection. Physical Review Letters, 63, 2563-2565.

[51] Takayasu, H., Nishikawa, I. & Tasaki, H. 1988. Power-Law Mass-Distribution Of Aggregation Systems With Injection. Physical Review A, 37, 3110-3117.

[52] Takayasu, H., Takayasu, M., Provata, A. & Huber, G. 1991. Statistical Properties Of Aggregation With Injection. Journal Of Statistical Physics, 65, 725-745.

[53] Treherne, J. E. & Foster, W. A. 1981. Group transmissin of predator avoidance behaviour ina marine insect: the Trafalgar effect. Animal Behaviour, 28.

[54] Weimerskirch, H., Martin, J., Clerquin, Y., Alexandre, P. & Jiraskova, S. 2001. Energy saving in flight formation - Pelicans flying in a 'V' can glide for extended periods using the other birds' air streams. Nature, 413, 697-698.

Collective animal behaviour: moving together

David J. T. Sumpter
Mathematics Department, Uppsala University

These notes are taken from chapter 5 of the book, Collective Animal Behaviour, forthcoming from Princeton Univeristy Press.

Some of the most mesmerizing examples of collective behaviour are seen overhead every day. V-shaped formations of migrating geese, starlings dancing in the evening sky and hungry seagulls swarming over a fish market, are just some of the wide variety of shapes formed by bird flocks. Fish schools also come in many different shapes and sizes: stationary swarms; predator avoiding vacuoles and flash expansions; hourglasses and vortices; highly aligned cruising parabolas, herds and balls. These dynamic spatial patterns often provide the examples that first come in to our heads when we think of animal groups.

While the preceding three chapters described the dynamics of animal groups, they did not explicitly describe the spatial patterns generated by these groups. For example, the decision-making of insects and fish was studied in situations where individuals have only two or a small number of alternative sites to choose between. In models of these phenomena, space is represented as the number of individuals who have taken each of these alternatives. This approach often simplifies our understanding of the underlying dynamics of these groups, but in doing so it can fail to capture the spatial structure that characterizes them. As a simple consequence of the fact these groups move, we need to give careful consideration to how they change position in space as well as time.

The main tool I will use in describing the dynamics of flocking are self-propelled particle (SPP) models (Vicsek et al., 1995; Czirok & Vicsek, 2000; Okubo, 1986). In SPP models 'particles' move in a one, two or three dimensional space. Each particle has a local interaction zone within which they respond to other particles. The exact form of this interaction varies between models but typically, individuals are repulsed by, attracted to, and/or aligned with other individuals within one or more different zones. These models allow us to investigate the conditions under which collective patterns are produced by spatially local interactions.

5.1 Attraction

Before animals can create spatial patterns they must first come together. In chapter 2, I discussed how and why animal groups form without specific reference to spatial structure. A good starting point for explicitly representing space comes from Niwa (2004). His model, which is an extension of a non-spatial model described in chapter 2, describes groups of individuals that are constrained to move on a lattice (see Box 5.A). Each group performs a random walk and when groups meet they merge. Groups split with a fixed probability per time step. Figure 5.1a shows an example of how composition of these groups changes through time and space. Over time groups 'clump' together. Sites containing large groups are usually located near to other sites containing large groups, while sites with few individuals are surrounded by other sites with few individuals. The position of these clumps changes through time as the groups move according to a random walk.

Box 5.A. Niwa's spatial merge and split model

The basic assumptions of this model are the same as in box 2.B. A total of m individuals are initially randomly distributed across s sites, and n_i represents the number of individuals on site i. The key difference in the spatial model is how the groups move. Here we assume that groups move on a d dimensional lattice of discrete sites, such that each site has $2d$ neighbouring sites, e.g. in one dimension each site has neighbours to the left and right and in two dimensions each site has neighbours to the north, east, south and west. The lattice is structured so that individuals moving off, for example, the north edge of the lattice reappear at the south. Thus the lattice is a circle in one dimension and a torus in two dimensions. On each time step, each group either moves to one of the neighbouring sites, each chosen with equal probability $1/2d$, or with probability p the group splits into two groups, one which stays on the same site and the other which moves to a randomly chosen neighbouring site. When a group splits the size of the two components is chosen uniformly at random, so that all group sizes are equally likely. If two groups of size n_i and n_j meet at site k, then they form a new group $n_k = n_i + n_j$. Thus, groups always merge when they meet. The same rule applies if three or more groups meet.

Figure 5.1a shows a simulation of the above model in one dimension ($d = 1$). From an initial distribution where each individual occupies one site, larger groups quickly form. These groups perform a random walk and increase in size as they meet other groups. After 1000 time steps there are around five or six large groups and a number of smaller groups. Figure 1b shows the distribution of group sizes at a randomly chosen site over 100,000 time steps of the simulation. Niwa (2004) went on to show that the distribution of group sizes in these simulations is characterized by exactly the same curve as in his earlier non-spatial model (Box 2.B). By finding the mean group size experienced by an individual it is possible to give an expression for the entire distribution of group sizes.

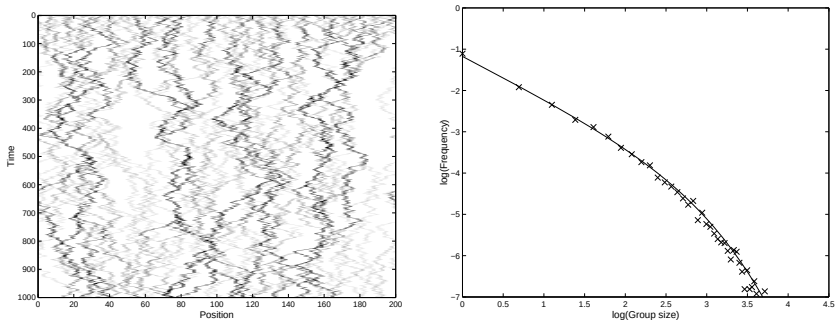


Figure 5.1. Simulation of Niwa's spatial merge and split model. Simulation of model described in Box 5.A with $s = m = 200$ sites/individuals and split probability $p = 0.05$. Initially each site contains a single individual, i.e. a group of size 1. (a) The time evolution of the number of individuals across the sites. Darker shading indicates larger groups at a particular site; white indicates sites containing no individuals. (b) Shows the distribution of the number of individuals in a randomly chosen site over 100,000 simulation time steps. The solid line is equation 2.1 with $\langle N \angle_P$ estimated directly from the simulation.

The unit of description in Niwa's model is the group. The model defines rules for how groups merge and split. The strength of this approach is that it reproduces the empirical distribution of fish school sizes (compare figure 5.1b and figure 2.6). The main limitation of this model is that it does not describe how between-individual interactions produce group dynamics. Establishing such a connection is often the central question in the study of flocking. It is here that self-propelled particle models play an important role.

In the simplest SPP model the only interaction between individual 'particles' is attraction (Box 5.B). Figure 5.2a shows the outcome of a one dimensional SPP model in which individuals are attracted to other individuals within a fixed distance. As in Niwa's model, relatively stable clusters of individuals quickly form. Unlike Niwa's model, larger clusters move slower than solitary individuals. This is because individuals on the edge of the cluster are attracted inwards, resulting in a constant pull towards the centre of the cluster's mass. As clusters increase in size they move less and less, while solitary individuals and smaller groups move and eventually join the clusters (Okubo, 1986). After some time a small number of large stationary clusters form.

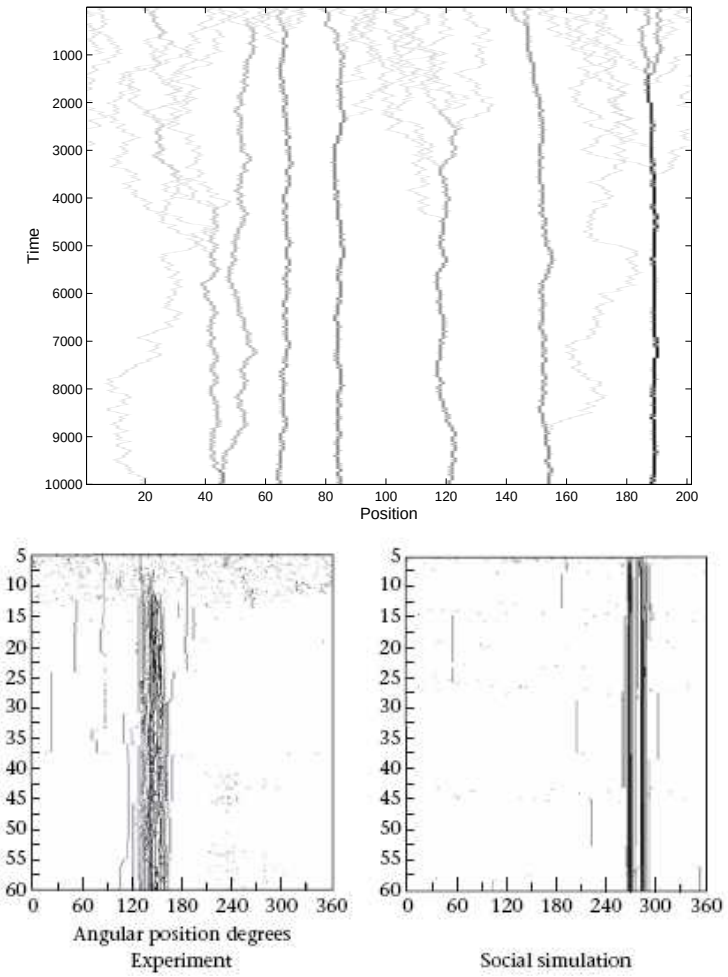


Figure 5.2. Outcome of (a) simple attraction model in Box 5.B compared to (b) experiments on cockroach aggregation and (c) Jeanson et al.'s (2005) detailed individual-based model.

Box 5.B. Self-propelled particle models

The term self-propelled particle (SPP) was introduced by Vicsek et al. (1995), but the idea of building models where individuals interact through zones of repulsion, attraction and alignment had been proposed independently by a number of authors (Aoki, 1982; Reynolds, 1987; Okubo, 1986; Gueron et al., 1996; Helbing & Molnar, 1995). This box presents some of the simplest of these models, including a model of aggregation and Vicsek and co-workers original SPP model of alignment, as well as a more detailed model by Couzin et al. (2002) including repulsion, attraction and alignment.

The general SPP model involves a group of N particles in a d dimensional space. Let the vectors x_i and u_i represent the position and velocity of individual i. Let r represent the interaction radius of the individuals. On each time step t, all individuals update their position and velocity as follows:

$$x_i(t+1) = x_i(t) + v_0 u_i(t+1)$$

$$u_i(t+1) = \alpha u_i(t) + (1-\alpha)s + e,$$

where v_0 is a constant determining a baseline distance which individuals move per time step and a is the inertia of an individual (i.e. its tendency to keep the same direction as on the previous time step). The vectors s and e are determined on each time step for each individual. s is a vector (usually a unit vector) with a direction that depends on the position and velocity of the set of particles, R_i, which are within distance r of individual, excluding itself. e is a random vector incorporating noise into the movement of the individual and may also be a function of the position and velocity of i's neighbours.

Attraction: To model individuals which are attracted to one another the vector s should point towards the average position of an individual's neighbours. In one dimension we can set

$$s = \frac{1}{|R_i|} \sum_{j \in R_i} sign\{x_i(t) - x_j(t)\}.$$

The function $sign\{a\}$ returns 1 if $a > 1$, -1 if $a < 1$, and 0 if $a = 0$. We set e to be a number selected uniformly at random from a range $[-\eta/2, \eta/2]$, where η is a constant. Figure 5.2a shows a simulation of this model on a one-dimensional ring. In this model aggregations form and move more slowly as their size increases.

Alignment: Individuals align by adopting the same direction as their neighbours. In one dimension, Czirok et al. (1999) use

$$s = G\left(\frac{1}{|R_i|} \sum_{j \in R_i} u_j(t)\right),$$

where

$$G(u) = \begin{cases} (u+1)/2 & \text{for } u > 0 \\ (u-1)/2 & \text{for } u < 0 \end{cases},$$

and e as in the attraction model above. The function G ensures that velocities of individuals equilibrate around either -1 or 1. Figure 5.4 gives examples of simulations of this model for different numbers of individuals. As density increases collective motion emerges in the form of a single large group of individuals all going in the same direction.

In two dimensions, Vicsek et al. (1995) lets $s + e$ be a unit vector with direction given by the average angle of the vectors plus some random term. Specifically,

$$s + e = \begin{pmatrix} \cos(\sum_{j \in R_i} \theta_j(t) + \epsilon) \\ \sin(\sum_{j \in R_i} \theta_j(t) + \epsilon) \end{pmatrix}$$

where the θ_j are the directions of i's neighbours and e is chosen uniformly at random from a range $[-\eta/2, \eta/2]$. Unlike the two models above, in Vicsek's model $\alpha = 0$, but the individual i is always included in the set R_i of neighbours. Thus each individual includes itself as a neighbour when averaging velocities. Figure 5.7 gives snapshots of simulations of this model for different magnitudes of noise. Noise plays the opposite role of density: for higher noise motion is less ordered.

Repulsion, attraction, alignment and blind angles: Couzin et al.'s (2002) model involves three zones of interaction: an inner zone of repulsion, an intermediate zone of orientation and an outer zone of attraction (figure 5.8a). The individuals have a blind angle behind them within which they do not respond to individuals which would otherwise be in their orientation or attraction zone. The rule for repulsion is simply that individuals move directly away from nearby individuals. The rules for attraction and alignment are similar to those described for the two simple models. Figure 5.8 investigates a three dimensional version of this model for different sizes of orientation zones. Provided there is a sufficiently large blind angle, the group goes through a transition from swarm to milling torus to a highly aligned group.

Such aggregation dynamics are seen in cockroach groups (Jeanson et al., 2005). Cockroaches interact via antennal contact and are attracted to other cockroaches through physical contact. Thus, relative to the size of their environment, their zone of attraction is small. Jeanson et al. (2005) placed small groups of cockroaches in a circular arena and watched their aggregation behaviour. Since cockroaches are strongly

attracted to walls, most of their movement is constrained to the edge of this arena. In effect, the attraction to the arena edge means that movements of the cockroaches take place in one dimension and the aggregation process can be visualised by plotting the angular position of the cockroaches through time (figure 5.2b). In experiments where cockroaches were initially placed at random within the arena, a cluster quickly formed containing nearly all of the cockroaches. As in the SPP model, cockroaches within the cluster move much less than those outside of it.

Jeanson et al. (2005) developed a parameterised model based on experiments on groups of two to four cockroaches. The principle underlying this model was similar to the simple aggregation SPP model, but it included more detail of walking trajectories in different parts of the two-dimensional arena, probabilities of individuals starting and stopping walking, and the effect of collisions from different directions such as front and behind. The model showed that local contacts alone were sufficient for the rapid aggregation observed in experiments (figure 5.2c).

Whether animals aggregate depends on their environmental context (Krause, 1994; Krause & Ruxton, 2002). Larger groups provide dilution from predator attack and individuals in smaller groups get a larger share of food discoveries (chapter 2). Hoare et al. (2004) found killifish group sizes were significantly smaller in the presence of food odour and larger in the presence of an alarm odour. To explain the behavioural mechanisms that produced these observations they used an SPP model of fish interactions, with terms for repulsion, attraction and alignment. They showed that the observed change in group size distribution could be explained solely by a change in the size of the interaction zone. The distance at which a fish is attracted to another fish decreases in the presence of food and increases in the presence of a predator. This study provides a nice link between mechanism and function: the regulation of group sizes to perceived risk results directly from a change in interaction radius.

The mechanisms underlying spatial aggregation have been studied for a range of species: from midges (Okubo & Chiang, 1974) and bark beetles (Deneubourg et al., 1990) to primates (Hemelrijk, 2000). More than twenty years since its publication, the review by Okubo (1986) still provides the best synthesis of mathematical and empirical aspects of aggregation.

5.2 Alignment

Attraction alone cannot explain the dynamics of most animal flocks. In particular, the aggregative clusters formed by between-individual attraction move slower as cluster size increases (figures 5.1a and 5.2a,c). These observations are in direct contrast to those of fish schools, locust swarms and migratory birds that, while remaining a cohesive group, move rapidly in the same direction. Indeed, it is the rapid propagation of directional information that characterises these groups, and poses the greatest challenge to our understanding of them (Couzin & Krause, 2003). How is it that a bird flock or a fish school can apparently turn in unison such that all members almost simultaneously change direction?

It was the pioneering experimental work by Radakov (1973) that first showed how changes in direction can be rapidly propagated by local interactions alone. He used an artificial stimulus to frighten only a small part of a school of silverside fish. The fish nearest to the stimulus changed direction to face directly away from it. As these fish changed direction they stimulated others nearby, but further away from the artificial stimulus, to also change direction. A "wave of agitation" spread away from the artificial stimulus (figure 5.3). This propagation of directional information was much more rapid than the displacement of the fish. The fish nearest to the stimulus moved less than 5cm in the same time it took every fish within 150cm of the stimulus to change direction to face away from the stimulus. Changes in direction propagated at speeds of up to 11.8 - 15.1 metres per second over distances of between 30 and 300cm.

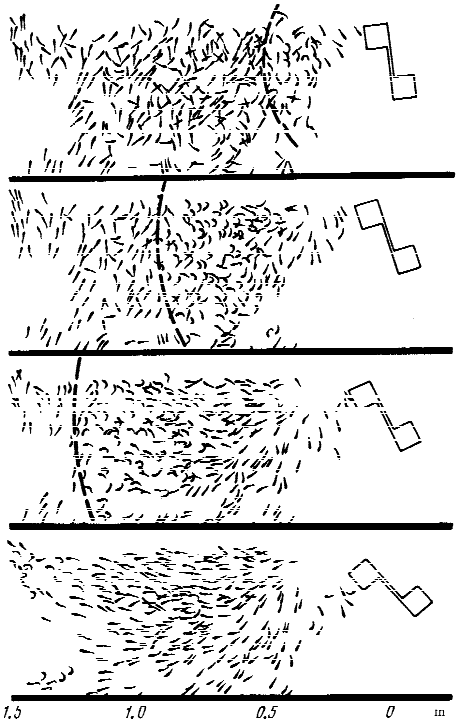


Figure 5.3. Example of Radakov's experiment where fish schools are presented with a fright stimulus. The position of fish was filmed and projected on a wall so that a picture could be made of the position and orientation of the fish. Reproduced from Radakov 1973.

While not directly inspired by Radakov's work, the transfer of directional information was the key ingredient in the self-propelled particle models of Vicsek et al. (1995). In fact, Vicsek's model has only two ingredients determining the direction particles move in: alignment to nearby particles and noise (Box 5.B). Figure 5.4a-c shows examples of these simulations in one dimension for different particle densities. A central prediction of Vicsek's model is that as the density of particles increases, a transition occurs from disordered movement to highly aligned collective motion (Vicsek et al., 1995; Czirok et al., 1999; Czirok et al., 1997). Figure 5.4d-f show how the mean direction, or the degree of alignment, of particles changes through time in a one dimensional version of the model from Box 5.B for three different particle densities. At low densities, the alignment remains close to zero (figure 5.4a,d). At intermediate densities, all particles adopt a common direction for a period of time but this direction switches at random intervals (figure 5.4b,e). At high densities, particles adopt a common direction which persists for a long period of time (figure 5.4c,f). The transition from disorder (random motion) to order (aligned motion) occurs at a critical density, below which alignment is zero and above which absolute alignment increases with group size (Czirok et al., 1999).

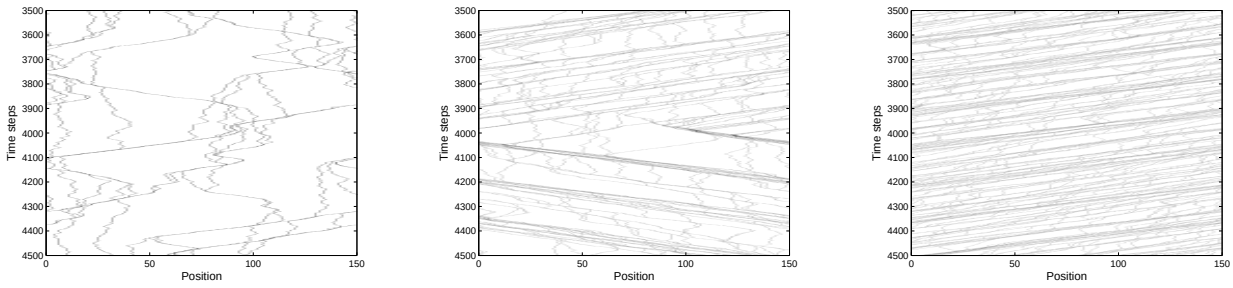


Figure 5.4.

Such a transition from disordered to ordered motion is seen in the collective motion of locusts. Buhl et al. (2006) looked at the alignment of various densities of locusts in an experimental ring-shaped arena. This setup effectively confined the locusts to one dimension and the degree of alignment could be measured as the average direction of movement relative to the centre of the arena. For small populations of locusts in the arena there was a low incidence of alignment among individuals. Where alignment did occur, it did

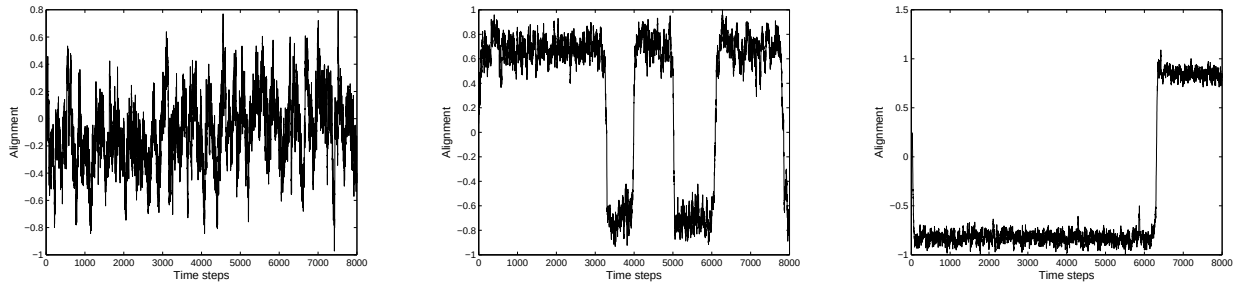


Figure 5.4. Example simulations from one-dimensional SPP models. Simulation of the SPP model of alignment in one dimension. The change in particle density through time for (a)$N = 10$ (b) $N = 50$ and (c) $N = 100$ particles. The alignment at time t is defined as

$$\frac{1}{n}\sum_{i=1}^{n} u_i(t).$$

The average direction. The alignment is given for (d) $N = 10$ (e) $N = 50$ and (f) $N = 100$ particles. Other parameters are $L = 150, r = 1, v = 1, \alpha = 0.66$ and $\eta = 0.8$.

so only after long initial periods of disordered motion (figure 5.5a). Intermediate-sized populations were characterized by long periods of collective rotational motion with rapid spontaneous changes in direction (figure 5.5b). At large arena populations, spontaneous changes in direction did not occur within the time scale of the observations, and the locusts quickly adopted a common and persistent direction (figure 5.5). As predicted by Vicsek's model, alignment of locusts becomes non-zero above a critical density (figure 5.6). The simplicity of Vicsek's SPP model suggests that phase transitions should be a universal feature of moving groups (Buhl et al., 2006). Similar transitions are observed in fish (Becco et al., 2006) and in tissue cells (Szabo et al., 2006).

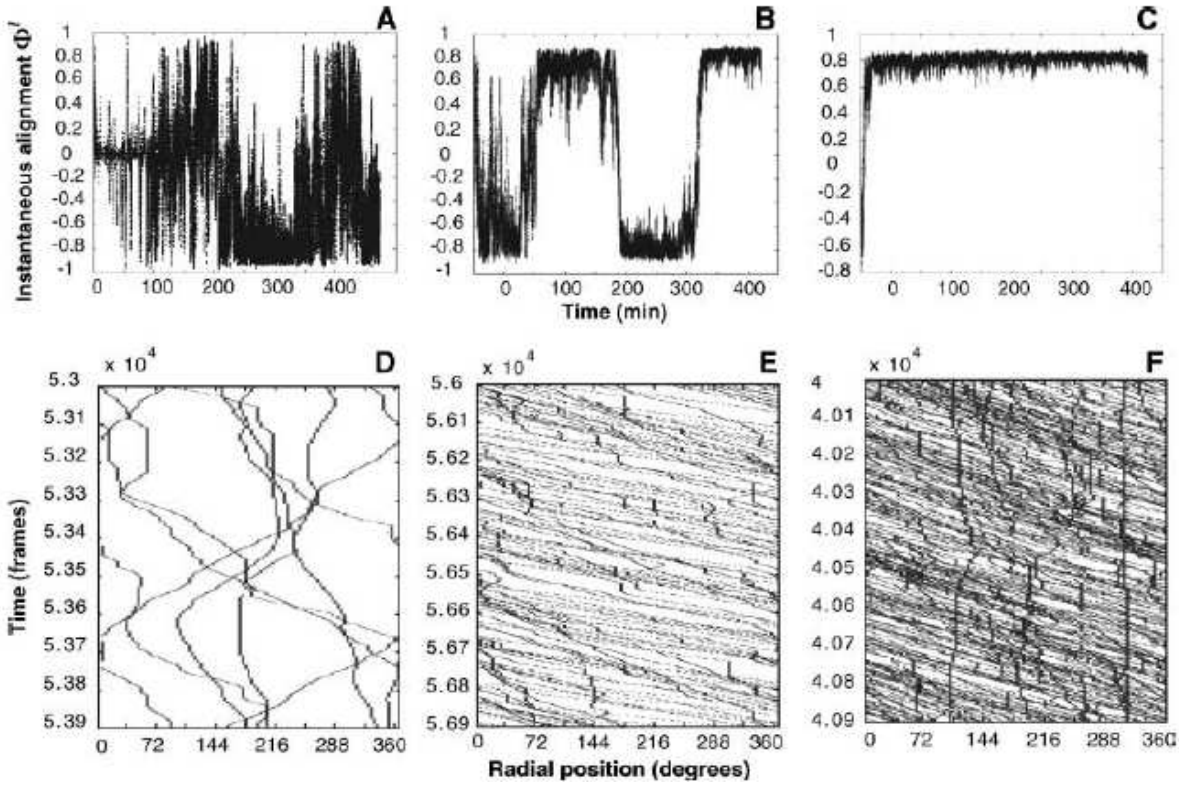


Figure 5.5. Experiments on locusts in a ring. The alignment over the experiment of (a) 7 locusts, (b) 20 locusts and (c) 60 locusts. (d to f) Corresponding samples of time-space plots (3 min), where the x-axis

represents the individuals' angular coordinates relative to the centre of the arena, and the y-axis represents time. Reproduced from Buhl et al (2006).

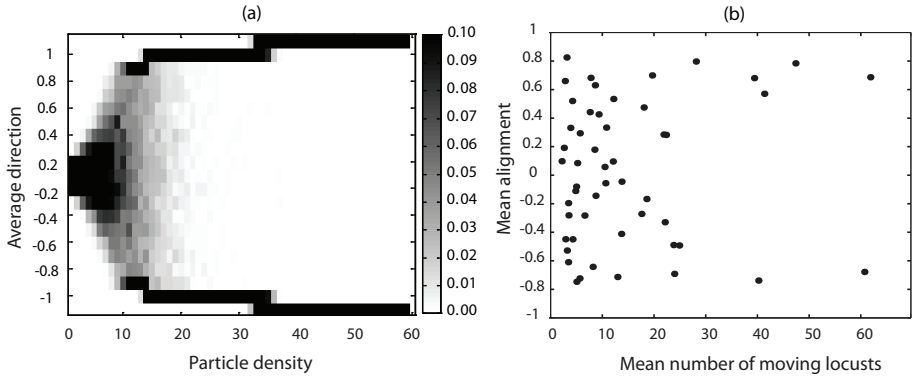


Figure 5.6. Comparison of the mean alignment in the (a) SPP model and (b) the locust data as a function of the number of particles (or locusts). Reproduced from Buhl et al (2006).

When extended to two or three dimensions, Vicsek's model generates spectacular dynamical patterns that are highly reminiscent of the movement of flocks (figure 5.7). Again the two dimensional model undergoes a phase transition where alignment becomes non zero above a critical particle density or below a critical nose level (Vicsek et al. 1995).

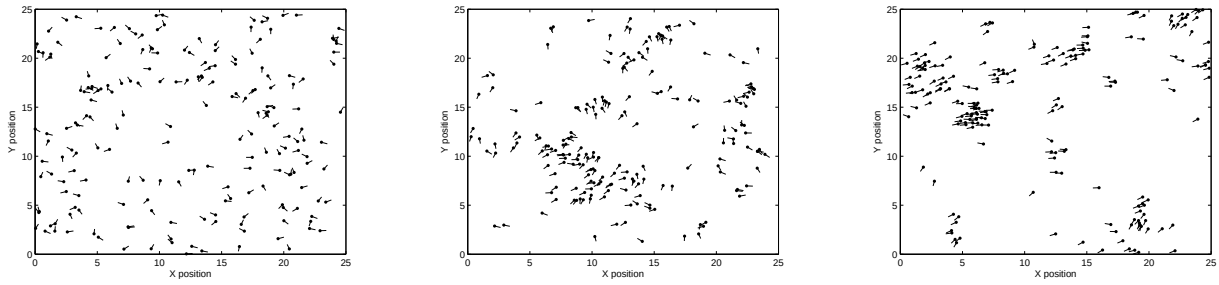


Figure 5.7. Example of patterns from the two-dimensional SPP model with alignment. Model is as described in Box 5.B. Parameters are $n = 200, v_0 = 0.5, L = 25, andr = 1$. The noise is varied between simulations (a) $\eta = 3$, (b) $\eta = 1.5$ and (c) $\eta = 0.5$.

While reproducing many of the characteristics of animal flocks, Vicsek's model is by no means sufficient to explain all aspects of flocking. To start with, it does not contain an attraction term of the type discussed in the previous section. In fish, attraction between individuals has long been viewed as having equal importance to alignment in determining group dynamics (Partridge, 1982). The omission of attraction from Vicsek's model means that a bounded group cannot form. In an SPP model without an attraction term, a large group of particles moving in the same direction spreads out and particles will 'escape' from the back of the group (Gregoire et al., 2003). When confined to a small space this diffusion will not lead to a significant breakup of the group because stragglers are picked up when they meet the large group again, but in an infinite (or large) space the group will eventually break apart.

A cohesive moving group can form if both attraction and alignment terms are included in an SPP model. Gregoire et al. (2003) drew a phase diagram for a two-dimensional SPP model which included terms for attraction, alignment and noise. They found that when attraction was weak relative to alignment, particles behaved as either a disordered or moving 'gas', similar to those seen in the two-dimensional Viscek model (figure 5.7). This gas was characterised by the proportion of particles that were members of the largest group being less than one. When attraction was increased the proportion of particles within the largest group tended to one, and Gregoire et al. classified this state as a liquid 'droplet'. Within this droplet two close together particles diffused away from each other through time while remaining within this large group. Compared to the gas in figure 5.7, in which groups split apart and reform, individuals moved around within the single droplet but did not leave it. As the attraction term was further increased, the liquid turned in to a solid 'crystal' and the particles remain at a fixed position within the crystal through time. Provided

alignment was sufficiently large relative to noise, both liquids and solid exhibited cohesive collective motion where all particles moved as a group in the same direction.

A number of aspects of Gregoire et al.'s model resemble the motion of animal flocks. Moving crystals and droplets both exhibit periods of ballistic flight, where the mean square displacement of the group was proportional to $(\text{time})^2$, i.e. groups fly in a straight line. Furthermore, the lengths of these ballistic flights increased with the size of the group. This is in contrast to the non-moving phases where attraction is dominant, e.g. as in figure 5.2a. In this case, the mean square displacement of the group was proportional to time and the lengths of ballistic flights decreased inversely proportionally to group size. Crystals and droplets both resemble various forms of moving animal groups: crystals look roughly like highly parallel groups of fish or birds, while the droplets possibly resemble flying locust swarms. Particularly interesting is the existence of mesoscopic "hydrodynamical" structures, such as jets, vortices, etc., within droplets (Gregoire et al., 2003). It is this dynamical patterning on a meso-scale within a generally coherent motion on the scale of the entire group that might be said to best characterise the collective motion of many flocking animals. However, the 'zoology' of these meso-scale shapes has not been fully investigated and compared to empirical observations.

5.3 Rules of motion

The attraction and alignment models discussed in the previous sections have not been calibrated against real data of how fish, birds or locusts interact with one another. Instead, the philosophy of these models is to provide as simple as possible model for the interaction of animals that reproduces the key features of flocks. This philosophy is aimed at ensuring that model outcomes are not dependent on some particular biological feature, but reveal universal properties of all flocks. The approach is also to some degree unavoidable. Empirical determination of the detailed interactions of fish or birds is technically difficult. These groups move in two or three dimensions and often come in close contact with each other, making automated or even manual tracking difficult (Hale, 2008).

There are, however, a number of high quality studies of fish interactions, most notable those of Partridge in the early 1980s. Studies of the structure of schools of saithe, cod and herring show that fish maintain a minimum distance between each other, supporting evidence for local repulsion (Partridge et al., 1980). By tracking individual fish, Partridge (1981) established that saithe match their swimming direction and speed to their two nearest neighbours, but probably not to more distant neighbours. Partridge & Pitcher (1980) found that 'blindfolded' saithe continued to match short term changes in velocity of their neighbours using their lateral line (the motion detecting sense organ which runs down fish bodies). Vision was however important in maintaining between neighbour distance, with blind fish having increased nearest neighbour distances. Fish which had their lateral line disabled compensated by changing postion so they could see direction changes by neighbours. In general, the lateral line appears to determine alignment, while vision determines attraction and repulsion.

An impressive step forward in the understanding of both the global structure of groups moving in three dimensions and the behaviour of individuals within these groups is the Starflag project (Cavagna et al., 2008a; Ballerini et al., 2008a; Cavagna et al., 2008b). Using multiple cameras these researchers were able to determine the position of most of the starlings in flocks consisting of thousands of birds. Like fish, the starlings maintain a minimum distance from each other, i.e. have a zone of repulsion (Ballerini et al., 2008a). Starlings are also less likely to have neighbours behind or in front of them than to have neighbours on either side. As distance from a focal bird increases this spatial organisation disappears, so that birds further away from a focal bird are equally likely to be at any angle.

Local spatial structure is not simply a function of distance but rather a function of neighbour number. The nearest neighbour is much more likely to be to the side of than directly in front of or behind a focal bird. This tendency then decreases for the second neighbour then the third neighbour and so on. After the sixth or seventh neighbour the spatial structure vanishes and these neighbours are equally likely to be at any angle relative to the focal bird (Ballerini et al., 2008b). This relationship is less robust when considering only the distance between neighbours. Even when the flock is more tightly packed spatial correlations are seen only between a fixed number of neighbours. The relationship would suggest that instead of interacting with all or some birds within a certain fixed radius, as is assumed in most models, starlings interact with their 6 or 7 nearest neighbours.

5.4 Complex moving patterns

The shapes of bird flocks, fish schools and locust swarms are not limited to groups of aggregated or aligned individuals. Some of these shapes can emerge from simple interactions of repulsion, attraction and alignment alone. For example, Couzin et al. (2002) proposed a model in which individual animals have three zones-repulsion, alignment and attraction-of increasing size, so that individuals are attracted to neighbours over

a larger range than they align, but decrease in priority, so that an individual always moves away from neighbours in the repulsion zone (figure 5.8a). These individuals also have a rear blind zone within which they cannot sense others.

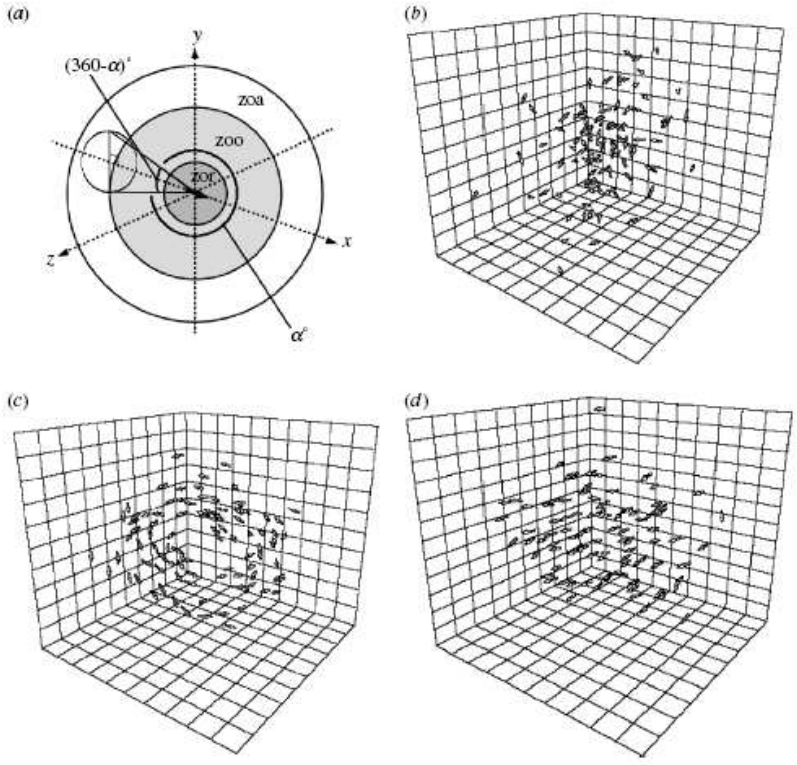


Figure 5.8. Transition from swarm to torus to alignment. (a) Illustration of the rules governing an individual in the fish model. The individual is centred at the origin: zor, zone of repulsion; zoo, zone of orientation; zoa, zone of attraction. The possible 'blind volume' behind an individual is also shown, a, field of perception. Collective behaviours exhibited by the model: (b) swarm, (c) torus and (d) dynamic parallel group.

Keeping the repulsion and attraction radii constant, Couzin found that as the alignment radius increased, individuals would go from a loosely packed stationary swarm (figure 5.8b), to a torus where individuals circle round their centre of mass (figure 5.8c) and, finally, to a parallel group moving in a common direction (figure 5.8d). This transition from milling to torus to departure is typical of the motion of real fish schools. The model shows that these three very different collective patterns self-organise in response to small adjustments to one factor: the radius over which individuals align with each other.

Other patterns seen in animal flocks may be more difficult to produce from models of identical 'memoryless' self-propelled particles interacting in a homogeneous environment. For example, Radakov (1973) reports "feeler" structures in silverside fish during their evening migration away from the shore. A few fish swim away from the group forming a ribbon-like structure as others follow. The leading group then reduces speed and starts feeding, at which point a "neck" builds up as more and more fish are drawn from the main group. In some cases this neck leads the whole group to the new feeding ground, while in others the neck breaks off and a sub -group separates from the main group. Overall, the process gives the impression of the school making a tentative investigation of whether it is worth moving feeding grounds.

Another common pattern in fish schools is the fountain response to the approach of a predator towards a group of prey (Pitcher, 1985). In this response the fish fan out in front of a predator and circle round behind it. Self-propelled particle models can reproduce this type of group response to predators (Iawad, 2001 and see section 5.6). However, Hall et al. (1986) argue that a fountain response can occur simply by each individual prey moving away from the predator while keeping it at the edge of its field of view. Fish have a blind angle of roughly 60°, so by keeping the predator behind them at an angle of 150° the fish are moving away from the predator as rapidly as possible without losing sight of it. This argument appears consistent with experimental data on the response of shoals of juvenile whiting (Hall et al., 1986), but it is not entirely clear whether social interactions may also play a role in creating the fountain effect.

Determining the degree to which simple rules for attraction and alignment capture the shapes produced by real animal groups remains a key problem (Parrish et al., 2002). No detailed statistical comparison has

been made between the motion of and within real flocks and those predicted by SPP models. For example, Uvarov (1977), describes the marching bands of locusts as having a dense front and columns that go through an otherwise diffuse cloud of individuals. These observations have little in common with the shapes arising from, for example, Gregoire et al.'s (2003) model. Similarly, Ballerini et al.'s (2008) observation that starling flocks have a dense boundary and a sparser interior directly contradicts most SPP models, which predict either homogeneous density within a group or a density which decreases with distance from the group's centre. Explaining the emergence of complex moving structures will require greater consideration of the rules adopted by individuals, of how individuals interact with the environment and of between-individual differences.

5.5 Decisions on the move

When navigating, animals in moving groups usually have access to two types of information, their own experience or internal compass information and the direction taken by other group members. A central problem faced by animals travelling in these groups is how to integrate this information, especially when members cannot assess which individuals are best informed. In the context of avian navigation, two alternative schemes have been proposed (Wallraff, 1978). The "many wrongs" hypothesis, which is described in more detail in section 4.3, is that individuals average their preferred direction, leading to a compromise in route choice. The average of these many wrongs should lead to an improvement in navigational performance. Wallraff's alternative to the many wrongs hypothesis is the 'leadership' hypothesis. Under this hypothesis, one or a small number of the animals takes a leading role and the others follow.

Neither the many wrongs nor the leadership hypothesis accounts for how information is transferred between group members through local interactions. Indeed, the many wrongs hypothesis leads to the paradox, discussed in section 4.4, that for information to be transferred some individuals must follow others but at the same time too much following will reduce the success of the averaging. To bypass this limitation, Biro et al. (2006) developed a mechanistic model of navigational conflict between pairs of individuals. In the model (described in Box 5.C), individuals interact according to two hypothesized forces: attraction to its own target position (own information) and attraction to the partner's current position (social information).

Box 5.C. Model of paired navigational decision-making

We consider a dynamic model for decision-making, where two individuals, X and Y, each decide on a real-valued 'position', starting from initial positions $x(0)$ and $y(0)$. These individuals come to a final position as a result of a combination of two forces: predisposition to move toward a target position and local attraction towards the other individual's current position.

Predisposition to target: X, respectively Y, is attracted to a target position with value 0, respectively d. The rate at which an individual moves toward its predisposed choice initially increases with distance from the target, but above a point of maximum attraction the rate decreases. For individual X, we model this rate with the function

$$-x\exp(-x/r_a) \quad \text{(equation 5.C.1)},$$

where x is the current position and r_a is the point at which the attractive force towards the target reaches a maximum. Individuals further from the target than r_a have a weaker attraction towards it due to difficulties in perceiving the target, while individuals nearer than r_a have a decreasing but positive attractive force, modelling an increasing degree of 'comfort' with decreasing distance to the target.

Between-individual attraction: We model this with the function

$$(x-y)\exp(-(\frac{(x-y)}{\sqrt{2}r_b})^2) \quad \text{(equation 5.C.2)},$$

where x and y are the current positions of the two individuals and r_b is the point of maximum attraction to other individuals. Attraction only occurs locally, so that once individuals move out of the range of perception, the rate of attraction quickly decreases. We combine the two forces acting on the individuals to give a differential equation model of how the individuals change position:

$$dx/dt = -x\exp(-x/r_a) - \alpha(x-y)\exp\left(-\left(\frac{(x-y)}{\sqrt{2}r_b}\right)^2\right) \quad \text{(equation 5.C.3)}$$

$$dy/dt = -\beta(d-y)\exp(-(d-y)/r_a) - \alpha(x-y)\exp\left(-\left(\frac{(x-y)}{\sqrt{2}r_b}\right)^2\right) \quad \text{(equation 5.C.4)}$$

The parameter α determines the ratio of the maximum between-individual attraction over the maximum attraction to the target. β determines the ratio $(Y:X)$ of the strength of the individuals' attraction to their targets. Figure 5.9a shows the equilibrium solutions to the model equations as a function of the distance d between the individuals' targets.

Figure 5.9a shows, for the model in Box 5.C, the effect of varying the distance between the individuals' targets, d, on the final decision reached. The model predicts that at small distances between established routes, individuals average, with their position equilibrating at $d/2$. At a critical between-route distance, of approximately twice the range at which individuals are maximally attracted to their established routes, a bifurcation occurs. For d larger than this critical value, both individuals move closer to that of one of the individuals. A third possible outcome is splitting, where each individual moves exclusively towards its own target. Such outcomes occur over a wide range of d but always result from initial differences in the individuals' positions.

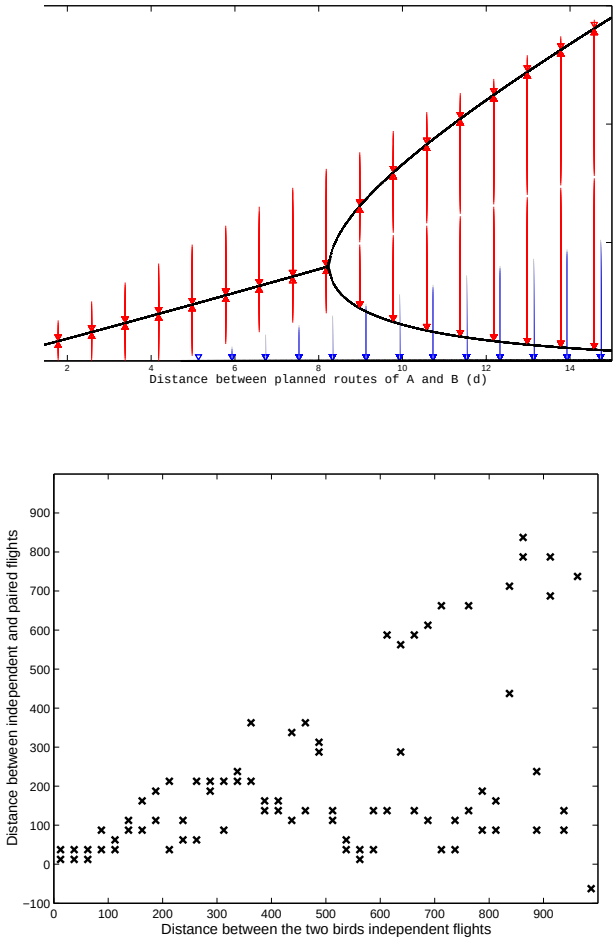


Figure 5.9. Outcome of decision-making in pairs. (a) Prediction model in Box 5C. Equilibrium solutions of equations 5.C.3 and 5.C.4 as a function of the distance between the individuals' targets, d. The arrows show how different initial positions of bird X lead to different equilibriums. The initial position of bird Y is always $d/2$. The parameter values $r_a = 400, rb = 80$ were chosen to reflect the perception ranges of real pigeons. The other parameters $\alpha = 1$ and $\beta = 1$ assume no intrinsic difference between the birds (b) Outcome of pigeon experiments. Point by point distances between each bird's established route and its taken route when in a pair are made into a histogram. The largest and the second largest modes of the data are then plotted.

While the model in Box 5.C provides an abstract representation of navigational decision-making, it was designed specifically with the behaviour of homing pigeons in mind. Predisposition to a target models the phenomenon of route recapitulation and route loyalty by homing pigeons and between-individual attraction models social cohesion between birds. We tested the model's predictions against data we collected on homing pigeons (Biro et al., 2006). We first allowed homing pigeons to each establish their own route home from a release site. Once individuals had learnt their own routes they were released in pairs. In these paired releases instances of many wrongs compromise and of leadership were observed, even within a single journey of a single pair of birds.

In order to test how the distance between the birds' 'target' routes affected the outcome of their paired flight, we looked point-by-point through the whole flight at how the distance between the birds' independent flights affected the distance between their routes. Figure 5.9b shows the largest and second largest modes of distances between routes taken by individuals during their paired flight and the immediately preceding

single (established) route as a function of distance between the birds' established routes at the corresponding point of the journey. We see a similar bifurcation in this data as we see in the model prediction (figure 5.9a). As the distance between the birds' targets increases a bifurcation occurs from compromise to leadership.

Our model is limited because it deals with only two individuals and abstracts away possibly important aspects of spatial interactions. Couzin et al. (2005) proposed an SPP model where individual particles move in a two dimensional space according to rules of attraction, alignment and repulsion. In this model a large group of 'uninformed' individuals interacts with two small groups of informed individuals which each move toward different targets. As the angle between the targets increases there is a bifurcation where the group goes from taking a direction intermediate to the two small leading groups to taking the direction preferred by one of the two groups.

5.6 Leading the swarm

An interesting prediction of the Couzin et al. (2005) model is that a small number of informed individuals can lead a large group. In these simulations groups of 200 uninformed individuals were almost always successfully led to a target by groups of less than 10 leaders. Thus observations of large numbers of birds, fish or insects moving in the same direction do not imply that even a majority of individuals know where they are going or even know which individuals know where they are going. The Couzin et al. (2005) model thus suggests a 'subtle guide' mechanism: a largely uninformed group can be led by a small group of informed 'leaders' even when the identity of the leaders is unknown.

One of the most impressive examples of a large group of uninformed individuals being led by a small group is the flight of honey bee swarms from their temporary bivouac on a tree branch to a new nest site (see section 9.3). Up to around 10,000 bees of which only 2 or 3% are informed of the location of the nest site fly as a single swarm to the site. How does such a small group lead such a large group to a small nest site? Lindauer (1955) hypothesised that the informed individuals repeatedly 'streak' through the swarm in order to inform the other bees of the direction of the nest. Janson et al. (2005) formalised this hypothesis in an SPP model and showed that 150 'streaker bees' could lead a swarm of 3,000 uninformed bees, and these swarms could avoid obstacles in their path without splitting. While streaking might help guide a swarm, the 'subtle guide' hypothesis presented above suggests that streaking is not a requirement for a small number of individuals to lead a large swarm. A further alternative to the 'subtle guide' or 'streaker bee' hypotheses is a 'vapour trail', where the informed bees move to the front of the swarm and release a chemical pheromone creating a gradient which the other bees follow (Avitabile et al., 1975).

Beekman et al. (2006) tested the 'vapour trail' hypothesis by sealing, in the bees, the glands which release pheromone and comparing the flight of sealed gland colonies with control colonies. Gland sealing had no significant effect on the flight speed of the swarm nor on the time it took the swarm to reach a nest box, contradicting hypotheses based on pheromones. Beekman et al. (2006) noted that some bees in the swarm were moving at maximum speed (9-10m/s) while the swarm as a whole moved at only 2-3 m/s, providing evidence for the 'streaker bee' hypothesis. Schultz et al. (2008) provided stronger evidence of streaking by filming a swarm from below. They found that bees in a top portion of the swarm flew quickly in the direction of the nest site and these fast moving bees were observed at the front, middle and back of the swarm. However, while it appears clear that some bees streak along the top of the swarm and then return through it at slower speeds, there is still no direct link between these fast flying bees and the scouts.

5.7 Evolution of flocking

Hamilton (1971) and Vine (1971) were the first researchers to look at how the geometry of an animal group might be shaped by natural selection. They both proposed 'selfish herd' models in which individuals in the group are motivated to move in to the centre of the group by the risk of predation. In Hamilton's model, individuals live on a one-dimensional lattice and follow the rule: if the site an individual occupies has a larger population than those to the left and right then it stays there, otherwise it moves to the neighbouring site that is occupied by the largest number of other individuals. In contrast to the mechanistic model of aggregation described in Box 5.A, Hamilton's model is motivated by functional considerations. However, the outcome of both models is similar: tightly packed clumps of individuals emerge (as they do in figure 5.2a). Vine and Hamilton both expand on this initial model and find similar results: tight aggregations are a consequence of selfish individuals' attempt to use other individuals as cover.

The geometrical predictions of selfish herd models hold for a wide range of species that form stationary groups (Krause, 1994; Krause & Ruxton, 2002; Quinn & Cresswell, 2006; Rayor & Uetz, 1990). Individuals near the centre of these groups are less likely to be attacked than those on the edge. Several studies have revealed that when there is a predation risk, fish move closer together (Tien et al., 2004; Krause, 1993). On the other hand, Focardi & Pecchioli (2005) found that the foraging success of deer increased with distance

from the centre of the group. There is thus a trade-off between increased food intake on the outside of the group and increased safety in the centre. We might then expect position in a group to be determined by nutritional state, with well fed individuals near the centre and hungry individuals on the outside.

In moving groups it is less clear how the position in a group relates to safety from predation. Parrish (1989) showed in laboratory experiments that while grouping silverside fish are attacked less often by sea bass than stragglers which have recently departed from the group, if the group is attacked it is the fish in the centre that are the subject of these attacks. Parrish suggested that this is because the predators attack the centre of the group, which then splits in two leaving central individuals exposed. This interpretation is supported by simulations of SPP models (Inada & Kawachi, 2002). Parrish's study is limited however by the fact that very few attacks by the predators were successful: only five group members were killed throughout all experiments, three of which were in the centre and two on the periphery.

The complex dynamic patterns generated by flocking animals should convince us that a selfish desire to be shielded by others is not the only evolutionary force that has shaped them. Group membership may also allow individuals to gain information about the location of food (Pitcher et al., 1982) and of predators (Treherne & Foster, 1981), to benefit in terms of energetic efficiency (Weimerskirch et al., 2001) and even to hunt co-operatively (Partridge et al., 1983). A problem however is disentangling functional and mechanistic explanations for dynamic patterns. Many patterns may be a consequence of the interactions between individuals and have little or no adaptive significance (Parrish et al. 2002). For example, the transition from disorder to order in locust marching appears to be a fundamental property of SPP models, suggesting that rather than resulting from the fine tuning of natural selection it is simply a necessary aspect of all grouping animals (Grunbaum, 2006). Similarly, it would be wrong to conclude that a moving fish torus has evolved to signal between group members that departure is imminent, but rather it could be an unavoidable consequence of all members increasing their tendency to align with each other (Couzin & Krause, 2003).

Behaviours which produce flocking patterns are in some cases themselves subject to natural selection. For example, one intrinsic property of SPP models is dynamic instability. Such instability was seen at intermediate densities in experiments on locusts, with changes in direction rapidly spreading through the entire group (figure 5.5e). If a small number of locusts spontaneously change direction, the others rapidly change their direction in response. This spread of directional information is reminiscent of Radakov's (1973) experiments on fish. Information about the presence of a stimulus is rapidly transmitted through the entire group.

Several modelling studies have investigated how the rules governing the alignment, repulsion and attraction of self-propelled particles might be optimised so as to allow the particles to avoid predation (Inada & Kawachi, 2002; Lee, 2006; Lee et al., 2006; Zheng et al., 2005). In these studies a predator particle that is introduced into the simulation attempts to attack the group of prey particles. Inada & Kawachi (2002) varied the maximum number of neighbouring individuals with which each prey aligned. They showed that if prey aligned with only one nearest neighbour then group movements were uncoordinated in response to a predator, but if they interacted with two or three the group was able to effectively align away from the predator. However, if prey individuals align with larger numbers of neighbours then the group would change direction slowly in response to a predator, because the minority of individuals that had sensed the predator and begun to move away from it would be ´outvoted´ by the uninformed majority that continue in their previous direction. Zheng et al. (2005) obtained similar results to Iwada by changing a different model parameter. They showed that there is an optimal weighting that individuals should put on aligning with other prey relative to orienting away from the predator. By aligning with each other rather than purely away from a predator, the prey avoid costly collisions. The collective outcome is a confusion effect, where the predator repeatedly changes target.

Most modelling studies of predator avoidance have looked at group success, measured in terms of number of group members captured by a predator, as a function of model parameters. From a functional viewpoint, however, the question is how individuals regulate their propensity to align, or their interaction range, or other aspects of their behaviour so as to minimise their own probability of being caught by the predator. While aligning with others may increase the confusion effect for the predator, the best strategy for a focal individual may be to move directly away from the predator. As a result a social parasitism dilemma arises: while co-operating individuals can generate a pattern which optimises group success, a defecting individual surrounded by co-operators can benefit to the greatest degree by not participating in the pattern. The pattern is then not evolutionarily stable (see chapter 10).

Wood et al. (2007) investigated the evolutionary stability of self-propelled particles to predation. They used the same model for particle movements as Couzin et al. (2002) but allowed the particles to evolve their interaction zones in response to predation. The main parameters governing the interaction zones are the relative size of the attraction, R_a and orientation zones, R_o, as well as the angle θ over which the particles

can 'see' their neighbours. The total area over which a particle could monitor its neighbours, i.e. $\theta\pi R_a^2$ was fixed to a constant for all particles. This constraint means that their viewing area is restricted to a local neighbourhood of constant area. On the first generation a population of 80 individuals each with its own values of R_a, R_o, and θ was simulated for a sufficient number of time steps so as to allow a dynamic pattern to form. A predator, which attempted to capture the prey individuals, was then introduced into the simulation. After a fixed number of time steps those surviving individuals, i.e. those which had not been caught by the predator, went on to the next generation and those individuals that were caught were replaced by 'offspring' of the surviving individuals. These offspring were subject to small mutations in the parameter values so that individuals with new values for R_a, R_o, and θ entered into the population.

There was a clear pattern in the evolution of the parameters. Firstly, the angle over which the particles could see evolved to be large, $\theta \approx 280°$ leaving a blind angle of 80°. This is reasonably close to the blind angle of 60° of many species of fish (Hall et al., 1986). The evolution of the small blind angle constrained the attraction radius R_a within which the orientation radius R_o was then free to evolve. Two evolutionary outcomes were possible for R_o, evolving either to be close to, but slightly larger than, 0 or to be close to, but slightly smaller than, R_a. In the first case the particles formed a slow moving milling group (figure 5.10a) while in the second they formed a fast moving dynamic group (figure 5.10b). Which of these outcomes evolves depends on the initial values of R_o within the population and the rate of mutation during selection. If R_o was initially large a dynamic group would evolve and if it was initially small a slow moving mill would evolve.

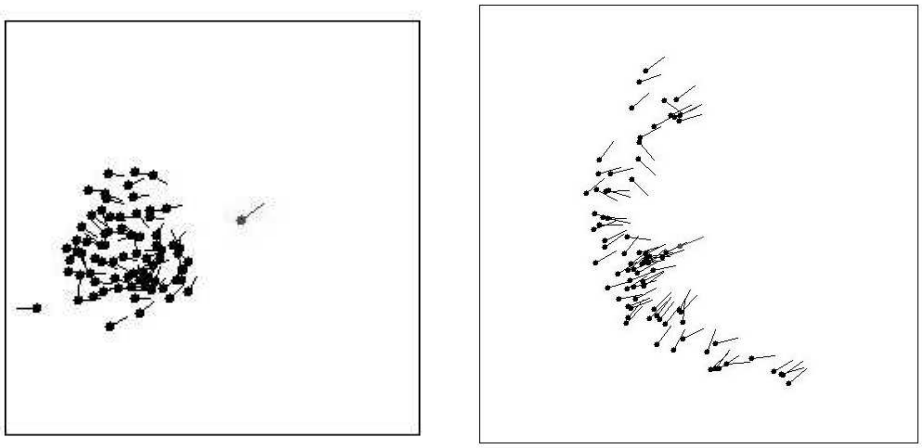


Figure 5.10. Typical example of the two types of evolutionarily stable flock types in the Wood et al. model. Each flock is shown before and during the attack of a predator. (a) is a compact milling torus that responds relatively slowly to the predator while (b) is a dynamic parallel group with a high degree of alignment but only loose between individual attraction. When a predator attacks, the group fans out to avoid it. Prey heads are marked with a circle and the line indicates their current velocity. Predators are larger and marked with an arrow.

While both evolving through 'natural selection', the dynamic group was more efficient than the slow moving mill at avoiding predation. The dynamic group had similar responses to predators as the optimised groups of Inada & Kawachi (2002) and of Zheng et al. (2005). It produced a confusion effect and split to avoid predation in 60-70% of cases. On the other hand, the predator was almost always successful in catching prey when faced with a slow moving mill. Wood et al.'s (2007) study is important because it provides evidence that complex collective level phenomena can evolve between 'selfish' individuals without the need to invoke arguments based on kin selection or repeated interactions between individuals.

References

[1] Aoki, I. 1982. A Simulation Study on the Schooling Mechanism in Fish. Bulletin of the Japanese Society of Scientific Fisheries, 48, 1081-1088.

[2] Avitabile, A., Morse, R. A. & Boch, R. 1975. Swarming honey bees guided by pheromones. . Annals of the Entomological Society of America, 6, 1079-1082.

[3] Ballerini, M., Cabibbo, N., Candelier, R., Cavagna, A., Cisbani, E., Giardina, I., Orlandi, A., Parisi, G., Procaccini, A., Viale, M. & Zdravkovic, V. 2008a. Empirical investigation of starling flocks: a benchmark study in collective animal behaviour. Animal Behaviour, 76, 201-215.

[4] Ballerini, M., Calbibbo, N., Candeleir, R., Cavagna, A., Cisbani, E., Giardina, I., Lecomte, V., Orlandi, A., Parisi, G., Procaccini, A., Viale, M. & Zdravkovic, V. 2008b. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. Proceedings of the National Academy of Sciences of the United States of America, 105, 1232-1237.

[5] Becco, C. V., N., Delcourt, J. & Poncin, P. 2006. Experimental evidences of a structural and dynamical transition in fish school. Physica A, 367, 487-493.

[6] Beekman, M., Fathke, R. L. & Seeley, T. D. 2006. How does an informed minority of scouts guide a honeybee swarm as it flies to its new home? Animal Behaviour, 71, 161-171.

[7] Biro, D., Sumpter, D. J. T., Meade, J. & Guilford, T. 2006. From compromise to leadership in pigeon homing. Current Biology, 16, 2123-2128.

[8] Buhl, J., Sumpter, D. J. T., Couzin, I. D., Hale, J. J., Despland, E., Miller, E. R. & Simpson, S. J. 2006. From disorder to order in marching locusts. Science, 312, 1402-1406.

[9] Cavagna, A., Cimarelli, A., Giardina, I., Orlandi, A., Parisi, G., Procaccini, A., Santagati, R. & Stefanini, F. 2008a. New statistical tools for analyzing the structure of animal groups. Mathematical Biosciences, 214, 32-37.

[10] Cavagna, A., Giardina, I., Orlandi, A., Parisi, G., Procaccini, A., Viale, M. & Zdravkovic, V. 2008b. The STARFLAG handbook on collective animal behaviour: 1. Empirical methods. Animal Behaviour, 76, 217-236.

[11] Couzin, I. D. & Krause, J. 2003. Self-organization and collective behavior in vertebrates. Advances in the Study of Behavior, 32, 1-75.

[12] Couzin, I. D., Krause, J., Franks, N. R. & Levin, S. A. 2005. Effective leadership and decision-making in animal groups on the move. Nature (London), 433, 513-516.

[13] Couzin, I. D., Krause, J., James, R., Ruxton, G. D. & Franks, N. R. 2002. Collective memory and spatial sorting in animal groups. Journal Of Theoretical Biology, 218, 1-11.

[14] Czirok, A., Barabasi, A. L. & Vicsek, T. 1999. Collective motion of self-propelled particles: Kinetic phase transition in one dimension. Physical Review Letters, 82, 209-212.

[15] Czirok, A., Stanley, H. E. & Vicsek, T. 1997. Spontaneously ordered motion of self-propelled particles. Journal Of Physics A-Mathematical And General, 30, 1375-1385.

[16] Czirok, A. & Vicsek, T. 2000. Collective behavior of interacting self-propelled particles. Physica A, 281, 17-29.

[17] Deneubourg, J. L., Grégoire, J. C. & Le Fort, E. 1990. Kinetics of larval gregarious behavior in the bark beetle Dendroctonus micans (Coleoptera: Scolytidae). Journal of Insect Behavior, 3, 169-182.

[18] Focardi, S. & Pecchioli, E. 2005. Social cohesion and foraging decrease with group size in fallow deer (Dama dama). Behavioral Ecology and Sociobiology, 59, 84-91.

[19] Gregoire, G., Chate, H. & Tu, Y. H. 2003. Moving and staying together without a leader. Physica D-Nonlinear Phenomena, 181, 157-170.

[20] Gueron, S., Levin, S. A. & Rubenstein, D. I. 1996. The dynamics of herds: From individuals to aggregations. Journal Of Theoretical Biology, 182, 85-98.

[21] Hale, J. J. 2008. Automated tracking and collective behaviour in locusts and humans. In: Zoology Department: University of Oxford.

[22] Hall, S. J., Wardle, C. S. & MacLennan, D. N. 1986. Predator evasion in a fish school: test of a model for the fountain effect. Marine biology, 143-148.

[23] Hamilton, W. D. 1971. Geometry for the selfish herd. Journal of Theoretical Biology, 31, 295-311.

[24] Helbing, D. & Molnar, P. 1995. Social Force Model for Pedestrian Dynamics. Physical Review E, 51, 4282-4286.

[25] Hemelrijk, C. K. 2000. Towards the integration of social dominance and spatial structure. Animal Behaviour, 59, 1035-1048.

[26] Hoare, D. J., Couzin, I. D., Godin, J. G. J. & Krause, J. 2004. Context-dependent group size choice in fish. Animal Behaviour, 67, 155-164.

[27] Inada, Y. & Kawachi, K. 2002. Order and flexibility in the motion of fish schools. Journal of Theoretical Biology, 214, 371-387.

[28] Janson, S., Middendorf, M. & Beekman, M. 2005. Honeybee swarms: how do scouts guide a swarm of uninformed bees? Animal Behaviour, 70, 349-358.

[29] Jeanson, R., Rivault, C., Deneubourg, J. L., Blanco, S., Jost, C. & Theraulaz, G. 2005. Self-organized aggregation in cockroaches. Animal Behaviour, 69, 167-180.

[30] Krause, J. 1993. The Effect of Schreckstoff on the Shoaling Behavior of the Minnow - a Test of Hamiltons Selfish Herd Theory. Animal Behaviour, 45, 1019-1024.

[31] Krause, J. 1994. Differential Fitness Returns in Relation to Spatial Position in Groups. Biological Reviews of the Cambridge Philosophical Society, 69, 187-206.

[32] Krause, J. & Ruxton, G. D. 2002. Living in groups. Oxford ; New York: Oxford University Press.

[33] Lee, S. H. 2006. Predator's attack-induced phase-like transition in prey flock. Physics Letters A, 357, 270-274.

[34] Lee, S. H., Pak, H. K. & Chon, T. S. 2006. Dynamics of prey-flock escaping behavior in response to predator's attack. Journal of Theoretical Biology, 240, 250-259.

[35] Lindauer, M. 1955. Schwarmbienen auf Wohnungssuche. Zeitschrift für vergleichende Physiologie, 37, 263-324.

[36] Niwa, H. S. 2004. Space-irrelevant scaling law for fish school sizes. Journal of Theoretical Biology, 228, 347-357.

[37] Okubo, A. 1986. Dynamical aspects of animal grouping. . Advances in Biophysics, 22, 1-94.

[38] Okubo, A. & Chiang, H. C. 1974. An analysis of the kinematics of swarming ofAnarete pritchardi kim (Diptera: Cecidomyiidae). Population Ecology, 1-42.

[39] Parrish, J. K. 1989. Reexamining The Selfish Herd - Are Central Fish Safer. Animal Behaviour, 38, 1048-1053.

[40] Parrish, J. K., Viscido, S. V. & Grunbaum, D. 2002. Self-organized fish schools: An examination of emergent properties. Biological Bulletin, 202, 296-305.

[41] Partridge, B. L. 1981. Internal Dynamics and the Interrelations of Fish in Schools. Journal of Comparative Physiology, 144, 313-325.

[42] Partridge, B. L. & Pitcher, T. J. 1980. The sensory basis of fish schools: relative roles of lateral line and vision. Journal of Comparative Physiology, 135, 315-325.

[43] Partridge, B. L., Pitcher, T. J., Cullen, M. J. & Wilson, J. 1980. The three-dimensional structure of fish schools. Behavioral Ecology and Sociobiology, 6.

[44] Partridge, B. L. J. 1982. The structure and function of fish schools. Scientific American, 246.

[45] Partridge, B. L. J., Johansson, J. & Kalisk, J. 1983. Structure of schools of giant bluefin tuna in Cape Cod Bay. Environmental Biology of Fishes, 9, 253-262.

[46] Pitcher, T. J., Magurran, A. E. & Winfield, I. J. 1982. Fish in Larger Shoals Find Food Faster. Behavioral Ecology and Sociobiology, 10, 149-151.

[47] Quinn, J. L. & Cresswell, W. 2006. Testing domains of danger in the selfish herd: sparrowhawks target widely spaced redshanks in flocks. Proceedings of the Royal Society B-Biological Sciences, 273, 2521-2526.

[48] Radakov, D. V. 1973. Schooling in the ecology of fish. New York: John Wiley & Sons.

[49] Rayor, L. S. & Uetz, G. W. 1990. Trade-Offs in Foraging Success and Predation Risk with Spatial Position in Colonial Spiders. Behavioral Ecology and Sociobiology, 27, 77-85.

[50] Reynolds, C. W. 1987. Flocks, herds and schools: A distributed behavioural model. Computer Graphics, 21, 25-33.

[51] Schultz, K. M., Passino, K. M. & Seeley, T. D. 2008. The mechanism of flight guidance in honeybee swarms: subtle guides or streaker bees? Journal of Experimental Biology, 211, 3287-3295.

[52] Szabo, B., Szollosi, G. J., Gonci, B., Juranyi, Z., Selmeczi, D. & Vicsek, T. 2006. Phase transition in the collective migration of tissue cells: Experiment and model. Physical Review E, 74.

[53] Tien, J. H., Levin, S. A. & Rubenstein, D. I. 2004. Dynamics of fish shoals: identifying key decision rules. Evolutionary Ecology Research, 6, 555-565.

[54] Treherne, J. E. & Foster, W. A. 1981. Group transmissin of predator avoidance behaviour ina marine insect: the Trafalgar effect. Animal Behaviour, 28.

[55] Uvarov, B. P. 1977. Grasshoppers and Locusts. Vol. 2.

[56] Wallraff, H. G. 1978. Social Interrelations Involved in Migratory Orientation of Birds - Possible Contribution of Field Studies. Oikos, 30, 401-404.

[57] Weimerskirch, H., Martin, J., Clerquin, Y., Alexandre, P. & Jiraskova, S. 2001. Energy saving in flight formation - Pelicans flying in a 'V' can glide for extended periods using the other birds' air streams. Nature, 413, 697-698.

[58] Vicsek, T., Czirok, A., Benjacob, E., Cohen, I. & Shochet, O. 1995. Novel Type of Phase-Transition in a System of Self-Driven Particles. Physical Review Letters, 75, 1226-1229.

[59] Vine, I. 1971. Risk of visual detection and pursuit by a predator and the selective advantage of flocking behaviour. Journal of Theoretical Biology, 30, 405-422.

[60] Wood, A. J. & Ackland, G. J. 2007. Evolving the selfish herd: emergence of distinct aggregating strategies in an individual-based model. Proceedings of the Royal Society B-Biological Sciences, 274, 1637-1642.

[61] Zheng, M., Kashimori, Y., Hoshino, O., Fujita, K. & Kambara, T. 2005. Behavior pattern (innate action) of individuals in fish schools generating efficient collective evasion from predation. Journal of Theoretical Biology, 235, 153-167.

Evolution of body condition-dependent dispersal under kin competition

Margarete Utz[*], *Eva Kisdi and Mats Gyllenberg*

Abstract

We present a model for the evolution of dispersal when dispersal probability is a function of individual body condition. The biological motivation of this model is found in the huge amount of empirical research that investigates body condition-dependent dispersal and in the lack of satisfying theories to explain especially the puzzling phenomenon of dispersal of strong individuals.

Under the given assumptions, which include that stronger individuals are better competitors, dispersal of strong individuals seems to be a common outcome of our model. Our model thus marks a first theoretical approach to gain more insight into mechanisms that shape body condition-dependent dispersal strategies.

This work is supported by the Graduate School in Computational Biology, Bioinformatics and Biometry (ComBi) of the Ministry of Education in Finland and by the Academy of Finland.

The model and results presented in this article are described in detail in references [1] and [2].

1 Introduction

Dispersal plays a crucial role in the dynamics of populations and in species persistence and expansion. There is a huge and diverse body of literature exploring the evolution of this important trait. Clearly, dispersers are not a random subset of the population. One significant parameter that influences dispersal behaviour is individual body condition. Differences in body condition between dispersers and non-dispersers are observed across species and in many instances, e.g. survival during dispersal or competitive ability depend on body condition.

The best known verbal hypothesis concerning body condition-dependent dispersal is the social dominance hypothesis. It states that stronger individuals that dominate socially suppress weaker individuals by e.g. defeating them in fights or denying them access to resources, whereupon the weaker individuals are forced to leave the local territory. On the other hand, dispersal of strong individuals is barely understood.

Based on the model of Hamilton and May for dispersal under kin competition, we present a model where the probability of dispersal is a function of individual body condition. Body condition is defined such that stronger individuals survive dispersal with higher probability and have higher competitive ability than weaklings.

2 The model

The underlying biological assumptions of the model are as follows. The habitat is spatially structured with patches of different environmental qualities $y \,\epsilon\, (-\infty, +\infty)$ with $\phi(y)$ being the probability density of patch qualities. Each patch can support one (female) individual. In the beginning of the year, each patch is typically occupied by one juvenile individual. An individual survives until maturity with probability s. The species is semelparous, i.e., individuals die immediately after reproduction. The average number of offspring of one individual is B. We do not consider maternal effects such that the body condition $z \,\epsilon\, (-\infty, +\infty)$ of offspring depends only on the quality of the natal patch and offspring body condition in one patch follows a distribution $\beta(z, y)$ with the patch quality y as its mean. For an offspring individual with body condition z, let $p(z, y)$ be the probability that it disperses from its natal patch that has quality y, and let $\Pi(z)$ denote the probability of surviving dispersal such that $\Pi(z)$ is a non-decreasing function of z. There is no cost to staying in the home patch. Dispersal is global such that dispersers are distributed uniformly over patches. After dispersal, immigrants and (if present) local non-dispersers compete in each patch and one individual establishes itself in the patch whereas all others die. Competitive ability depends on body condition such that stronger individuals establish themselves with higher probability. At the end of the season we randomly

[*]Department of Mathematics and Statistics, 00014 University of Helsinki, Finland, margarete.utz@helsinki.fi

reassign patch qualities, assuming that the environmental quality of the habitat fluctuates temporally and locally such that patch qualities are independent of the past and independent of one another.

We will investigate the fate of a rare mutant that occurs in the population when the resident population is in dynamical equilibrium by applying the theory of Adaptive Dynamics. The resident population is automatically in equilibrium because offspring body condition is determined only by patch qualities and therefore, at reproduction, the fraction of offspring individuals with body condition z that are born in patches of quality y is the same every season.

A mutant offspring with body condition z disperses from a patch of quality y with probability $p_m(z,y)$, whereas the resident uses the strategy $p(z,y)$. There is no other difference between the mutant and the resident. The mutant fitness is

$$W(p_m) = \int_{-\infty}^{+\infty} \phi(y) \int_{-\infty}^{+\infty} s\,B\,\beta(z,y)\Big(p_m(z,y)\Pi(z)R(z) + (1-p_m(z,y))P_m(z,y)\Big)\,dz\,dy \tag{1}$$

where $R(z)$ is the probability that one mutant offspring with body condition z establishes itself in a resident patch given it disperses and survives dispersal. A disperser can immigrate into two kinds of patches: (i) patches where the individual that occupied the patch in the year before survived until reproduction. In these patches, competition is carried out between local non-dispersers and immigrants. The fraction of such patches among all patches is s. (ii) patches where the individual in the previous year died before maturation. There, only immigrants compete. The fraction of those patches is $1-s$. Let $P_1(z,y)$ be the probability that an immigrant with body condition z wins competition in a patch of type (i) with environmental quality y, and let $P_2(z)$ be the probability that an immigrant with body condition z wins competition in a patch of type (ii) (in such patches, patch quality does not play a role since no local non-dispersers exist). Then,

$$R(z) = \int_{-\infty}^{+\infty} \phi(y)\Big(sP_1(z,y) + (1-s)P_2(z)\Big)dy \tag{2}$$

$P_m(z,y)$ in (1) is the probability that a non-dispersing mutant with body condition z retains the natal patch, which is of quality y.

The particular forms of $P_1(z,y)$, $P_2(z)$ and $P_m(z,y)$ depend on the mechanism that determines competition. In the following section we will investigate different competition scenarios.

We are interested in finding dispersal strategies $\hat{p}(z,y)$ that are evolutionarily stable strategies (ESS), i.e., strategies that maximize the fitness W when $p(z,y) = p_m(z,y) = \hat{p}(z,y)$.

Applying the calculus of variations, Euler's equation is a necessary condition for the functional W to have an extremal at $p = p_m = \hat{p}$. In our case, Euler's equation implies that

$$\Pi(z)\,R(z) = P_m(z,y) \tag{3}$$

has to be satisfied for all z and y such that $0 < \hat{p}(z,y) < 1$. This condition is known as the marginal value theorem and balances the expected fitness of an individual with body condition z born in a patch of quality y if it successfully dispersed (left hand side of (3)) and if it stayed in the natal patch (right hand side).

3 Examples for competition

3.1 Offspring body condition corresponds to the quality of the natal patch

Let us first assume that all offspring born in a patch have the same body condition that corresponds to the environmental quality of the patch, i.e., the offspring condition distribution in a patch of quality y is the point mass $B\delta(\cdot - y)$ concentrated at y. The ESS condition (3) assumes then the form

$$\Pi(z)\,R(z) = P_m(z) \tag{4}$$

3.1.1 Weighted lottery competition

The most common way to model asymmetric competition is a weighted lottery. Let $g(z)$ be a non-decreasing weight function. Then

$$P_1(z,y) = \frac{g(z)}{B\int_{-\infty}^{+\infty} g(z')u_1(z',y)\,dz'} \tag{5}$$

$$P_2(z) = \frac{g(z)}{B\int_{-\infty}^{+\infty} g(z')u_2(z')\,dz'} \tag{6}$$

and

$$P_m(z,y) = \frac{g(z)}{B\int_{-\infty}^{+\infty} g(z')\left(\beta(z',y)\left(1-p_m(z',y)\right)+u_2(z')\right)\,dz'} \tag{7}$$

where $u_1(z,y)$ and $u_2(z)$ are the body condition distribution of the resident after dispersal in a patch of quality y and of type (i) and type (ii), respectively,

$$u_1(z,y) = \beta(z,y)\left(1-p(z,y)\right)+u_2(z) \tag{8}$$

$$u_2(z) = \int_{-\infty}^{+\infty} \phi(y')\, s\, \beta(z,y')\, p(z,y')\, \Pi(z)\, dy' \tag{9}$$

If $g(z) = e^z$, patch quality distribution is $\phi(y) = (1/\sqrt{2\pi})\exp(-y^2/2)$, probability of survival until maturation is $s = 0.9$ and probability of survival during dispersal is $\Pi(z) = 0.6$, then the evolutionarily stable dispersal strategy is an increasing function of body condition as shown in Figure 1(a).

3.1.2 Strongest offspring wins

Another suggestive mechanism for local competition is that the strongest competitor wins. We assume that first condition-independent mortality reduces the number of individuals to k competitors in every patch, and then the strongest individual among these establishes itself in the patch. These k individuals are chosen at random, and k may vary between patches. The probabilities that a mutant with body condition z wins competition in a patch of quality y are as follows.

$$P_1(z,y) = \sum_{k=0}^{N} Pr(X=k)\frac{k}{B\int_{-\infty}^{+\infty} u_1(z',y)\,dz'}\left(\frac{\int_{-\infty}^{z} u_1(z',y)\,dz'}{\int_{-\infty}^{+\infty} u_1(z',y)\,dz'}\right)^{k-1} \tag{10}$$

$$P_2(z) = \sum_{k=0}^{N} Pr(X=k)\frac{k}{B\int_{-\infty}^{+\infty} u_2(z')\,dz'}\left(\frac{\int_{-\infty}^{z} u_2(z')\,dz'}{\int_{-\infty}^{+\infty} u_2(z')\,dz'}\right)^{k-1} \tag{11}$$

$$\begin{aligned} P_m(z,y) \;=\; & \sum_{k=0}^{N} Pr(X=k)\frac{k}{B\int_{-\infty}^{+\infty}\left(\beta(z',y)\left(1-p_m(z',y)\right)+u_2(z')\right)\,dz'}\cdot \\ & \cdot\sum_{i=0}^{k-1}\binom{k-1}{i}\left(\frac{\int_{-\infty}^{z} u_2(z')\,dz'}{\int_{-\infty}^{+\infty}\left(\beta(z',y)\left(1-p_m(z',y)+u_2(z')\right)\right)\,dz'}\right)^{i}\cdot \\ & \cdot\left(\frac{\int_{-\infty}^{z}\beta(z',y)\left(1-p_m(z',y)\right)dz'}{\int_{-\infty}^{+\infty}\left(\beta(z',y)\left(1-p_m(z',y)\right)+u_2(z')\right)\,dz'}\right)^{k-1-i} \end{aligned} \tag{12}$$

Figure 1(b) shows the ESSs for different choices for the distribution of the parameter k when $\phi(y) = (1/\sqrt{2\pi})\exp(-y^2/2)$, $s = 0.9$ and $\Pi(z) = 0.6$.

3.1.3 Mixture of weighted and fair lottery competition

Let us now assume that condition-independent mortality occurs before competition only in a fraction μ of the patches. In these patches exactly one randomly chosen individual survives as in a fair lottery, which then establishes itself in the patch. In all other patches a weighted lottery determines competition as in the first example.

Figure 1(c) shows the ESS for $\phi(y) = (1/\sqrt{2\pi})\exp(-y^2/2)$, $g(z) = e^z$, $s = 0.9$, $\Pi(z) = 0.6$ and $\mu = 0.1$.

3.2 Offspring body condition is distributed around the quality of the natal patch

When offspring body condition in a patch follows e.g. a normal distribution with the patch quality as its mean, $\beta(z,y) = (1/\sqrt{2\pi})\exp(-(z-y)^2/2)$, the ESS condition has the form as in (3), which can generally not be satisfied for all z and y, as both sides of the equation in (3) are functions of different dimensions. Therefore, the fitness W is maximized for a $\hat{p}$ at the boundaries of the interval $[0,1]$, and $\hat{p}$ is a step function that assumes the values 0 and 1.

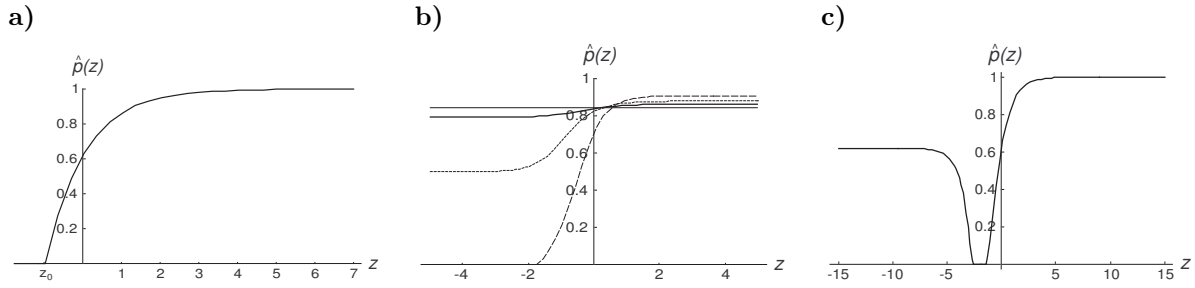


Figure 1: Evolutionarily stable dispersal strategies $\hat{p}(z)$ for different mechanisms of competition when offspring body condition corresponds to the environmental quality of the natal patch, $\beta(z, y) = \delta(z - y)$. (a) weighted lottery (b) strongest wins: thick line: k follows a truncated Poisson distribution with $\lambda = 0.5$, such that $\Pr(k \geq 3) = 0$; horizontal thin line: $\Pr(k = 0) = 2/3$, $\Pr(k = 1) = 1/3$ and $\Pr(k \geq 2) = 0$; dotted line: $\Pr(k = 0) = 0.47$, $\Pr(k = 1) = 0.23$, $\Pr(k = 2) = 0.3$ and $\Pr(k \geq 3) = 0$; dashed line: k follows a truncated Poisson distribution with $\lambda = 2.5$ such that $\Pr(k \geq 7) = 0$. (c) mixed weighted and fair lottery.

3.2.1 Lottery Competition

When local competition is determined by a (weighted or fair) lottery (cf. (5)–(7)), and patch qualities follow a standard normal distribution ($\phi(y) = (1/\sqrt{2\pi})\exp(-y^2/2)$), then the ES dispersal strategy is

$$\hat{p}(z, y) = \begin{cases} 0 & \text{if } z < z_0(y) \\ 1 & \text{otherwise} \end{cases} \tag{13}$$

At which value of z_0 the jump from 0 to 1 happens in each patch, depends on the respective patch quality y.

When all patches have the same quality (e.g. $\phi(y) = \delta(y)$), then

$$\hat{p}(z) = \begin{cases} 0 & \text{if } z < z_0 \\ 1 & \text{otherwise} \end{cases} \tag{14}$$

As an example, assume that the survival probability during dispersal is the sigmoid function

$$\Pi(z) = 0.2 + \frac{0.6}{1 + e^{-z}} \tag{15}$$

and the probability that an individual survives until maturation is $s = 0.6$.

With the weight function $g(z) = e^z$, the ESS is

$$\hat{p}(z) = \begin{cases} 0 & \text{if } z < 0.105 \\ 1 & \text{otherwise} \end{cases} \tag{16}$$

If we assume a fair lottery for competition, i.e. $g(z) = 1$, then the ESS is

$$\hat{p}(z) = \begin{cases} 0 & \text{if } z < -0.754 \\ 1 & \text{otherwise} \end{cases} \tag{17}$$

4 Discussion

Our model explains dispersal of strong individuals, which is a puzzling phenomenon given that strong individuals could easily retain the natal patch and yet are exposed to risky dispersal. A common outcome of the present model seems to be an increasing dispersal fraction, or at least a strategy where weaklings do not disperse at all but strong individuals surely disperse. But we also presented a mechanism for competition that yields a non-monotone ESS.

In all examples, except in the mixture of weighted and fair lottery (see Figure 1(c)), very weak individuals stay all in the natal patch, because their chances to survive dispersal and establish themselves in another patch are negligible, whereas they increase the probability that one family member retains the home patch if all offspring stay in the patch. In examples shown in Figure 1(a) and (b), with increasing body condition, $\hat{p}(z)$ increases, as an individual's chance to win a patch after dispersal increases with body condition, and in families with stronger and stronger individuals less and less non-dispersers are necessary to maintain a high probability of retaining the home patch.

When offspring body condition follows a distribution around the natal patch quality, the ESS is a step function. As a (weighted) lottery determines competition, it is important to keep a certain weighted mass of offspring at home, and since stronger individuals do better during dispersal and competition than weaklings, they are sent away and weaklings are kept at home.

The intriguing non-monotonic shape of the example depicted in Figure 1(c) is due to the fact that, in a pure weighted lottery, selection is very weak on very weak individuals. With the introduction of a fair lottery in a small fraction of patches, very weak individuals take the chance to disperse and hope to be able to establish themselves in a patch where a fair lottery determines competition, i.e., where body condition does not play a role. For intermediate and high values of z, the shape of $\hat{p}(z)$ is not affected much by the fact that competition is random in a small part of the patches and is thus very similar to the ESS in a pure weighted lottery (Figure 1(a)).

References

[1] M. Gyllenberg, E. Kisdi, M. Utz. Evolution of condition-dependent dispersal under kin competition. *J. Math. Biol.* (57):285–307, 2008

[2] M. Gyllenberg, E. Kisdi, M. Utz. Effects of differences in body condition of kin on dispersal behaviour. *in preparation*

Implementation Models of ICT in Teaching of Complex Numbers

Anton Vrdoljak
Faculty of Civil Engineering, University of Mostar, Mostar, Bosnia and Herzegovina

Abstract

The objective of this paper is to report my educational research project based on the advantage of learning and teaching with ICT, focusing on teaching of complex numbers. Some examples of motivational impacts of ICT existed already within a wide literature on educational uses and outcomes of ICT in classrooms, as well as out of school settings. The project was designed to build on such models and investigate the issues in more detail, and to explore ways to enhance motivational impacts and outcomes.

Keywords: mathematics; teaching mathematics; learning mathematics; complex numbers; educational technology; ICT; e-Learning; Web-based learning; motivational impacts of ICT; educational potential.

1 Introduction

The teaching of complex numbers has always been a challenge for teachers of mathematics, because it is very important to understand the reason why complex numbers were invented. Next, it is extremely important that pupils become introduced into the abstract mathematical system of complex numbers, and learn concepts which can be successfully applied later in other mathematical topics. It is well known that the "reason" which leads to complex numbers concerns solutions of equations. More exact, we can accept that the complex numbers were invented to provide solutions to polynomial equations. On the other hand, complex numbers are useful abstract quantities that can be used in calculations that result in physically meaningful solutions [13]. Today, the complex number system is so deeply rooted in physical theory (e.g. quantum mechanics) that one could argue that the complex number system is a more "real" description of the world than the real number system [2]. The famous physicist Roger Penrose wrote an essay to this effect, "Nature is complex". However, recognition of this fact is one that took a long time for mathematicians to accept.

Usually pupils are afraid of complex numbers; they say it is a hard subject with a lot of difficult problems and lot of calculations. Even when the lectures have been supported by powerful digital technology, it is possible that pupils are still passive observers [5]. It is not easy to suggest teaching methods, especially in comparison to traditional lectures, which would be effective and would engage pupils actively and generate stimulating learning. Anyway, we should be aware of several different roles that technology, especially information and communication technology (ICT), can play in instructions: from eliminating computational drudgery in realistic applications to providing environments for active exploration of the properties of mathematical structures and objects, or to getting a variety of experience using different ICT tools. So, if we want to obtain an effective teaching of mathematics in our schools, one of things what we need is to focus our future work (or research) on implementation models of ICT in mathematics education. ICT and "Computer Aided Teaching" have become an important part of life today, and are widely used to improve teaching and learning techniques. Learning with implementation models of ICT, learning and teaching become more interesting for both, pupils and teachers. Pupils can learn and practice their knowledge at home also, without repeating things that are already familiar to them, so they can concentrate on what they don't know [7].

A model of ICT is a concept that carries a lot of potential, but only if it's implemented in an efficient way. It should be more than just another way of delivering information to pupils with inserting a couple of simple animations. It should be an environment developed in highly interactive way allowing pupils to receive knowledge, develop creative thinking and reasoning skills. For better fulfilment of these conditions and for developing more robust online content, for example, software Macromedia Flash MX can be used. This is currently the best choice because this software is considered to be a leader in creating various online elements and it's moving from being just a tool for animations, that are inserted in HTML websites, to a tool for developing complete websites which are modern, interactive, interesting and attract more attention [7]. There are more specific reasons why to use Flash:

- diversity — allowing the use of different graphical tools along with applying programming for complex tasks, suitable for all elements from small animations to whole web-sites with high level of dynamics
- presence — flash player has become a necessity for a computer user, approximately 96% of all web browsers have Flash Player installed
- suitable for any bandwidth — Flash player is suitable for any kind of bandwidth because of more than one reason: it uses vector based graphics allowing efficient storage of the images; it has a high-performance compression facility enables Flash files to run quickly; multimedia components are downloaded partially when needed so the bandwidth is not wasted on elements that are not used

- platform independent — Flash files are standalone and they don't have problems with running on any platform
- easy usage — the software itself is very easy to use with lots of beneficial features: user-friendly interface, a variety of templates (quizzes, presentations, photo slideshows etc.), pre-built components (checkboxes, combo boxes, push buttons, scroll bars etc.).

On the other hand, one has to be aware of the limitations of this software [7]. Concerning development of contents, limitations of Flash are, for example:

- Flash developers must manually build support for features such as back button and book marking, otherwise these features are not available for users
- Flash does not use browser settings for font size so text may appear tiny for some users. However, good characteristics of Flash prevailed in making choose of software for developing mathematical contents.

Regarding to all this I have tried to implement models of ICT in teaching of complex numbers (make e-learning content regarding complex numbers) for pupils and teachers as well. It is important to write here that I did it through the educational research project *Implementation models of ICT in Mathematics Education*, which is the main scientific framework for my master thesis. As my master thesis is still underway, models (contents) are being added frequently. In other words, the main object of this thesis is to actually provide assistance to pupils in understanding contents that have not been fully grasped during regular classes as well as provide more for those eager to further expand their knowledge. In recent time we earned a lot of experience and feedbacks from our own work using implemented models of ICT, as well as from other colleagues' work. Using a lot of animations and with a variety of problems I've prepared teaching and learning content on complex numbers covering the whole topic about complex numbers in the secondary school curriculum. I've also tried to find problems form real world to make this closer to pupils.

The contents on complex numbers are based on the Mathematics curriculum of Bosnia and Herzegovina. Pupils are introduced to complex numbers in the second grade of secondary school for the first time. Topics for that grade include the set of complex numbers based on the extended set of real numbers. Pupils learn the definition of imaginary and complex numbers, the standard form of a complex number, equality of complex numbers, arithmetic operations with complex numbers, absolute value of complex numbers and the Gaussian (complex) plane. In the forth grade, the complex number notation is extended to the polar form of a complex number and definitions are based on trigonometric functions.

An important one of things that I was thinking about was what kind of language to use, in a meaning weather be more precise or try to explain things as simply as I could. Should I use mathematical or everyday language making my content? Complex numbers is a topic that has a lot formulas that pupils need to remember, so it is hard to escape formal language. At the end I decided to use something between.

2 The structure of contents

My content is written in Croatian, one of the three official languages in Bosnia and Herzegovina, and it is an integral part of a web-page which is located on the server of my Faculty [12]. So far on the web-page we can choose:

- lecture,
- interactive lessons,
- test and quizzes.

This choice is based on a similar example described by group of authors [7], and because the fact that I am a member of this group (team). This web-page and content on it can be used for both learning and teaching. Because the teachers should be able to help pupils use ICT to acquire the skills of searching for, managing, analyzing, integrating, and evaluating information, the content and activities based on that content are designed in a way that engage pupils in collaborative problem solving, research, or artistic creation. The lessons in the lecture part are similar to the lessons in the textbooks. Every lesson is followed with solved examples. This is good for pupils, because they can find formulas and information about things they do in school in one place. And teachers can use this content to make classes a little bit different than in the traditional way of teaching, and also it is easier for them, because in this case there is no necessity to write all the facts onto the blackboard.

The lecture content is followed by interactive lessons. In other words, they are covering the lecture content but with a different approach. In this part pupils are not just supposed to read lessons, but also to fill in the blanks. These lessons are made like a conversation between pupils and computer. The pupils go through the lesson answering questions. In this way information is not just served to pupils like in a lecture part, but they have to search for information. They learn about some specific topic giving a correct answer. All problems have a button *Check*, where pupils can see if their solution is correct or not, and some of them have a button *Solution* that shows the correct solution with explanation. This part of the web-page teachers can use to take a variety of problems and examples. Also using this way of teaching, there is no difficulty of going further for pupils that know more and to those who know less. Pupils can go through the page and learn by themselves on the level that they are. Next, according to some educational researches teachers indicated that pupils were more able to reach their highest potential level because they were less limited in terms of experiences that they could gain [9].

Test and quizzes are made for revising the knowledge from one specific topic. Here pupils can see if their solution is correct or not, without any explanation. Teachers can use this part to check the pupils' knowledge from each topic.

3 The contents of complex numbers

The contents of complex numbers are based on the Mathematics curriculum of Bosnia and Herzegovina (see Table 1). Already is described when pupils are introduced complex numbers in the secondary school, and here is shortly explained about lectures, interactive lessons, games, test and quizzes, so here it will be explained more the content (implemented models of ICT in teaching of complex numbers).

Table 1: Contents of mathematics curriculum regarding complex numbers (B&H).

Grade	**Topic**
2^{nd}	Imaginary unit and imaginary numbers
2^{nd}	Set of complex numbers
2^{nd}	Complex numbers and the Gaussian (complex) plane
2^{nd}	Standard form of a complex number
2^{nd}	Equality of complex numbers
2^{nd}	The modulus of a complex number
2^{nd}	The arithmetic of complex numbers
2^{nd}	Addition and subtraction of complex numbers
2^{nd}	Multiplication of complex numbers
2^{nd}	Conjugates of complex numbers
2^{nd}	Division of complex numbers
4^{th}	The polar form of a complex number
4^{th}	Multiplication and division of complex numbers in the polar form
4^{th}	De Moivre's theorem
2^{nd} & 4^{th}	Geometric interpretations of complex numbers

3.1 The Arithmetic of Complex Numbers: Multiplication of complex numbers

After this lecture the user should be able to define one of the basic arithmetic operations of complex numbers — multiplication, know the main properties and some special cases of this operation and apply it in problem solving exercises, which is the main aim of this lecture. Next, lecture should contain the definition of multiplication and should explain the technique for multiplying two complex numbers. Because the complex multiplication is a more difficult operation to understand than either an algebraic or geometric one, we will do it algebraically first. The user will be later introduced to some special cases of multiplication, for example, multiplying a complex number by a real number, multiplying a complex number by i, multiplication and absolute value, and the main properties of this operation, for example commutation. At the end of the lecture, user is led to the interactive lessons to discover techniques for multiplying two complex numbers, some special cases of multiplication, main properties of this operation, and finding square roots of a complex number. In addition, the user will later be able to use few bottomless worksheets of multiplying complex numbers.

The interactive lesson starts with an example, in witch the user (pupil) should enter two complex numbers, and later be able to see techniques for multiplying these two complex numbers, or submit and check his answer. This example also contains a bottomless worksheet.

The second example is similar to the previous one, but this time the user should determine real and imaginary parts in a product of two random complex numbers. So, in this example he should additionally use knowledge about equality of complex numbers. This example also contains a bottomless worksheet, worksheet number 2.

The third example is again similar to the previous two examples. The aim of these three examples is that a user by using dynamic techniques discovers a technique for multiplying two complex numbers, and its use in various situations. The approach is based on self discovery through observation of dynamic techniques and conclusions based on the new lecture as well as on the previous knowledge of complex numbers, for example, equality of complex numbers.

The fourth example first introduces a special case of multiplication, multiplying a complex number by a real number, and later another special case, multiplication and absolute value. This example contains a simple questionnaire for the observation data.

The fifth example introduces a second special case of multiplication, multiplying a complex number by i, and later, again, another special case, multiplication and absolute value. Also, this example contains a simple questionnaire for the observation data. The questionnaire is divided into a few parts. In the first part, the user uses dynamic multiplying by i to realize that multiplying a random complex number z by i, leads to a counter clockwise rotation of the point z by 90° around the origin to a new point zi. Later, the user will be asked to determine the effect of multiplying a complex number by i and again by i, effect of multiplying complex number by $-i$, and effect of multiplying by i^3. In this part, the user uses dynamic multiplying by i and again by i to realize that multiplying a random complex number z by i and again by i leads to a counter clockwise rotation of the point z by 180° around the origin to a new point $zi^2 = -z$, in other words, the user should realize that this multiplying leads to the opposite number of number z, and get that i squared is -1. Effects of multiplying complex numbers by $-i$ and by i^3 will be explained in details. At the end of the questionnaire user is led to discover that points z, $zi, -z, zi^3$ lie on the same circumference.

The aim of these two examples is that a user by using dynamic multiplying by a real number, dynamic multiplying by i and dynamic multiplying by i and again by i, discovers that these multiplications lead to scaling and rotation. Also, a user should discover a relationship between multiplication and absolute value. Approach is based on the self discovery through observation of graphical data and conclusions based on the new lecture as well as on the previous knowledge of complex numbers.

The sixth example, also the last example, should help the user in understanding of finding the square roots of complex numbers. This is the most complex of the examples. It includes solving a system of previously obtained equations and the equality of complex numbers. The test includes application of learned contents both from the lecture and interactive lessons. It should contain exercises with multiplication of two complex numbers, determining of real and imaginary parts of complex numbers, opposite numbers, comparing and determining the modulus of some special complex numbers, and finding the square roots of a complex number. In this part the user shouldn't use dynamic technique or dynamic multiplying as it is a test of knowledge.

3.2 Geometric Interpretations of the Triangle Inequality

Here is another one lecture as a part of the developed contents based on the advantage of learning and teaching with ICT, especially web-based learning and teaching. I chose this example because here it is very easy to see how the availability of easy-to-use FLASH media files highlights the role of graphical representations of dots and vectors, as well as the triangle inequality. The dynamic plane has two mobile dots, which are at the ends of vectors z & w (see Figure 1 below). By dragging them, it's possible to dynamically change lengths of vectors and the values of complex numbers which are represented numerically and graphically. At the same time, the values of sums are given as a feedback or information. This is the most important function of this interactive lesson, because giving of the values of sums immediately has the potential to change the pupil's activities in a very impressive manner.

I expect here that the learner finds, by exploring and observing numerical data and graphic representation, the geometric interpretation of triangle inequality based on his or hers previous knowledge, or that there are no occasions where it is possible in a triangle for one side of a triangle to be larger than the sum of the lengths of the other two sides.

Teaching experiments, later kept in one secondary school in Bosnia and Herzegovina, shows that pupils were able to accept triangle inequality very well and were enthusiastic throughout the lesson [14, 16]. Evident difference of pupil's activity was observed as they were motivated to actively participate during the lesson. After the experiment teachers noted that the following issues are important pedagogically/mathematically:

- the possibility of scaling
- the possibility of meeting surprises and investigating their origins
- the possibility of a dynamic change of parameters to better appreciate a triangle inequality.

Also, they concluded that these kinds of contents properly combined with traditional methods of teaching and pupil learning could improve pupils' interest in mathematics, as well as their understanding of different mathematical concepts.

4 Conclusion

The brief description of given examples in last chapter points only some of the educational potentials and possibilities afforded with an implemented models of ICT. Although the general findings of my educational research are not completed in this moment, because my master thesis is still underway, but the core findings of my work have shown that:

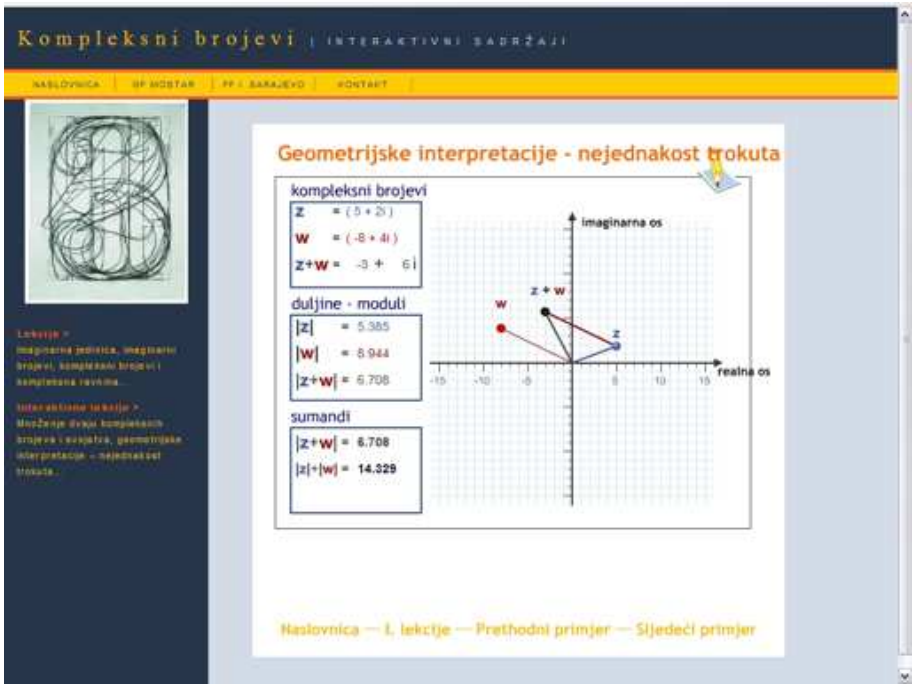


Figure 1: Red and blue dots can be dragged and the change of values is visible.

- motivation has been enhanced most positively in school situations where models of ICT are implemented within a framework that considers impacts upon learning and teachings,
- pupils have positive attitudes to the use implemented models of ICT,
- implemented models of ICT are seen by pupils as a learning aid rather than a means to gain a competitive advantage over other pupils in the class,
- ICT enables a multi-sensory approach to both teaching and learning, and many children need visual, auditory and kinaesthetic stimulation in order to enable learning,
- ICT supports independence of pupil working and pupils feel more in control of their learning when it is used appropriately,
- ICT can be used to extend the teaching day (by enabling pupils to work on tasks outside classrooms and in ways that they could not do without using ICT), and supporting communications between teachers and pupils is motivating for both teachers and pupils.

These core findings are in accordance with results of some other referred authors. It is obvious that for this result we can thank mainly the integration and implementation of ICT in teaching of complex numbers. Technology brings to pupils and their teachers the opportunity to individualize learning — to generate illustrative examples, as well as a dynamic and rich presentation of a given subject, to follow interesting topics to the desire depth, to choose their own problems and appropriate tools for solving them. Because awareness regarding huge transformations of the workplace for teaching and learning of mathematics with ICT, the next goal in my educational research will be to examine did developed learning and teaching contents fully improve the advantage of learning and teaching with ICT, and to determine how to maximise the motivational impact from ICT. With these findings I will be able to close (conclude) my master thesis, and others will be able to consider the ways in which teachers could enhance motivational impacts of ICT for pupils, especially for those disaffected with traditional forms of learning.

References

[1] A.W. Bates and G. Poole, *Effective teaching with technology in higher education: Foundation for success*, Jossey-Bass, San Francisco, 2003.

[2] M. Boyle, Complex numbers and series, *Notes regarding Calculus II course*, Department of Mathematics, University of Maryland, 2006.

[3] N. A. Buzzetto-More, Principles of effective online teaching, *Informing Science Institute, Santa Rosa, CA*, 2007.

[4] N. A. Buzzetto-More, Advanced principles of effective e-learning, *Informing Science Institute, Santa Rosa, CA*, 2007.

[5] Lj. Dikovic, Interactive learning and teaching of linear algebra by web technologies: some examples, *The Teaching of Mathematics 2007*, Vol. X, 2, pp. 109-116.

[6] S. Egenfeldt-Nielsen, *Beyond Edutainment: Educational Potential of Computer Games*, Continuum International Publishing Group, London, 2007.

[7] Group of authors, Development of learning content with Information Communication Technology (ICT) and e-Learning Environment for Informatics and Mathematics, *Reports No. 1, No. 2 and No. 3 by JICA Trainees from Bosnia and Herzegovina*, Center for Research on International Cooperation In Educational Development (CRICED) University of Tsukuba, Tsukuba, 2005, 2006, 2007.

[8] A. Oldknow and R. Taylor, *Teaching Mathematics using Information and Communication Technology*, 2nd Edition, Continuum International Publishing Group Ltd., London, 2004.

[9] D. Passey & ass., The Motivational Effect of ICT on Pupils, *Department of Educational Research, University of Lancaster*, 2004.

[10] K. Pjanic & ass., Development of Learning Contents with ICT on Mathematics and Informatics, *Proceedings of the 29th annual meeting of JSSE (Japan Society for Science Education)*, 29 (2005), pp. 601-602, Gifu, Japan, 2005.

[11] R. Sutherland, *Teaching for Learning Mathematics*, Mc Graw-Hill, Maidenhead, 2006.

[12] URL: `http://www.gfmo.ba/kompleksni_brojevi/`

[13] URL: `http://mathworld.wolfram.com/ComplexNumber.html`

[14] A. Vrdoljak, K. Aoyama, H. Yahara and M. Isoda, Development of Mathematics Learning Contents with ICT Focused on Complex Number Contents of "Interactive Lessons", *Proceedings of the 30th annual meeting of JSSE (Japan Society for Science Education)*, 30 (2006), pp. 265-266, Tsukuba, Japan, 2006.

[15] A. Vrdoljak and N. Bouz-Asal, Learning portal: MATHEMATICS ONLINE, *The First Petra International Conference on Mathematics, PICOM 2007*, Al-Hussein Bin Talal University, Ma'an, Jordan, 2007.

[16] A. Vrdoljak and N. Bouz-Asal, The educational potential of learning portals: The case of Mathematics ONLINE, *The Third International Conference on Mathematical Science, ICM 2008*, United Arab Emirates University, Al-Ain, UAE, 2008.

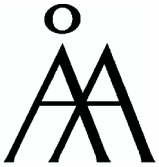
Turku Centre for Computer Science
TUCS General Publications

33. **Peter Selinger (Editor)**, Proceedings of the 2nd International Workshop on Quantum Programming Languages
34. **Kai Koskimies, Johan Lilius, Ivan Porres and Kasper Østerbye (Eds.)**, Proceedings of the 11th Nordic Workshop on Programming and Software Development Tools and Techniques, NWPER'2004
35. **Kai Koskimies, Ludwik Kuzniarz, Johan Lilius and Ivan Porres (Eds.)**, Proceedings of the 2nd Nordic Workshop on the Unified Modeling Language, NWUML'2004
36. **Franca Cantoni and Hannu Salmela (Eds.)**, Proceedings of the Finnish-Italian Workshop on Information Systems, FIWIS 2004
37. **Ralph-Johan Back and Kaisa Sere**, CREST Progress Report 2002-2003
38. **Mats Aspnäs, Christel Donner, Monika Eklund, Ulrika Gustafsson, Timo Järvi and Nina Kivinen (Eds.)**, Turku Centre for Computer Science, Annual Report 2004
39. **Johan Lilius, Ricardo J. Machado, Dragos Truscan and João M. Fernandes (Eds.)**, Proceedings of MOMPES'05, 2nd International Workshop on Model-Based Methodologies for Pervasive and Embedded Software
40. **Ralph-Johan Back, Kaisa Sere and Luigia Petre**, CREST Progress Report 2004-2005
41. **Tapio Salakoski, Tomi Mäntylä and Mikko Laakso (Eds.)**, Koli Calling 2005 - Proceedings of the Fifth Koli Calling Conference on Computer Science Education
42. **Petri Paju, Nina Kivinen, Timo Järvi and Jouko Ruissalo (Eds.)**, History of Nordic Computing - HiNC2
43. **Tero Harju and Juhani Karhumäki (Eds.)**, Proceedings of the Workshop on Fibonacci Words 2006
44. **Michal Kunc and Alexander Okhotin (Eds.)**, Theory and Applications of Language Equations, Proceedings of the 1st International Workshop, Turku, Finland, 2 July 2007
45. **Mika Hirvensalo, Vesa Halava and Igor Potapov, Jarkko Kari (Eds.)**, Proceedings of the Satellite Workshops of DLT 2007
46. **Anne-Maria Ernvall-Hytönen, Matti, Jutila, Juhani Karhumäki and Arto Lepistö (Eds.)**, Proceedings of Conference on Algorithmic Number Theory 2007
47. **Ralph-Johan Back and Ion Petre (Eds.)**, Proceedings of COMPMOD 2008
48. **Elena Troubitsyna (Editor)**, Proceedings of Doctoral Symposium held in conjunction with Formal Methods 2008
49. **Reima Suomi and Sanna Apiainen (Eds.)**,Promoting Health in Urban Living: Proceedings of the Second International Conference on Well-being in the Information Society (WIS 2008)
50. **Aulis Tuominen, Jussi Kantola, Arho Suominen and Sami Hyrynsalmi (Eds.)**, NEXT 2008 - Proceedings of the Fifth International New Exploratory Techniques Conference
51. **Tapio Salakoski, Dietrich Reibholz-Schuhmann and Sampo Pyysalo (Eds.)**, Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)
52. **Helena Karsten, Barbro Back, Tapio Salakoski, Sanna Salanterä and Hanna Suominen (Eds.)**,The Proceedings of the First Conference on Text and Data Mining of Clinical Documents (Louhi'08)
53. **Anne-Maria Ernvall-Hytönen and Camilla Hollanti (Eds.)**,Proceedings of the 3rd Nordic EWM Summer School for PhD Students in Mathematics
54. **Terry Rout, Ivan Porres, Risto Nevalainen and Beatrix Barafort (Eds.)**, Software Process Improvement and Capability Determination 9th International Conference, SPICE 2009, Turku, Finland, June 2009 Proceedings

Turku
Centre *for*
Computer
Science

Joukahaisenkatu 3-5 B, 20520 Turku, Finland | www.tucs.fi

University of Turku
- Department of Information Technology
- Department of Mathematics



Åbo Akademi University
- Department of Information Technologies

Turku School of Economics
- Institute of Information Systems Sciences

ISBN 978-952-12-2279-5
ISSN 1239-1905

Ernvall-Hytönen, Hollanti

Proceedings of the 3rd Nordic EWM Summer School for PhD Students in Mathematics

Ernvall-Hytönen, Hollanti

Proceedings of the 3rd Nordic EWM Summer School for PhD Students in Mathematics