ADVANCED SERIES IN MATHEMATICAL PHYSICS

Published

Vol. 10: Yang-Baxter Equations in Integrable Systems
edited by M Jimbo

Vol. 11: New Developments in the Theory of Knots
edited by T Kohno

Vol. 12: Soliton Equations and Hamiltonian Systems
by L A Dickey

Vol. 13: The Variational Principles of Dynamics
by B A Kupershmidt

Vol. 14: Form Factors in Completely Integrable Models of Quantum Field Theory
by F A Smirnov

Vol. 15: Non-Perturbative Quantum Field Theory – Mathematical Aspects and Applications
by J Fröhlich

Vol. 16: Infinite Analysis – Proceedings of the RIMS Research Project 1991
edited by A Tsuchiya, T Eguchi and M Jimbo

Vol. 17: Braid Group, Knot Theory and Statistical Mechanics (II)
edited by C N Yang and M L Ge

Vol. 18: Exactly Solvable Models and Strongly Correlated Electrons
by V Korepin and F H L Ebler

Vol. 19: Under the Spell of the Gauge Principle
by G 't Hooft

Vol. 20: The State of Matter
edited by M Aizenman and H Araki

Vol. 21: Multidimensional Hypergeometric Functions and Representation Theory of Lie Algebras and Quantum Groups
by A Varchenko

Vol. 22: *W*-Symmetry
by P Bouwknegt and K Schoutens

Vol. 23: Quantum Theory and Global Anomalies
by R A Baadhio

Vol. 25: Basic Methods in Soliton Theory
by I Cherednik

Advanced Series in Mathematical Physics
Vol. 24

THE MATHEMATICAL BEAUTY OF PHYSICS

A Memorial Volume For Claude Itzykson

Saclay, France *5–7 June 1996*

Editors

J M Drouffe
J B Zuber

CEN-Saclay
Service de Physique Théorique
France

World Scientific
Singapore • New Jersey • London • Hong Kong

CONTENTS

THE MATHEMATICAL BEAUTY OF PHYSICS

Published by

World Scientific Publishing Co. Pte. Ltd.

P O Box 128, Farrer Road, Singapore 912805

USA office: Suite 1B, 1060 Main Street, River Edge, NJ 07661

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data

The mathematical beauty of physics : in memory of Claude Itzykson :
Saclay, 5–7 June 1996 / edited by J.M. Drouffe and J.B. Zuber.
p. cm. -- (Advanced series in mathematical physics : vol. 24)
ISBN 9810228074 (alk. paper)
1. Mathematical physics -- Congresses. 2. Itzykson, Claude.
I. Itzykson, Claude. II. Drouffe, Jean-Michel. III. Zuber, Jean
Bernard. IV. Series.
QC19.2.M367 1997
530.15--dc21 97-3545
CIP

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

This book is printed on acid-free paper.

Printed in Singapore by Uto-Print

FOREWORD

To honour the memory of Claude Itzykson, who passed away on May 22, 1995, the Service de Physique Théorique at Saclay organised a conference entitled *The Mathematical Beauty of Physics.* It took place in Saclay on June 5–7, 1996 and was attended by more than 140 participants. It is intended to be the first of a series of annual meetings, dedicated to the memory of our distinguished friend and colleague.

The variety of interests of Claude Itzykson was reflected in the broad range of topics from mathematical physics and mathematics covered during the conference. The meeting consisted of seventeen lectures, fifteen of which are presented here. The order of presentation follows that of the proceedings. J. Fröhlich was unfortunately unable to provide us with written versions of his beautiful lecture. The proceedings also contain a contribution from E. Witten, who could not attend the conference, but kindly provided a text written in collaboration with N. Seiberg.

The organisors want to express their gratitude to all those who made this conference possible. We would like to thank Monsieur Robert Dautray, Haut Commissaire à l'Energie Atomique, who presided over the meeting and opened the first session of the conference. We also thank Madame C. Cesarsky, Directeur of the Direction des Sciences de la Matière, for her support of the project. The success of the conference was of course in large part due to the beautiful presentations. We would thus like to wholeheartedly thank all the invited speakers, together with N. Seiberg and E. Witten. Finally we want to thank the staff of the Service de Physique Théorique, A.-M. Arnold, J. Delouvrier, L. Dumets, M. Féron, B. Savelli and S. Zaffanella for the smooth running of the conference and M. Gingold for the preparation of these proceedings.

J.-M. Drouffe and J.-B. Zuber

Claude Itzykson
1938–1995

Claude Itzykson died of cancer on 22 May 1995 in Paris. French theoretical physics has lost one of its leaders and most flamboyant representatives.

He was born on 11 April 1938 in Paris. After the death of his father in a concentration camp during the Second World War, he was educated in an orphan's institution near Paris. His devouring passion for reading already impressed his friends there. Brilliant studies at the Lycée Condorcet, Paris, opened for him the doors to the Ecole Polytechnique, which he entered at the age of 19. There he graduated from the prestigious Corps des Mines. Having thus the opportunity of being elevated to a post in the higher french Civil Service, he declined and followed his passion for basic science joining the Commissariat à l'Energie Atomique, where he became a member of the Saclay theory group in 1963.

Itzykson's first research works, under the guidance of Maurice Jacob and Raymond Stora, dealt with particle physics, in the framework of the $SU(3)$ symmetry and of current algebra, leading to a thesis (1967) on non leptonic hyperon decays. Very quickly he demonstrated his deep knowledge of group theory, writing an article on the representations of unitary groups (still a very useful reference) and two beautiful papers on hidden symmetries of the hydrogen atom. In quantum electrodynamics he studied the problem of bound states and pair creation in a strong field. In a way quite characteristic of his style, these works start from a practical physical problem, and develop the appropriate mathematical framework in the most elegant manner.

From the middle of the seventies on, his work united in the most fruitful way concepts of quantum field theory and statistical mechanics. He immediately realized the fundamental and practical importance of the lattice discretization of gauge theories proposed by Wilson, exploring its implications by a variety of methods; mean field approximation and high and low temperature expansions. Simultaneously he investigated other non-perturbative approaches to quantum field theory: finding a characterization of large order behaviour in quantum electrodynamics and producing his seminal work on the "large N limit" of matrix field theory which was to pave the way for a major breakthrough, ten years later, in the understanding of two-dimensional quantum gravity.

Itzykson's active interest in disordered systems covered the geometry of random lattices and random surfaces, field theory on a random lattice, the

localization problem, the density of states and supersymmetry properties of electrons in a strong magnetic field in the presence of impurities. The comparison between spectra of chaotic and integrable systems lead him, also, to questions in number theory.

Most of his activity in the last ten years was focused on the study of conformal invariant quantum field theories in two dimensions and related mathematical issues. There he made numerous contributions to the classification of universality classes of two-dimensional systems, and to the study of the conformal (Virasoro) algebra, and its representations and extensions. He recently returned to integrals over large matrices, applying them to problems as diverse as classical integrable systems, the fractional quantum Hall effect and questions in "enumerative geometry", a branch of nineteenth century mathematics in which modern quantum field theory has recently led to unexpected and spectacular progress.

The majority of his more than 150 papers were written in collaborations in which he was always a major driving force, and in which his impetus and enthusiasm played a decisive role.

Itzykson's wide ranging knowledge and interests, and his passionate ability to communicate to students and young researchers, produced the classic text-book "Quantum Field Theory", McGraw Hill, a standard reference to almost a generation of young theorists. This was later complemented by the two volumes of "Statistical Field Theory", Cambridge University Press, which presented applications of field theory to statistical mechanics. Throughout his life he lectured in innumerable french and foreign institutions, and for this he was awarded the title of Chevalier de l'Ordre des Palmes Académiques. Itzykson was also awarded the Prix Langevin (1972) and Robin (1988) of the Société Française de Physique and the Prix Ampère (1995) of the Académie des Sciences.

The importance, elegance and depth of his work, as well as the diversity of themes are what make his contribution to science so remarkable. His vast scientific knowledge and intuition alongside his brilliant technical ability enabled him to find fruitful relationships between problems that at first sight seemed far apart. He also played a major role in bringing the French physical and mathematical communities closer together. His interests also included history and literature: he was particularly fond of eighteenth century French writers. Claude Itzykson was a man of immense scientific talent and great integrity, with a warm and charming personality, who inspired respect and admiration to the whole physics community. He will be greatly missed.

The Mathematical Beauty of Physics in memory of Claude Itzykson

Saclay, 5-7 June 1996

Wednesday June 5th, 1996

Chairman M. Jacob

9:30 Opening session, in memory of Claude Itzykson

10:30 *Pause*

10:45 E. Brézin, (Physique Théorique, E.N.S. Paris) *Random matrices in an external matrix source*

11:45 Ph. Di Francesco, (SPhT, Saclay) *Meanders*

Chairman M. Bershadsky

14:30 R. Stora, (LAPP, Annecy) *Exercises of equivariant cohomology and topological models*

15:30 T. Eguchi, (Tokyo) *Study of 4-dimensional superconformal field theories*

16:30 *Pause*

16:45 D. Zagier, (Max Planck Institut, Bonn) *Moduli spaces of curves*

Thursday June 6th,1996

Chairman A. Martin

9:30 G. Parisi, (La Sapienza, Rome) *Random matrices and the replica method*

10:30 J. Cardy, (Oxford) *Renormalisation group approach to reaction-diffusion problems*

11:30 *Pause*

11:45 B. Derrida, (Physique Statistique, E.N.S. Paris) *Exact solution of one-dimensional growth models*

Chairman I. Singer

14:30 A. Beauville, (Mathématiques, E.N.S. Paris) *Towards a Verlinde formula for non simply connected groups*

15:30 M. Bauer, (SPhT, Saclay) *Would Galois have liked conformal field theories?*

16:30 *Pause*

16:45 L. Alvarez-Gaumé, (Cern) *Softly broken $N = 2$ Quantum Chromodynamics*

Friday June 7th, 1996

Chairman D. Zwanziger

9:30 J.-M. Luck, (SPhT, Saclay) *From rational polygons to self-similar tilings of the plane*

10:30 M. Gutzwiller, (IBM, New York) *Physics and Arithmetic Chaos in the Fourier Transform*

11:30 *Pause*

11:45 M. Berry, (Bristol) *Arithmetic optics: the Talbot effect*

Chairman C. De Dominicis

14:30 H. Saleur, (USC, Los Angeles) *Quantum impurity problems in 1+1 dimensions*

15:30 M. Kontsevich, (IHES, Bures-sur-Yvette) *On Lyapunov exponents and Hodge theory*

16:30 *Pause*

16:45 J. Fröhlich, (ETH, Zurich) *What light and (non-relativistic) matter teach us about renormalization, differential topology and differential geometry.*

17:45 *Conclusion*

DYSON'S UNIVERSALITY IN GENERALIZED ENSEMBLES OF RANDOM MATRICES

E. BRÉZIN
Laboratoire de Physique Théorique, Ecole Normale Supérieure[a]
24 rue Lhomond 75231, Paris Cedex 05, France

To Claude, the physicist, the unforgettable friend, with grief

We consider generalisations of ensembles of random matrices in which the Hamiltonian H is the sum of a deterministic part H_0 and of a Gaussian random potential V. The standard methods of the theory of random matrices, such as the method of orthogonal polynomials, are not available for such cases. We first analyse the density of levels; then the level correlations and verify that, at short distance, they are independent of the spectrum of H_0. This is another aspect of the universality discussed by Dyson (for zero H_0) who conjectured that these correlations were independent of the probability distribution of V. We follow in this work a method introduced by Kazakov, relying on the Itzykson-Zuber integral, which leads to a representation of the correlation functions for finite $N \times N$ matrices in terms of contour integrals over a finite number of variables. This article is based on joint work with Hikami[1].

1 Introduction

Let us first recall the results for the correlations between two eigenvalues for the simple unitary ensemble, in which the full Hamiltonian is treated as random. In the simplest Gaussian ensemble (GUE) one considers $N \times N$ random Hermitian matrices H with probability distribution

$$P(H) = \frac{1}{Z}\exp(-\frac{N}{2}\mathrm{Tr}H^2) \tag{1}$$

The density of eigenvalues and the two-level correlation function are defined as

$$\rho(\lambda) = < \frac{1}{N}\mathrm{Tr}\delta(\lambda - H) > \tag{2}$$

and

$$\rho^{(2)}(\lambda,\mu) = < \frac{1}{N}\mathrm{Tr}\delta(\lambda - H)\frac{1}{N}\mathrm{Tr}\delta(\mu - H) > \tag{3}$$

The correlation function, when λ and μ are arbitrary, has a complicated oscillatory behavior, even for the simplest Gaussian distribution. It simplifies

[a]Unité propre du Centre National de la Recherche Scientifique, Associée à l'Ecole Normale Supérieure et à l'Université de Paris-Sud

when $\lambda - \mu$ is small, N is large, and the scaling variable

$$x = \pi N(\lambda - \mu)\rho(\frac{1}{2}(\lambda + \mu)) \tag{4}$$

is held finite. Then one finds[2]

$$\begin{aligned} \rho_c^{(2)}(\lambda, \mu) &\simeq \frac{1}{N}\delta(\lambda - \mu)\rho(\lambda) - \rho(\lambda)\rho(\mu)\frac{\sin^2 x}{x^2} \\ &\simeq \frac{1}{N}\delta(\lambda - \mu)\rho(\lambda) - \frac{1}{\pi^2 N^2}\frac{\sin^2[\pi N(\lambda - \mu)\rho(\frac{\lambda+\mu}{2})]}{(\lambda - \mu)^2} \end{aligned} \tag{5}$$

$$\rho(\lambda) = < \frac{1}{N}\mathrm{Tr}\delta(\lambda - H) > \tag{6}$$

and

$$\rho^{(2)}(\lambda, \mu) = < \frac{1}{N}\mathrm{Tr}\delta(\lambda - H)\frac{1}{N}\mathrm{Tr}\delta(\mu - H) > \tag{7}$$

For a non-Gaussian probability distribution for H, the density of eigenvalues is no longer given by a semi-circle law; for the correlations between two levels two kinds of universal correlations between eigenvalues are known to be present : a) a short-distance universal oscillatory behavior; b) a finite distance universality of smoothed correlations.

Let us review these two properties. a) in the scaling regime defined by (4) one recovers universally the result (5). b) Away from this short-distance region, for arbitrary λ and μ, the correlations simplify only if one "smooths" the oscillations. This is what one usually does, if one lets N go to infinity first in the resolvent, before returning to the real axis. The result, which is known to be universal, is [7,3]

$$\rho_c^{(2)}(\lambda, \mu) = -\frac{1}{2N^2\pi^2}\frac{1}{(\lambda - \mu)^2}\frac{(a^2 - \lambda\mu)}{[(a^2 - \lambda^2)(a^2 - \mu^2)]^{1/2}} \tag{8}$$

where a is an end point of the support.

There are many equivalent derivations of the property b). They are based either on orthogonal polynomials[3], or on summing over planar diagrams [4,5], or solving an integral equation [6,7]; however the property a) is known only through the orthogonal polynomials approach [3]. For the generalization that we have in mind here, in which the "unperturbed" part of the Hamiltonian is deterministic, if again for b) a diagrammatic approach still works [4,5,8,9], we are not aware of any method which would allow us to study whether a) still holds. To this effect we shall generalize a method, introduced by Kazakov[10], to the study of correlation functions. It consists of introducing an external matrix

source. It leads to an exact representation of the correlation function for finite N in terms of contour integrals over two variables [1]. From now on we shall consider a Hamiltonian $H = H_0 + V$, where H_0 is deterministic and V is a random $N \times N$ matrix. The Gaussian distribution P is given by

$$\begin{aligned} P(H) &= \frac{1}{Z} e^{-\frac{N}{2}\mathrm{Tr}V^2} \\ &= \frac{1}{Z} e^{-\frac{N}{2}\mathrm{Tr}(H^2 - 2H_0H + H_0^2)} \end{aligned} \tag{9}$$

We are thus simply dealing with a Gaussian unitary ensemble modified by a matrix source H_0. Up to a factor the probability distribution for H is thus

$$P(H) = \frac{1}{Z}\exp(-\frac{N}{2}\mathrm{Tr}H^2 + N\mathrm{Tr}H_0H) \tag{10}$$

2 Density of states

Let us first show how one deals with the density of states $\rho(\lambda)$. It is the Fourier transform of the average "evolution" operator

$$U(t) = < \frac{1}{N}\mathrm{Tr}e^{itH} > \tag{11}$$

and $\rho(\lambda)$ is

$$\rho(\lambda) = \int_{-\infty}^{+\infty} \frac{dt}{2\pi} e^{-it\lambda} U(t) \tag{12}$$

We integrate first over the unitary matrix ω which diagonalizes H, and without loss of generality we may assume that H_0 is a diagonal matrix with eigenvalues $(\epsilon_1, \cdots, \epsilon_N)$. This is done by the well-known Itzykson-Zuber integral [12],

$$\int d\omega \exp(\mathrm{Tr}A\omega B\omega^\dagger) = \frac{\det(\exp(a_i b_j))}{\Delta(A)\Delta(B)} \tag{13}$$

where $\Delta(A)$ is the Van der Monde determinant constructed with the eigenvalues of A:

$$\Delta(A) = \prod_{i<j}^{N} (a_i - a_j) \tag{14}$$

We are then led to

$$\begin{aligned} U(t) = & \quad \frac{1}{Z\Delta(H_0)} \frac{1}{N} \sum_{\alpha=1}^{N} \int d\lambda_1 \cdots d\lambda_N e^{it\lambda_\alpha} \Delta(\lambda_1, \cdots, \lambda_N) \\ & \times \quad \exp(-\frac{N}{2}\sum \lambda_i^2 + N \sum \epsilon_i \lambda_i) \end{aligned} \tag{15}$$

The normalization is

$$U(0) = 1 \tag{16}$$

The integration over the λ_i may be done easily, if we note that

$$\begin{aligned} \int \; & d\lambda_1 \cdots d\lambda_N \Delta(\lambda_1, \cdots, \lambda_N) \exp(-\frac{N}{2} \sum \lambda_i^2 + N \sum a_i \lambda_i) \\ = \; & \Delta(a_1, \cdots, a_N) e^{\frac{N}{2} \sum a_i^2} \end{aligned} \tag{17}$$

If we use this, with

$$a_i = \epsilon_i + \frac{it}{N} \delta_{\alpha,i} \tag{18}$$

we obtain

$$U(t) = \frac{1}{N} \sum_{\alpha=1}^{N} \prod_{\gamma \neq \alpha} \left(\frac{\epsilon_\alpha - \epsilon_\gamma + \frac{it}{N}}{\epsilon_\alpha - \epsilon_\gamma} \right) e^{-\frac{t^2}{2N} + it\epsilon_\alpha} \tag{19}$$

The sum over N terms may be replaced by a contour-integral in the complex u plane,

$$U(t) = \frac{1}{it} \oint \frac{du}{2\pi i} \prod_{\gamma=1}^{N} \left(\frac{u - \epsilon_\gamma + \frac{it}{N}}{u - \epsilon_\gamma} \right) e^{itu - \frac{t^2}{2N}} \tag{20}$$

The contour of integration encloses all the eigenvalues ϵ_γ. Note that we would recover the simple Wigner ensemble if we let all the ϵ_γ go to zero; we then obtain

$$U_0(t) = \frac{1}{it} e^{-\frac{t^2}{2N}} \oint \frac{du}{2\pi i} e^{itu} (1 + \frac{it}{Nu})^N \tag{21}$$

From this exact representation (21) for finite N, it is immediate to recover all the well-known properties, the semi-circle law, or the more subtle edge behavior of the density of states. Let us do that here as simple preliminary exercices.

Semi circle law.
For large N, finite t, $U_0(t)$ has the limit:

$$U_0(t) = \frac{1}{it} \oint \frac{du}{2\pi i} e^{itu + \frac{it}{u}} \tag{22}$$

i.e.

$$U_0(t) = \frac{1}{it} \int_{-\pi}^{+\pi} \frac{d\alpha}{2\pi} e^{i\alpha + 2it\cos\alpha} \tag{23}$$

From there we obtain immediately

$$\rho_0(\lambda) = \int_{-\infty}^{+\infty} \frac{dt}{2\pi} U_0(t) e^{-it\lambda} = -\int_{-\pi}^{+\pi} \frac{d\alpha}{2\pi} e^{i\alpha} \Theta(\lambda - 2cos\alpha) \qquad (24)$$

We thus recover the semi-circle law

$$\rho_0(\lambda) = \frac{1}{\pi} \Theta(4 - \lambda^2) \sqrt{1 - \frac{\lambda^2}{4}} \qquad (25)$$

Edge cross-over.

We can easily recover from the above representation the behaviour of the density of eigenvalues near the edge of the semi-circle. In (25) N has gone to infinity first and ρ_0 vanishes outside the semi-circle. This limit is approached with exponentially small corrections, of the type e^{-N}, for $\lambda^2 > 4$, and with $1/N^2$-corrections for $\lambda^2 < 4$. Near the edge, $\lambda = 2$ for instance, there is a cross-over region of size $N^{-2/3}$ between these two regimes. The characterization of this cross-over is obtained easily with (21) which leads to

$$\rho_0(\lambda) = \int \frac{dt}{2\pi} \frac{1}{it} e^{-it\lambda} \oint \frac{du}{2\pi i} e^{itu - \frac{t^2}{2N}} (1 + \frac{it}{Nu})^N. \qquad (26)$$

We change t to Nt, then t to $t + iu$, and find

$$\frac{d\rho_0(\lambda)}{d\lambda} = -\int \frac{dt}{2\pi} e^{-NS} \qquad (27)$$

with

$$S = \frac{t^2}{2} + \frac{u^2}{2} + it\lambda - u\lambda - Log(\frac{it}{u}). \qquad (28)$$

The large N limit is thus given by a saddle-point in the t-u plane; however it is easy to see that for $\lambda = 2$, two saddle-points merge at $u = 1, t = -i$ and the expansion near the saddle-point has to go beyond Gaussian order. Defining

$$\begin{aligned} \lambda &= 2 + N^{-2/3}\delta \\ u &= 1 + N^{-1/3}x \\ t &= -i + N^{-1/3}y \end{aligned} \qquad (29)$$

we find that the result is proportional to the square of an Airy function:

$$\frac{d\rho_0(\lambda)}{d\lambda} = -N^{-2/3}(\int_0^\infty \frac{dx}{\pi} cos(\delta x + \frac{x^3}{3}))^2 = -N^{-2/3}(Ai(\delta))^2 \ . \qquad (30)$$

Large N limit of the density of states

For arbitrary H_0, the density of state $\rho(\lambda)$ was first found, in the large N limit, by Pastur[13]. The result may be easily recovered by summing planar diagrams, which are here simple"rainbow" diagrams. It follows immediately that the self-energy is proportional to the Green function itself in the large N limit[4], and this leads at once to Pastur's result. From the contour-integral representation (20), let us show how to recover this result . The average resolvent $G(z)$ is written in terms of the evolution operator as

$$\begin{aligned} G(z) &= < \frac{1}{N} \mathrm{Tr} \frac{1}{z-H} > \\ &= i \int_0^{+\infty} dt e^{-itz} U(t) \end{aligned} \tag{31}$$

We substitute (20) for $U(t)$ and replace the product

$$\prod_{\gamma=1}^{N} \left(1 + \frac{it}{N(u-\epsilon_\gamma)}\right) = \exp \sum_{\gamma=1}^{N} \mathrm{Log} \left(1 + \frac{it}{N(u-\epsilon_\gamma)}\right) \tag{32}$$

by its leading term in the large N limit, namely

$$\exp(\frac{it}{N} \sum_{\gamma=1}^{N} \frac{1}{u-\epsilon_\gamma}) \tag{33}$$

If we define the density of states of the unperturbed matrix H_0

$$\rho_0(\epsilon) = \frac{1}{N} \sum_{\alpha=1}^{N} \delta(\epsilon - \epsilon_\alpha) \tag{34}$$

we may write this expression (32) as

$$\exp(it \int d\epsilon \frac{\rho_0(\epsilon)}{u-\epsilon}) \tag{35}$$

We then easily obtain

$$\frac{\partial G}{\partial z} = \oint \frac{du}{2\pi i} \frac{1}{u + G_0(u) - z} \tag{36}$$

We have now to specify the contour of integration in the complex u-plane. It surrounds all the eigenvalues of H_0 and we have to determine the location

of the zeroes of the denominator with respect to this contour. Let us return to the discrete form for the equation

$$u + G_0(u) = z \tag{37}$$

i.e.

$$u + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{u-\epsilon_i} = z \tag{38}$$

which possesses $(N+1)$ real or complex roots in the u-plane. For z real and large, N of these roots are close to the ϵ_i and one, which will be denoted $\hat{u}(z)$, goes to infinity with z as

$$\hat{u}(z) = z - \frac{1}{z} + O\left(\frac{1}{z^2}\right) \tag{39}$$

Therefore, for large z, the contour encloses all the roots of (38) except $\hat{u}(z)$. When z decreases the contour should not be crossed by any other root of the equation, therefore it is defined by the requirement that only one root remains at its exterior. Therefore it is easier to calculate the integral (36) by taking the residues of the singularities outside of the contour, rather than the N poles enclosed by this contour. There are two of them outside; one is $\hat{u}(z)$ and the other one is at infinity (since for large u, $G_0(u)$ vanishes). Taking these two singularities we obtain

$$\begin{aligned}\frac{\partial G}{\partial z} &= 1 - \frac{1}{1+\frac{dG_0}{d\hat{u}(z)}} \\ &= 1 - \frac{d\hat{u}(z)}{dz}\end{aligned} \tag{40}$$

The integration gives

$$G(z) = z - \hat{u}(z) \tag{41}$$

(there is no integration constant since $G(z)$ vanishes for z large; note that it does behave as it should as $\frac{1}{z}$ for z large). This combined with (37) gives Pastur's self-consistent relation

$$G(z) = G_0(z - G(z)) \tag{42}$$

3 Two-level correlation function

For the two-level correlation function, $\rho^{(2)}(\lambda,\mu)$ is obtained from the Fourier transform $U(t_1,t_2)$,

$$\rho^{(2)}(\lambda,\mu) = \int\int \frac{dt_1 dt_2}{(2\pi)^2} e^{-it_1\lambda - it_2\mu} U(t_1,t_2) \tag{43}$$

where $U(t_1, t_2)$ is

$$U(t_1, t_2) = < \frac{1}{N} \mathrm{Tr} e^{it_1 H} \frac{1}{N} \mathrm{Tr} e^{it_2 H} > \tag{44}$$

The normalization conditions are

$$\begin{aligned} U(t_1, t_2) &= U(t_2, t_1) \\ U(t_1, 0) &= U(t_1) \\ U(0) &= 1 \end{aligned} \tag{45}$$

Dealing with $U(t_1, t_2)$ is also simple. After performing the Itzykson-Zuber integral over the unitary group as in (13), we obtain through the same procedure,

$$U^{(2)}(t_1, t_2) = \frac{1}{N^2} \sum_{\alpha_1, \alpha_2 = 1}^{N} \int \prod_{i=1}^{N} dr_i \frac{\Delta(r)}{\Delta(H_0)} e^{-N \sum (\frac{1}{2} r_i^2 + r_i \epsilon_i) + i(t_1 r_{\alpha_1} + t_2 r_{\alpha_2})} \tag{46}$$

After integration over the r_i, we obtain

$$U(t_1, t_2) = \frac{1}{N^2} \sum_{\alpha_1, \alpha_2} \frac{\prod_{i<j} (\epsilon_i - \epsilon_j + \frac{it_1}{N}(\delta_{i,\alpha_1} - \delta_{j,\alpha_1}) + \frac{it_2}{N}(\delta_{i,\alpha_2} - \delta_{j,\alpha_2}))}{\prod_{i<j} (\epsilon_i - \epsilon_j)}$$

$$\times e^{-it_1 \epsilon_{\alpha_1} - it_2 \epsilon_{\alpha_2} - \frac{t_1^2}{2N} - \frac{t_2^2}{2N} - \frac{t_1 t_2}{N} \delta_{\alpha_1, \alpha_2}} \tag{47}$$

The terms of this double sum in which $\alpha_1 = \alpha_2$ are written as a single contour integral and their sum is simply $\frac{1}{N} U(t_1 + t_2)$ of (19). The Fourier transform of this term becomes

$$\frac{1}{N(2\pi)^2} \int \int dt_1 dt_2 e^{-it_1 \lambda - it_2 \mu} U(t_1 + t_2) = \frac{1}{N} \delta(\lambda - \mu) \rho(\lambda) \tag{48}$$

The remaining part, after the subtraction of the disconnected part, becomes

$$\begin{aligned} U_c(t_1, t_2) \quad &= -\frac{1}{N^2} \oint \frac{du dv}{(2\pi i)^2} e^{-\frac{t_1^2}{2N} - \frac{t_2^2}{2N} - it_1 u - it_2 v} \frac{1}{(u - v + \frac{it_1}{N})(u - v - \frac{it_2}{N})} \\ &\times \prod_{\gamma=1}^{N} (1 + \frac{it_1}{N(u - \epsilon_\gamma)})(1 + \frac{it_2}{N(v - \epsilon_\gamma)}) \end{aligned} \tag{49}$$

where the contours are taken around $u = \epsilon_\gamma$ and $v = \epsilon_\gamma$. If we include also the contour-integration around the pole, $v = u + \frac{it_1}{N}$, this gives precisely the term $U(t_1 + t_2)$ of (27), which contributes to the delta-function part. This coincidence had already been noticed for the Laguerre ensemble [11].

4 Dyson's universality

We now consider the correlation function in the large N limit for nearby levels. In the integral representation (49) we may neglect the terms t^2/N in the large N limit and replace the products as in (35). This gives the large N limit of $U^{(2)}(t_1,t_2)$ as

$$U^{(2)}(t_1,t_2) = -\frac{1}{N^2}\oint \frac{dudv}{(2\pi i)^2}\frac{1}{(u-v)^2} e^{-it_1(u+\int \frac{\rho_0(\epsilon)}{u-\epsilon}d\epsilon)-it_2(v+\int \frac{\rho_0(\epsilon)}{v-\epsilon}d\epsilon)} \quad (50)$$

Noting that

$$\frac{\partial^2}{\partial z_1 \partial z_2}\ln[\hat{u}(z_1)-\hat{u}(z_2)] = \frac{1}{(\hat{u}(z_1)-\hat{u}(z_2))^2}\frac{d\hat{u}}{dz_1}\frac{d\hat{u}}{dz_2} \quad (51)$$

we obtain, through identical steps, the connected two-particle Green function

$$\begin{aligned} G_c^{(2)}(z_1,z_2) &= < \frac{1}{N}\mathrm{tr}\frac{1}{z_1-H}\frac{1}{N}\mathrm{tr}\frac{1}{z_2-H} >_c \\ &= -\frac{1}{N^2}\frac{\partial^2}{\partial z_1 \partial z_2}\ln[\hat{u}(z_1)-\hat{u}(z_2)] \end{aligned} \quad (52)$$

This result was derived earlier by diagrammatic methods [4], and was used to show that the singularity of the correlations, obtained when z_1 and z_2 approach the real axis with opposite imaginary parts, is universal.

However if we want to study the correlation function in the short-distance limit, we cannot use the resolvent any more (since we need to let the imaginary parts of z_1, z_2 go to zero before N goes to infinity).

Returning then to (28), and making the shifts, $t_1 \to t_1 + iuN$, and $t_2 \to t_2 + ivN$, the two-level correlation function is remarkably factorized since,

$$\begin{aligned} \rho_c(\lambda_1,\lambda_2) &= \int \frac{dt_1}{2\pi}\oint \frac{dv}{2\pi i}\prod_{\gamma=1}^{N}(\frac{\epsilon_\gamma+\frac{it_1}{N}}{v-\epsilon_\gamma})\frac{1}{v+\frac{it_1}{N}} e^{-\frac{N}{2}v^2-\frac{t_1^2}{2N}-it_1\lambda_1-Nv\lambda_2} \\ &\times \int \frac{dt_2}{2\pi}\oint \frac{du}{2\pi i}\prod_{\gamma=1}^{N}(\frac{\epsilon_\gamma+\frac{it_2}{N}}{u-\epsilon_\gamma})\frac{1}{u+\frac{it_2}{N}} e^{-\frac{N}{2}u^2-\frac{t_2^2}{2N}-it_2\lambda_2-Nu\lambda_1} \\ &= -K_N(\lambda_1,\lambda_2)K_N(\lambda_2,\lambda_1) \end{aligned} \quad (53)$$

This kernel $K_N(\lambda_1,\lambda_2)$ is further simplified by the shift $t_1 \to t + ivN$,

$$K_N(\lambda_1,\lambda_2) = \int \frac{dt}{2\pi}\oint \frac{dv}{2\pi i}\frac{1}{it}\prod_{\gamma=1}^{N}(1-\frac{it}{N(v-a_\gamma)}) e^{-\frac{t^2}{2N}-ivt-it\lambda_1+Nv(\lambda_1-\lambda_2)} \quad (54)$$

Note that $K_N(\lambda_1, \lambda_1)$ reduces to the density of states. We replace again the product in (53) by its large N limit, neglect $\frac{t^2}{N}$ and integrate over t, leading to

$$\frac{\partial K_N}{\partial \lambda_1} = \frac{1}{\pi} \mathrm{Im} \oint \frac{du}{2\pi i} \frac{1}{u + G_0(u) - \lambda_1 + i\epsilon} e^{-uy} \tag{55}$$

with $y = N(\lambda_1 - \lambda_2)$. Therefore

$$\begin{aligned} \frac{\partial K_N}{\partial \lambda_1} &= \frac{1}{\pi} \mathrm{Im} \frac{d\hat{u}}{d\lambda_1} e^{-y\hat{u}(\lambda_1 - i\epsilon)} \\ &= -\frac{1}{\pi y} \frac{\partial}{\partial \lambda_1} \mathrm{Im} \left(e^{-y\hat{u}(\lambda_1 - i\epsilon)} \right) \end{aligned} \tag{56}$$

Since, from (41),

$$\hat{u}(\lambda_1 - i\epsilon) = \lambda_1 - \mathrm{Re}G(\lambda_1) - i\pi\rho(\lambda_1) \tag{57}$$

we obtain

$$K_N(\lambda_1, \lambda_2) = -\frac{1}{\pi y} e^{-y[\lambda_1 - \mathrm{Re}G(\lambda_1)]} \sin[\pi y \rho(\lambda_1)] \tag{58}$$

Repeating this calculation for $K_N(\lambda_2, \lambda_1)$ we end up, in the large N, finite y limit, with

$$\rho_c(\lambda_1, \lambda_2) = -\frac{1}{\pi^2 y^2} \sin^2[\pi \rho(\frac{\lambda_1 + \lambda_2}{2}) y] \tag{59}$$

Note that this result is independent of H_0 (apart from the scale factor present in the densitty of states). In the case in which H_0 vanishes it is also independent of the probability distribution of V [3].

It is thus natural to conjecture that Dyson's short-distance universality with respect to the probability distribution of V remains true for H_0 non-zero as well, but we do not know how prove it.

1. E. Brézin and S.Hikami, preprint
2. M. L. Mehta, Random matrices, 2nd ed. (Academic Press, New York 1991).
3. E. Brézin and A. Zee, Nucl. Phys. **B 402** (1993) 613.
4. E. Brézin and A. Zee, Phys. Rev. **E 49** (1994) 2588.
5. E. Brézin, S. Hikami and A. Zee, Phys. Rev **E 51** (1995) 5442.
6. C. W. J. Beenakker, Nucl. Phys. **B 422**, 515 (1994).
7. J. Ambjorn and Yu. M. Makeenko, Mod. Phys. Lett. **A 5**, 1753 (1990).
8. E. Brézin and A. Zee, Nucl. Phys. **B 453** (1995) 531.

9. A. Zee, a preprint NSF-ITP-96-12, cond-mat/9602146.
10. V. A. Kazakov, Nucl. Phys. **B 354** (1991) 614.
11. E. Brézin, S. Hikami and A. Zee, Nucl. Phys. **B 464** (1996) 411.
12. C. Itzykson and J. -B. Zuber, J. Math. Phys. **21** (1980) 411.
13. L. A. Pastur, Theor. Math. Phys. (USSR) **10**, 67 (1972).

MEANDERS

P. Di FRANCESCO,

O. GOLINELLI

and

E. GUITTER*,

CEA/Saclay, Service de Physique Théorique
F-91191 Gif sur Yvette, France

1. Introduction

The meander problem is one of these apparently very simple problems which resist all attempts to solve them. A fascinating problem which could not go unnoticed with Claude Itzykson. He indeed kept encouraging us at the early stage of this work, even providing us with some mathematical references which were the real starting point of our study. This note is intended as an account of the earlier and latest developments towards a solution of the problem, yet to be invented.

The meander problem is a simply stated combinatorial question: count the number of configurations of a closed non-self-intersecting road crossing an infinite river through a given number of bridges. Despite its apparent simplicity, this problem still awaits a solution, if only for asymptotics when the number of bridges is large. The problem emerged in various contexts ranging from mathematics to computer science [1]. In particular, Arnold re-actualized it in connection with Hilbert's 16th problem, namely the enumeration of ovals of planar algebraic curves [2], and it also appears in the classification of 3-manifolds [3].

Remarkably, the meander problem can be rephrased in the physical language of critical phenomena, through its equivalence with a particular problem of Self-Avoiding Walks: the counting of the compact foldings of a linear chain.

Several techniques have been applied to this problem: direct combinatorial approaches [4] [5], random matrix model techniques [6] [7] [8], an algebraic approach using the Temperley-Lieb algebra and Restricted Solid-On-Solid models [9].

* e-mails: philippe,golinel,guitter@spht.saclay.cea.fr

This note is organized as follows. In Sect.2, we define precisely the meander (resp. semi-meander) counting problems, arising in the context of closed (resp. open) chain-folding, and solve them in some simple cases. Sect.3 is an overview of various reformulations of the problem in physical or mathematical terms: the matrix model formulation, which provides us with a complete recursive scheme to compute the meander and semi-meander partition functions, including their higher genus generalizations; the symmetric group formulation, which eventually leads to some compact expressions in terms of the symmetric group characters; the Temperley-Lieb algebra formulation, which gives yet another, completely algebraic viewpoint on the problem. Sect.4 is dedicated to a more direct *enumerative* approach and a thorough analysis of its results in the spirit of critical phenomena. The semi-meander problem is generalized to include the case of several non-intersecting but possibly interlocking roads with a weight q per road, and crossing the river through a total of n bridges. The corresponding generating functions are analyzed as functions of q, through large n extrapolations, and through their large q asymptotic expansion in powers of $1/q$, for $n \to \infty$. Evidence is given for a phase transition for semi-meanders at a value of $q = q_c \simeq 2$ between a low-q and a large-q regimes, discriminated by the relevance of winding of the roads around the source. The large-q expansion provides an accurate description of the whole $q > q_c$ phase. We gather conclusions and a few conjectures in Sect.5.

2. The meander problem

2.1. Definitions, observables

A *meander* of order n is a planar configuration of a non-self-intersecting loop (road) crossing a line (river), through a given number $2n$ of points (bridges). We consider as equivalent any two configurations which may be continuously deformed into each other, keeping the river fixed (this is therefore a topological equivalence). The number of inequivalent meanders of order n is denoted by M_n. For instance, we have $M_1 = 1$, $M_2 = 2$, $M_3 = 8$... More numbers can be found in [6] [7] [12].

We stumbled on the meander problem by trying to enumerate the distinct *compact folding* configurations of a closed polymer, i.e. the different ways of folding a closed chain of $2n$ identical constituents onto itself. The best image of such a closed polymer is that of a closed strip of $2n$ identical stamps, attached by their edges, serving as hinges in the

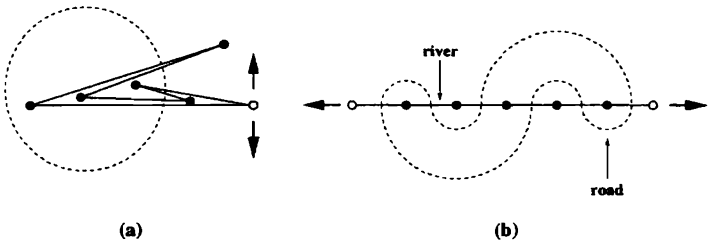


Fig. 1: The mapping between compactly folded closed strip of stamps and meanders. We display a compact folding configuration (a) of a closed strip with $2n = 6$ stamps. To transform it into a meander, first draw a (dotted) line through the centers of the stamps and close it to the left of the picture. Then cut the bottom right hinge (empty circle) and pull its ends apart as indicated by the arrows, so as to form a straight line (b): the straight line forms the river, and the dashed line the road of the resulting meander.

folding process: a compactly folded configuration of the strip is simply a folded state in which all the stamps are piled up on top of one of them.

Such a compactly folded configuration is easily identified with a meander configuration as depicted in Fig.1. Draw a closed line (road) passing though the centers (bridges) of all the piled-up monomers, then open one hinge of the polymer (we choose to always open the bottom right one) and pull the stamps apart so as to form a straight line: the latter is identified with the river, whereas the distorted line becomes the road of the resulting meander.

Fig. 2: The 4 inequivalent foldings of a strip of 3 stamps. The fixed stamp is indicated by the empty circle: it is attached to a support (shaded area). The other circles correspond to the edges of the stamps.

When the strip of stamps is open (see Fig.2), we decide to attach the first stamp to a support, preventing the strip from winding around it, while the last stamp has a free extremal edge. In this case, a slightly generalized transformation maps any compactly folded open configuration of $(n-1)$ stamps to what we will call a *semi-meander* configuration of order n, in the following manner.

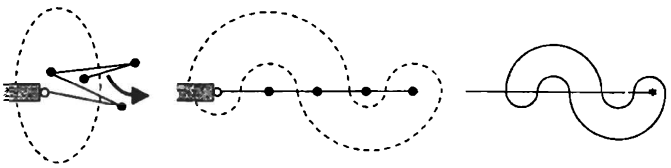


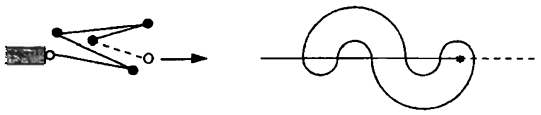
Fig. 3: The mapping of a compactly folded configuration of 4 stamps onto a semi-meander of order 5. (a) draw a (dashed) curve through the pile of stamps and the (shaded) support. (b) pull the free edge of the last stamp to form a half-line (the river with a source). (c) the result is a semi-meander configuration of order 5, namely that of a road, crossing a semi-infinite river through 5 bridges (the source of the river, around which the road is free to wind, is indicated by an asterisk).

As shown in Fig.3, draw a curve (road) though the $(n-1)$ centers (bridges) of all the piled-up stamps, then close this curve across the support (this last intersection is the n-th bridge), and pull the free edge of the last stamp in order to form a straight half-line (river with a source). The resulting picture is a configuration of a road (the curve) crossing a semi-infinite river (stamps and support) through n bridges: this is called a semi-meander configuration of order n. Note that the road in a semi-meander may wind freely around the source of the river, and that consequently the number of bridges may be indifferently even or odd, as opposed to meanders. The number of distinct semi-meanders of order n is denoted by $\bar{M}_n$. For instance, we have $\bar{M}_1 = 1$, $\bar{M}_2 = 1$, $\bar{M}_3 = 2$, $\bar{M}_4 = 4$... More numbers can be found in [4] [7] and in appendix A.

Through its compact folding formulation, the semi-meander problem is a particular reduction of the two-dimensional self-avoiding walk problem, in which only topological constraints are retained. It is therefore natural to define, by analogy with self-avoiding walks the connectivity $\bar{R}$ per stamp and the configuration exponent γ which determine the large n behavior of the semi-meander numbers as follows[1]

$$\bar{M}_n \sim \bar{c}\,\frac{\bar{R}^n}{n^\gamma} \tag{2.1}$$

[1] That the semi-meander numbers $\bar{M}_n$ actually have these leading asymptotics may be proved by deriving upper and lower bounds on $\bar{R}$. See [7] for further details.



(a) (b)

Fig. 4: The "end-to-end distance" of the folded strip of stamps (a) is the number ($w = 1$ here) of stamps to be added to the strip (the added stamp is represented in dashed line), so that the new free end (empty circle) is in contact with the infinity to the right. This coincides with the "winding" of the corresponding semi-meander (b), namely the number of bridges to be added if we continue the river to the right of its source (dashed line).

The connectivity $\tilde{R}$ may be interpreted as the average number of possibilities of adding one stamp to the folded configurations. The exponent γ is characteristic of the (open) boundary condition on the strip of stamps.

A natural observable for self-avoiding walks is the end-to-end distance. The corresponding notion for a compactly folded open strip of stamps is the "distance" between the free end of the strip and, say the support. This distance should also indicate how far the end of the strip is buried inside the folded configuration. It is defined as the minimal length w of a strip of stamps to be attached to the free end, such that a resulting folding with $n - 1 + w$ stamps has its free end outside of the folding, namely can be connected to the infinity to the right of the folding by a half-line which does not intersect any stamp. Indeed, the infinity to the right can be viewed as the nearest topological neighbor of the support, hence w measures a distance from the free end of the strip to the support. This is illustrated in Fig.4(a), with $n = 5$ and $w = 1$. In the semi-meander formulation (see Fig.4(b)), this distance w is simply the *winding* of the road around the source of the river, namely the number of bridges to be added if we continue the river to the right of its source. By analogy with self-avoiding walks, we expect the average winding over all the semi-meanders of order n to have the leading behavior

$$\langle w \rangle_n \equiv \frac{1}{\tilde{M}_n} \sum_{\text{semi-meanders}} w \sim n^{\nu} \tag{2.2}$$

where ν is some positive (end-to-end) exponent $0 \leq \nu \leq 1$, as w is always smaller or equal to n.

In this language, a meander of order n is simply a semi-meander of order $2n$ with winding $w = 0$. By analogy with closed (as compared to open) self-avoiding walks, we expect the asymptotics

$$M_n \sim c \frac{R^{2n}}{n^{\alpha}} \tag{2.3}$$

where the connectivity per bridge R is the same as that for semi-meanders (2.1), $R = \bar{R}$, and the configuration exponent $\alpha \neq \gamma$ is characteristic of the closed boundary condition on the strip of stamps.

In the following, we will mainly focus our study on the semi-meander numbers.

2.2. Arches and connected components

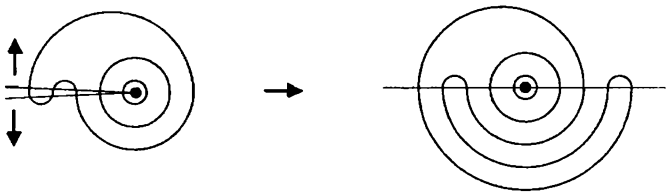


Fig. 5: A semi-meander viewed as a particular meander: the semi-infinite river must be opened up as indicated by the arrows. This doubles the number of bridges in the resulting meander, hence the order is conserved ($n = 5$ here). By construction, the lower arch configuration of the meander is always a rainbow arch configuration of same order.

Any semi-meander may be viewed as a particular meander by opening the semi-infinite river as indicated by the arrows on Fig.5. In the process, the number of bridges is doubled, hence the order is conserved. The resulting meander however is very peculiar. Note that in general a meander is made of an upper (resp. lower) configuration consisting of non-intersecting arches (arcs of road) connecting the bridges by pairs above (resp. below) the river. In the present case the lower configuration is fixed: it is called the rainbow arch configuration of order n (the bridge i is connected to the bridge $(2n-i+1)$, $i = 1, 2, ..., n$). On the other hand, the upper arch configuration may take any of the $\bar{M}_n$ values leading to semi-meanders of order n.

There are however

$$c_n = \frac{(2n)!}{n!(n+1)!} \tag{2.4}$$

distinct arch configurations of order n [7], as is readily proved by recursion ($c_{n+1} = \sum_{0 \leq j \leq n} c_j c_{n-j}$, with $c_0 = 1$, hence $c_1 = 1$, $c_2 = 2$, $c_3 = 5$, $c_4 = 14$,...: the c_n are called the Catalan numbers). Hence not all upper arch configurations, once supplemented by a lower rainbow arch configuration of same order, lead to an opened semi-meander ($\bar{M}_n < c_n$). This is because, in general, the corresponding object will have $k \geq 1$ connected components: we call it a semi-meander of order n with k connected components. Indeed, if the river is folded back into a semi-infinite one, we are simply left with a collection of k possibly interlocking semi-meanders of respective orders n_1, n_2,..., n_k, with $n_1 + n_2 + ... + n_k = n$. We always have $1 \leq k \leq n$, and $k = n$ only for the superposition of an upper and a lower rainbow configurations, leading to $2n$ concentric circles in the open river picture. We denote by $\bar{M}_n^{(k)}$ the number of inequivalent semi-meanders of order n with k connected components. In particular, we have $\bar{M}_n^{(1)} = \bar{M}_n$ and $\bar{M}_n^{(n)} = 1$ for all n.

The direct numerical study of the asymptotics of the numbers $\bar{M}_n^{(k)}$ turns out to be delicate, as the natural scaling variable of the problem is the ratio $x = k/n$, which depends on n and takes only a discrete set of values. To circumvent this problem, we will study the generating function $\bar{m}_n(q)$ for these numbers, also referred to as the *semi-meander polynomial.*

$$\bar{m}_n(q) \;=\; \sum_{k=1}^{n} q^k \; \bar{M}_n^{(k)} \tag{2.5}$$

This quantity makes it possible to study the large n asymptotics of the $\bar{M}_n^{(k)}$ in a global way, by use of extrapolation techniques for all real values of q. The semi-meander polynomial (2.5) may be viewed as the partition function of a statistical assembly of multicomponent semi-meanders of given order n, with a fugacity q per connected component. As such, it is expected to have an extensive large n behavior, namely

$$\bar{m}_n(q) \;\sim\; \bar{c}(q) \; \frac{\bar{R}(q)^n}{n^{\gamma(q)}} \tag{2.6}$$

where $\bar{R}(q)$ is the partition function per bridge, $\gamma(q)$ is a possibly varying exponent and $\bar{c}(q)$ a function independent of n. For $q \to 0$ ($k = 1$), we must recover the connected semi-meanders, namely that $\bar{m}_n(q)/q \to \bar{M}_n$, i.e.

$$\bar{R}(q) \to R \qquad \gamma(q) \to \gamma \qquad \bar{c}(q)/q \to \bar{c} \tag{2.7}$$

(c.f. (2.1)). The notion of winding is well-defined for multi-component semi-meanders as well, as the sum of the individual windings of each connected component, namely the *total*

number of times the various roads forming the semi-meander wind around the source of the river. Therefore we define

$$\langle w \rangle_n(q) \;=\; \frac{1}{\bar{m}_n(q)} \sum_{\substack{\text{multicomp.} \\ \text{semi-meanders}}} w\; q^k \sim n^{\nu(q)} \tag{2.8}$$

where $\nu(q)$ is the generalized winding exponent for multi-component semi-meanders, satisfying $0 \leq \nu(q) \leq 1$.

Analogously, we define multi-component meanders of order n, as configurations of k non-intersecting roads ($1 \leq k \leq n$) crossing the river through a total of $2n$ bridges, and denote by $M_n^{(k)}$ their number. We also define the *meander polynomial*

$$m_n(q) \;=\; \sum_{k=1}^{n} q^k\; M_n^{(k)} \tag{2.9}$$

This is nothing but the restriction of (2.5) with $n \to 2n$, to semi-meanders with zero winding $w = 0$. We therefore expect the asymptotics for large n

$$m_n(q) \;\sim\; c(q) \frac{R(q)^{2n}}{n^{\alpha(q)}} \tag{2.10}$$

In this estimate, the partition function per bridge $R(q)$ is expected to be identical to that of semi-meanders $\bar{R}(q)$ only if the winding is irrelevant, namely if $\nu(q)$ is strictly less than 1

$$R(q) \;=\; \bar{R}(q) \quad \text{iff} \quad \nu(q) < 1 \tag{2.11}$$

Otherwise, the fraction of semi-meanders with zero winding may be exponentially small, and we only expect that $R(q) < \bar{R}(q)$ if $\nu(q) = 1$.

2.3. Exact results for large numbers of connected components ($q = \infty$)

For very large q, we simply have

$$\bar{m}_n(q) \;\sim\; q^n \tag{2.12}$$

as the meander polynomial is dominated by the $k = n$ term, corresponding to the unique semi-meander of order n made of n concentric circular roads, each crossing the semi-infinite river only once. The winding of this semi-meander is clearly $w = n$, hence we have, for $q \to \infty$

$$\bar{R}(q) \to q \qquad \gamma(q) \to 0 \qquad \bar{c}(q) \to 1 \qquad \nu(q) \to 1 \tag{2.13}$$

As to meanders, the only way to build a meander of order n with the maximal number n connected components is that each component be a circle, crossing the river exactly twice. This is readily done by taking any upper arch configuration and completing it by reflection symmetry w.r.t. the river. This leads to $M_n^{(n)} = c_n$ (c.f. (2.4)) meanders with n connected components. By Stirling's formula, we find that when $q \to \infty$ the meander polynomial behaves as

$$\begin{aligned} m_n(q) &\sim c_n \, q^n \\ &\sim \frac{1}{\sqrt{\pi}} \frac{(2\sqrt{q})^{2n}}{n^{3/2}} \end{aligned} \tag{2.14}$$

hence, when $q \to \infty$

$$R(q) \to 2\sqrt{q} \qquad \alpha(q) \to 3/2 \qquad c(q) \to 1/\sqrt{\pi} \tag{2.15}$$

This confirms the abovementioned property (2.11) that $R(q) < \bar{R}(q)$ when $\nu(q) = 1$, as $2\sqrt{q} < q$ for large q.

2.4. Exact results for random walks on a half-line ($q = 1$)

When $q = 1$ in (2.5), $\bar{m}_n(1)$ simply counts all the multi-component semi-meanders, irrespectively of their number of connected components. This simplifies the problem drastically, as we are simply left with a purely combinatorial problem which can be solved exactly. The multicomponent semi-meanders are obtained by superimposing any arch configuration of order n with the rainbow of order n, hence

$$\bar{m}_n(1) \;=\; c_n \;\sim \frac{1}{\sqrt{\pi}} \frac{4^n}{n^{3/2}} \tag{2.16}$$

by use of Stirling's formula for large n. This gives the values

$$\bar{R}(1) = 4 \qquad \gamma(1) = 3/2 \qquad \bar{c}(1) = 1/\sqrt{\pi} \tag{2.17}$$

The study of the winding at $q = 1$ is more transparent in the formulation of arch configurations of order n as random walks of $2n$ steps on a semi-infinite line. For each arch configuration of order n, let us label by 1, 2,...,$2n-1$ each segment of river in-between two consecutive bridges, and 0 the leftmost semi-infinite portion, $2n$ the rightmost one. Let $h(i)$, $i = 0, 1, ..., 2n$ denote the number of arches passing at the vertical of the corresponding segment i. By definition, $h(0) = h(2n) = 0$. More generally, going along the river from

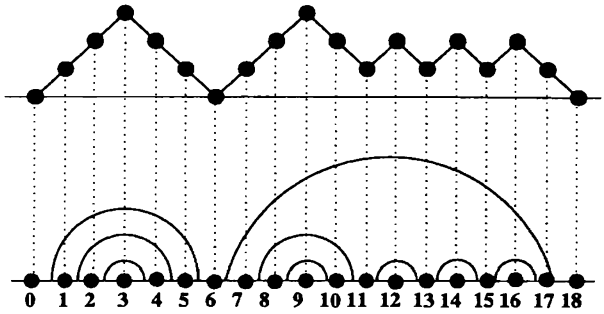


Fig. 6: A walk diagram of 18 steps, and the corresponding arch configuration of order 9. Each dot corresponds to a segment of river. The height on the walk diagram is given by the number of arches intersected by the vertical dotted line.

left to right, we have $h(i) = h(i-1) + 1$ (resp. $h(i) = h(i-1) - 1$) if an arch originates from the bridge i (resp. terminates at the bridge i).

The function h satisfies $h(i) \geq 0$, for all i, and may be interpreted as a "height" variable, defined on the segments of river, whose graph is nothing but a walk of $2n$ steps as shown in Fig.6. This may be seen as the two-dimensional extent of a brownian motion of $2n$ steps on a half-line, originating and terminating at the origin of the line. This interpretation makes the leading behavior $c_n \sim 2^{2n}$ of (2.16) clear: it corresponds to the 2 possible directions (up or down) that the motion may take at each step. The exponent $3/2$ in (2.16) is characteristic of the boundary condition, namely that the motion is closed and takes place on a half-line (other boundary conditions would lead to different values of γ, e.g. for a closed walk on a line, we would have a behavior $\binom{n}{2n} \sim 2^{2n}/\sqrt{n}$).

In this picture, the winding is simply given by the height $w = h(n)$ of the middle point. Let us evaluate more generally the average height of a point i over the arch configurations of order n. It is given by

$$\langle h(i) \rangle_n = \frac{1}{c_n} \sum_{h \geq 0} h A_{n,i}(h) \tag{2.18}$$

where $A_{n,i}(h)$ denotes the number of arch configurations of order n such that $h(i) = h$. A simple calculation [9] shows that

$$A_{n,i}(h) = \left(\binom{i}{\frac{i+h}{2}} - \binom{i}{\frac{i+h}{2}+1} \right) \left(\binom{2n-i}{n-\frac{i-h}{2}} - \binom{2n-i}{n-\frac{i-h}{2}+1} \right) \tag{2.19}$$

as the $A_{n,i}(h)$ walks are simply obtained by gluing two independent walks of i and $2n-i$ steps linking the origin to the height h.

In the case of the winding, $w = h(i = n)$, (2.18) leads to a more compact formula, according to the parity of n

$$\begin{aligned} n = 2p: \qquad \langle w \rangle_{2p} &= \frac{\binom{2p}{p}^2}{c_{2p}} - 1 \\ n = 2p+1: \qquad \langle w \rangle_{2p+1} &= 2\frac{\binom{2p}{p}\binom{2p+1}{p}}{c_{2p+1}} - 1 \end{aligned} \tag{2.20}$$

For large n, this gives the following expansion

$$\boxed{\langle w \rangle_n = 2\sqrt{\frac{n}{\pi}} - 1 + \frac{5}{4\sqrt{\pi n}} + O(1/n^{3/2})} \tag{2.21}$$

irrespectively of the parity of n. This implies that

$$\nu(q=1) = 1/2 \tag{2.22}$$

This is the well-known result for the Brownian motion, for which the extent of the path scales like $n^{1/2}$ for large n. It is instructive to note that, thanks to (2.21), the observable $w+1$ is less sensitive than w to the finite size effects at $q = 1$. This will be useful in the forthcoming numerical estimates for arbitrary q where we observe that the numerical extrapolations are improved by considering $w+1$ instead of w. Using (2.19), we may now compute the probability distribution $P_n(w)$ for an arch configuration of order n to have winding $h(n) = w$, which takes for large n the scaling form

$$P_n(w) = \frac{1}{c_n} A_{n,n}(w) \sim \frac{1}{\langle w \rangle_n} f\left(\frac{w}{\langle w \rangle_n}\right) \tag{2.23}$$

with a scaling function f independent of n for large n, readily obtained by use of Stirling's formula, upon writing $w = 2\sqrt{n/\pi}\,\xi$ for large n. This gives

$$\boxed{f(\xi) = \frac{32}{\pi^2}\xi^2 e^{-\frac{4}{\pi}\xi^2}} \tag{2.24}$$

for all $\xi > 0$.

The meanders of order n are the semi-meanders of order $2n$ with winding $w = h(2n) = 0$. They are therefore built as the juxtaposition of two independent walks of length $2n$. Hence

$$m_n(1) \;=\; (c_n)^2 \;\sim\; \frac{1}{\pi}\frac{4^{2n}}{n^3} \tag{2.25}$$

or, in other words

$$R(1) = \bar{R}(1) = 4 \qquad \alpha(1) = 3 \qquad c(1) = 1/\pi \tag{2.26}$$

This is again in agreement with (2.11), as $\nu(1) = 1/2 < 1$, i.e. the winding is irrelevant at $q = 1$.

3. Various formulations of the meander problem

This section is an overview of some very different formulations of the meander problem, each resorting to different mathematical objects (graphs, groups, algebras). The subsequent section will be devoted to yet another approach, dealing with direct enumeration.

3.1. Matrix model

Field theory, as a computational method, involves expansions over graphs weighted by combinatorial factors. In this subsection, we present a particular field theory which precisely generates planar graphs with a direct meander interpretation. The planarity of theses graphs is an important requirement, which ensures that the arches of the meander do not intersect each other, when drawn on a planar surface. The topology of the graphs is best taken into account in matrix models, where the size N of the matrices governs a topological expansion in which the term of order N^{2-2h} corresponds to graphs with genus h. The planar graphs (with $h = 0$) are therefore obtained by taking the large N limit of matrix models (see for instance [14] for a review on random matrices).

The enumeration of (planar) meanders is very close to that of 4-valent (genus 0) graphs made of two self-avoiding loops (say one black and one white), intersecting each other at simple nodes [6]. The white loop stands for the river, closed at infinity. The black loop is the road. Such a graph will be called a black and white graph. An example is given in Fig.7. The fact that the river becomes a loop replaces the order of the bridges by a cyclic order, and identifies the regions above the river and below it. Hence the number of meanders M_n is $2 \times 2n$ (2 for the up/down symmetry and $2n$ for the cyclic symmetry)

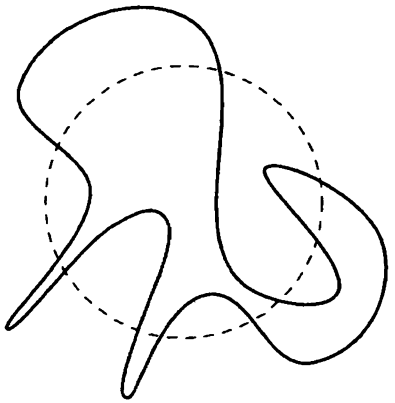


Fig. 7: A sample black and white graph. The white loop is represented in thin dashed line. There are 10 intersections.

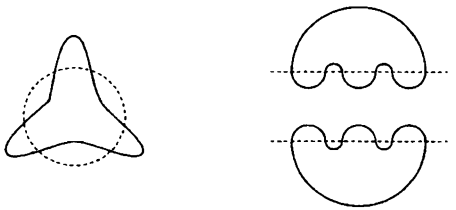


Fig. 8: A particular black and white graph with 6 intersections, and its two associated meanders. The automorphism group of the black and white graph is $\mathbb{Z}_6$.

times that of inequivalent black and white graphs with $2n$ intersections, weighed by the symmetry factor $1/|\mathrm{Aut}(\Gamma)|$ (the inverse of the order of the symmetry group of the graph). The same connection holds between $M_n^{(k)}$ and the black and white graphs where the black loop has k connected components.

For illustration, we display a particular black and white graph Γ in Fig.8, together with its two corresponding meanders of order 3. The automorphism group of this black and white graph is $\mathbb{Z}_6$, with order $\mathrm{Aut}(\Gamma) = |\mathbb{Z}_6| = 6$. The two meanders come with an overall factor $1/(2 \times 6)$, hence contribute a total $2 \times 1/12 = 1/6$, which is precisely the desired symmetry factor.

A simple way of generating black and white graphs is the use of the multi-matrix

integral (with $m+n$ hermitian matrices of size N denoted by B and W)

$$Z(m,n,c,N) = \frac{1}{\kappa_N}\int \prod_{\alpha=1}^{m} dB^{(\alpha)} \prod_{\beta=1}^{n} dW^{(\beta)} e^{-N\mathrm{Tr}\, P(B^{(\alpha)},W^{(\beta)})} \tag{3.1}$$

where the matrix potential reads

$$P(B^{(\alpha)},W^{(\beta)}) = \sum_{\alpha}\frac{(B^{(\alpha)})^2}{2} + \sum_{\beta}\frac{(W^{(\beta)})^2}{2} - \frac{c}{2}\sum_{\alpha,\beta} B^{(\alpha)}W^{(\beta)}B^{(\alpha)}W^{(\beta)} \tag{3.2}$$

The measure of integration is the usual Haar measure for hermitian matrices, and the normalization constant κ_N is such that $Z(m,n,c=0,N)=1$. In the following, the α and β indices will be referred to as color indices.

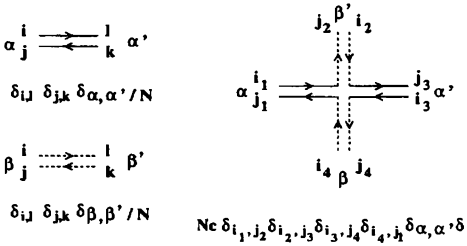


Fig. 9: The Feynman rules for the black and white matrix model. Solid (resp. dashed) double–lines correspond to black (resp. white) matrix elements, whose indices run along the two oriented lines. An extra color index α (resp. β) indicates the number of the matrix in its class, $B^{(\alpha)}$, $\alpha = 1,2,...,m$ (resp. $W^{(\beta)}$, $\beta = 1,2,...,n$). The only allowed vertices are 4–valent, and have alternating black and white edges: they describe simple intersections of the black and white loops.

The logarithm of the function (3.1) can be evaluated pertubatively as a power series of c. A term of order V in this expansion is readily evaluated as a Gaussian multi–matrix integral. It can be obtained as a sum over 4–valent connected graphs (the logarithm performs the necessary subtractions to go from disconnected to connected graphs), whose V vertices have to be connected by means of the two types of edges

$$\begin{aligned} &\text{black edges} \quad \langle [B^{(\alpha)}]_{ij}[B^{(\alpha')}]_{kl}\rangle = \frac{\delta_{il}\delta_{jk}}{N}\delta_{\alpha\alpha'} \\ &\text{white edges} \quad \langle [W^{(\beta)}]_{ij}[W^{(\beta')}]_{kl}\rangle = \frac{\delta_{il}\delta_{jk}}{N}\delta_{\beta\beta'} \end{aligned} \tag{3.3}$$

which have to alternate around each vertex. The corresponding Feynman rules are summarized in Fig.9. This is an exact realization of the desired connected black and white graphs, except that any number of loops[2] of each color is allowed. In fact, each graph receives a weight

$$N^{2-2h}\ c^V\ m^b\ n^w \tag{3.4}$$

where we have identified the Euler characteristic of the graph as $2-2h = V-E+L$ (V vertices with weight N each, E edges with weight $1/N$ each and L loops over which we have to sum the matrix indices, resulting in a weight N each) and b (resp. w) denote the total numbers of black (resp. white) loops.

A simple trick to reduce the number of say white loops w to one is to send the number n of white matrices W to 0, and to retain only the contributions of order 1 in n. Hence

$$f(m,c,N) \;=\; \lim_{n\to 0}\frac{1}{n}\mathrm{Log}\ Z(m,n,c,N) \;=\; \sum_{\substack{\text{b.\&w. conn. graphs } \Gamma \\ \text{with one } w \text{ loop}}} N^{2-2h}\ c^V\ m^b\ \frac{1}{|\mathrm{Aut}(\Gamma)|} \tag{3.5}$$

If we restrict this sum to the leading order N^2, namely the genus 0 contribution ($h=0$), we finally get a relation to the meander numbers in the form

$$\boxed{\begin{aligned} f_0(m,c) \;&=\; \lim_{N\to\infty}\frac{1}{N^2}f(m,c,N) \\ &=\; \sum_{p=1}^{\infty}\frac{c^{2p}}{4p}\sum_{k=1}^{p}M_p^{(k)}m^k \end{aligned}} \tag{3.6}$$

where the abovementioned relation between the numbers of black and white graphs and multi-component meanders has been used to rewrite the expansion (3.5).

The particular form of the matrix potential (3.2) allows one to perform the exact integration over say all the W matrices (the dependence of P on W is Gaussian), with the result

$$Z(m,n,c,N) \;=\; \frac{1}{\theta_N}\int\prod_{\alpha=1}^{m}dB^{(\alpha)}\ \det\left[\mathrm{I}\otimes\mathrm{I}-c\sum_{\alpha}B^{(\alpha)\ t}\otimes B^{(\alpha)}\right]^{-n/2}e^{-N\mathrm{Tr}\sum_{\alpha}\frac{(B^{(\alpha)})^2}{2}} \tag{3.7}$$

[2] The reader must distinguish between these loops, made of double-lines of a definite color, from the oriented loops along which the matrix indices run.

where $\mathbf{I}$ stands for the $N \times N$ identity matrix, $\otimes$ denotes the usual tensor product of matrices, and the superscript t stands for the usual matrix transposition. The prefactor θ_N is fixed by the condition $Z(m, n, c = 0, N) = 1$. With this form, it is easy to take the logarithm and to let n tend to 0, with the result

$$\begin{aligned} f(m,c,N) &= -\frac{1}{2\theta_N}\int \prod_{\alpha=1}^{m} dB^{(\alpha)} \mathrm{Tr}(\mathrm{Log}[\mathbf{I}\otimes\mathbf{I} - c\sum_{\alpha} B^{(\alpha)\,t}\otimes B^{(\alpha)}])e^{-N\mathrm{Tr}\sum_{\alpha}\frac{(B^{(\alpha)})^2}{2}} \\ &= \sum_{p=1}^{\infty}\frac{c^p}{2p}\langle \mathrm{Tr}(\sum_{\alpha=1}^{m} B^{(\alpha)\,t}\otimes B^{(\alpha)})^p\rangle_{\mathrm{Gauss}} \\ &= \sum_{p=1}^{\infty}\frac{c^p}{2p}\sum_{1\le\alpha_1,\ldots,\alpha_p\le m}\langle\left|\mathrm{Tr}(B^{(\alpha_1)}\ldots B^{(\alpha_p)})\right|^2\rangle_{\mathrm{Gauss}} \end{aligned} \tag{3.8}$$

where we still use the notation $\langle ... \rangle_{\mathrm{Gauss}}$ for the multi–Gaussian average over the matrices $B^{(\alpha)}$, $\alpha = 1, 2, ..., m$. The modulus square simply comes from the hermiticity of the matrices $B^{(\alpha)}$, namely

$$\mathrm{Tr}(\prod B^{(\alpha_i)\,t}) = \mathrm{Tr}(\prod B^{(\alpha_i)\,*}) = \mathrm{Tr}(\prod B^{(\alpha_i)})^* \tag{3.9}$$

Taking the large N limit (3.6), it is a known fact [14] that correlations should factorize, namely

$$\langle\left|\mathrm{Tr}(\prod_{i=1}^{p} B^{(\alpha_i)})\right|^2\rangle_{\mathrm{Gauss}} \underset{N\to\infty}{\longrightarrow} \left|\langle \mathrm{Tr}(\prod_{i=1}^{p} B^{(\alpha_i)})\rangle_{\mathrm{Gauss}}\right|^2 \tag{3.10}$$

By parity, we see that only even p's give non–vanishing contributions, and comparing with (3.6) we find a closed expression for the meander numbers of order n with k connected components

$$\boxed{\sum_{k=1}^{n} M_n^{(k)} m^k = \sum_{1\le\alpha_1,\ldots,\alpha_{2n}\le m}\left|\lim_{N\to\infty}\langle\frac{1}{N}\mathrm{Tr}(\prod_{i=1}^{2n} B^{(\alpha_i)})\rangle_{\mathrm{Gauss}}\right|^2} \tag{3.11}$$

This expression is only valid for integer values of m, but as it is a polynomial of degree n in m (with vanishing constant coefficient), the n first values $m = 1, 2, ..., n$ of m determine it completely. So we only have to evaluate the rhs of (3.11) for these values of m to determine all the coefficients $M_n^{(k)}$.

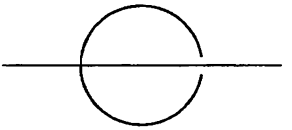


Fig. 10: The connected toric meander of order 1: it has only 1 bridge.

The relation (3.11) suggests to introduce higher genus meander numbers, denoted by $M_p^{(k)}[h]$, with $M_{2n}^{(k)}[0] = M_n^{(k)}$ (note that the indexation is now by the number of intersections, or bridges), through the generating function

$$\sum_{h=0}^{\infty}\sum_{k=1}^{\infty} M_p^{(k)}[h] m^k N^{2-2h} = \sum_{1\leq\alpha_1,\ldots,\alpha_p\leq m} \langle \left|\mathrm{Tr}(\prod_{i=1}^{p} B^{(\alpha_i)})\right|^2 \rangle_{\mathrm{Gauss}} \tag{3.12}$$

which incorporates the contribution of all genera in the Gaussian averages. Note that the genus h is that of the corresponding black and white graph and not that of the river or the road alone. In particular, the river (resp. the road) may be contractible or not in meanders of genus $h > 0$. As an example the $M_1^{(1)} = 1$ toric meander is represented in Fig.10.

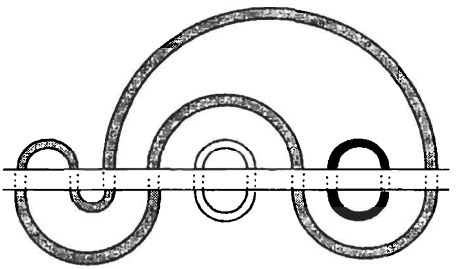


Fig. 11: A typical graph in the computation of the rhs of (3.12). The two p-valent vertices corresponding to the two traces of words are represented as racks of p double legs ($p = 10$ here). The connected components of the resulting meander (of genus $h = 0$ on the example displayed here) correspond to loops of matrices $B^{(\alpha)}$. This is indicated by a different coloring of the various connected components. Summing over all values of α_i yields a factor m per connected component, hence m^3 here.

The relation (3.12) can also be proved directly as follows. Its rhs is a sum over correlation functions of the traces of certain words (products of matrices) with themselves. More precisely, using the hermiticity of the matrices $B^{(\alpha)}$, the complex conjugate of the trace $\mathrm{Tr}(\prod_{1\leq i\leq 2n} B^{(\alpha_i)})$ can be rewritten as

$$\mathrm{Tr}(\prod_{1\leq i\leq 2n} B^{(\alpha_i)})^* = \mathrm{Tr}(\prod_{1\leq i\leq 2n} B^{(\alpha_{2n+1-i})}) \tag{3.13}$$

i.e. in the form of an analogous trace, with the order of the B's reversed. According to the Feynman rules of the previous section in the case of only black matrices, such a correlation can be computed graphically as follows. The two traces correspond to two p–valent vertices, and the Gaussian average is computed by summing over all the graphs obtained by connecting pairs of legs (themselves made of pairs of oriented double–lines) by means of edges. Re–drawing these vertices as small racks of p legs as in Fig.11, we get a sum over all multi–component, multi–genera meanders. More precisely, the edges can only connect two legs with the *same* matrix label α, which can be interpreted as a color: indeed, we have to sum over all colorings of the graph by means of m colors. But this coloring is constrained by the fact that the colors of the legs of the two racks have to be identified two by two (the color of both first legs is α_1,...,of both p-th legs is α_p). This means that each connected component of the resulting meander is painted with a color $\alpha \in \{1, 2, ..., m\}$. A graph of genus h comes with the usual weight N^{2-2h}. Summing over all the indices $\alpha_1, ..., \alpha_p = 1, 2, ..., m$, we get an extra factor of m for each connected component of the corresponding meander, which proves the relation (3.12).

In the genus 0 case, we must only consider planar graphs, which correspond to genus 0 meanders by the above interpretation. Due to the planarity of the graph, the two racks of $p = 2n$ legs each are connected to themselves through n edges each, and are no longer connected to each other: they form two disjoint arch configurations of order n. This explains the factorization mentioned in eq.(3.10), and shows that the genus 0 meanders are obtained by the superimposition of two arch configurations. The beauty of eq.(3.11) is precisely to keep track of the number of connected components k in this picture, by the m–coloring of the connected components.

This last interpretation leads to a straightforward generalization of (3.12) to semi-meanders, in the form

$$\boxed{\begin{array}{c}\displaystyle\sum_{k=1}^{n}\bar{M}_n^{(k)}m^k \;=\; \sum_{1\leq\alpha_1,\ldots,\alpha_n\leq m} \\ \displaystyle\lim_{N\to\infty}\frac{1}{N}\langle \mathrm{Tr}(B^{(\alpha_1)}B^{(\alpha_2)}\ldots B^{(\alpha_n)}B^{(\alpha_n)}B^{(\alpha_{n-1})}\ldots B^{(\alpha_1)})\rangle_{\mathrm{Gauss}}\end{array}} \tag{3.14}$$

To get this expression, we have used the m–coloring of the matrices to produce the correct rainbow-type connections between the loops of matrices.

All the above expressions for the various meander and semi-meander numbers reduce to the computation of multi-matrix Gaussian averages of traces of words, i.e. of products of matrices. This is readily done by using the so-called loop equations for the Gaussian matrix model (see [7] for all the details), with the following result.

The most general average of trace of word in m matrices in the large N limit is denoted by

$$\begin{array}{l}\gamma^{(m)}_{p_1,p_2,\ldots,p_{mk}} \;= \\ \displaystyle\lim_{N\to\infty}\frac{1}{N}\langle \mathrm{Tr}((B^{(1)})^{p_1}(B^{(2)})^{p_2}\ldots(B^{(m)})^{p_m}(B^{(1)})^{p_{m+1}}\ldots(B^{(m)})^{p_{mk}})\rangle_{\mathrm{Gauss}}\end{array} \tag{3.15}$$

In the above, some powers p_j may be zero, but no m consecutive of them vanish (otherwise the word could be reduced by erasing the m corresponding pieces). Of course $2p=\sum_i p_i$ has to be an even number for (3.15) to be non–zero, by parity. For $m=1$, we easily compute

$$\lim_{N\to\infty}\frac{1}{N}\langle \mathrm{TrB}^p\rangle_{\mathrm{Gauss}} \;=\; \gamma^{(1)}_p \;=\; \left\{\begin{array}{ll} c_n & \text{if } p=2n \\ 0 & \text{otherwise}\end{array}\right. \tag{3.16}$$

where c_n is the catalan number (2.4). If $\omega=\exp(2i\pi/m)$ denotes the primitive m–th root of unity, then we have the following recursion relation between large N averages of traces of words, for $m\geq 2$

$$\boxed{\gamma^{(m)}_{p_1,p_2,\ldots,p_{mk}} \;=\; -\sum_{j=1}^{mk-1}\omega^j\,\gamma^{(m)}_{p_1,\ldots,p_j}\,\gamma^{(m)}_{p_{j+1},\ldots,p_{mk}}} \tag{3.17}$$

When j is not a multiple of m, it is understood in the above that the multiplets $(p_1,\ldots,p_j)$ and $(p_{j+1},\ldots,p_{km})$ have to be completed by zeros so as to form sequences of m-uplets.

For instance, we write $\gamma_3^{(3)} = \gamma_{3,0,0}^{(3)} = \gamma_{0,0,3}^{(3)}$. Note also that if only $r < m$ matrices are actually used to write a word, the corresponding $\gamma^{(m)}$ can be reduced to a $\gamma^{(r)}$ by erasing the spurious zeros (for instance, $\gamma_{3,0,0}^{(3)} = \gamma_3^{(1)}$). Together with the initial condition $\gamma_{0,\ldots,0,2n+1,0,\ldots,0}^{(m)} = \gamma_{2n+1}^{(1)} = 0$ and $\gamma_{0,\ldots,0,2n,0,\ldots,0}^{(m)} = \gamma_{2n}^{(1)} = c_n$, this gives a compact recursive algorithm to compute all the large N averages of traces of words in any multi-Gaussian matrix model, and henceforth to evaluate the meander and semi-meander numbers.

3.2. Symmetric group

Each arch configuration of order n is naturally labelled by permutation $\mu \in S_{2n}$, the symmetric group over $2n$ objects, in such a way that if we label the bridges of the arch configuration $1, 2, ..., 2n$, the permutation μ indicates the pairs of bridges linked by arches, namely, for any $i = 1, 2, ..., 2n$, $\mu(i)$ is the bridge linked to i by an arch. By definition, μ is made of n cycles of length 2, it is therefore an element of the class $[2^n]$ of S_{2n}. Note that an element of this class generally does not lead to an arch configuration, because the most general pairing of bridges has intersecting arches. A permutation $\mu \in [2^n]$ will be called **admissible** if it leads to an arch configuration.

Fig. 12: An arch configuration of order 3 and the corresponding interpretation as a ribbon graph, with $V = 1$ six-valent vertex and $E = 3$ edges. On the intermediate diagram, the arches have been doubled and oriented. These oriented arches indicate the pairing of bridges, i.e. represent the action of μ. Similarly, the oriented horizontal segments indicate the action of the shift permutation σ. Each oriented loop corresponds to a cycle of the permutation $\sigma\mu$.

Let us write the admissibility condition explicitly. This condition states that arches do not intersect each other, namely that the ribbon graph (see Fig.12) with only one $2n$-valent vertex (the $2n$ bridges), whose legs are connected according to the arch configuration, is *planar*, i.e. of genus $h = 0$. This graph has $V = 1$ vertex, and $E = n$ edges (arches). Let us compute its number L of oriented loops in terms of the permutation μ. Let σ denote the "shift" cyclic permutation, namely $\sigma(i) = i + 1$, $i = 1, 2, ..., 2n - 1$ and $\sigma(2n) = 1$. Then an oriented loop in the ribbon graph is readily seen to correspond to a *cycle* of the

permutation $\sigma\mu$. Indeed, the total number of loops is $L = \text{cycles}(\sigma\mu)$, the number of cycles of the permutation $\sigma\mu$. The admissibility condition reads

$$\begin{aligned} \chi \;=\; 2 = L - E + V \;&=\; 1 - n + \text{cycles}(\sigma\mu) \\ \Leftrightarrow \quad \text{cycles}(\sigma\mu) \;&=\; n+1 \end{aligned} \tag{3.18}$$

Note that if we demand that the ribbon graph be of genus h, the above condition becomes

$$\text{cycles}(\sigma\mu) \;=\; n + 1 - 2h \tag{3.19}$$

Given an admissible permutation $\mu \in [2^n]$, let us now count the number of connected components of the corresponding semi-meander of order n. Let τ be the "rainbow" permutation $\tau(i) = 2n + 1 - i$. Note that τ changes the parity of the bridge label. On the other hand, the admissible permutation μ is readily seen to also change the parity of the bridge labels. As a consequence, the permutation $\tau\mu$ preserves the parity of bridge labels. In other words, even bridges are never mixed with odd ones. The successive iterations of the permutation $\tau\mu$ describe its cycles. The corresponding meander will be connected iff these cycles are maximal, namely μ has two cycles of length n (one for even bridges, one for odd bridges), i.e. $\tau\mu \in [n^2]$. We get a purely combinatorial expression for connected semi-meander numbers

$$\boxed{\bar{M}_n \;=\; \text{card}\{\mu \in [2^n] \mid \text{cycles}(\sigma\mu) = n+1, \text{ and } \tau\mu \in [n^2]\}} \tag{3.20}$$

More generally, the semi-meander corresponding to μ will have k connected components iff $\tau\mu$ has exactly k pairs of cycles of equal length (one over even bridges, one over odd ones).

The above conditions on various permutations are best expressed in terms of the characters of the symmetric group. Denoting by $[i^{\nu_i}]$ the class of permutations with ν_i cycles of length i, and labelling the representations of S_{2n} by Young tableaux Y with $2n$ boxes as customary, the characters can be expressed as

$$\chi_Y([i^{\nu_i}]) \;=\; \det\left(p_{i+\ell_i-j}(\theta_.)\right)\Big|_{t_\nu} , \tag{3.21}$$

where the Young tableau has ℓ_i boxes in its i-th line, counted from the top, $t_\nu = \prod_i \frac{\theta_i^{\nu_i}}{\nu_i!}$, $p_m(\theta_.)$ is the m-th Schur polynomial of the variables θ_1, θ_2,...

$$p_m(\theta_.) \;=\; \sum_{\substack{k_i \geq 0, i=1,2,\ldots \\ \sum i k_i = m}} \prod_i \frac{\theta_i^{k_i}}{k_i!} , \tag{3.22}$$

and we used the symbol $f(\theta_.)|_{t_\nu}$ for the coefficient of the monomial $\prod_i \frac{\theta_i^{\nu_i}}{\nu_i!}$ in the polynomial $f(\theta_.)$. As group characters, the χ_Y's satisfy the orthogonality relation

$$\sum_Y \chi_Y([\lambda])\chi_Y([\mu]) \;=\; \frac{(2n)!}{|[\lambda]|}\,\delta_{[\lambda],[\mu]} \tag{3.23}$$

where the sum extends over all Young tableaux with $2n$ boxes, $[\lambda]$ denotes the class of a permutation $\lambda \in S_{2n}$, and $|[\lambda]|$ the order of the class. The order of the class $[i^{\nu_i}]$ is simply

$$|[i^{\nu_i}]| \;=\; \frac{(2n)!}{\prod_i i^{\nu_i}\nu_i!} \tag{3.24}$$

The orthogonality relation (3.23) provides us with a means of expressing any condition on classes of permutations in terms of characters. It leads to the following compact expression for the connected semi-meander numbers

$$\begin{aligned}\bar{M}_n \;&=\; \sum_{\substack{[i^{\lambda_i}]\in S_{2n}\\ \Sigma\lambda_i = n+1}} \sum_{\mu\in[2^n]} \delta_{[\sigma\mu],[i^{\lambda_i}]}\;\delta_{[\tau\mu],[n^2]} \\ &=\; \sum_{\substack{[i^{\lambda_i}]\in S_{2n}\\ \Sigma\lambda_i = n+1}} \sum_{\mu\in S_{2n}} \sum_{Y,Y',Y''} \frac{|[2^n]|\;|[i^{\lambda_i}]|\;|[n^2]|}{((2n)!)^3} \\ &\quad\times \chi_Y([\mu])\chi_Y([2^n])\chi_{Y'}([\sigma\mu])\chi_{Y'}([i^{\lambda_i}])\chi_{Y''}([\tau\mu])\chi_{Y''}([n^2])\end{aligned} \tag{3.25}$$

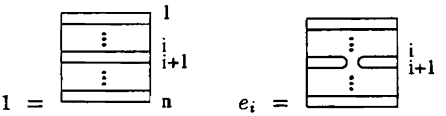
Analogous expressions hold for (higher genus) semi-meanders with k connected components and for meanders as well.

3.3. Temperley-Lieb algebra

The Temperley-Lieb algebra of order n and parameter q, denoted by $TL_n(q)$, is defined through its n generators $1, e_1, e_2, ..., e_{n-1}$ subject to the relations

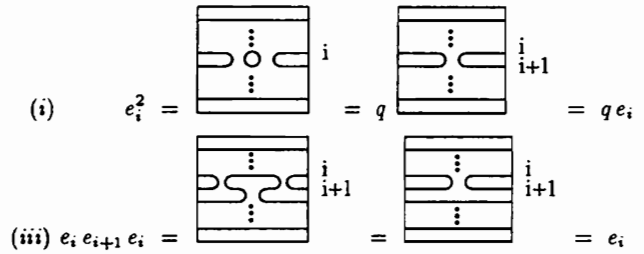
$$\begin{aligned}&(i) && e_i^2 \;=\; q\,e_i \quad i = 1,2,...,n-1 \\ &(ii) && [e_i, e_j] \;=\; 0 \quad \text{if } |i-j|>1 \\ &(iii) && e_i\,e_{i\pm1}\,e_i \;=\; e_i \quad i = 1,2,...,n-1\end{aligned} \tag{3.26}$$

This definition becomes clear in the "braid" pictorial representation, where the generators act on n parallel strings as follows:



(3.27)

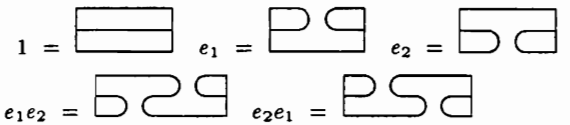
and a product of elements is represented by the juxtaposition of the corresponding braid diagrams, like dominos. The relation (ii) expresses the locality of the e's, namely that the e's commute whenever they involve distant strings. The relations (i) and (iii) read respectively



(3.28)

In the relation (i), the loop has been erased, but affected the weight q. The relation (iii) is simply obtained by stretching the $(i+2)$-th string.

The algebra $TL_n(q)$ is built out of arbitrary products of generators e_i. Up to numerical factors depending on q, any such product can be reduced by using the relations (i)-(iii). The algebra $TL_n(q)$, as a real vector space, is therefore naturally endowed with the basis formed by all the distinct reduced elements of the algebra. For illustration, the reduced elements of $TL_3(q)$ read



(3.29)

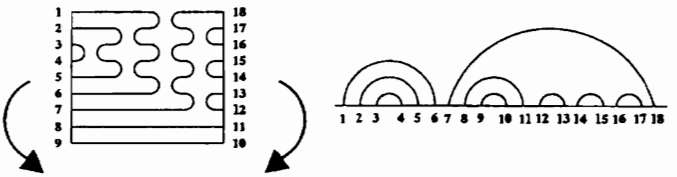


Fig. 13: The transformation of a reduced element of $TL_9(q)$ into an arch configuration of order 9. The reduced element reads $e_3e_4e_2e_5e_3e_1e_6e_4e_2$.

Let us now show that the reduced elements of $TL_n(q)$ are in one to one correspondence with arch configurations of order n. This is most clearly seen by considering the braid pictorial representation of a reduced element. Such a diagram has no internal loop (by virtue of (i)), and all its strings are stretched (using (iii)). As shown in Fig.13, one can construct a unique arch configuration of order n by deforming the diagram so as to bring the $(2n)$ ends of the strings on a line. This deformation is invertible, and we conclude that, as a vector space, $TL_n(q)$ has dimension

$$\dim(TL_n(q)) \;=\; c_n \tag{3.30}$$

This identification allows us to denote the elements of the basis of reduced elements of $TL_n(q)$ by the corresponding arch configurations a of order n.

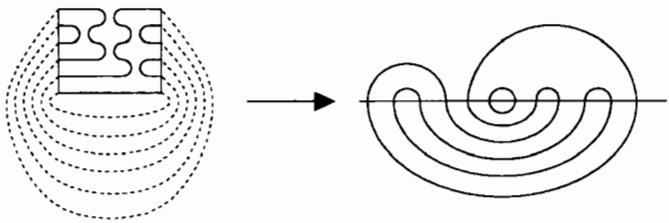


Fig. 14: The trace of an element $e \in TL_6(q)$ is obtained by identifying the left and right ends of its strings (dashed lines). In the arch configuration picture, this amounts to closing the upper configuration by a rainbow of order 6. The corresponding semi-meander has 3 connected components, hence $\mathrm{Tr}(e) = q^3$.

A scalar product on $TL_n(q)$ is defined as follows. First one introduces a trace over $TL_n(q)$. From the relation (i) of (3.26), we see that in any element of $TL_n(q)$ each closed loop may be erased and replaced by a prefactor q. Taking the trace of a reduced basis element a corresponds to identifying the left and right ends of each string as in Fig.14, and assigning an analogous factor to each closed loop, which results in a factor

$$\mathrm{Tr}(a) \;=\; q^{c(a)} \tag{3.31}$$

where $c(a)$ is precisely the number of connected components of the closure of a by a rainbow of order n: indeed, the rainbow connects the i-th bridge to the $(2n+1-i)$-th, which

exactly corresponds to the above identification of string ends. This makes the connection with meander problems clear. In particular, this permits to identify the semi-meander polynomial as

$$\bar{m}_n(q) \;=\; \sum_{\text{arch configs. } a} q^{c(a)} \;=\; \mathrm{Tr}\Big(\sum_{\text{red. basis}} a\Big) \tag{3.32}$$

We also define the transposition on $TL_n(q)$, by its action on the generators $e_i^t = e_i$, and the relation $(ab)^t = b^t a^t$ for any $a, b \in TL_n(q)$. In the arch configuration picture, this corresponds to the reflection $i \to (2n+1-i)$ of the bridges. It may also be viewed as the reflection w.r.t. the river.

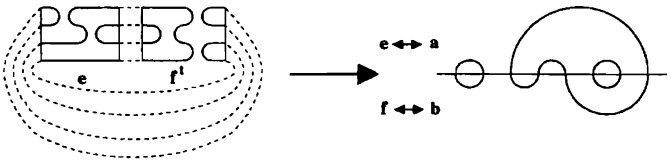


Fig. 15: The scalar product (e, f) is obtained by first multiplying e with f^t, and then identifying the left and right ends of the strings (by the dashed lines). Here we have $(e, f) = q^3$. The corresponding meander is obtained by superimposition of the upper arch configuration a corresponding to e and lower arch configuration b corresponding to f (the transposition of f is crucial to recover b as lower arch configuration). Here the meander has $c(a, b) = c(e, f) = 3$ connected components.

For any two elements e and $f \in TL_n(q)$, the scalar product is defined as

$$(e, f) \;=\; \mathrm{Tr}(e\, f^t) \tag{3.33}$$

This has a simple interpretation in terms of meanders. We have indeed

$$(e, f) \;=\; q^{c(e,f)} \;=\; q^{c(a,b)} \tag{3.34}$$

where $c(e, f) = c(a, b)$ is the number of connected components of the meander obtained by superimposing the a and b arch configurations corresponding respectively to e and f (see Fig.15 for an example).

The Gram matrix $\mathcal{G}_n(q)$ of the reduced basis of $TL_n(q)$ is the $c_n \times c_n$ symmetric matrix with entries equal to the scalar products of the basis elements, namely

$$\boxed{\big[\mathcal{G}_n(q)\big]_{a,b} \;=\; \mathrm{Tr}(ab^t) \;=\; q^{c(a,b)}} \tag{3.35}$$

For instance, $\mathcal{G}_3(q)$ reads, in the basis (3.29), with the order $(e_2e_1, e_1e_2, e_2, e_1, 1)$:

$$\mathcal{G}_3(q) = \begin{pmatrix} q^3 & q^2 & q^2 & q & q^2 \\ q^2 & q^3 & q & q^2 & q \\ q^2 & q & q^3 & q^2 & q \\ q & q^2 & q^2 & q^3 & q^2 \\ q^2 & q & q & q^2 & q^3 \end{pmatrix} \tag{3.36}$$

The meander and semi-meander polynomials are easily expressed in terms of the Gram matrix. Arranging the elements of basis 1 by growing winding (in particular, the unit 1 is the last element), and defining the c_n-dimensional vectors

$$\vec{u} = (1, 1, 1, \cdots, 1) \qquad \vec{v} = (0, 0, \cdots, 0, 1) \tag{3.37}$$

we have

$$\begin{aligned} m_n(q) &= \vec{u} \cdot \mathcal{G}_n(q)\vec{u} \\ \bar{m}_n(q) &= \vec{v} \cdot \mathcal{G}_n(q)\vec{u} \end{aligned} \tag{3.38}$$

where $\vec{x} \cdot \vec{y}$ denotes the ordinary Euclidian scalar product of $\mathbb{R}^{c_n}$. Moreover, we also have

$$m_n(q^2) = \mathrm{tr}(\mathcal{G}_n(q)^2) \tag{3.39}$$

The Gram matrix $\mathcal{G}_n(q)$ contains therefore all the information we need about meanders and semi-meanders. In [10], using the representation theory of the Temperley-Lieb algebra [15], we have computed exactly the determinant of the Gram matrix (3.35), with the simple result

$$\boxed{\begin{aligned} D_n(q) &= \det\left(\mathcal{G}_n(q)\right) = \prod_{i=1}^{n} U_i(q)^{a_{n,i}} \\ a_{n,i} &= \binom{2n}{n-i} - 2\binom{2n}{n-i-1} + \binom{2n}{n-i-2} \end{aligned}} \tag{3.40}$$

where $U_i(q)$ are the Chebishev polynomials of the second kind ($U_j(x) = xU_{j-1}(x) - U_{j-2}(x)$, $U_0(x) = 1$, $U_1(x) = x$). We have also used the convention that $\binom{j}{k} = 0$ if $j < 0$. For instance, the determinant of the matrix $\mathcal{G}_3(q)$ (3.36) reads

$$D_3(q) = U_1(q)^4\, U_2(q)^4\, U_3(q) = q^5\,(q^2-1)^4\,(q^2-2) \tag{3.41}$$

A remarkable fact is that $D_n(q)$ has only real zeros z, with $|z| < 2$. Actually, the representation theory of $TL_n(q)$ enables one to orthogonalize the Gram matrix (3.35) explicitly. This in turn translates into new "RSOS-type" expressions for the semi-meander and meander polynomials through (3.38) and (3.39) (see [10] for details). These expressions display drastic differences according to whether $|q|$ is larger or smaller than 2, a critical value which will re-emerge in the subsequent section. Hopefully these will enable one to study the large n asymptotics of the corresponding polynomials.

4. Exact enumeration and its analyses: the winding transition

In this section, we present results of an exact enumeration of $\bar{M}_n^{(k)}$ for small n ($n \leq 29$), and analyze their large n extrapolation. The enumeration is performed by implementing on a computer a recursive algorithm which describes all the semi-meanders up to some given order. Clearly, the complexity is proportional to the Catalan numbers ($c_n \sim 4^n$) hence the limitation on n.

This data is then used to derive a large q expansion of the semi-meander polynomial large n asymptotics, thanks to some remarkable property of the semi-meander numbers with large number of connected components.

The main result of this study is a strong evidence for a winding transition from a low-q $< q_c$ phase of irrelevant winding to a large-q $> q_c$ phase of relevant winding for semi-meanders.

4.1. The main recursion relation

We derive now a recursion relation generating all the semi-meanders of order $(n+1)$ from those of order n.

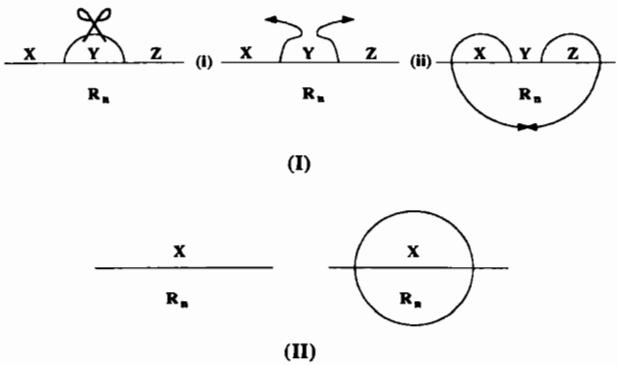


Fig. 16: The construction of all the semi–meanders of order $n+1$ with arbitrary number of connected components from those of order n. Process (I): (i) pick any exterior arch and cut it (ii) pull its edges around the semi–meander and paste them below. The lower part becomes the rainbow configuration $\mathcal{R}_{n+1}$ of order $n+1$. This process preserves the number of connected components $k \to k$. Process (II): draw a circle around the semi–meander of order n. This process adds one connected component $k \to k+1$.

We start from any semi-meander of order n with k connected components, in the open-river picture. We may construct a semi-meander of order $(n+1)$ in either following way (denoted (I) or (II)), as illustrated in Fig.16

(I) Pick any exterior arch, i.e. any arch with no other arch passing above it. Cut it and pull its ends all the way around the others (in order to add two bridges), and reconnect them below, by creating an extra concentric lower arch for the rainbow. In this process, we have $n \to n+1$, but the number of connected components has not changed: $k \to k$. Another way of picturing this transformation is the following: one simply has pulled the exterior arch all the way around the semi-meander and brought it below the figure, creating two new bridges along the way. As no cutting nor pasting is involved, the number of connected components is clearly preserved.

(II) Draw a circle around the semi-meander. This adds a lower concentric semi-circle which increases the order of the rainbow to $(n+1)$, and also adds one connected component to the initial semi-meander $k \to k+1$.

These two possibilities exhaust all the semi-meanders of order $(n+1)$, as the transformation is clearly invertible, by pulling back up the lower external arch of the rainbow. Note that by construction, there are as many possibilities for the process (I) as exterior arches, and the transformation is therefore one-to-many.

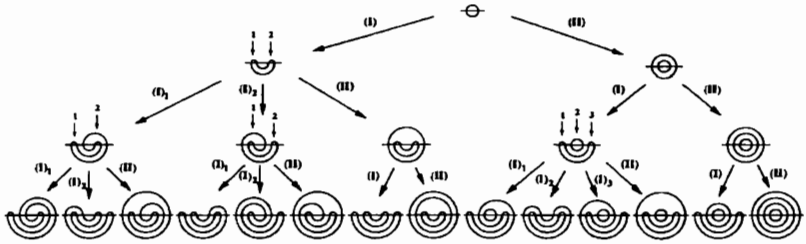


Fig. 17: The tree of semi-meanders down to order $n = 4$. This tree is constructed by repeated applications of the processes (I) and (II) on the semi-meander of order 1 (root). We have indicated by small vertical arrows the multiple choices for the process (I), each of which is indexed by its number. The number of connected components of a given semi-meander is equal to the number of processes (II) in the path going from the root to it, plus one (that of the root).

We may now construct a tree of all the semi-meanders, generated recursively from that of order 1 (root), as displayed in Fig.17. Note that we have adopted the open-river formulation to represent them.

Keeping track of the connected components, this translates into the following relation between the semi-meander polynomials

$$\bar{m}_{n+1}(q) \;=\; \bar{m}_n(q)\langle \text{ext.arch.}\rangle_n(q) + q\bar{m}_n(q) \tag{4.1}$$

where we denoted by $\langle \text{ext.arch.}\rangle_n(q)$ the average number of exterior arches in a semi-meander of order n, weighed by q^k, k its number of connected components. In (4.1), the first term corresponds to all the processes (I), whereas the second term corresponds to (II).

Taking the large n limit in (4.1), this permits to interpret

$$\bar{R}(q) - q \;=\; \langle \text{ext.arch.}\rangle_\infty(q) \tag{4.2}$$

as the limit when $n \to \infty$ of the average number of exterior arches in semi-meanders of order n, weighed by an activity q per connected component. For large q, we get the limit

$$\bar{R}(q) - q \;\to\; 1 \tag{4.3}$$

as the corresponding leading semi-meander has only one exterior arch. We also find for $q = 1$ that there is an average of $3 = 4 - 1$ exterior arches in arbitrary arch configurations of order n. Finally, for $q = 0$, the partition function per bridge $\bar{R}(0)$ is interpreted as the average number of exterior arches in connected semi-meanders.

4.2. Numerical analysis

By implementing the above recursion on a computer, we have been able to enumerate the semi-meander numbers up to $n = 29$ bridges, and the expectation values of various observables up to $n = 24$ bridges. Many of these results can be found in [7] [10]. For illustration, we give below a typical Fortran program, usable on any computer, for the enumeration of the connected semi-meanders.

```
      PARAMETER (nmax = 14)                    ! maximal order
      INTEGER A(-nmax+1:nmax)                  ! arch representation
      INTEGER Sm(nmax)                         ! semi-meander counter
      INTEGER n                                ! current depth (or order)
      INTEGER j                                ! next branch to visit
      DATA n, Sm /0, nmax*0/                   ! n and Sm initialized to 0
      A(0) = 1                                 ! single-arch semi-meander
      A(1) = 0
    2 n = n + 1                                ! a new node is visited
      Sm(n) = Sm(n) + 1
      j = -n + 1                               ! leftmost (exterior) arch
    1 IF((n.EQ.nmax).OR.(j.EQ.n+1)) GOTO 3     ! up or down ?
      A(A(j)) = n+1                            ! go down with process (I)
      A(n+1) = A(j)
      A(j) = -n
      A(-n) = j
      GOTO 2
    3 A(A(-n+1)) = A(n)                        ! going up
      A(A(n)) = A(-n+1)
      j = A(n)+1                               ! next arch to break
      n = n - 1
      IF (n .GT. 1) GOTO 1
      PRINT '(i3, i15)', (n, Sm(n), n = 1, nmax)
      END
```

This program lists the numbers $Sm(n) = \tilde{M}_n$ for $n = 1, ..., nmax$.

This data was further analyzed by large n extrapolation, and we now present a few results.

The results for $\tilde{R}(q)$ and $R(q)$ are displayed in Fig.18. The two functions are found to coincide in the range $0 \leq q \leq q_c$ with $q_c \simeq 2$, and to split into $\tilde{R}(q) > R(q)$ for $q > q_c$. As explained before, the comparison between $\tilde{R}(q)$ and $R(q)$ determines directly whether $\nu(q)$ is 1 or not. The result of Fig.18 is therefore the signal of a phase transition at $q = q_c$ between a low-q regime where the winding is essentially irrelevant ($\nu(q) < 1$) and a large-q phase with relevant winding ($\nu(q) = 1$).

This is compatible with the direct extrapolation for $\nu(q)$ displayed in Fig.19, which is however less reliable in the region around $q = 2$, due to its sub-leading (and probably discontinuous) character.

The configuration exponent for semi-meanders $\gamma(q)$ is represented in Fig.20, for two different orders in our extrapolation scheme. The extrapolation proves to be stable for $0 < q < 2$. For $q > 2$, it develops oscillations around a mean value, estimated to vanish ($\gamma(q) \sim 0$) for q large enough.

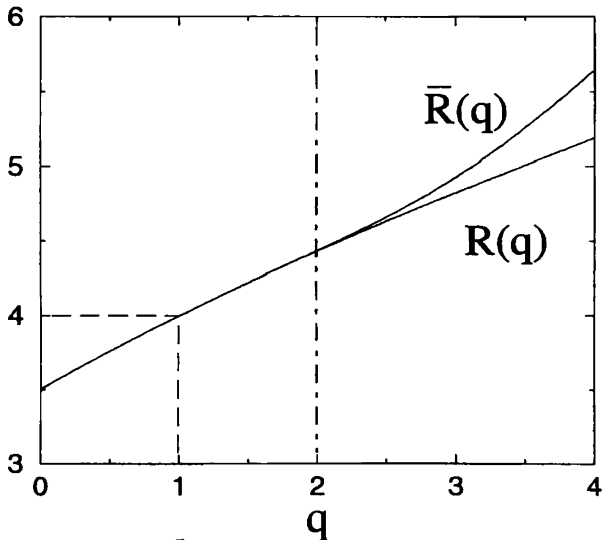


Fig. 18: The functions $\bar{R}(q)$ and $R(q)$ for $0 \leq q \leq 4$ as results of large n extrapolations. The two curves coincide for $0 \leq q \leq 2$ and split for $q > 2$ with $\bar{R}(q) > R(q)$. Apart from the exact value $\bar{R}(1) = R(1) = 4$, we find the estimates $\bar{R}(0) = 3.50(1)$, $\bar{R}(2) = 4.44(1)$, $\bar{R}(3) = 4.93(1)$ and $\bar{R}(4) = 5.65(1)$.

By analogy with critical phenomena, in addition to the scaling behaviors (2.6), (2.10) and (2.8) involving the critical exponents $\gamma(q)$, $\alpha(q)$ and $\nu(q)$, we expect to find more refined scaling laws involving scaling functions. A particular example of such scaling functions has been derived for $q = 1$ (2.23), for the probability distribution $P_n(w)$ of the winding w among arch configurations of order n. It involves the scaling function (2.24). For $q = 0$ we expect the same behavior for the corresponding probability distribution

$$P_n^{(0)}(w) \;=\; \frac{\bar{M}_n^{(1)}(w)}{\bar{M}_n^{(1)}} \tag{4.4}$$

of winding w among connected semi-meanders of order n. We expect the scaling behavior

$$P_n^{(0)}(w) \;\sim\; \frac{1}{\langle w \rangle_n(0)} f^{(0)}\left(\frac{w}{\langle w \rangle_n(0)}\right) \tag{4.5}$$

This is precisely what we observe in Fig.21, where we plot $\langle w+1 \rangle_n(0)\, P_n^{(0)}(w)$ as a function of the reduced variable $\xi = (w+1)/\langle w+1 \rangle_n(0)$ for different values of n. Indeed,

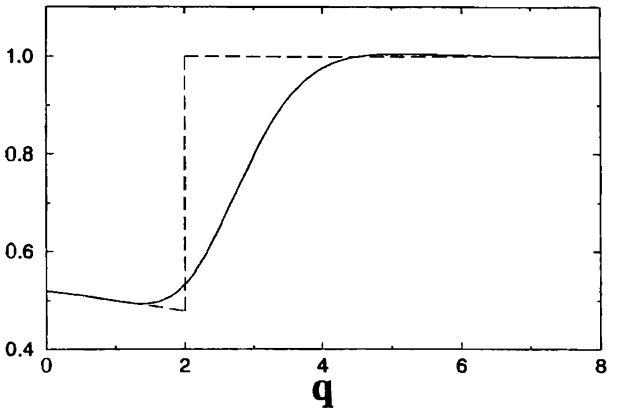


Fig. 19: The winding exponent $\nu(q)$ for $0 \leq q \leq 8$, as obtained from a large n extrapolation. We observe a drastic change of behavior between low q's and large q's, with an intermediate regime where the extrapolation fails, hence is not reliable. The dashed line indicates a possible scenario for the exact function $\nu(q)$, compatible with a transition at $q_c \simeq 2$. Apart from the exact value $\nu(1) = 1/2$, we read $\nu(0) = 0.52(1)$.

as already explained in the $q = 1$ case, we have taken the variable $(w + 1)$ instead of w to improve the convergence. All the data accumulate on a smooth curve, which represents the scaling function $f^{(0)}(\xi)$. The shape of this function is reminiscent of that of the end-to-end distribution for polymers. By analogy, we expect a certain power law behavior for small ξ

$$f^{(0)}(\xi) \sim \xi^{\theta} \tag{4.6}$$

where θ satisfies the relation

$$\alpha - \gamma = \nu(1 + \theta) \tag{4.7}$$

obtained by identifying

$$P_{2n}^{(0)}(0) \propto \frac{1}{n^{\nu}} f^{(0)}\left(\frac{1}{n^{\nu}}\right) \tag{4.8}$$

to

$$\frac{\bar{M}_{2n}^{(1)}(0)}{\bar{M}_{2n}^{(1)}} = \frac{M_n}{\bar{M}_{2n}} \propto n^{\gamma-\alpha} \tag{4.9}$$

For large ξ, we expect a behavior $f^{(0)}(\xi) \sim \exp(-\text{const.}\,\xi^{\delta})$ with a possible Fisher-law behavior $\delta = 1/(1-\nu)$. The observed function of Fig.21 is compatible with these limiting behaviors, although we cannot extract reliable estimates of the exponents θ and δ.

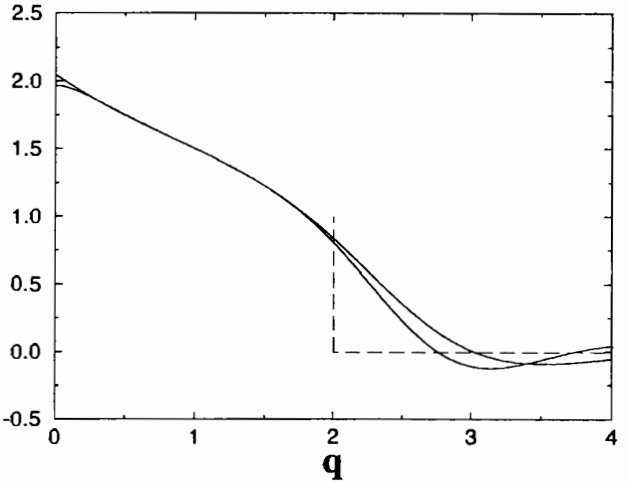


Fig. 20: The configuration exponent $\gamma(q)$ for $0 \leq q \leq 4$, from two different large n extrapolations. Apart from the exact value $\gamma(1) = 3/2$, we estimate $\gamma(0) \simeq 2$.

4.3. Large q asymptotic expansions

In the previous subsection, we have observed two regimes for the semi-meander polynomials, namely a low-q regime in which the winding is irrelevant and a large-q regime where the winding is relevant, separated by a transition at a value of $q = q_c \simeq 2$. On the other hand, we have already exhibited an exact solution of the problem at $q = \infty$ (2.13), and a first correction thereof for large q in (4.3). It is therefore tempting to analyze the large q phase by a systematic expansion in $1/q$.

Let us write the large q expansion of the semi-meander polynomial $\bar{m}_n(q)$ of eq.(2.5) as

$$\bar{m}_n(q) \;=\; q^n \;(\bar{M}_n^{(n)} + \frac{\bar{M}_n^{(n-1)}}{q} + \frac{\bar{M}_n^{(n-2)}}{q^2} + \cdots) \tag{4.10}$$

involving the semi-meander numbers in the form $\bar{M}_n^{(n-k)}$, $k = 0, 1, 2, \ldots$ Remarkably, these numbers display some polynomial structure.

When $k = 0$, there is a unique semi-meander of order n with n connected components, namely that made of n concentric circular roads, each intersecting the river through one bridge, and therefore winding once around the source. Hence we identify the first polynomial p_0 of degree 0 as

$$p_0(n) \;=\; \bar{M}_n^{(n)} \;=\; 1 \tag{4.11}$$

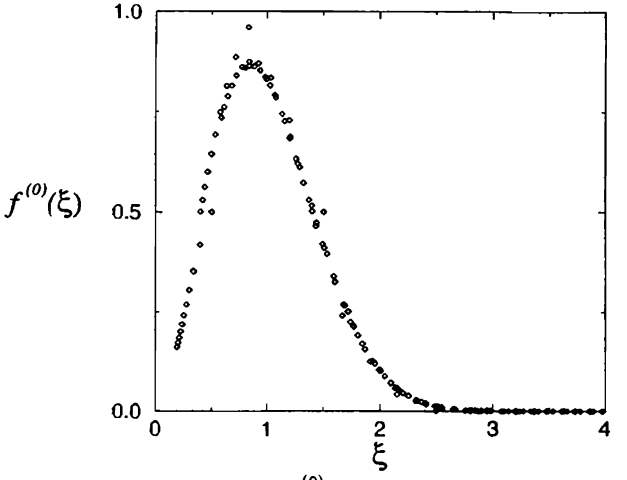


Fig. 21: Plot of $(w+1)_n(0)\, P_n^{(0)}(w)$ as a function of the reduced variable $\xi = (w+1)/(w+1)_n(0)$ for $n = 2, 3, ..., 24$. The points accumulate to a smooth scaling function $f^{(0)}(\xi)$. The erratic points correspond to small values of n, which have not reached the asymptotic regime.

When $k = 1$, all semi-meanders of order n with $n-1$ connected components are made of $n-2$ concentric circles intersecting the river once each, plus one loop, drawn in-between two consecutive circles, which intersects the river through two bridges and has no winding. There are $n-1$ available positions for this extra loop, resulting in

$$p_1(n) \;=\; \bar{M}_n^{(n-1)} \;=\; n-1 \tag{4.12}$$

where we have identified the result as a polynomial p_1 of degree 1 in n.

More generally, using the recursive construction of the previous section, one can prove the following proposition: the number $\bar{M}_n^{(n-k)}$ is equal to a polynomial $p_k(n)$ of degree k in n, for all $k \geq 0$ and $n \geq 2k-1$. The proof is purely combinatorial, and to just give a flavor of it let us compute the leading coefficient of $p_k(n)$. The $\bar{M}_n^{(n-k)}$ semi-meanders of order n with $n-k$ components are generated in the tree 17, starting from the root, by exactly k applications of the process (I) and $n-1-k$ applications of the process (II). This leads to $\binom{n-1}{k} \sim n^k/k! \sim p_k(n)$ possible choices for $n >> k$. The choices are however not independent, as consecutive applications of the process (I) may lead to more possibilities. Those are included in the lower order coefficients of $p_k(n)$, gathering

lower order combinatorial factors. When $n \leq 2k-2$, some non-polynomial corrections emerge, signaling the break-down of the large q phase of semi-meanders. In the latter, the polynomial $\bar{m}_n(q)$ is asymptotic to the series

$$q^n \sum_{k=0}^{\infty} p_k(n) q^{-k} \tag{4.13}$$

which must display an asymptotic behavior of the form (2.6). This induces strong constraints on the polynomials $p_k(n)$, which allow for their complete determination up to $k = 18$, out of their first values for small n, which were enumerated exactly up to $n = 27$ (the polynomials p_k are listed in [10] for $k = 0, 1, ..., 18$). In turn, these values of p_k yield the following large q expansions of $\bar{R}(q)$ and $\bar{c}(q)$

$$\begin{aligned}
\bar{R}(q) &= q+1+\frac{2}{q}+\frac{2}{q^2}+\frac{2}{q^3}-\frac{4}{q^5}-\frac{8}{q^6}-\frac{12}{q^7}-\frac{10}{q^8}-\frac{4}{q^9}+\frac{12}{q^{10}}+\frac{46}{q^{11}} \\
&\quad +\frac{98}{q^{12}}+\frac{154}{q^{13}}+\frac{124}{q^{14}}+\frac{10}{q^{15}}-\frac{102}{q^{16}}+\frac{20}{q^{17}}-\frac{64}{q^{18}}+O(\frac{1}{q^{19}}) \\
\bar{c}(q) &= 1-\frac{1}{q}-\frac{4}{q^2}-\frac{4}{q^3}+\frac{14}{q^5}+\frac{44}{q^6}+\frac{56}{q^7}+\frac{28}{q^8}-\frac{82}{q^9}-\frac{252}{q^{10}}-\frac{388}{q^{11}} \\
&\quad -\frac{588}{q^{12}}-\frac{772}{q^{13}}-\frac{620}{q^{14}}+\frac{1494}{q^{15}}+\frac{5788}{q^{16}}+\frac{7580}{q^{17}}-\frac{690}{q^{18}}+O(\frac{1}{q^{19}})
\end{aligned} \tag{4.14}$$

Moreover, due to the intrinsic polynomial character of the large q expansion (4.13), we find that

$$\gamma(q) = 0 \tag{4.15}$$

This result is expected to hold as long as the corrections to the polynomial behavior of the $\bar{M}_n^{(n-k)}$ are negligible. This condition defines precisely the large q phase $q > q_c$. Therefore the exponent $\gamma(q)$ vanishes identically over the whole phase $q > q_c$.

It is interesting to compare the result of these large q expansions to the previous direct large n extrapolations. As far as $\bar{R}(q)$ is concerned, we find a perfect agreement for the values $q \geq 2$, down to $q = 2$, where we find $\bar{R}(2) \simeq 4.442(1)$ using (4.14), in perfect agreement with the previous estimate. The precision of (4.14) increases with q, leading to far better estimates than before: $\bar{R}(3) \simeq 4.92908(1)$, $\bar{R}(4) \simeq 5.6495213(1)$...

As to $\gamma(q)$, our prediction that $\gamma(q) = 0$ for all $q > 2$ is compatible with the previous extrapolation of Fig.20, where this value is represented in dashed line (indeed, the large q

expansions give precise results down to $q \sim 2$ where the extrapolations become dubious). We therefore expect $\gamma(q)$ to have a discontinuity at $q = 2$, where it goes from a non-zero $\gamma(q = 2^-)$ value to zero.

This is further confirmed by a refined analysis of the average winding (2.2) in the large q phase. This requires a refined study of the semi-meander numbers $\bar{M}_n^{(n-k)}(w)$ with fixed winding w, which display a similar polynomial structure as the $\bar{M}_n^{(n-k)}$. As a result, we find that

$$\boxed{\langle w \rangle_n(q) \;=\; \lambda(q)n + \mu(q)} \tag{4.16}$$

hence $\nu(q) = 1$ throughout the large q phase, and the coefficients $\lambda(q)$ and $\mu(q)$ have the following large q expansions up to order 14 in $1/q$

$$\begin{aligned}
\lambda(q) \;&=\; 1 - \frac{2}{q} - \frac{2}{q^2} + \frac{2}{q^3} + \frac{2}{q^4} + \frac{2}{q^5} + \frac{10}{q^6} - \frac{6}{q^7} - \frac{14}{q^8} - \frac{10}{q^9} \\
&\quad + \frac{22}{q^{10}} + \frac{86}{q^{11}} - \frac{58}{q^{12}} - \frac{222}{q^{13}} - \frac{118}{q^{14}} + O(\frac{1}{q^{15}}) \\
\mu(q) \;&=\; \frac{2}{q} + \frac{10}{q^2} + \frac{22}{q^3} + \frac{54}{q^4} + \frac{134}{q^5} + \frac{246}{q^6} + \frac{622}{q^7} + \frac{1434}{q^8} + \frac{3178}{q^9} \\
&\quad + \frac{6834}{q^{10}} + \frac{13786}{q^{11}} + \frac{30834}{q^{12}} + \frac{66590}{q^{13}} + \frac{140582}{q^{14}} + O(\frac{1}{q^{15}})
\end{aligned} \tag{4.17}$$

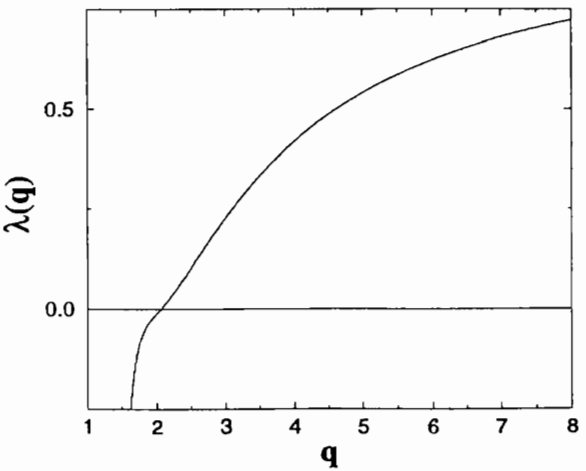


Fig. 22: The series $\lambda(q)$ (4.17) of $1/q$ up to order 14, for $1 < q < 8$. The curve seems to vanish precisely at $q = 2$.

The plot of the function $\lambda(q)$ is displayed in Fig.22. Remarkably, this coefficient seems to vanish at the point $q = 2$ with an excellent precision. Since this coefficient must be positive, we deduce that our large q formulas break down for $q < 2$. We interpret this as yet another evidence of the drastic change of behavior of the average winding $\langle w \rangle$, which is no longer linear in n below q_c, and we find $q_c = 2$ with an excellent precision.

In conclusion, we gave strong evidence for the existence of a winding transition of the semi-meander partition function in the large n limit, taking place at a value $q_c = 2$ which we conjecture to be exact. The order parameter for this transition is clearly

$$\boxed{\lim_{n\to\infty} \frac{1}{n}\langle w \rangle_n = \begin{cases} \lambda(q) & \text{for } q > q_c \\ 0 & \text{for } q < q_c \end{cases}} \tag{4.18}$$

which vanishes for $q < q_c$ (irrelevant winding, i.e. $\nu(q) < 1$) and is nonzero for $q > q_c$ (relevant winding, i.e. $\nu(q) = 1$). With the order parameter (4.18), the transition is found to be continuous, as the leading coefficient $\lambda(q)$ (4.17) vanishes at $q = q_c$. As argued before, The low-q phase is characterized by a meander-type behavior of the semi-meander polynomial, where $\bar{R}(q) = R(q)$. The smooth character of the transition is also visible from the fact that $\bar{R}(q)$ approaches $R(q)$ tangentially at $q = q_c$ (c.f. Fig.18).

5. Conclusion

We must admit that none of the compact expressions (matrix model and symmetric group) for the meander and semi-meander numbers, although conceptually interesting (beautiful?), give an efficient way of computing them. There is always some lengthy process involved, such as evaluating Gaussian averages of traces of words or writing the group characters, which render the evaluation in fact untractable. The Temperley-Lieb algebra connection is maybe one of the most promising approaches towards exact asymptotics, but we have no definite answer to this day.

In the direct enumerative approach, we have analyzed the meander problem in the language of critical phenomena, by analogy with Self-Avoiding Walks. In particular, we have displayed various scaling behaviors, involving both scaling exponents and scaling functions. We have presented strong evidence for the existence of a phase transition for semi-meanders weighed by a factor q per connected component (road).

In a large-q regime ($q > q_c$), the winding is found to be relevant, with a winding exponent $\nu(q) = 1$, while the configuration exponent $\gamma(q) = 0$. In this regime, the partition function per bridge for semi-meanders $\bar{R}(q)$ is strictly larger than that of meanders $R(q)$. The particular form of its large q series expansion in $1/q$ (4.14) with slowly alternating integer coefficients, which furthermore grow very slowly with the order, suggests a possible re-expression in terms of modular forms of q, yet to be found. It is striking to notice that our numerical estimate for $\bar{R}(2)$ agrees up to the third digit with the value

$$\pi\sqrt{2} \;=\; 4\prod_{k=1}^{\infty}\frac{(4k)^2}{(4k+1)(4k-1)} \;=\; 4.4428... \qquad (5.1)$$

suggesting maybe an infinite product form for $\bar{R}(q)$, which is still to be found.

In a low-q regime $q < q_c$, $\bar{R}(q)$ and $R(q)$ coincide, in agreement with an irrelevant winding $\nu(q) < 1$. The exponent $\gamma(q)$ is no longer 0, but a strictly positive function of q. We have estimated the value of the transition point $q_c \simeq 2$ with an excellent precision, and we conjecture that $q_c = 2$ exactly. This special value of q has actually been singled out in the algebraic study of the meander problem, in connection with the Temperley-Lieb algebra as sketched in Sect.3.3. Indeed, as shown in [9], one can re-express the meander and semi-meander partition functions as that of some Restricted Solid-On-Solid model, whose Boltzmann weights are positive precisely iff $q \geq 2$, indicating very different behaviors for $q < 2$ and $q > 2$.

There still remains to find the varying exponents $\gamma(q)$ and $\nu(q)$ in the $q < 2$ regime, as well as the precise value of $R(q) = \bar{R}(q)$. Although we improved our numerical estimates, we are limited to conjectures. For $q = 0$, we confirm a previous conjecture [7] that $\gamma = 2$, and that [6] $\alpha = 7/2$. We also conclude from the numerical analysis that $\nu(0) \simeq 0.52(1)$ is definitely not equal to the trivial random-walk exponent $1/2$.

References

[1] K. Hoffman, K. Mehlhorn, P. Rosenstiehl and R. Tarjan, *Sorting Jordan sequences in linear time using level-linked search trees*, Information and Control **68** (1986) 170-184.

[2] V. Arnold, *The branched covering of $CP_2 \to S_4$, hyperbolicity and projective topology*, Siberian Math. Jour. **29** (1988) 717-726.

[3] K.H. Ko, L. Smolinsky, *A combinatorial matrix in 3-manifold theory*, Pacific. J. Math **149** (1991) 319-336.

[4] J. Touchard, *Contributions à l'étude du problème des timbres poste*, Canad. J. Math. **2** (1950) 385-398.

[5] W. Lunnon, *A map-folding problem*, Math. of Computation **22** (1968) 193-199.

[6] S. Lando and A. Zvonkin, *Plane and Projective Meanders*, Theor. Comp. Science **117** (1993) 227-241, and *Meanders*, Selecta Math. Sov. **11** (1992) 117-144.

[7] P. Di Francesco, O. Golinelli and E. Guitter, *Meander, folding and arch statistics*, to appear in Journal of Mathematical and Computer Modelling (1996).

[8] Y. Makeenko, *Strings, Matrix Models and Meanders*, proceedings of the 29th Inter. Ahrenshoop Symp., Germany (1995); Y. Makeenko and H. Win Pe, *Supersymmetric matrix models and the meander problem*, preprint ITEP-TH-13/95 (1996); G. Semenoff and R. Szabo *Fermionic Matrix Models* preprint UBC/S96/2 (1996).

[9] P. Di Francesco, O. Golinelli and E. Guitter, *Meanders and the Temperley-Lieb algebra*, Saclay preprint T96/008 (1996).

[10] P. Di Francesco, O. Golinelli and E. Guitter, *Meanders: a direct enumeration approach*, Saclay preprint T96/062 (1996).

[11] H. Temperley and E. Lieb, *Relations between the percolation and coloring problem and other graph-theoretical problems associated with regular planar lattices: some exact results for the percolation problem*, Proc. Roy. **A322** (1971) 251-280.

[12] N. Sloane, *the on-line encyclopedia of integer sequences*, e-mail: sequences@research.att.com

[13] R. Baxter, *Exactly solved models in statistical mechanics*, Academic Press, London (1982).

[14] E. Brézin, C. Itzykson, G. Parisi and J.-B. Zuber, *Planar Diagrams*, Commun. Math. Phys. **59** (1978) 35-51; P. Di Francesco, P. Ginsparg and J. Zinn-Justin *2D Gravity and Random Matrices*, Phys. Rep. **254** (1995) 1-133.

[15] P. Martin, *Potts models and related problems in statistical mechanics*, World Scientific (1991).

EXERCISES IN EQUIVARIANT COHOMOLOGY AND TOPOLOGICAL THEORIES

R. STORA
Laboratoire de Physique Théorique ENSLAPP[a], B.P. 110,
F-74941 Annecy-le-Vieux Cedex, France
and
Theory Division, CERN, CH-1211, Geneva 23, Switzerland.

Equivariant cohomology is suggested as an alternative algebraic framework for the definition of topological field theories constructed by E. Witten circa 1988. It also enlightens the classical Faddeev Popov gauge fixing procedure.

1 Introduction

Before going into the subject of this talk, I would like to describe some concrete exercises done by Claude and I which represent a very small portion of the numerous discussions we had, mostly by exchange of letters. We happened to be both guests of the CERN theory division during the academic year 1972-1973.

The perturbative renormalization of gauge theories was still a hot subject, and, whereas most of our colleagues considered the problem as solved we were both still very innocent. I happened to be scheduled for a set of lectures for the "Troisième cycle de la Suisse Romande" in the spring 1973, on the subject "Models with renormalizable Lagrangians: Perturbative approach to symmetry breaking", and I decided to conclude those lectures with a summary of the known constructions related to gauge theories, mostly at the classical level, except for a heuristic derivation of the now called [1] Slavnov Taylor identities, taking seriously the Faddeev Popov ghost and antighost as local fields. What had to be done was indicated in A. Slavnov's preprint which I had remarked: perform a gauge transformation of parameter $m^{-1}\bar{\xi}$ where m is the Faddeev Popov operator and $\bar{\xi}$ the source of the antighost field. That strange trick was due to E.S. Fradkin and I.V. Tyutin as indicated in Slavnov's preprint. At the time, I was not aware of J.C. Taylor's paper which came to my attention much later. Anyway, Claude and I carried out that calculation whose result is reported in the notes, with details in an appendix for which the authors (A. Rouet and I) thank Claude Itzykson for generous help [2]. It is that form of the identity which, a few months later drew Carlo Becchi and Alain Rouet's atten-

[a] URA 1436 du CNRS, associée à l'Ecole Normale Supérieure de Lyon et à l'Université de Savoie.

References

[1] K. Hoffman, K. Mehlhorn, P. Rosenstiehl and R. Tarjan, *Sorting Jordan sequences in linear time using level-linked search trees*, Information and Control **68** (1986) 170-184.

[2] V. Arnold, *The branched covering of $CP_2 \to S_4$, hyperbolicity and projective topology*, Siberian Math. Jour. **29** (1988) 717-726.

[3] K.H. Ko, L. Smolinsky, *A combinatorial matrix in 3-manifold theory*, Pacific. J. Math **149** (1991) 319-336.

[4] J. Touchard, *Contributions à l'étude du problème des timbres poste*, Canad. J. Math. **2** (1950) 385-398.

[5] W. Lunnon, *A map-folding problem*, Math. of Computation **22** (1968) 193-199.

[6] S. Lando and A. Zvonkin, *Plane and Projective Meanders*, Theor. Comp. Science **117** (1993) 227-241, and *Meanders*, Selecta Math. Sov. **11** (1992) 117-144.

[7] P. Di Francesco, O. Golinelli and E. Guitter, *Meander, folding and arch statistics*, to appear in Journal of Mathematical and Computer Modelling (1996).

[8] Y. Makeenko, *Strings, Matrix Models and Meanders*, proceedings of the 29th Inter. Ahrenshoop Symp., Germany (1995); Y. Makeenko and H. Win Pe, *Supersymmetric matrix models and the meander problem*, preprint ITEP-TH-13/95 (1996); G. Semenoff and R. Szabo *Fermionic Matrix Models* preprint UBC/$\mathcal{S}$96/2 (1996).

[9] P. Di Francesco, O. Golinelli and E. Guitter, *Meanders and the Temperley-Lieb algebra*, Saclay preprint T96/008 (1996).

[10] P. Di Francesco, O. Golinelli and E. Guitter, *Meanders: a direct enumeration approach*, Saclay preprint T96/062 (1996).

[11] H. Temperley and E. Lieb, *Relations between the percolation and coloring problem and other graph-theoretical problems associated with regular planar lattices: some exact results for the percolation problem*, Proc. Roy. **A322** (1971) 251-280.

[12] N. Sloane, ***the on-line encyclopedia of integer sequences***, **e-mail: sequences@research.att.com**

[13] R. Baxter, *Exactly solved models in statistical mechanics*, Academic Press, London (1982).

[14] E. Brézin, C. Itzykson, G. Parisi and J.-B. Zuber, *Planar Diagrams*, Commun. Math. Phys. **59** (1978) 35-51; P. Di Francesco, P. Ginsparg and J. Zinn-Justin *2D Gravity and Random Matrices*, Phys. Rep. **254** (1995) 1-133.

[15] P. Martin, *Potts models and related problems in statistical mechanics*, World Scientific (1991).

EXERCISES IN EQUIVARIANT COHOMOLOGY AND TOPOLOGICAL THEORIES

R. STORA
Laboratoire de Physique Théorique ENSLAPP[a], *B.P. 110,*
F-74941 Annecy-le-Vieux Cedex, France
and
Theory Division, CERN, CH-1211, Geneva 23, Switzerland.

Equivariant cohomology is suggested as an alternative algebraic framework for the definition of topological field theories constructed by E. Witten circa 1988. It also enlightens the classical Faddeev Popov gauge fixing procedure.

1 Introduction

Before going into the subject of this talk, I would like to describe some concrete exercises done by Claude and I which represent a very small portion of the numerous discussions we had, mostly by exchange of letters. We happened to be both guests of the CERN theory division during the academic year 1972-1973.

The perturbative renormalization of gauge theories was still a hot subject, and, whereas most of our colleagues considered the problem as solved we were both still very innocent. I happened to be scheduled for a set of lectures for the "Troisième cycle de la Suisse Romande" in the spring 1973, on the subject "Models with renormalizable Lagrangians: Perturbative approach to symmetry breaking", and I decided to conclude those lectures with a summary of the known constructions related to gauge theories, mostly at the classical level, except for a heuristic derivation of the now called [1] Slavnov Taylor identities, taking seriously the Faddeev Popov ghost and antighost as local fields. What had to be done was indicated in A. Slavnov's preprint which I had remarked: perform a gauge transformation of parameter $m^{-1}\bar{\xi}$ where m is the Faddeev Popov operator and $\bar{\xi}$ the source of the antighost field. That strange trick was due to E.S. Fradkin and I.V. Tyutin as indicated in Slavnov's preprint. At the time, I was not aware of J.C. Taylor's paper which came to my attention much later. Anyway, Claude and I carried out that calculation whose result is reported in the notes, with details in an appendix for which the authors (A. Rouet and I) thank Claude Itzykson for generous help [2]. It is that form of the identity which, a few months later drew Carlo Becchi and Alain Rouet's atten-

[a] URA 1436 du CNRS, associée à l'Ecole Normale Supérieure de Lyon et à l'Université de Savoie.

tion, leading them to the remark that the gauge fixed Faddeed Popov action possesses a symmetry naturally called the Slavnov symmetry. A year later, when the paper by E.S. Fradkin and G.A. Vilkovisky on the quantization of canonical systems with constraints came out, Claude and I had a conversation on the telephone and we found we had both noticed that paper. I suggested that the action they proposed possessed a Slavnov symmetry. A couple of days later, Claude called me back and gave me the formula -at least in the case of gauge constraints- which I immediately forgot. When I met E.S. Fradkin in Moscow in the fall 1976, I told him about Claude's finding, and there followed the first article by I.A. Batalin and G.A. Vilkovisky who unfortunately thank me for suggesting the problem, and do not mention Claude at all.

These are only two examples of the innumerable discussions we had on physics and other things as well, mostly in writing, because life did not make our trajectories intersect so often. The last long series of discussions I had with him took place in Turku, Finland, at the meeting of the spring 1991. Almost every evening, we were ambulating around the big lawn in front of the dining room, trying to reconstruct, at his request, the arguments which produce the existence of 27 straight lines on an unruled third degree surface. That was a prelude to his later work on enumerative geometry.

Generous, he was; intelligent he was; cultivated he was; we remain deprived of patiently gathered wisdom, a rather rare item.

Returning to technicalities I will now try to describe a few facts about the Lagrangian formulation of topological -more precisely cohomological- field theories, constructed by E. Witten from 1988 on, in as much as they are relevant to our poor understanding of gauge theories. That is to say I will insist on the field theory aspects in particular, the distinction between fields and observables, even though a host of beautiful results and conjectures have been obtained otherwise.

Equivariant cohomology is roughly forty five years old, and yet, does not belong to most theoretical physicists' current mathematical equipment. The easy parts, namely, definitions, terminology, elementary properties are described in the appendix whose content is freely used throughout the text.

Section 2 is devoted to a reminder on dynamical gauge theories and a formal description of the Faddeev Popov gauge fixing procedure in terms of notions belonging to the theory of foliations[3].

Section 3 describes some aspects of "cohomological" topological theories with emphasis on some of the features which distinguish them from dynamical theories at the algebraic level provided by the Lagrangian descriptions.

2 Formal aspects of dynamical gauge theories

Here are a few considerations on formal aspects of the Faddeev Popov gauge fixing procedure which allowed to handle, thanks to the very strong consequences of locality, the ultraviolet difficulties found in the perturbative treatment of theories of the Yang Mills type. This can be found in most textbooks and usually proceeds via factoring out of the relevant functional integral the infinite volume of the gauge group produced by the gauge invariance of the functional measure. There is a more satisfactory strategy sketched in J. Zinn Justin's book [4] which avoids this unpleasant step, and fits more closely mathematical constructions now classical in the theory of foliations [3].

The set up is as follows:

M_4 is a smooth space time manifold, which one may choose compact without boundary, in euclidean field theory. $P(M,G)$ is a principal G bundle over M_4, $\bigcup_i (U_i \times G)$ modulo glueing maps above $U_i \cap U_j$, where $\{U_i\}$ is an open covering of M). G is a compact Lie group referred to as the structure group. $\mathcal{A}$ is the set of principal connections a on $P(M,G)$ (Yang Mills fields). On M_4

$$a_M = \sum_\alpha a_\mu^\alpha(x) dx^\mu e_\alpha \qquad e_\alpha : \text{basis of Lie } G \tag{1}$$

On $P(M,G)$, locally,

$$a = g^{-1} a_M \; g + g^{-1} dg \quad (x,g) \text{ local coordinates in } U \times G \tag{2}$$

$$F(a) = da + \frac{1}{2}[a,a] \tag{3}$$

is the curvature of a (the field strength).

$\mathcal{A}$ is acted upon by $\mathcal{G}$, the gauge group, i.e. the group of vertical automorphisms of $P(M,G)$ ("gauge transformations"). Upon suitable restrictions, $\mathcal{A}$ is a principal $\mathcal{G}$ bundle over $\mathcal{A}/\mathcal{G}$, the set of gauge orbits.

Dynamical gauge theories are models in which the fields are the a's (and, possibly matter fields), and the observables are gauge invariant functions of the a's (or functions on $\mathcal{A}/\mathcal{G}$).

For historical as well as technical reasons related to locality, one chooses models specified by a local gauge invariant action

$$S_{YM}(a) = \frac{1}{4g^2} \int_{M_4} tr F \wedge *F. \tag{4}$$

Heuristically, one considers the $\mathcal{G}$ invariant measure on $\mathcal{A}$

$$\Omega_{YM} = e^{-S_{YM}(a)} \underbrace{\wedge \, \delta a}_{\mathcal{D}a} \tag{5}$$

If $\{X_\alpha\}$ denotes a basis of fundamental vertical vector fields representing the action of Lie $\mathcal{G}$ on $\mathcal{A}$, one constructs the Ruelle Sullivan[5] current

$$\Omega_{RS} = i(\bigwedge_\alpha X_\alpha)\Omega_{YM} \tag{6}$$

which is closed and horizontal, therefore basic: (cf. Appendix A)

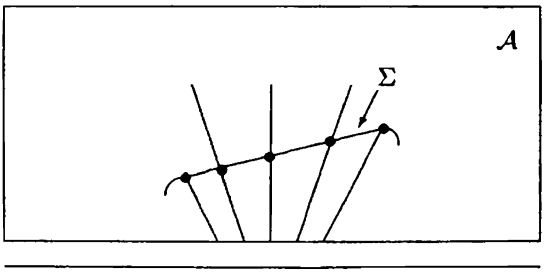
$$\begin{aligned} \delta\Omega_{RS} &= 0 \\ i(X_\alpha)\Omega_{RS} &= 0 \end{aligned} \tag{7}$$

hence

$$\ell(X_\alpha)\Omega_{RS} = 0 \tag{8}$$

It follows in particular that Ω_{RS} is invariant under field dependent gauge transformations.

Given a gauge invariant observable $\mathcal{O}(a)$, the question is to integrate it against Ω_{RS}, or rather to integrate its image as a function on $\mathcal{A}/\mathcal{G}$ against the image of Ω_{RS} as a top form on $\mathcal{A}/\mathcal{G}$.



Choose a local section Σ (transverse to the fibers) with local equations

$$g(a) = 0 \tag{9}$$

and corresponding local coordinates $\dot{a}$ so that a local parametrization of $\mathcal{A}$ is given by

$$a = \dot{a}^g \tag{10}$$

i.e. all a's are, locally gauge transforms of points on the chosen transversal manifold.

One can represent the transverse measure associated with the chosen section as follows:

$$< \mathcal{O} >_\Omega = \int_\Sigma \mathcal{O}(\dot{a})\Omega_{RS|\Sigma} = \int_\Sigma \mathcal{O}(\dot{a})\Omega_{RS|\Sigma} \underbrace{\int_{fiber} \delta(g) \wedge \delta g}_{=1}$$

$$= \int \mathcal{O}(a)\Omega_{RS}\ \delta(g) det m(\wedge g^{-1}\delta g) \qquad (11)$$

where the volume $\wedge g^{-1}\delta g$ is chosen so that

$$i(\wedge_{\dot{\alpha}} X_{\dot{\alpha}})(\wedge g^{-1}\delta g) = 1 \qquad (12)$$

and m is given by

$$m = \frac{\delta g}{\delta a} D_a \qquad (13)$$

Thus

$$\Omega_{RS}(\wedge g^{-1}\delta g) = \Omega_{YM} \qquad (14)$$

and the result follows:

$$< \mathcal{O} >_{\Omega} = \int \mathcal{O}(a)\ \delta(g)\ \det m\ \Omega_{YM} \qquad (15)$$

This, of course only holds if $\mathcal{O}(a)$ has its support inside the chosen chart. By construction, the result is independent of the choice of a local section, two local sections differing by a field dependent gauge transformation.

The final outcome is to replace Ω_{YM} by

$$\Omega_{YM\Phi\Pi} = \Omega_{YM}\ \Omega_{\Phi\Pi} \qquad (16)$$

where

$$\Omega_{\Phi\Pi} = \int \mathcal{D}\bar{\omega}\mathcal{D}\omega\mathcal{D}b\ e^{i<b,g(a)>+<\bar{\omega},m\omega>} \qquad (17)$$

where we have used the Stueckelberg Nakanishi Lautrup Lagrange multiplier b, the Faddeev Popov fermionic ghost ω, the Faddeev Popov fermionic Lagrange multiplier (antighost) $\bar{\omega}$. The modern reading of the exercise done with Claude is that not only $\Omega_{YM\Phi\Pi}$ is invariant under the operation s

$$\begin{aligned} sa &= -\mathcal{D}_a\omega \\ s\omega &= -\frac{1}{2}[\omega,\omega] \quad s^2 = 0 \\ s\bar{\omega} &= -ib \\ sb &= 0 \end{aligned} \qquad (18)$$

but, thanks to the introduction of the b-field,

$$i < b, g > + < \bar{\omega}, m\omega >= s(- < \bar{\omega}, g >) \qquad (19)$$

This allows to discuss perturbative renormalization using all the power of locality. The useful part involves the local cohomology of Lie $\mathcal{G}$ in terms of which the observables can be defined and which also classifies obstructions to gauge invariance due to quantum deformations (i.e. anomalies).

We shall see in the next section that the cohomology involved in topological theories is different !

Of course the above discussion is local over orbit space, and a constructive procedure to glue the charts is missing. This is the Gribov problem.

3 Cohomological Theories

E. Witten's 1988 paper [6] contains several things. First, invoking "twisted $N = 2$ supersymmetry" E. Witten gets an action $S(a, \psi, \varphi; ...)$ where ψ resp φ is a 1 resp 0 form with values in Lie G and the dots represent a collection of Lagrange multiplier fields. Then it is observed that

$$QS = 0 \tag{20}$$

with

$$\begin{array}{cc} Qa = \psi & \text{infinitesimal} \\ Q\psi = D_a\varphi & Q^2 = \text{gauge transformation} \\ Q\varphi = 0 & \text{of parameter } \varphi \end{array} \tag{21}$$

Furthermore there is an identity of the form

$$\int trF \wedge F = S - Q\chi(a, \psi, \varphi; ...) \tag{22}$$

where χ is gauge invariant.

The observables are classified according to the gauge invariant cohomology of Q, with the example

$$\begin{aligned} Q\ tr\ F \wedge F &= -d\ tr\ 2F\psi \\ Q\ tr\ 2F\psi &= -d\ tr\ (\psi \wedge \psi + 2F\varphi) \\ Q\ tr(\psi_\wedge \psi + 2F\psi) &= -d(2\psi\varphi) \\ Q\ tr\ 2\psi\varphi &= -d\ tr\ \varphi^2 \\ Q\ tr\ \varphi^2 &= 0 \end{aligned} \tag{23}$$

It follows that integrating the polynomials exhibited in these descent equations over cycles of the correct dimensions yields (non trivial !) elements of the cohomology of Q whose correlation functions are conjectured to reproduce Donaldson's polynomials.

Very soon after the appearance of E. Witten's article, L. Baulieu and I.M. Singer [7] remarked that Eq.(22) can be rewritten as

$$S = \int tr\ F \wedge F + Q\chi(a, \psi, \varphi; ...) \tag{24}$$

so that this action looks like the gauge fixing of a topological invariant. Furthermore, at the expense of introducing a Faddeev Popov ghost ω, Q can be replaced by s:

$$\begin{aligned} sa &= \psi - \mathcal{D}_a\omega \\ s\psi &= -D_a\Omega + [\psi, \omega] \qquad s^2 \equiv 0 \\ s\omega &= \Omega - \frac{1}{2}[\omega, \omega] \\ s\Omega &= -[\omega, \Omega] \end{aligned} \tag{25}$$

(For homogeneity in the notations, we have replaced φ by Ω).

This has however a defect, namely, s has no cohomology and therefore is not adequate to describe the physics of the model.

Inspired by an article by J. Horne [8], devoted to a supersymmetric formulation of this model, S. Ouvry, R. S. and P. van Baal [9] solved that difficulty by phrasing J. Horne's observation as follows: S and χ are not only gauge invariant but also are independent of ω !

In other words they are invariant under

$$\begin{array}{lll} I(\lambda), L(\lambda), & \lambda \in \text{Lie } \mathcal{G} & \\ I(\lambda)\omega = \lambda & I(\lambda) \text{ other} = 0 & \\ L(\lambda)\omega = [\lambda, \omega] & L(\lambda) \text{ other} = & \text{infinitesimal gauge} \\ & & \text{transformation of parameter } \lambda \end{array} \tag{26}$$

and, one can verify that

$$L(\lambda) = [I(\lambda), s]_+ \tag{27}$$

The cohomology that defines the physics of the model is the basic cohomology of s for the operation $\{I(\lambda), L(\lambda)\}$. This is not empty and coïncides with that of Q. Looking into that direction was suggested during a seminar by P. Braam at the CERN theory division in the spring 1988. There it was stated that the subject was the equivariant cohomology of $\mathcal{A}$ (restricted to $F = *F$). Further geometrical interpretations of $\psi\omega\Omega$ were given by L. Baulieu and I.M. Singer [7] and the general set up was precisely phrased in terms of equivariant cohomology by J. Kalkman [10] who developed the algebraic equipment

further. Two general types of equivariant cohomology classes are involved in the present models:

- Mathaï Quillen [11] representatives of Thom class of vector bundles (Gaussian deformations of covariant δ functions). Those occur in the action.

- Equivariant characteristic classes of vector bundles. They are expressed in terms of an arbitrary invariant connection[12]. They provide the known topological observables. In the case where the manifold to be quotiented is a principal bundle, Cartan's "theorem 3"[13] transforms equivariant cohomology classes into basic cohomology classes, by the substitution $\omega \rightarrow \tilde{\omega}, \Omega \rightarrow \tilde{\Omega}$, where $\tilde{\omega}$ is a connection and $\tilde{\Omega}$ its curvature. It is expressible in terms of another identity in which integral representation of both bosonic and fermionic δ functions provides other terms in the action:

$$\int \mathcal{D}\omega\mathcal{D}\Omega \; \delta(\omega - \tilde{\omega}) \; \delta(\Omega - \tilde{\Omega}) = 1 \tag{28}$$

This can only be understood if ω is introduced, although it does not always appear in the action.

We shall now illustrate these general recipes in the case of topological Yang Mills theories (YM_4^{top}).

The observables are constructed as universal cohomology classes of $\mathcal{A}/\mathcal{G}$ as follows: consider the G bundle $P(M,G) \times \mathcal{A}$ and, on it, the $\mathcal{G}$ invariant G connection a (a zero form on $\mathcal{A}$, a one form on $P(M,G)$).

The equivariant curvature of a, in the intermediate scheme (see appendix A) is

$$R_{int}^{eq.} = F(a) + \psi + \Omega \tag{29}$$

with

$$\psi = \delta a. \tag{30}$$

In the Weil scheme, we are interested in

$$R_{w}^{eq.} = F(a) + \psi + \Omega \tag{31}$$

with

$$\psi = \delta a + \mathcal{D}_a \omega. \tag{32}$$

This is the object first considered by L. Baulieu, I.M. Singer [7].

The equivariant characteristic class $tr(R_w^{eq.})^2$ fulfills

$$(d + \delta) \; tr(R_w^{eq.})^2 = 0 \tag{33}$$

which provides the descent equations (Eq.23). Replacing ω by $\tilde{\omega}, \Omega$ by $\tilde{\Omega}$, where $\tilde{\omega}$ is a $\mathcal{G}$ connection on $\mathcal{A}$, provides a basic form on $P(M,G) \times \mathcal{A}$.

One may choose[7,11]

$$\tilde{\omega} = -D_a^* \frac{1}{D_a^* D_a} \delta a \tag{34}$$

provided reducible connections are excluded.

Let now $\mathcal{O}_i(a, \psi, \omega, \Omega)$ be equivariant classes of $\mathcal{A}$ obtained by integration over cycles in M with the proper dimension. We want to find an integral representation in terms of fields of the form on $\mathcal{A}/\mathcal{G}$ corresponding to a basic form $\mathcal{O} = \prod_i \mathcal{O}_i$ and, in the case of a form of maximal degree ("top form") of its integral.

Let $\tilde{a}$ be coordinates of a local section Σ

$$g(\tilde{a}) \equiv 0 \quad \frac{\delta g}{\delta a} \delta \tilde{a} \equiv \frac{\delta g}{\delta a} (\tilde{\psi} - D_{\tilde{a}} \tilde{\omega}) \equiv 0 \tag{35}$$

We have

$$\mathcal{O}(a, \psi, \tilde{\omega}, \tilde{\Omega})_{|\Sigma} = \mathcal{O}\left(\tilde{a}, \delta\tilde{a} + D_{\tilde{a}}\tilde{\omega}_{|\Sigma}, \tilde{\omega}_{|\Sigma}, \tilde{\Omega}_{|\Sigma}\right) \tag{36}$$

This defines a cohomology class on $\mathcal{A}/\mathcal{G}$, independently of the choice of Σ, because of the basicity of $\mathcal{O}$. The expression at hand can be expressed through the introduction of a collection of δ-functions.

First, in the case of YM_4^{top}, one has to restrict to $F = *F$, which goes through a δ function or a smeared gaussian thereof according to the Mathaï Quillen formula (cf. Ref.[11] and appendix A).

The replacement $\omega \to \tilde{\omega}$ $\Omega \to \tilde{\Omega}$ can be carried out using the δ functions of Eq.(28):

$$\begin{aligned} &\int \delta(\omega - \tilde{\omega}) \delta(\Omega - \tilde{\Omega}) \mathcal{D}\omega \mathcal{D}\Omega \\ &= \int \mathcal{D}\bar{\omega} \mathcal{D}\bar{\Omega} \mathcal{D}\omega \mathcal{D}\Omega \; e^{(s+\delta)(\bar{\Omega}(\omega - \tilde{\omega}))} \end{aligned} \tag{37}$$

where s is extended to

$$\begin{aligned} s\bar{\Omega} &= \bar{\omega} - [\omega, \bar{\Omega}] \\ s\bar{\omega} &= [\Omega, \bar{\Omega}] - [\omega, \bar{\omega}] \end{aligned} \tag{38}$$

If $\tilde{\omega}$ is the solution of a local equation e.g.

$$D_a^* \tilde{\Psi} = D_a^* (\delta a + D_a \tilde{\omega}) \tag{39}$$

this can be rewritten, thanks to the cancellation of determinants, as:

$$\int \mathcal{D}\omega \mathcal{D}\Omega \mathcal{D}\bar{\omega} \mathcal{D}\bar{\Omega} \; e^{s(\bar{\Omega} D^* \Psi)} \tag{40}$$

Other local choices can be made, e.g. the flat connection determined by the local section Σ[14], but, in this case, a change of local section produces a change of representative in the cohomology class under consideration due to the associated change of connection.

Finally, the restriction to Σ goes via the insertion of the δ function identity

$$\int \delta(a-\tilde{a})\delta(\psi-\tilde{\psi})\mathcal{D}a\mathcal{D}\psi = 1 \tag{41}$$

This can be rewritten as

$$\int \mathcal{D}a\mathcal{D}\psi \int \mathcal{D}\bar{\alpha}\mathcal{D}\bar{\psi}\; e^{(s+\delta)(\bar{\psi}(a-\tilde{a}))} = 1 \tag{42}$$

with

$$\begin{aligned} s\bar{\psi} &= \bar{\alpha} - [\omega, \bar{\psi}] \\ s\bar{\alpha} &= [\Omega, \bar{\psi}] - [\omega, \bar{\alpha}] \end{aligned} \tag{43}$$

Integrating over all a's and Ψ's yields a field theory representation of forms on orbit space, as advocated in ref.[14]. Integrating over the superfiber (the tangent bundle of a fiber with Grassmann variables on the vectorial part) yields a formal field theory representation of the integral over orbit space of a basic top form. In terms of the local equations Eq.(35), this can be rewritten as

$$\int \mathcal{D}a\mathcal{D}\psi \int \mathcal{D}\bar{\alpha}\mathcal{D}\bar{\psi}\; e^{s(\bar{\gamma}g(a))} = 1 \tag{44}$$

with

$$\begin{aligned} s\bar{\gamma} &= \beta + \omega \cdot \bar{\gamma} \\ s\beta &= -\Omega \cdot \bar{\gamma} + \omega.\beta \end{aligned} \tag{45}$$

where the dot denotes the action of $\mathcal{G}$ on the bundle over $\mathcal{A}$ of which g is a section.

If $\mathcal{O}$ is a top form, integration transforms the integration over the fiber, in Eqs (42, 43) into integration over $\mathcal{A}$, after localizing $\mathcal{O}$ inside the domain of Σ. The result is then a functional integral of the exponential of an action of the form $s\chi$. If this representation involves ultraviolet problems one may conjecture that, besides the necessity to include in χ all terms consistent with power counting the gauge fixing term in Eq.(44) has to be written in the form $sW\chi$ where W is another operation which anticommutes with s and involves a Faddeev Popov ghost field, its graded partner, and the corresponding antighosts. This however is still waiting for confirmation.

In support of the relevance of these constructions, one may give a few examples:

i) The equivariant curvature Eq.(31),(33) precisely yields the observables constructed by E. Witten via the interpretation given by L. Baulieu, I.M. Singer. The same method yields the observables constructed by C. Becchi, R. Collina, C Imbimbo[14] in the case of 2-d topological gravity (see also L. Baulieu, I.M. Singer[7]).

ii) Recent work by M. Kato[15] and collaborators remarking the equivalence of some pairs of topological conformal models through similarity transformations of the form e^R is interpretable by $R = i_M(\omega)$, in J. Kalkman's language[10].

iii) The identification in topological actions of terms which fix a choice of connection is an additional piece of evidence[6],[14].

4 Conclusion

The formalism of equivariant cohomology provides an elegant algebraic set up for topological theories of the cohomological type. Its relationship with $N = 2$ supersymmetry via twisting is still mysterious and may still require some refinements before it provides some principle of analytic continuation. At the moment, it is still a question whether topological theories can be treated as field theories according to strict principles[14] or whether the formal integral representations they provide can at best suggest mathematical conjectures to be mathematically proved or disproved.

Acknowledgments

I wish to thank C. Becchi and C. Imbimbo for numerous discussions about their work on 2d topological gravity. I also wish to thank R. Zucchini for discussions about his recent work.

Appendix A

Equivariant Cohomology

Example 1.

M is a smooth manifold with a smooth action of a connected Lie group $\mathcal{G}$; $\Omega^*(M)$ is the exterior algebra of differential forms on M, d_M the exterior differential; $\lambda \in \text{Lie } \mathcal{G}$ is represented by a vector field $\underline{\lambda} \in \text{Vect}M . i_M(\lambda) = i(\underline{\lambda})$ operates on $\Omega^*(M)$ by contraction with $\underline{\lambda}$; the Lie derivative is defined by

$$\ell_M(\lambda) = \ell(\underline{\lambda}) = [i(\underline{\lambda}), d_M]_+ \tag{46}$$

One has

$$\begin{aligned} [i_M(\lambda), i_M(\lambda')]_+ &= 0 \\ [\ell_M(\lambda), i_M(\lambda')]_- &= i_M([\lambda, \lambda']) \\ [\ell_M(\lambda), \ell_M(\lambda')]_- &= \ell_M([\lambda, \lambda']) \end{aligned} \tag{47}$$

Forms $\omega \in \Omega^*(M)$ such that

$$i_M(\lambda)\omega = 0 \quad \forall \lambda \in \text{Lie } \mathcal{G} \tag{48}$$

are called horizontal.

Forms $\omega \in \Omega^*(M)$ such that

$$\ell_M(\lambda)\omega = 0 \quad \forall \lambda \in \text{Lie } \mathcal{G} \tag{49}$$

are called invariant.

Forms which are both horizontal and invariant are called basic.

The basic de Rham cohomology is the cohomology of d_M restricted to basic forms.

Generalization.

E is a graded commutative differential algebra with differential d_E and two sets of graded derivations $i_E(\lambda)$ (of grading -1) $\ell_E(\lambda)$ (of grading 0) fulfilling Eq.(47), with M replaced by E. The notions of horizontal and invariant elements similarly generalize as well as that of basic cohomology.

Example 2: The Weil algebra of $\mathcal{G}$: $W(\mathcal{G})$.

$$W(\mathcal{G}) = \wedge(\text{Lie } \mathcal{G})^* \otimes S\left((\text{Lie } \mathcal{G})^*\right) \tag{50}$$

whose factors are generated by ω, of grading $1, \Omega$ of grading 2, with values in Lie $\mathcal{G}$. We define the differential d_w by

$$\begin{aligned} d_W\ \omega &= \Omega - \frac{1}{2}[\omega, \omega] \\ d_W\ \Omega &= [\omega, \Omega] \end{aligned} \tag{51}$$

$i_W(\lambda), \ell_W(\lambda)$ by

$$\begin{aligned} i_W(\lambda)\omega &= \lambda \quad i_W(\lambda)\Omega = 0 \\ \ell_W(\lambda) &= [i_W(\lambda), d_W]_+ : \\ \ell_W(\lambda)\omega &= [\lambda, \omega] \\ \ell_W(\lambda)\Omega &= [\lambda, \Omega] \end{aligned} \tag{52}$$

Definition: The equivariant cohomology of M is the basic cohomology of $W(\mathcal{G})\otimes\Omega^*(M)$ for the differential d_W+d_M and the action $i_W(\lambda)+i_M(\lambda), \ell_W(\lambda)+\ell_M(\lambda)$.

This is the Weil model of equivariant cohomology.

One can define the intermediate model according to J. Kalkman[10] by applying the algebra automorphism

$$x \to e^{-i_M(\omega)}x \tag{53}$$

which transforms the differential into

$$d_{int} = d_W + d_M + \ell_M(\omega) - i_M(\Omega) \tag{54}$$

and the operation into

$$\begin{aligned} i_{int}(\lambda) &= i_W(\lambda) \\ \ell_{int}(\lambda) &= \ell_W(\lambda) + \ell_M(\lambda) \end{aligned} \tag{55}$$

From this one easily sees that the equivariant cohomology is that of $[\Omega^*(M)\otimes S((\text{Lie } \mathcal{G})^*)]^{\mathcal{G}}$ with the differential

$$d_C = d_M - i_M(\Omega) \tag{56}$$

where the superscript $\mathcal{G}$ denotes $\mathcal{G}$-invariant elements. This is the Cartan model[13], [10]. If M is a principal $\mathcal{G}$ bundle with a connection $\tilde{\omega}$, the mapping

$$\omega \to \tilde{\omega} \quad \Omega \to \tilde{\Omega} \tag{57}$$

where $\tilde{\Omega}$ is the curvature of $\tilde{\omega}$, maps isomorphically the equivariant cohomology of M into its basic cohomology, independently of the choice of $\tilde{\omega}$. This is Cartan's theorem 3[13].

There are two standard ways to produce non trivial equivariant cohomology classes:

i)[12] If the action of $\mathcal{G}$ can be lifted to a principal bundle $P(M,K)$ with structure group K, and Γ is a $\mathcal{G}$ invariant connection on $P(M,K)$, the intermediate equivariant curvature is defined as

$$R^{eq}_{int}(\Gamma) = D_{int}\Gamma + \frac{1}{2}[\Gamma,\Gamma] = R(\Gamma) - i_P(\Omega)\Gamma \tag{58}$$

One has

$$\begin{aligned} i_{int}(\lambda)\ R^{eq}_{int}(\Gamma) &= 0 \\ \ell_{int}(\lambda)\ R^{eq}_{int} &= [\lambda, R^{eq}_{int}(\lambda)] \end{aligned} \tag{59}$$

It follows that any K invariant polynomial of Lie K, P_{inv} yields an equivariant "characteristic" cohomology class. This can be written in the Weil model using Kalkman's automorphism and is at the root of the construction of topological observables [6], [14].

ii) If $E(X,V)$ is a vector bundle over the manifold X, reducible to $\mathcal{G}$, one may write

$$E(X,V) = P(X,\mathcal{G}) \otimes_{\mathcal{G}} V \tag{60}$$

where P is the associated frame bundle.

There is a basic cohomology class, the universal Thom class obtained as follows [11]:

$$\tau_0 \equiv \delta(v) \wedge dv = N_0 \int db\ d\bar{\omega}\ e^{i<b,v>+<\bar{\omega},dv>} \tag{61}$$

for some normalization constant N_0 where b and $\bar{\omega} \in V^*$, the dual of V, $\int d\bar{\omega}$ means Berezin integration, and $<\ ,\ >$ denotes the duality pairing. Introducing s by

$$\begin{aligned}
s\ v &= dv + \omega v \equiv \psi + \omega v \\
s\ dv &= -\Omega v + \omega dv \\
s\ \omega &= \Omega - \frac{1}{2}[\omega,\omega] \\
s\ \Omega &= -[\omega,\Omega] \\
s\ \bar{\omega} &= -ib - \bar{\omega}\omega \\
s\ ib &= -ib\omega + \bar{\omega}\Omega
\end{aligned} \tag{62}$$

One may write

$$\tau_0 = \delta(v)(\wedge dv) = N_0 \int db\ d\bar{\omega}\ e^{s<\bar{\omega},V>} \tag{63}$$

It is easy to prove that

$$\tau = N_0 \int db\ d\bar{\omega}\ e^{s[<\bar{\omega},v>-i(\bar{\omega},b)]} \tag{64}$$

where $(\bar{\omega},b)$is a $\mathcal{G}$ invariant bilinear form on $\mathcal{G}^*$, is an equivariant class of V, with fast decrease. Replacing ω by $\bar{\omega}$, a connection on $P(X,\mathcal{G})$, yields a basic class of $E(X,V)$, once written in the Weil scheme ($\psi_{Weil} = dv - \omega v$, whereas $\psi_{int} = dv$). The extension of the s-operation to the integration variables brings a substantial simplification to the original calculations.

The substitution of v by a section $v(x)$ transforms τ into the cohomology class associated with the submanifold of X defined by $v(x) = 0$.

Formula 64 gives the Mathaï Quillen representative of the Thom class of $E(X,V)$ and leads to a gaussianly spread Dirac current of the submanifold in question.

As a last example, used in the text, let us describe the Ruelle Sullivan[3,5] class associated with an invariant closed form ω on M:

$$\omega_{RS} = i(\wedge_\alpha e_\alpha)\omega \tag{65}$$

where e_α is a basis of Lie $\mathcal{G}$.

That ω_{RS} is both closed and invariant follows from the closedness and invariance of ω, and horizontality is trivial ($i(e_\alpha)i(e_\alpha) = 0$).

References

1. Bibliographical documentation can be found, e.g., in: BRS Symmetry, M. Abe, N. Nakanishi, Iojima eds, Universal Academy Press, Tokyo, Japan, 1996.
2. The corresponding pages of these notes are available from the author upon request.
3. A. Connes, Non Commutative geometry Academic Press New York USA, 1994, p. 59-71.
4. J. Zinn-Justin, "Quantum field theory and critical phenomena", Oxford Science Publications, Clarendon Press, Oxford 1989, p. 485.
5. D. Ruelle, D. Sullivan, Topology **14** (1975), 319-327.
6. E. Witten, C.M.P. **117** (1988) 353.
7. L. Baulieu, I.M. Singer, *Nucl. Phys.* B **15**, 12 (1988) (Proc. Suppl.); C.M.P. **135** (1991) 253.
8. J.H. Horne, *Nucl. Phys.* B **318**, 22 (1989).
9. S. Ouvry, R. Stora, P. van Baal *Phys. Lett.* B **220**, 159 (1989).
10. J. Kalkman, C.M.P. **153** (1993) 447.
11. V. Mathaï, D. Quillen, Topology **25** (1986) 85; M.F. Atiyah, L. Jeffrey, J.G.P.**7** (1990) 119; S. Cordes, G Moore, S. Rangoolam, Les Houches Lectures 1994.
12. N. Berline, E. Getzler, M. Vergne, Heat Kernels and Dirac Operators, Grundlehren des Mathematischen Wissenschaft 298 Springer Verlag Berlin Heidelberg (1992).; R. Stora, F. Thuillier, J.C. Wallet, Lectures at the 1st Caribbean Spring School of Mathematics and Theoretical Physics, Saint-François, Guadeloupe, May 30-June 5, 1995.
13. H. Cartan, Colloque de Topologie, (Espaces Fibrés), Bruxelles 1950 CBRM, 15-56.

14. C. Becchi, C. Imbimbo, in Ref. [1].
15. in ref. [1].

$N = 2$ SUPERCONFORMAL FIELD THEORIES IN 4 DIMENSIONS AND A-D-E CLASSIFICATION

T. EGUCHI
Department of Physics, Faculty of Science, University of Tokyo, Tokyo 113, Japan

K. HORI*
Institute for Nuclear Study, University of Tokyo, Tokyo 188, Japan

Making use of the exact solutions of the $N = 2$ supersymmetric gauge theories we construct new classes of superconformal field theories (SCFTs) by fine-tuning the moduli parameters and bringing the theories to critical points. In the case of SCFTs constructed from pure gauge theories without matter $N = 2$ critical points seem to be classified according to the A-D-E classification as in the two-dimensional SCFTs.

Recently there have been some major advancements in our understanding of the strong coupling dynamics of 4-dimensional supersymmetric gauge theories [1–4]. In the case of $N = 2$ supersymmetry exact results for the low-energy effective Lagrangians have been obtained for a large class of gauge groups and matter couplings [1,2,5–15]. It turned out that the prepotential of the effective theory develops singularities in the strong coupling region when some of the solitons become massless. The behavior of the theory around the strong coupling singularities can be determined by rewriting the theory in terms of the dual 'magnetic' variables. Information on the strong coupling behavior of the theory together with its known behavior in the weak coupling region leads to a complete determination of the prepotential in the whole range of the moduli space.

In this paper we make use of these exact solutions of $N = 2$ supersymmetric gauge theories and construct systematically new classes of $N = 2$ superconformal field theories (SCFTs) in 4-dimensions. We use the approach of refs.[16,17] where the parameters of the moduli space of the theory (expectation values of the scalar field of the $N = 2$ vector multiplet) and masses of matter hypermultiplets are adjusted so that massless solitons with mutually non-local charges coexist. When solitons of mutually non-local charges are present, the system is necessarily at a critical point and one obtains a superconformal field theory.

*Address after September 1996: Department of Physics, University of California, Berkeley, CA 94720

$N = 2$ SUPERCONFORMAL FIELD THEORIES IN 4 DIMENSIONS AND A-D-E CLASSIFICATION

T. EGUCHI
Department of Physics, Faculty of Science, University of Tokyo, Tokyo 113, Japan

K. HORI*
Institute for Nuclear Study, University of Tokyo, Tokyo 188, Japan

Making use of the exact solutions of the $N = 2$ supersymmetric gauge theories we construct new classes of superconformal field theories (SCFTs) by fine-tuning the moduli parameters and bringing the theories to critical points. In the case of SCFTs constructed from pure gauge theories without matter $N = 2$ critical points seem to be classified according to the A-D-E classification as in the two-dimensional SCFTs.

Recently there have been some major advancements in our understanding of the strong coupling dynamics of 4-dimensional supersymmetric gauge theories [1–4]. In the case of $N = 2$ supersymmetry exact results for the low-energy effective Lagrangians have been obtained for a large class of gauge groups and matter couplings [1,2,5–15]. It turned out that the prepotential of the effective theory develops singularities in the strong coupling region when some of the solitons become massless. The behavior of the theory around the strong coupling singularities can be determined by rewriting the theory in terms of the dual 'magnetic' variables. Information on the strong coupling behavior of the theory together with its known behavior in the weak coupling region leads to a complete determination of the prepotential in the whole range of the moduli space.

In this paper we make use of these exact solutions of $N = 2$ supersymmetric gauge theories and construct systematically new classes of $N = 2$ superconformal field theories (SCFTs) in 4-dimensions. We use the approach of refs.[16,17] where the parameters of the moduli space of the theory (expectation values of the scalar field of the $N = 2$ vector multiplet) and masses of matter hypermultiplets are adjusted so that massless solitons with mutually non-local charges coexist. When solitons of mutually non-local charges are present, the system is necessarily at a critical point and one obtains a superconformal field theory.

*Address after September 1996: Department of Physics, University of California, Berkeley, CA 94720

In the following we first describe our recent work [18] and discuss in detail the $SU(N)$ gauge theories coupled with matter hypermultiplets and find new classes of non-trivial SCFTs. We locate superconformal points and determine the critical exponents of scaling operators. We shall see that the nature of the SCFTs is controlled by the unbroken sub-group of $SU(N)$ and the global flavor symmetry. We then study SCFTs based on $SO(N), Sp(2N)$ gauge theories.

It turns out that in the case of pure gauge theories without matter hypermultiplets SCFTs constructed from $SU(N+1), SO(2N+1)$ and $Sp(2N)$ gauge theories are identical and form a single universality class. On the other hand, SCFTs constructed from $SO(2N)$ are distinct and form another universality class. We call these as A_N-type and D_N-type SCFTs, respectively. Critical exponents of their scaling fields are expressed by a formula $2(e_i+1)/(h+2)$, $i = 1, 2, \cdots, N$ where e_i and h are Dynkin exponents and dual-Coxeter numbers of A_N and D_N algebras, respectively.

This suggests the possibility of the $A-D-E$ classification of $N=2$ four-dimensional SCFTs. As for the case of the E_N gauge theories, we will analyze their critical behaviors by making use of the string-theoretic construction given recently in [21] which reproduces elliptic curves and differential forms of known $N=2$ gauge theories starting from K_3-fibered Calabi-Yau manifolds. In the case of SCFTs based on E_N groups we again find the formula for their critical exponents $2(e_i+1)/(h+2)$ where e_i and h represent the Dynkin exponents and dual-Coxeter numbers of E_6, E_7 and E_8 algebras.

Thus we conjecture that there exists a $A-D-E$ classification behind the $N=2$ 4-dimensional SCFTs: the classification originates from that of the degeneration of K_3 surfaces which appear in the K_3-fibered Calabi-Yau manifolds in the study of heterotic-type II string duality [20,22,23].

SCFTs from gauge theories coupled to matter

Let us first briefly recall the results of ref.[17] on $SU(2)$ gauge theory. In the case of the gauge group $SU(2)$ it is possible to introduce matter hypermultiplets (in the vector representation) up to $N_f = 4$ without loosing the asymptotic freedom. In [17] authors considered the case of a common mass $m = m_i$ $(i = 1, \cdots, N_f)$ for all N_f flavors. This is the case when the highest criticality is reached for each value of N_f. Parameters of the theory are then given by $u = \frac{1}{2}\mathrm{Tr}\phi^2$ and m where ϕ denotes the scalar field of the $N=2$ vector multiplet.

Let us discuss the case of $N_f = 2$ for the sake of illustration. We first

recall that the exact solution of the theory is described using an elliptic curve

$$\mathcal{C}: \quad y^2 = (x^2 - u + \frac{\Lambda^2}{8})^2 - \Lambda^2(x+m)^2, \tag{1}$$

where Λ is the dynamical mass scale of the theory. The discriminant of the curve is given by

$$\Delta = \frac{1}{16}\Lambda^4(8m^2 - 8u + \Lambda^2)^2\Delta_m, \tag{2}$$

$$\Delta_m = (8u - 8\Lambda m + \Lambda^2)(8u + 8\Lambda m + \Lambda^2). \tag{3}$$

The power 2 of the factor $8m^2 - 8u + \Lambda^2$ in Δ means that the singularity at $u = u^* = m^2 + \Lambda^2/8$ has a multiplicity 2 and belong to the $\underline{2}$ representation of the flavor symmetry group $SU(N_f = 2)$. When the mass m becomes large, this zero of the discriminant moves out to ∞ as $u \approx m^2$ and becomes the massless squark singularity with its bare mass $\pm m$ being canceled by the vacuum value $a = \pm m$ of the scalar field ϕ. Let us call such a singularity as the squark singularity although in the strong coupling region it represents a massless solitonic state carrying a magnetic charge.

In order to locate the superconformal point one first sets the value of u at the squark singularity u^*. Then the curve and Δ_m become

$$y^2 = (x+m)^2(x-m-\Lambda)(x-m+\Lambda), \quad \Delta_m = 4\Lambda^4(2m+\Lambda)^2(2m-\Lambda)^2. \tag{4}$$

One then adjusts the value of m at $m^* = \pm\Lambda/2$ so that Δ_m vanishes. Then the squark singularity collides with the singularity of the monopole (or dyon) and we generate a critical behavior

$$y^2 = (x \pm \frac{\Lambda}{2})^3(x \mp \frac{3\Lambda}{2}). \tag{5}$$

It is straightforward to analyze perturbations around this critical point and determine scaling dimensions $[u], [m]$ of the parameters u, m given by $[m] = 2/3, [u] = 4/3$. Results of ref.[17] are summarized in Table 1.

N_f	m	u	C_2	C_3
1	4/5	6/5		
2	2/3	4/3	2	
3	1/2	3/2	2	3

Table 1: Universality Classes of $N = 2$ SCFTs based on $SU(2)$ Gauge Theory

We note that in the cases $N_f \geq 2$ there appear Casimir operators C_j associated with the global flavor symmetry group $SU(N_f)$ with the dimensions $[C_j] = j$.

Let us now turn to the case of the $SU(N_c)$ theory and start presenting our results. We consider the case of N_f matter hypermultiplets in vector representations with a common mass m. (We may add extra flavors with different masses, however, at critical points this amounts only to shifting the rank N_c of the group). The curve is given by [9]

$$\mathcal{C}: \qquad y^2 = C(x)^2 - G(x), \tag{6}$$

$$C(x) = x^{N_c} + s_2 x^{N_c-2} + \cdots + s_{N_c} + \frac{\Lambda^{2N_c-N_f}}{4} \sum_{i=0}^{N_f-N_c} x^{N_f-N_c-i} \begin{pmatrix} N_f \\ i \end{pmatrix} m^i, \tag{7}$$

$$G(x) = \Lambda^{2N_c-N_f}(x+m)^{N_f}, \tag{8}$$

where the terms proportional to $\Lambda^{2N_c-N_f}$ in $C(x)$ are absent in the case $N_f < N_c$. The meromorphic 1-form is given by

$$\lambda = x d \log \frac{C-y}{C+y}. \tag{9}$$

Expectation values of the scalar field ϕ are obtained by integrating the 1-form around suitable homology cycles of the curve.

When $N_f \geq 2$, it is possible to show that the discriminant of the curve $\mathcal{C}$ has a factorized form

$$\Delta = \Delta_s \Delta_m, \tag{10}$$

$$\Delta_s = \left(C(x=-m)\right)^{N_f} \tag{11}$$

((10),(11) may be shown in the following way. In the case of even flavors $N_f = 2n$ we can split the curve as $y^2 = y_+(x)y_-(x)$ with $y_\pm(x) = C(x) \pm \Lambda^{N_c-n}(x+m)^n$. Then the discriminant becomes $\Delta = (\mathit{resultant}(y_+, y_-))^2 \times \mathit{resultant}(y_+, y_+') \times \mathit{resultant}(y_-, y_-')$. Here $'$ means the derivative in x. If one uses the formula $\mathit{resultant}(y_+, y_-) = C(-m)^n$, one recovers (11). The N_f=odd case may be treated in a similar way). The factor Δ_s carries the power N_f and represents the squark singularity. In our search for superconformal points let us first set $\Delta_s = 0$. This fixes the value of s_{N_c} at $s_{N_c}^* = -(-m)^{N_c} - s_2(-m)^{N_c-2} \cdots$. The function $C(x)$ becomes divisible by $x+m$ and expressed as $C(x) = (x+m)C_1(x)$ with a polynomial $C_1(x)$ of order $N-1$. The curve becomes $y^2 = (x+m)^2\Big(C_1(x)^2 - \Lambda^{2N_c-N_f}(x+m)^{N_f-2}\Big)$. It turns out that the value of Δ_m at $s_N = s_N^*$ factors as $\Delta_1\Delta_{2m}$ where Δ_1 is (a power of) the resultant of $C_1(x)$ and $x+m$. We next set $\Delta_1 = 0$ by adjusting the

parameter s_{N-1} to its critical value s^*_{N-1}. $C_1(x)$ then becomes divisible by $x+m$ and is written as $C_1(x) = (x+m)C_2(x)$ with a polynomial $C_2(x)$ of order $N-2$. The curve now has a 4-th order degeneracy at $x=-m$, $y^2 = (x+m)^4(C_2(x)^2 - \Lambda^{2N_c-N_f}(x+m)^{N_f-4})$ and describes a critical theory.

We can iterate this procedure. We extract powers of $x+m$ from $C(x)$ by adjusting parameters $s_N, s_{N-1}, s_{N-2}, \cdots$ successively and bring the curve to higher criticalities. As far as the extracted power ℓ of $x+m$ from $C(x)$ does not exceed $N_f/2$, the order of degeneracy of $C(x)^2$ is lower than that of $G(x)$ and the curve acquires a degeneracy of order $2\ell, y^2 \approx (x+m)^{2\ell}$.

When ℓ becomes greater than $N_f/2$, the order of degeneracy of $C(x)^2$ exceeds that of $G(x)$ and one can not necessarily increase the criticality of the curve by extracting higher powers out of $C(x)$. As we shall see below, when N_f=odd, the highest criticality of the curve is given by N_f, $y^2 \approx (x+m)^{N_f}$ while in the case of even flavor $N_f = 2n$ it is given by N_c+n, $y^2 \approx (x+m)^{N_c+n}$.

We classify critical points of $SU(N_c)$ theories into 4 groups;

$$1.\quad y^2 \approx (x+m)^{2\ell}; \qquad \begin{cases} 2\ell \le 2n-2, & N_f = 2n \\ 2\ell \le 2n, & N_f = 2n+1 \end{cases} \tag{12}$$

$$2.\quad y^2 \approx (x+m)^{N_f}; \qquad N_f = 2n+1 \tag{13}$$

$$3.\quad y^2 \approx (x+m)^{N_f}; \qquad N_f = 2n \tag{14}$$

$$4.\quad y^2 \approx (x+m)^{p+N_f}; \qquad 0 < p \le N_c - n, \quad N_f = 2n \tag{15}$$

It turns out that theories of the class 1 above are free field theories. In order to construct non-trivial theories we have to bring the criticality of the curve at least as high as N_f as in (13),(14),(15). We first analyze the SCFTs of the class 1 above and then turn to the discussion of the non-trivial SCFTs given by the classes 2, 3 and 4.

Class 1

In these theories the $G(x)$ term in the curve (6) has a higher criticality than $C(x)^2$ and may be ignored when we analyze the theory at the critical point. We may also ignore terms with higher power of $x+m$ in $C(x)$ than $(x+m)^\ell$. Then $y^2 \approx C(x)^2 = (x+m)^{2\ell}$. We apply perturbations to this critical point as

$$C(x) = (x+m)^\ell - t_j(x+m)^j, \qquad 0 \le j \le \ell - 1. \tag{16}$$

Perturbation splits the ℓ-fold zeros of $C(x)$ at $x=-m$ into j-fold zeros. In order to describe the removal of the degeneracy we make a change of variable

$$x = -m + t_j{}^{\frac{1}{\ell-j}} z. \tag{17}$$

Then

$$C(x) = t_j^{\frac{\ell}{\ell-j}}(z^\ell - z^j). \tag{18}$$

Integration contours of the 1-form are given by the paths connecting $(\ell - j)$-th roots of unity in the complex z-plane. It is possible to show that the term $d\log(C-y)/(C+y)$ in λ (9) does not produce a factor dependent on t_j. Only a power of t_j appears from the factor x in front of $d\log(C-y)/(C+y)$ under the change of variable (16). Thus periods behave as $a_i, a_i^D \approx t_j^{\frac{1}{\ell-j}}$. By requiring the dimensions of the periods to be unity [17], we find that $[t_j] = \ell - j$. Integral values of the dimensions indicate that this is a free field theory.

In addition to the above perturbations removing the degeneracy of the roots of $C(x)$ we may consider perturbations which remove the degeneracy of the masses m_i, $i = 1, \cdots, N_f$ of the hypermultiplets. Under perturbation $G(x)$ is replaced by

$$\Lambda^{2N_c - N_f}\Big((x+m)^{N_f} + \sum_{i=2}^{N_f} C_i (x+m)^{N_f - i}\Big). \tag{19}$$

Proceeding as in the previous case we can easily determine the exponents of the fields $C_i(x+m)^{N_f - i}$ as

$$[C_i] = i, \qquad i = 2, \cdots, N_f. \tag{20}$$

These are in fact the Casimir operators of the $SU(N_f)$ flavor symmetry.

A basic feature of the superconformal points of the class 1 is that the value of the mass m is left arbitrary at the critical point and the critical values of the tuned parameters $s_{N_c}^*, s_{N_c-1}^*, \cdots, s_{N_c-\ell+1}^*$ become simultaneously large as the mass m is increased. Values of the other parameters $s_2, \cdots, s_{N_c-\ell}$ are not fixed at the critical point and we may also take them to be large. Thus the critical points of class 1 stretch out to the "exterior region" of the moduli space. When the values of the moduli parameters are all much larger than Λ, we may adopt the semi-classical reasoning. If we ignore instanton effects and put $\Lambda = 0$, the curve becomes classical $y^2 = C(x)^2 = (x^{N_c} + s_2 x^{N_c-2} + \cdots + s_N)^2$ and its discriminant is given by a classical expression. Then the condition of degeneracy of the function $C(x) \approx (x+m)^\ell$ becomes the condition of the degeneracy of the eigenvalues of the field ϕ, i.e. ℓ of the eigenvalues of ϕ have to coincide. This implies that the $SU(\ell)$ sub-group of $SU(N_c)$ is left unbroken at class 1 superconformal points.

Thus near the region of the critical points of the class 1 theories we have effectively an $SU(\ell)$ gauge theory coupled with N_f hypermultiplets. Since $2\ell < N_f$, the theory is in the asymptotically non-free regime. We expect that

class 1 theories are at trivial fixed points. Let us denote the class 1 theory with the behavior $y^2 \approx (x+m)^{2\ell}$ as $M_{2\ell}^{N_f}$.

Class 2

Let us now turn to the class 2 theories which are intrinsic to the odd flavor case $N_f = 2n+1$. Class 2 theories are obtained by further extracting powers of $x+m$ from the function $C(x)$. Once the extracted power ℓ exceeds n, $C(x)^2$ term becomes irrelevant at the critical point and the curve is dominated by the term $G(x)$, $y^2 \approx (x+m)^{N_f}$. The perturbations on the eigenvalues of ϕ are given by

$$y^2 = ((x+m)^\ell - t_j(x+m)^j)^2 - \Lambda^{2N_c-N_f}(x+m)^{N_f} \tag{21}$$

$$\approx t_j^2(x+m)^{2j} - \Lambda^{2N_c-N_f}(x+m)^{N_f}, \quad 0 \le j < N_f/2. \tag{22}$$

By making a change of variable as

$$x = -m + t_j^{\frac{2}{N_f-2j}} z \tag{23}$$

and using the fact that $C \approx y$, we find that again the only power of t_j comes from the factor x in front of the 1-form. The scaling dimensions are given by

$$[t_j] = \frac{N_f}{2} - j, \qquad j = 0, 1, \cdots, n. \tag{24}$$

These fields have half-integral dimensions. If one restricts oneself to relevant fields, there exist only two $[t_n] = 1/2, [t_{n-1}] = 3/2$.

There also exist Casimir operators associated with the $SU(N_f)$ symmetry and the operators carry integral dimensions

$$[C_j] = j \tag{25}$$

as in the class 1 case.

The special feature of the class 2 theories is that these superconformal points do not depend on N_c so far as $2N_c > N_f$. In fact the dimensions of the relevant operator $1/2, 3/2$ are exactly the same as in the case of $N_c = 2$ (see Table 1). Thus they represent a universality class of $N = 2$ SCFTs with the global $SU(N_f = \text{odd})$ symmetry. We denote this universality class as M_{2n+1}^{2n+1}.

As in class 1 theories class 2 conformal points extend to the semi-classical regions of the moduli space, i.e. all the adjusted parameters grow as m increases. (The case $N_f = 2N_c - 1$ is an exception. In this case the value of m is fixed to be in the strong coupling region $m^* \approx \Lambda$). The same argument as

in the class 1 theories applies and we have effectively an $SU(\ell)$ gauge theory interacting with N_f matter multiplets. In the present case, however, $2\ell > N_f$ and the system is in the asymptotically free regime. Thus the class 2 theories are non-trivial and are interacting superconformal models.

Class 3

Let us now go to class 3 theories of even flavors $N_f = 2n$. These theories are obtained by further adjusting the parameters of class 1 theories so that a power $(x+m)^n$ is extracted from $C(x)$. $C(x)$ is written as $C(x) = (x+m)^n C_n(x)$ with an $(N_c - n)$-th order polynomial $C_n(x)$ and the curve becomes

$$y^2 = (x+m)^{2n}(C_n(x)^2 - \Lambda^{2N_c - N_f}). \tag{26}$$

Without further tuning parameters the curve has the behavior $y^2 \approx (x+m)^{2n}$.

Class 3 theories also extend to the semi-classical regions $\{s_i\}, m \gg \Lambda$. At the class 3 critical point the unbroken subgroup is $SU(n)$ which is exactly the gauge group whose beta function vanishes in the presence of $N_f = 2n$ flavors. Class 3 theories are therefore expected to be in the same universality class as the known $N = 2$ SCFT [2,9,11,12] of $SU(n)$ gauge theory with $2n$ massless matter multiplets. In fact we may decompose each of the squark superfields Q^a, $a = 1, 2, \cdots, 2n$ (belonging to the vector representation of $SU(N_c)$) into a sum of vector and singlet representations of $SU(n)$. Then the bare masses of the squark superfields of the vector representation of $SU(n)$ are exactly canceled by the vacuum expectation values of the field ϕ (which has n degenerate eigenvalues $-m$). Thus there exist $2n$ massless squarks belonging to the vector representation of $SU(n)$. Hence the class 3 theories belong to the same universality class as the massless $N_c = n, N_f = 2n$ theory. We denote this universality class as M_{2n}^{2n}.

Class 4

Let us now turn to class 4 theories of even flavors $N_f = 2n$. We start from the curve of the class 3 theory (26) and enhance its criticality by adjusting the parameters of $C_n(x)$. We first note that the right-hand-side of (26) is given by a product of factors, $(x+m)^{2n}$ and $C_n(x)^2 - \Lambda^{2N_c - N_f}$. The first factor describes the curve of the M_{2n}^{2n} theory and the second factor describes that of the pure Yang-Mills theory of gauge group $SU(N_c - n)$ (without matter fields). Thus class 4 theories are interpreted as the coupled model of M_{2n}^{2n} and pure Yang-Mills theories.

Let us rewrite the curve as

$$y^2 = (x+m)^{2n}(C_n(x) + \Lambda^{N_c-n})(C_n(x) - \Lambda^{N_c-n}) \tag{27}$$

and expand $C_n(x)$ in powers of $(x+m)$, $C_n(x) = (x+m)^{N_c-n} - Nm(x+m)^{N_c-n-1} + \tilde{s}_2(x+m)^{N_c-n-2} + \cdots + \tilde{s}_{N_c-n}$. We can successively adjust the parameters as $\tilde{s}^*_{N_c-n} = -\Lambda^{N_c-n}, \tilde{s}^*_{N_c-n-1} = 0$ etc. and bring the curve to higher criticalities. The number of available parameters is $N_c - n$ and hence the highest criticality is given by $y^2 \approx (x+m)^{N_c+n}$. The parameters of the most singular curve are all fixed inside the strong coupling region and hence it represents an inherently strongly coupled field theory (at a lower criticality there are parameters which are left undetermined). We denote the universality class represented by a curve $y^2 \approx (x+m)^{p+2n}$ $(0 < p \leq N_c - n)$ as M^{2n}_{2n+p}.

Let us next analyze the perturbations of the most critical theory $M^{2n}_{N_c+n}$. Properties of perturbations of other theories are similar. The critical value of the mass m^* vanishes in $M^{2n}_{N_c+n}$ and we may set $m = 0$ from the start. (The case $N_f = 2n-2$ is an exception. In this case m^* is non-zero and is of the order of Λ). Then it is easy to locate the critical point

$$s^*_{N_c} = 0, s^*_{N_c-1} = 0, \cdots, s^*_{N_c-n} = \pm\Lambda^{N_c-n}, \cdots, s^*_3 = 0, s^*_2 = 0. \tag{28}$$

(If $N_f > N_c$, $s^*_{2N_c-N_f} = -\Lambda^{2N_c-N_f}/4$). The curve reads as $y^2 = x^{N_c+n}(x^{N_c-n} - 2\Lambda^{2N_c-N_f})$.

We then apply perturbations of the form $t_j x^j$ to $C(x)$,

$$y^2 \approx (x^{N_c} - t_j x^j)x^n, \qquad 0 \leq j \leq N_c - 1. \tag{29}$$

If we make a change of variable

$$x = t_j^{\frac{1}{N_c-j}} z, \tag{30}$$

we find $y \approx t_j^{(N_c+n)/2(N_c-j)}\sqrt{(z^{N_c} - z^j)z^n}, C \approx t_j^{n/(N_c-j)}z^n$ and $|y| \ll |C|$. The 1-form behaves as

$$\lambda = xd\log(1-y/C)/(1+y/C) \approx -2xd(y/C) \approx t_j^{(N_c+2-n)/2(N_c-j)}. \tag{31}$$

Hence the dimensions are given by

$$[t_j] = \frac{2(N_c-j)}{N_c+2-n}, \qquad 0 \leq j \leq N_c - 1. \tag{32}$$

Class 4 theories are strongly coupled conformal theories and the dimensions of the scaling fields (32) could become arbitrary small as N_c is increased.

It is easy to repeat the computation in the case of lower critical points M^{2n}_{2n+p} with $p < N_c - n$ and we find that the dimensions of the perturbations are given by (32) with $N_c - n$ being replaced by p. In fact the lower critical point of an $SU(N_c)$ gauge theory corresponds exactly to the most critical one of the gauge theory of a lower rank $SU(N_c' = n + p)$.

In summary, we have obtained the list of universality classes of Table 2.

class	name		dimensions
1	$M^{N_f}_{2\ell}$	$2\ell < N_f$	$\begin{cases} 1, 2, 3, \cdots, \ell \\ 2, 3, \cdots, N_f \end{cases}$
2	M^{2n+1}_{2n+1}		$\begin{cases} \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \cdots, n+\frac{1}{2} \\ 2, 3, \cdots, 2n+1 \end{cases}$
3	M^{2n}_{2n}		$\begin{cases} 1, 2, 3, \cdots, n \\ 2, 3, \cdots, 2n \end{cases}$
4	M^{2n}_{2n+p}	$0 < p \leq N_c - n$	$\begin{cases} \frac{2j}{p+2},\ j = 1, 2, \cdots, n+p \\ \frac{2(p+j)}{p+2},\ j = 2, 3, \cdots, 2n \end{cases}$

Table 2: Universality Classes of $N = 2$ SCFTs based on $SU(N_c)$ Gauge Theory

In the second row of dimensions in each universality class exponents of the Casimir operators of the global flavor symmetry are listed. Note that a part of the scaling dimensions of the M^{2n}_{2n+p} theory ($\frac{2j}{p+2}$, $j = 2, \cdots, p$) agree with those given by the pure $SU(p)$ gauge theory.

Let us next discuss SCFTs based on $N = 2$ gauge theories with gauge groups $SO(N_c)$ and $Sp(2N_c)$ coupled with matter in vector representations. We first recall that the curves of $SO(N_c)$ and $Sp(2N_c)$ theories are given by [7,8,12,13]

$$SO(2r): \quad \begin{aligned} y^2 &= C(x)^2 - \Lambda^{2(2r-2-N_f)} x^4 (x^2 - m^2)^{N_f}, \\ C(x) &\equiv x^{2r} + s_2 x^{2r-2} + \cdots + s_{2r-2} x^2 + \tilde{s}_r^2, \end{aligned} \tag{33}$$

$$SO(2r+1): \quad \begin{aligned} y^2 &= C(x)^2 - \Lambda^{2(2r-1-N_f)} x^2 (x^2 - m^2)^{N_f}, \\ C(x) &\equiv x^{2r} + s_2 x^{2r-2} + \cdots + s_{2r-2} x^2 + s_{2r}, \end{aligned} \tag{34}$$

$$Sp(2r): \quad \begin{aligned} x^2 y^2 &= C(x)^2 - \Lambda^{2(2(r+1)-N_f)} (x^2 - m^2)^{N_f}, \\ C(x) &\equiv x^{2r+2} + s_2 x^{2r} + \cdots + s_{2r} x^2 + \Lambda^{2(r+1)-N_f} m^{N_f}. \end{aligned} \tag{35}$$

(There is an extra term $P(x)$ which is a polynomial of order $2N_f - (N_c - 2)$ in x and m for $N_f \geq N_c/2 - 2$ ($N_f \geq N_c/2 - 3/2$) in $SO(N_c = \text{even})$

($SO(N_c = \text{odd})$ theories). The 1-forms read as

$$\lambda = x d \log \frac{C-y}{C+y}; \qquad SO(N_c), \tag{36}$$

$$\lambda = x d \log \frac{C-xy}{C+xy}; \qquad Sp(2N_c). \tag{37}$$

We can again adjust the moduli parameters and extract powers of $x^2 - m^2$ from $C(x)$ and bring theories to superconformal points. Depending on the power ℓ of $(x^2 - m^2)^\ell$ extracted from $C(x)$ and the parity of N_f we can construct the analogues of the class 1-4 theories.

It is easy to argue generally that the SCFTs built on $SO(N_c), Sp(2N_c)$ gauge theories with $m^* \neq 0$ are identical to those we have just constructed for $SU(N_c)$ gauge symmetry. In fact at the critical point $x^2 \approx m^{*2} \neq 0$ extra factors of x^4 and x^2 in (47)-(49) beome irrelevant and the curves and the 1-forms become exactly the same as those of the $SU(N_c)$ case. Thus the scaling dimensions of the theories are identical to those listed in Table 2.

We may also note that when the $m^* \neq 0$, the global flavor symmetry of the theory is $SU(N_f)$ irrespective of the gauge groups. Moreover, when $C(x)$ is divisible by $(x^2 - m^2)^\ell$, the unbroken subgroup of the gauge group is given by $SU(\ell)$. (The scalar field ϕ possesses ℓ pairs of degenerate eigenvalues $m, -m$ which breaks the gauge groups $SO(2r), Sp(2r)$ down to $SU(\ell)$, $\ell < r$). Thus the flavor group and the effective gauge group coincide with those of the $SU(N_c)$ theory and the SCFTs built on $SO(N_c), Sp(2N_c)$ agree with those of $SU(N_c)$. Note that, however, SCFTs based on $SO(N_c)$ and $Sp(2N_c)$ gauge theories with $m^* = 0$ have flavor symmetries $Sp(2N_f)$ and $SO(2N_f)$, respectively [12] and belong to different universality classes.

SCFTs constructed from pure gauge theories

So far we have assumed that $N_f \geq 2$ so that we can distinguish squarks from other singularities. Let us now turn to the discussion of the pure gauge case $N_f = 0$ and its universality classes. As we shall see below, $SO(2r+1)$ and $Sp(2r)$ theories have critical points at $x = x^* \neq 0$ and their SCFTs belong to the same universality as the $SU(r+1)$ theory. On the other hand, $SO(2r)$ theories have critical points at $x = x^* = 0$ and provide new universality classes.

In the case of $SU(r+1)$ and $SO(2r)$ one can easily locate the highest criticality of pure $N = 2$ Yang-Mills theories: $y^2 = x^{r+1}$ for $SU(r+1)$ and $y^2 = x^{2r+2}$ for $SO(2r)$. The moduli parameters are tuned to be of the order Λ and these are strongly coupled SCFTs. Let us denote their universality classes as MA_r and MD_r, respectively. Scaling dimensions are given by $2j/(r+3)$ ($j =$

$2, 3, \cdots, r+1$) for MA_r and j/r ($j = 2, 4, \cdots, 2r-2, r$) for MD_r, respectively. On the other hand, in the case of groups $SO(2r+1)$ and $Sp(2r)$ it is not easy to locate the highest criticality explicitly. By counting the number of parameters, however, we find that the singularity is of the form $y^2 = (x^2 - b^2)^{r+1}$ ($b \neq 0$ is of order Λ) and their SCFTs belong to the same universality class as MA_r. In summary, in Table 3 we present a list of universality classes in $N = 2$ pure Yang-Mills theories with classical gauge groups.

name	gauge groups	dimensions	
MA_r	$SU(r+1), SO(2r+1), Sp(2r)$	$2\dfrac{\mathrm{e}+1}{h+2}$	$\mathrm{e} = 1, 2, \cdots, r$ $h = r+1$
MD_r	$SO(2r)$	$2\dfrac{\mathrm{e}+1}{h+2}$	$\mathrm{e} = 1, 3, \cdots, 2r-3, r-1$ $h = 2r-2$

Table 3: SCFTs based on $N = 2$ pure Yang-Mills theories

We explicitly write down the dimensions for lower rank theories:

rank 2:	MA_2	4/5, 6/5
rank 3:	MA_3	2/3, 1, 4/3
rank 4:	MA_4	4/7, 6/7, 8/7, 10/7
	MD_4	1/2, 1, 3/2, 1

Note that there exist unique universality classes in rank 2 and 3 theories and they coincide with the $SU(2)$ gauge theory with $N_f = 1$ and $N_f = 2$ flavors, respectively (see Table 1). At rank 4, there appear two universality classes and one of them, MD_4, coincides with the $N_f = 3$, $SU(2)$ theory.

E_n Gauge Theories

In the above we have seen that the SCFTs based on pure gauge theories with gauge groups $SU(r+1), SO(2r+1), Sp(2r)$ coincide and give a universality class MA_r of SCFTs while those based on $SO(2r)$ are different and provide another universality class MD_r. Critical exponents of both series are expressed as

$$2\frac{e_i+1}{h+2}, \quad i = 1, 2, \cdots, r \tag{38}$$

where e_i and h are the Dynkin exponents and dual-Coxeter numbers of the algebras A_r and D_r. This result suggets the possibility of an $A-D-E$ type classification of $N = 2$ SCFTs.

We would like to now discuss critical behaviors of E_n gauge theories in order to test the idea of the A-D-E classification. At the moment an explicit solution of E_n gauge theories is not known although a method for their construction has been suggested in [19]. In the following we adopt a different approach making use of a string-theoretic derivation of the 4-dimensional Yang-Mills theories [20,21]. In [21] authors have streamlined the derivation of (hyperelliptic) curves and differential forms of the $N = 2$ Yang-Mills theory by considering the Calabi-Yau manifolds with a K_3 fibration [20,22,23]. For instance, we start from a Calabi-Yau hypersurface in a weighted projective space $WP^5_{1,1,2,8,12}$

$$W = \frac{1}{24}(x_1^{24} + x_2^{24}) + \frac{1}{12}x_3^{12} + \frac{1}{3}x_4^3 + \frac{1}{2}x_5^2 + \psi_0(x_1x_2x_3x_4x_5)$$
$$+\frac{1}{6}\psi_1(x_1x_2x_3)^6 + \frac{1}{12}(x_1x_2)^{12} = 0. \tag{39}$$

(39) is known to describe the $SU(3)$ pure gauge theory in the field theory limit [20]. By introduing new parameters $a = -\psi_0^6/\psi_1$, $b = \psi_2^{-2}$, $c = -\psi_2/\psi_1^2$ and change of variables $x_1/x_2 = \zeta^{\frac{1}{12}} b^{\frac{1}{24}}$, $x_1^2 = x_0\zeta^{\frac{1}{12}}$, W is rewritten as

$$W(\zeta, a, b, c) = \frac{1}{24}(\zeta + \frac{b}{\zeta} + 2)x_0^{12} + \frac{1}{12}x_3^{12} + \frac{1}{3}x_4^3 + \frac{1}{2}x_5^2 + \frac{1}{6\sqrt{c}}(x_0x_3)^6 + (\frac{a}{\sqrt{c}})^{\frac{1}{6}}x_0x_3x_4x_5. \tag{40}$$

For each fixed value of z, $W = 0$ describes a K_3 surface (we choose a patch $x_0 = 1$).

Discriminant of the Calabi-Yau manifold (40) is given by

$$\Delta \approx (b-1)((1-c)^2 - bc^2)(((1-a)^2 - c)^2 - bc^2) \tag{41}$$

The field theory limit is achieved by taking

$$a = -2\epsilon u^{3/2}/3\sqrt{3}, \quad b = \epsilon^2\Lambda^6, \quad c = 1 - \epsilon(-2u^{2/3}/3\sqrt{3} + v) \tag{42}$$

and letting $\epsilon \equiv (\alpha')^{3/2} \to 0$. Here u, v are gauge invariant Casimirs of $SU(3)$. After a suitable change of variables W takes the form

$$W = z + \frac{\Lambda^6}{z} + 2C_{A_2}(x, u, v) + s^2 + w^2 \tag{43}$$

where $C_{A_2}(x, u, v) = x^3 - ux - v$.

Now one considers the period integral of the holomorphic 3-form on the Calabi-Yau manifold

$$\omega = \int \Omega = \int \frac{dz}{z} \wedge \frac{ds \wedge dx}{\frac{\partial W}{\partial w}\Big|_{W=0}} \tag{44}$$

In the case of the A_2 singularity

$$\left.\frac{\partial W}{\partial w}\right|_{w=0} = 2\sqrt{z+\Lambda^6/z+2C_{A_2}(x,u,v)+s^2}. \tag{45}$$

The integral over s is trivial and equals 2π. The boundary of the s-integration is given by

$$z+\frac{\Lambda^6}{z}+2C_{A_2}(x,u,v)=0 \tag{46}$$

which describes the spectral curve[24,19]. If one shifts $z \to y - C_{A_2}$, one recovers the hyperelliptic curve

$$y^2 = C_{A_2}(x,u,v)^2 - \Lambda^6. \tag{47}$$

The period is rewritten as

$$\begin{aligned}\omega &= \pi \int \frac{dz}{z}\wedge dx = \pi \oint x\frac{d(y-C_{A_2})}{(y-C_{A_2})}, \qquad y^2 = C_{A_2}-\Lambda^6 \\ &= \pi \oint x d\log(y-C_{A_2}) = \frac{\pi}{2}\oint x d\log\frac{(y-C_{A_2})}{(y+C_{A_2})}\end{aligned} \tag{48}$$

which reproduces (9).

In the above we have considered the case when K_3 surface (ALE space) degenerates to have an A_2-type singularity. We may generalize this construction and consider the case when K_3 surface degenerates into more general types of ADE singularities. In the case of E_n singularities one can no longer perform the s-integral and reduce ω to an integral over a Riemann surface. It turns out, however, we can still analyze the critical behavior of ω and determine the exponents of the E_n gauge theories.

Let us first discuss the E_6 theory,

$$\begin{aligned}W &= z+\frac{\Lambda^6}{z}+2C_{E_6}(x,s)+w^2, \\ C_{E_6}(x,s) &= x^4+s^3+u_1x^2s+u_4xs+u_5x^2+u_7s+u_8x+u_{11}\end{aligned} \tag{49}$$

where u_i are the perturbations of the E_6 singularity. The period is given by

$$\omega = \int \frac{dz}{z}\wedge\frac{ds\wedge dx}{\sqrt{z+\frac{\Lambda^6}{z}+2C_{E_6}(x,s)}} \tag{50}$$

The critical point is located at

$$x^*=0,\quad s^*=0,\quad z^*=\Lambda^3,\quad u_{11}^*=-2\Lambda^3. \tag{51}$$

Let us first determine the exponent of the parameter u_1. By perturbing away from the critical point

$$z = \Lambda^3 + t^{1/2}\Lambda^{3/2}\tilde{z}, \quad x = t^{1/4}\tilde{x}, \quad s = t^{1/3}\tilde{s}, \quad u_1 = t^{1/6}\tilde{u_1} \tag{52}$$

we have

$$\omega \approx t^{1/2-1/2+1/4+1/3} \int d\tilde{z} \wedge \frac{d\tilde{s} \wedge d\tilde{x}}{\sqrt{\tilde{z}^2 + \tilde{x}^4 + \tilde{s}^3 + \tilde{u_1}\tilde{x}^2\tilde{s}}} \approx t^{7/12} \tag{53}$$

By introducing the unit of mass μ we find

$$t \approx \mu^{12/7}, \qquad u_1 \approx \mu^{2/7}. \tag{54}$$

Thus the perturbation u_1 has the exponent 2/7. We can similarly determine the exponents of the parameters u_i ($i = 4, 5, 7, 8, 11$). They read as

$$\frac{2(e_i + 1)}{14}, \quad i = 1, 4, 5, 7, 8, 11 \tag{55}$$

and hence again fit to the formula $2(e_i + 1)/(h + 2)$ where the dual-Coxeter number h of E_6 is 12.

We may also compute exponents for the E_7 and E_8 gauge theories. Singularities are described by the polynomials

$$\begin{aligned} C_{E_7}(x, s) &= x^3 + xs^3 + u_1 s^4 + u_5 s^3 + u_7 xs + u_9 s^2 + u_{11} x \\ &\quad + u_{13} s + u_{17}, \end{aligned} \tag{56}$$

$$\begin{aligned} C_{E_8}(x, s) &= x^5 + s^3 + u_1 x^3 s + u_7 x^2 s + u_{11} x^3 + u_{13} xs + u_{17} x^2 \\ &\quad + u_{19} s + u_{23} x + u_{29}. \end{aligned} \tag{57}$$

We find that their critical exponents are given by

$$\frac{2(e_i + 1)}{20}, \quad i = 1, 5, 7, 9, 11, 13, 17 \qquad \text{for} E_7 \tag{58}$$

$$\frac{2(e_i + 1)}{32}, \quad i = 1, 7, 11, 13, 17, 19, 23, 29 \qquad \text{for} E_8 \tag{59}$$

(58) (59) again fit to the formula (38).

It is easy to check that the above construction reproduces the exponents of Table 3 in the case of A_n and D_n singularities. Thus we have some considerable evidence for the A-E-D classification of SCFTs originating from pure $N = 2$ gauge theories. The pattern of the classification follows from that of the

degeneraion of K_3 surfaces which appear in the heterotic-type II duality based on K_3-fibered Calabi-Yau manifolds. More details will be discussed elsewhere.

We would like to thank K.Ito and S.K.Yang for discussions.

1. N. Seiberg and E. Witten, Nucl. Phys. **B426** (1994) 19.
2. N. Seiberg and E. Witten, Nucl. Phys. **B431** (1994) 484.
3. N. Seiberg, Nucl. Phys. **B435** (1994) 129.
4. K. Intriligator and N. Seiberg, *Lectures on Supersymmetric Gauge Theories and Electric-Magnetic Duality*, hep-th/9509066.
5. P.C. Argyres and A.E. Faraggi, Phys. Rev. Lett. **74** (1995) 3931.
6. A. Klemm, W. Lerche, S. Theisen and S. Yankielowicz, Phys. Lett. **B344** (1995) 169.
7. U.H. Danielsson and B. Sundborg, Phys. Lett. **B358** (1995) 273.
8. A. Brandhuber and K. Landsteiner, Phys. Lett. **B358** (1995) 73.
9. A. Hanany and Y. Oz, Nucl. Phys. **B452** (1995) 283.
10. P.C. Argyres, R.N. Plesser and A.D. Shapere, Phys. Rev. Lett. **75** (1995) 1699.
11. J.A. Minahan and D. Nemeschansky, *Hyperelliptic Curves for Supersymmetric Yang-Mills*, hep-th/9507032.
12. P.C. Argyres and A.D. Shapere, *The Vacuum Structure of $N = 2$ Super-QCD with Classical Gauge Groups*, hep-th/9509175.
13. A. Hanany, *On the Quantum Moduli Space of Vacua of $N = 2$ Supersymmetric Gauge Theories*, hep-th/9509176.
14. U.H. Danielsson and B. Sundborg, *Exceptional Equivalences in $N = 2$ Supersymmetric Yang-Mills Theory*, hep-th/9511180.
15. M. Alishahiha, F. Ardalan and F. Mansouri, *The Moduli Space of the $N = 2$ Supersymmetric G_2 Yang-Mills Theory*, hep-th/9512005.
16. P.C. Argyres and M. Douglas, Nucl. Phys. **B448** (1995) 93.
17. P.C. Argyres, R.N. Plesser, N. Seiberg and E. Witten, *New $N = 2$ Superconformal Field Theories in Four Dimensions*, hep-th/9511154.
18. T. Eguchi, K. Hori, K. Ito and S.K. Yang, Nucl. Phys. **B471** (1996) 430.
19. E. Martinec and N. Warner, Nucl. Phys. **B459** (1996) 97.
20. S. Kachcru and C. Vafa, Nucl. Phys. **B450** (1995) 69.
21. A. Klemm, W. Lerche, P. Mayr, C. Vafa and N. Warner, *Self-Dual Strings and $N = 2$ Supersymmetric Field Theory*, hep-th/9604034
22. V. Kaplunovsky, J. Louis and S.J. Theisen, Phy. Lett. **B357** (1995) 71.
23. A. Klemm, W. Lerche and P. Mayr, Phys. Lett. **357** (1995) 313.
24. A. Gorsky, I. Krichever, A. Marshakov, A. Mironov and A. Morozof, Phys. Lett. **B355** (1995) 466.

PERIOD FUNCTIONS AND THE SELBERG ZETA FUNCTION FOR THE MODULAR GROUP

J. LEWIS AND D. ZAGIER
Max Planck Institut für Mathematik,
Gottfried-Claren Str. 26, D-5300 Bonn 3

The Selberg trace formula on a Riemann surface X connects the discrete spectrum of the Laplacian with the length spectrum of the surface, that is, the set of lengths of the closed geodesics of on X. The connection is most strikingly expressed in terms of the Selberg zeta function, which is a meromorphic function of a complex variable s that is defined for $\Re(s) > 1$ in terms of the length spectrum and that has zeros at all $s \in \mathbb{C}$ for which $s(1-s)$ is an eigenvalue of the Laplacian in $L^2(X)$. We will be interested in the case when X is the quotient of the upper half-plane $\mathcal{H}$ by either the modular group $\Gamma_1 = \mathrm{SL}(2,\mathbb{Z})$ or the extended modular group $\Gamma = \mathrm{GL}(2,\mathbb{Z})$, where $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$ acts on $\mathcal{H}$ by $z \mapsto (az+b)/(cz+d)$ if $\det(\gamma) = +1$ and $z \mapsto (a\bar{z}+b)/(c\bar{z}+d)$ if $\det(\gamma) = -1$. In this case the length spectrum of X is given in terms of class numbers and units of orders in real quadratic fields, while the eigenfunctions of the Laplace operator are the non-holomorphic modular functions usually called Maass wave forms. (Good expositions of this subject can be found in [6] and [7]).

A striking fact, discovered by D. Mayer [4, 5] and for which a simplified proof will be given in the first part of this paper, is that the Selberg zeta function $Z_\Gamma(s)$ of $\mathcal{H}/\Gamma$ can be represented as the (Fredholm) determinant of the action of a certain element of the quotient field of the group ring $\mathbb{Z}[\Gamma]$ on an appropriate Banach space. Specifically, let $\mathbf{V}$ be the space of functions holomorphic in $\mathbb{D} = \{z \in \mathbb{C} \mid |z-1| < \frac{3}{2}\}$ and continuous in $\overline{\mathbb{D}}$. The semigroup $\{\gamma \in \Gamma \mid \gamma(\mathbb{D}) \subseteq \mathbb{D}\}$ acts on the right by $\pi_s\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) f(z) = (cz+d)^{-2s} f\left(\frac{az+b}{cz+d}\right)$. In particular, for all $n \geq 1$ the element $\left(\begin{smallmatrix} 0 & 1 \\ 1 & n \end{smallmatrix}\right)$, which can be written in terms of the generators $\sigma = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ and $\rho = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ of Γ as $\sigma^{n-1}\rho$, acts on $\mathbf{V}$. It turns out (cf. §2) that the formal expression

$$\mathcal{L} = (1-\sigma)^{-1}\rho = \sum_{n=1}^{\infty} \left[\begin{smallmatrix} 0 & 1 \\ 1 & n \end{smallmatrix}\right] \tag{1}$$

defines an operator $L_s = \pi_s(\mathcal{L})$ of trace class on $\mathbf{V}$ (first for $\Re(s) > \frac{1}{2}$, and then by analytic continuation to all s). This implies that the operator $1 - L_s$ has a determinant in the Fredholm sense; and the result then is:

Theorem 1. *The Selberg zeta function of $\mathcal{H}/\Gamma$ is given by*

$$Z_\Gamma(s) = \det(1 - L_s)\,. \tag{2}$$

degeneraion of K_3 surfaces which appear in the heterotic-type II duality based on K_3-fibered Calabi-Yau manifolds. More details will be discussed elsewhere.

We would like to thank K.Ito and S.K.Yang for discussions.

1. N. Seiberg and E. Witten, Nucl. Phys. **B426** (1994) 19.
2. N. Seiberg and E. Witten, Nucl. Phys. **B431** (1994) 484.
3. N. Seiberg, Nucl. Phys. **B435** (1994) 129.
4. K. Intriligator and N. Seiberg, *Lectures on Supersymmetric Gauge Theories and Electric-Magnetic Duality*, hep-th/9509066.
5. P.C. Argyres and A.E. Faraggi, Phys. Rev. Lett. **74** (1995) 3931.
6. A. Klemm, W. Lerche, S. Theisen and S. Yankielowicz, Phys. Lett. **B344** (1995) 169.
7. U.H. Danielsson and B. Sundborg, Phys. Lett. **B358** (1995) 273.
8. A. Brandhuber and K. Landsteiner, Phys. Lett. **B358** (1995) 73.
9. A. Hanany and Y. Oz, Nucl. Phys. **B452** (1995) 283.
10. P.C. Argyres, R.N. Plesser and A.D. Shapere, Phys. Rev. Lett. **75** (1995) 1699.
11. J.A. Minahan and D. Nemeschansky, *Hyperelliptic Curves for Supersymmetric Yang-Mills*, hep-th/9507032.
12. P.C. Argyres and A.D. Shapere, *The Vacuum Structure of $N = 2$ Super-QCD with Classical Gauge Groups*, hep-th/9509175.
13. A. Hanany, *On the Quantum Moduli Space of Vacua of $N = 2$ Supersymmetric Gauge Theories*, hep-th/9509176.
14. U.H. Danielsson and B. Sundborg, *Exceptional Equivalences in $N = 2$ Supersymmetric Yang-Mills Theory*, hep-th/9511180.
15. M. Alishahiha, F. Ardalan and F. Mansouri, *The Moduli Space of the $N = 2$ Supersymmetric G_2 Yang-Mills Theory*, hep-th/9512005.
16. P.C. Argyres and M. Douglas, Nucl. Phys. **B448** (1995) 93.
17. P.C. Argyres, R.N. Plesser, N. Seiberg and E. Witten, *New $N = 2$ Superconformal Field Theories in Four Dimensions*, hep-th/9511154.
18. T. Eguchi, K. Hori, K. Ito and S.K. Yang, Nucl. Phys. **B471** (1996) 430.
19. E. Martinec and N. Warner, Nucl. Phys. **B459** (1996) 97.
20. S. Kachcru and C. Vafa, Nucl. Phys. **B450** (1995) 69.
21. A. Klemm, W. Lerche, P. Mayr, C. Vafa and N. Warner, *Self-Dual Strings and $N = 2$ Supersymmetric Field Theory*, hep-th/9604034
22. V. Kaplunovsky, J. Louis and S.J. Theisen, Phy. Lett. **B357** (1995) 71.
23. A. Klemm, W. Lerche and P. Mayr, Phys. Lett. **357** (1995) 313.
24. A. Gorsky, I. Krichever, A. Marshakov, A. Mironov and A. Morozof, Phys. Lett. **B355** (1995) 466.

PERIOD FUNCTIONS AND THE SELBERG ZETA FUNCTION FOR THE MODULAR GROUP

J. LEWIS AND D. ZAGIER
Max Planck Institut für Mathematik,
Gottfried-Claren Str. 26, D-5300 Bonn 3

The Selberg trace formula on a Riemann surface X connects the discrete spectrum of the Laplacian with the length spectrum of the surface, that is, the set of lengths of the closed geodesics of on X. The connection is most strikingly expressed in terms of the Selberg zeta function, which is a meromorphic function of a complex variable s that is defined for $\Re(s) > 1$ in terms of the length spectrum and that has zeros at all $s \in \mathbb{C}$ for which $s(1-s)$ is an eigenvalue of the Laplacian in $L^2(X)$. We will be interested in the case when X is the quotient of the upper half-plane $\mathcal{H}$ by either the modular group $\Gamma_1 = \mathrm{SL}(2,\mathbb{Z})$ or the extended modular group $\Gamma = \mathrm{GL}(2,\mathbb{Z})$, where $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$ acts on $\mathcal{H}$ by $z \mapsto (az+b)/(cz+d)$ if $\det(\gamma) = +1$ and $z \mapsto (a\bar{z}+b)/(c\bar{z}+d)$ if $\det(\gamma) = -1$. In this case the length spectrum of X is given in terms of class numbers and units of orders in real quadratic fields, while the eigenfunctions of the Laplace operator are the non-holomorphic modular functions usually called Maass wave forms. (Good expositions of this subject can be found in [6] and [7]).

A striking fact, discovered by D. Mayer [4, 5] and for which a simplified proof will be given in the first part of this paper, is that the Selberg zeta function $Z_\Gamma(s)$ of $\mathcal{H}/\Gamma$ can be represented as the (Fredholm) determinant of the action of a certain element of the quotient field of the group ring $\mathbb{Z}[\Gamma]$ on an appropriate Banach space. Specifically, let $\mathbf{V}$ be the space of functions holomorphic in $\mathbb{D} = \{z \in \mathbb{C} \mid |z-1| < \frac{3}{2}\}$ and continuous in $\overline{\mathbb{D}}$. The semigroup $\{\gamma \in \Gamma \mid \gamma(\mathbb{D}) \subseteq \mathbb{D}\}$ acts on the right by $\pi_s\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) f(z) = (cz+d)^{-2s} f\left(\frac{az+b}{cz+d}\right)$. In particular, for all $n \geq 1$ the element $\left(\begin{smallmatrix} 0 & 1 \\ 1 & n \end{smallmatrix}\right)$, which can be written in terms of the generators $\sigma = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ and $\rho = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ of Γ as $\sigma^{n-1}\rho$, acts on $\mathbf{V}$. It turns out (cf. §2) that the formal expression

$$\mathcal{L} = (1-\sigma)^{-1}\rho = \sum_{n=1}^{\infty} \left[\begin{smallmatrix} 0 & 1 \\ 1 & n \end{smallmatrix}\right] \tag{1}$$

defines an operator $L_s = \pi_s(\mathcal{L})$ of trace class on $\mathbf{V}$ (first for $\Re(s) > \frac{1}{2}$, and then by analytic continuation to all s). This implies that the operator $1 - L_s$ has a determinant in the Fredholm sense; and the result then is:

Theorem 1. *The Selberg zeta function of $\mathcal{H}/\Gamma$ is given by*

$$Z_\Gamma(s) = \det(1 - L_s)\,. \tag{2}$$

(Actually, Mayer's result is that the Selberg zeta function of $\mathcal{H}/\Gamma_1$ equals $\det(1 - L_s^2)$, but everything works in much the same way for the two groups Γ and Γ_1. We will discuss both cases, but in our exposition have given precedence to the larger group Γ.)

On the other hand, as we already mentioned, the function $Z_\Gamma(s)$ has a meromorphic continuation with zeros corresponding to the eigenvalues of even Maass wave forms on $SL(2,\mathbb{Z})$. Formally, equation (2) says that these zeros correspond to the fixed points of L_s, i.e., to the functions $h \in \mathbf{V}$ such that $h(z) = \sum_{n=1}^{\infty}(z+n)^{-2s}h(1/(z+n))$. Adding $z^{-2s}h(1/z)$ to both sides we find that $h(z) + z^{-2s}h(1/z) = h(z-1)$, or equivalently, that the shifted function $\psi(z) = h(z-1)$ satisfies the *three-term functional equation*

$$\psi(z) = \psi(z+1) + z^{-2s}\,\psi(1+1/z)\,. \tag{3}$$

It is therefore natural to ask whether there is a direct connection between the spectrum of the Laplace operator Δ on $\mathcal{H}/\Gamma$ and the solutions of the three-term functional equation. Such a connection was discovered (independently of Mayer's work) in [2], whose main result, in a slightly strengthened form, can be stated as follows:

Theorem 2. *Let s be a complex number with $0 < \Re(s) < 1$. Then there is a canonical bijection between square integrable solutions of $\Delta u = s(1-s)\,u$ in $\mathcal{H}/\Gamma$ and holomorphic solutions of* (3) *in the cut plane $\mathbb{C}' = \mathbb{C} \setminus (-\infty, 0]$ satisfying the growth condition $\psi(x) = O(1/x)$ as $x \to \infty$.*

The formula for the correspondence $u \mapsto \psi$ in [2] was completely explicit (eq. (12) below), but its proof was indirect and did not make the reasons for its properties at all transparent. Other proofs and several other formulas for ψ in terms of u were found in [3], where it was also observed that this correspondence is exactly analogous to the relationship between a holomorphic modular form and its period polynomial in the sense of Eichler, Shimura, and Manin. We will call the function $\psi(z)$ the *period function* of the wave form u.

Taken together, these two theorems give another point of view on the Selberg trace formula: Theorem 1 relates the "length spectrum" definition of the Selberg zeta function to the fixed points of the operator L_s and hence, by implication, to the solutions of the functional equation (3), and Theorem 2 relates the solutions of (3) to the "discrete spectrum of the Laplacian" definition of Z_Γ. In this paper (which, except for the simplifications in the proof of Theorem 1, is mostly expository) we will discuss both aspects. Part I uses reduction theory to establish the connection between the Selberg zeta function and the operator L_s. In §1 we outline a proof of Theorem 1. The details (e.g. the proofs of various assertions needed from reduction theory, verification of convergence, etc.) are filled in in §2, while the following section gives various complements: the modifications when Γ is replaced by Γ_1, a reformulation of some of the ideas of the proof in terms of group algebras, and a brief description of Mayer's original approach via the symbolic dynamics of the continued fraction map. Part II describes the connection between the solutions of the functional equation (3) and the eigenfunctions of the Laplacian in $\mathcal{H}/\Gamma$. We will give several formulas for the $u \leftrightarrow \psi$ correspondence, sketch some the ideas involved in the proof, describe the analogy with the theory of periods of modular forms, and discuss some other properties of solutions of (3) on $\mathbb{C}'$ or on $\mathbb{R}^+$. Here we will give fewer details than in Part I and omit all proofs, referring the reader to [2] and [3] for more information.

§1. The formal calculation. The basic conjugacy invariants of an element $\gamma \in \Gamma$ are the numbers $\mathrm{Tr}(\gamma)$, $\det(\gamma)$, and $\Delta(\gamma) = \mathrm{Tr}(\gamma)^2 - 4\det(\gamma)$ (trace, determinant, discriminant). We will call γ *hyperbolic* if $\Delta(\gamma)$is positive and (to distinguish between γ and $-\gamma$, which act in the same way on $\mathcal{H}$) also $\mathrm{Tr}(\gamma) > 0$. If γ is hyperbolic, we set

$$\mathcal{N}(\gamma) = \left(\frac{\mathrm{Tr}(\gamma) + \Delta(\gamma)^{\frac{1}{2}}}{2}\right)^2, \qquad \chi_s(\gamma) = \frac{\mathcal{N}(\gamma)^{\frac{1}{2}-s}}{\Delta(\gamma)^{\frac{1}{2}}} = \frac{\mathcal{N}(\gamma)^{-s}}{1 - \det(\gamma)\,\mathcal{N}(\gamma)^{-1}} \qquad (s \in \mathbb{C})$$

and define $k(\gamma)$ as the largest integer k such that $\gamma = \gamma_0^k$ for some $\gamma_0 \in \Gamma$ (which is then hyperbolic and *primitive*, i.e. $k(\gamma_0) = 1$). The Selberg zeta function $Z_\Gamma(s)$ for Γ is defined by

$$Z_\Gamma(s) = \prod_{\substack{\{\gamma\} \text{ in } \Gamma \\ \gamma \text{ primitive}}} \prod_{m=0}^{\infty} \left(1 - \det(\gamma)^m\, \mathcal{N}(\gamma)^{-s-m}\right) \qquad (\Re(s) > 1),$$

where the notation "$\{\gamma\}$ in Γ" means that the product is taken over all (primitive hyperbolic) elements of Γ up to Γ-conjugacy. That the function $Z_\Gamma(s)$ extends meromorphically to all complex values of s is one of the standard consequences of the Selberg trace formula, which expresses its logarithmic derivative as a sum over the eigenvalues of the (hyperbolic) Laplacian in $\mathcal{H}/\Gamma$.

For $\Re(s) > 1$ we have the simple computation

$$\begin{aligned} -\log Z_\Gamma(s) &= \sum_{\substack{\{\gamma\} \text{ in } \Gamma \\ \gamma \text{ primitive}}} \sum_{m=0}^{\infty} \sum_{k=1}^{\infty} \frac{1}{k} \det(\gamma)^{km}\, \mathcal{N}(\gamma)^{-k(s+m)} \\ &= \sum_{\substack{\{\gamma\} \text{ in } \Gamma \\ \gamma \text{ primitive}}} \sum_{k=1}^{\infty} \frac{1}{k} \frac{\mathcal{N}(\gamma)^{-ks}}{1 - \det(\gamma)^k \mathcal{N}(\gamma)^{-k}} \\ &= \sum_{\substack{\{\gamma\} \text{ in } \Gamma \\ \gamma \text{ hyperbolic}}} \frac{1}{k(\gamma)}\, \chi_s(\gamma), \end{aligned} \tag{4}$$

where the last step just expresses the fact that every hyperbolic element of Γ can be written uniquely as γ^k with γ primitive and $k \geq 1$.

To get further we use a version of reduction theory. This theory is usually presented for quadratic forms, but is translatable into the language of matrices by the standard observation that there is a 1:1 correspondence between conjugacy classes of matrices of trace t and determinant n and equivalence classes of quadratic forms of discriminant $t^2 - 4n$. We define the set of *reduced* elements of Γ by

$$\mathrm{Red} = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma \;\middle|\; 0 \leq a \leq b, c \leq d \right\},$$

i.e. matrices with non-negative entries which are non-decreasing downwards and to the right. (We will explain in §3**B** where this definition comes from.) Then we have the following facts, whose proofs will be indicated in §2:

(**I**) Every reduced matrix can be written uniquely as a product $\left(\begin{smallmatrix}0&1\\1&n_1\end{smallmatrix}\right)\cdots\left(\begin{smallmatrix}0&1\\1&n_\ell\end{smallmatrix}\right)$ with $n_1,\dots,n_\ell \geq 1$ for a unique positive integer $\ell = \ell(\gamma)$.

(**II**) Every conjugacy classes of hyperbolic matrices in Γ contains reduced representatives γ; they all have the same value of $\ell(\gamma)$ and there are $\ell(\gamma)/k(\gamma)$ of them.

(**III**) If γ is reduced, then the operator $\pi_s(\gamma)$ is of trace class and $\operatorname{Tr}\pi_s(\gamma) = \chi_s(\gamma)$.

Combining these assertions with (4), we find

$$\begin{aligned} -\log Z_\Gamma(s) &\underset{(II)}{=} \sum_{\gamma\in\mathsf{Red}} \frac{1}{\ell(\gamma)}\chi_s(\gamma) \\ &\underset{(III)}{=} \operatorname{Tr}\Big(\sum_{\gamma\in\mathsf{Red}} \frac{1}{\ell(\gamma)}\pi_s(\gamma)\Big) \\ &\underset{(I)}{=} \operatorname{Tr}\Big(\sum_{\ell=1}^{\infty}\frac{1}{\ell}\Big(\sum_{n=1}^{\infty}\pi_s\left(\begin{smallmatrix}0&1\\1&n\end{smallmatrix}\right)\Big)^{\ell}\Big), \end{aligned} \tag{5}$$

and this is equivalent to (2) by the definition of $\mathcal{L}$ and the Fredholm determinant formula $\log\det(1-L_s) = -\sum_{\ell=1}^{\infty}\operatorname{Tr}(L_s^\ell)/\ell$.

§2. **Details.** In this section we verify the assertions (I)–(III) and establish the validity of the formal calculations of §1 for $\Re(s) > 1$; (2) then holds for all s by analytic continuation.

A. Proof of (I). Suppose that $\gamma = \left(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\right) \in \mathsf{Red}$. If $a = 0$ then $\gamma = \left(\begin{smallmatrix}0&1\\1&d\end{smallmatrix}\right)$ with $d \geq 1$ and we are already finished with $\ell(\gamma) = 1$. If $a > 0$, we set $n = \lceil d/b\rceil - 1$ (i.e. n is the unique integer $n < d/b \leq n+1$). One easily checks that this is the only $n \in \mathbb{Z}$ for which $\gamma = \left(\begin{smallmatrix}0&1\\1&n\end{smallmatrix}\right)\gamma^*$ with $\gamma^* \in \mathsf{Red}$. Moreover, the sum of the entries of γ^* is smaller than that of γ, so we can assume by induction that γ^* has the form claimed, and then so does γ with $\ell(\gamma) = \ell(\gamma^*) + 1$.

B. Proof of (II). This is essentially equivalent to the theory of periodic continued fractions: to each hyperbolic matrix $\gamma = \left(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\right)$ we associate the roots $\frac{a-d\pm\sqrt{\Delta(\gamma)}}{2c}$ of $\gamma x = x$, which are quadratic irrationalities; two γ's are conjugate if and only if the corresponding roots are Γ-equivalent; each quadratic irrationality has a continued fraction expansion $1/(m_1 + 1/(m_2 + 1/\cdots))$ which is eventually periodic; and if the fixed point of γ has a continued fraction expansion with period $(n_1,\dots,n_\ell)$ then γ is conjugate to the reduced matrix $\left(\begin{smallmatrix}0&1\\1&n_1\end{smallmatrix}\right)\cdots\left(\begin{smallmatrix}0&1\\1&n_\ell\end{smallmatrix}\right)$ with $n_1,\dots,n_\ell \geq 1$. However, one can also do the reduction procedure directly on the matrix level. We define a conjugacy class preserving map F from the set of hyperbolic matrices to itself by $\gamma = \left(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\right) \mapsto F(\gamma) = \left(\begin{smallmatrix}0&1\\1&n\end{smallmatrix}\right)^{-1}\gamma\left(\begin{smallmatrix}0&1\\1&n\end{smallmatrix}\right)$ where n is the unique integer for which the interval $[n, n+1]$ contains both d/b and c/a. (This definition must be modified slightly if $a = 0$.) Notice that if γ is reduced then this is the same n as was used in the proof of (I) and $F(\gamma)$ is simply $\gamma^*\left(\begin{smallmatrix}0&1\\1&n\end{smallmatrix}\right)$ in the notation

above. The effect of F on a reduced matrix $\gamma = \left(\begin{smallmatrix} 0 & 1 \\ 1 & n_1 \end{smallmatrix}\right) \cdots \left(\begin{smallmatrix} 0 & 1 \\ 1 & n_\ell \end{smallmatrix}\right)$ is thus simply to replace it by the cyclically permuted product $F(\gamma) = \left(\begin{smallmatrix} 0 & 1 \\ 1 & n_2 \end{smallmatrix}\right) \cdots \left(\begin{smallmatrix} 0 & 1 \\ 1 & n_\ell \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & 1 \\ 1 & n_1 \end{smallmatrix}\right)$. It is clear that under this "internal conjugation" the exact period of γ will be the number $\ell(\gamma)/k(\gamma)$. (*Proof.* If γ is the kth power of another matrix γ' with $k \geq 1$, then γ' is also reduced and hence also a product of matrices $\left(\begin{smallmatrix} 0 & 1 \\ 1 & n \end{smallmatrix}\right)$, and the cycle $(n_1, \ldots, n_\ell)$ for γ is just the k-fold concatenation of the cycle for γ'; and conversely if the cycle of γ is a k-fold concatenation then γ is a kth power. Hence $\ell(\gamma)/k(\gamma)$ is the exact period of the sequence of integers $\{n_i\}$.) The assertion of (II) is thus proved if we show that (i) iterating F often enough eventually sends an arbitrary hyperbolic element of Γ to an element of Red, and (ii) two elements of Red are Γ-conjugate only if they are already "internally" conjugate, i.e., if and only if one is mapped to the other by a power of F. Both steps are proved by a series of elementary inequalities which show that each application of F "improves things" (i.e. either makes a non-reduced matrix more nearly reduced in the sense that some positive integer measuring the failure of the inequalities defining Red gets smaller, or else reduces the size of the entries of the matrix conjugating one reduced γ into another). We omit the details, which are exactly parallel to the proofs of the corresponding assertions in the usual reduction theory of quadratic forms as carried out in standard books, e.g. in §13 of [8].

C. Proof of (III). If γ is reduced, then γ maps the closed interval $[-\frac{1}{2}, \frac{5}{2}]$ into the half-open interval (0,2] and hence maps the closed disk $\overline{\mathbb{D}}$ into the open disk $\mathbb{D}$. Standard results from the theory of composition operators on spaces of holomorphic functions (cf. [5], Thm. 7.9 and Lemma 7.10 and the papers cited there) then imply that the operator $\pi_s(g)$ is of trace class and that its trace equals $\chi_s(g)$.

D. Verification of convergence. The operator $L_s = \pi_s(\mathcal{L})$ sends $h \in \mathbf{V}$ to

$$(L_s h)(z) = \sum_{n=1}^{\infty} \frac{1}{(z+n)^{2s}} h\left(\frac{1}{z+n}\right).$$

Since h is holomorphic at 0, the sum converges absolutely for s in the half-plane $\Re(s) > \frac{1}{2}$ to a function which again belongs to $\mathbf{V}$, and the absolute convergence also implies that this operator is of trace class. We have to show that in the smaller half-plane $\Re(s) > 1$ all of the steps of the calculations in (4) and (5) are justified. But this follows from the calculations themselves: The absolute convergence of the product defining $Z_\Gamma(s)$ (and hence of the sum defining its logarithm) for $\Re(s) > 1$ is well-known, and since in (4) and (5) all terms are replaced by their absolute value when s is replaced by its real part, the various interchanges in the order of summation are automatically justified. The validity of the last line of the proof also follows, since the formula $\sum \mathrm{Tr}(A^\ell)/\ell = -\log\det(1-A)$ is true for any trace class operator A for which $\sum |\mathrm{Tr}(A^\ell)/\ell|$ converges. We can also run the calculation backwards (and hence verify the convergence of the infinite product for $Z_\Gamma(s)$ in the half-plane $\Re(s) > 1$) by showing directly that the sum $M_s := \sum_{\gamma \in \mathrm{Red}} |\chi_s(\gamma)|$ is convergent for $\Re(s) > 1$. Indeed, we have

$$M_s = \sum_{k \geq 3} \frac{c_k^+}{\sqrt{k^2-4}} \left(\frac{k+\sqrt{k^2-4}}{2}\right)^{1-2\Re(s)} + \sum_{k \geq 1} \frac{c_k^-}{\sqrt{k^2+4}} \left(\frac{k+\sqrt{k^2+4}}{2}\right)^{1-2\Re(s)},$$

with $c_k^{\pm} = \#\{\gamma \in \mathsf{Red} \mid \mathrm{Tr}(\gamma) = k, \det(\gamma) = \pm 1\}$, and the required convergence follows from the estimate $c_k^{\pm} \ll k^{1+\varepsilon}$ $(\forall \varepsilon > 0)$, which is obtained by straightforward estimates using the divisor function.

§3. Complements. In this section we discuss some further aspects of the proof given in the last two sections.

A. The Selberg zeta function for $SL(2,\mathbb{Z})$. In this subsection we treat the case when the group $\Gamma = GL(2,\mathbb{Z})$ is replaced by its subgroup $\Gamma_1 = SL(2,\mathbb{Z})$, the usual modular group. We denote by $Z(s)$ the Selberg zeta function for Γ_1, which is defined for $\Re(s) > 1$ by the same product expansion as before but with the product running over Γ_1-conjugacy classes of primitive hyperbolic elements of Γ_1. As mentioned in the introduction, the statement of Theorem 1 for Γ_1 is the identity

$$Z(s) = \det(1 - L_s^2)\,. \tag{6}$$

We indicate the changes that have to be made in the proof of §§1–2 in order to prove this.

The calculation (4) is unchanged except that now the summation is over Γ_1-conjugacy classes and the number $k(\gamma)$ must be replaced by $k_1(\gamma)$, the largest integer n such that γ is the nth power of an element in Γ_1. For the first line of (5) we needed that

$$\gamma \in \Gamma \quad \Rightarrow \quad \#\{\gamma' \in \mathsf{Red} \mid \gamma' \underset{\Gamma}{\sim} \gamma\} \;=\; \frac{\ell(\gamma)}{k(\gamma)}\,, \tag{7}$$

which followed from Statement (II). This must now be replaced by

$$\gamma \in \Gamma_1 \quad \Rightarrow \quad \#\{\gamma' \in \mathsf{Red} \mid \gamma' \underset{\Gamma_1}{\sim} \gamma\} \;=\; \frac{\ell(\gamma)}{2k_1(\gamma)}\,, \tag{8}$$

which we will prove in a moment. The first line in (5) then becomes

$$-\log Z(s) = \sum_{\gamma \in \Gamma_1 \cap \mathsf{Red}} \frac{2}{\ell(\gamma)}\, \chi_s(\gamma)\,, \tag{9}$$

and the restriction $\gamma \in \Gamma_1$ implies that in the last line of (5) we sum only over even ℓ.

It remains to prove (8). Write $\gamma = \gamma_0^k$ where $k > 0$ and γ_0 is primitive in Γ, and set $\ell_0 = \ell(\gamma_0)$. Then $k(\gamma) = k$, $\ell(\gamma) = k\ell_0$, and (7) expresses the fact that the Γ-conjugates to γ in Red correspond to the ℓ_0 possible "internal conjugates" of a reduced representative of this conjugacy class. We now distinguish two cases. If $\det(\gamma_0) = +1$, then $k_1(\gamma) = k$ (because $\gamma_0 \in \Gamma_1$ and is clearly primitive there), but ℓ_0 is even and the number of $\gamma' \in \mathsf{Red}$ which are Γ_1-conjugate to γ is $\ell_0/2$, because half of the ℓ_0 "internal" conjugations in our cycle are conjugations by elements of determinant -1 and hence are no longer counted. If on the other hand $\det(\gamma_0) = -1$, then k is even and $k_1(\gamma) = k/2$, because the element γ_0^2 is now primitive in Γ_1, but to make up for it the number of $\gamma' \in \mathsf{Red}$ which are Γ_1-conjugate to γ is now the full number ℓ_0, because there is no longer any distinction between internal conjugacies by elements of determinant $+1$ or -1. (Conjugating by the product of the first r matrices $\left(\begin{smallmatrix} 0 & 1 \\ 1 & n \end{smallmatrix}\right)$ of the cycle of γ_0 is the same as conjugating by the product of the last $\ell_0 - r$ of them, and r and $\ell_0 - r$ have opposite parities.) This establishes (8) in both cases.

B. Identities in the group ring. In this subsection we redo part of the calculation in §1 in a slightly different way in which the elements of Red are built up out of powers of a finite rather than an infinite sum; this also helps us to understand the structure of Red and permits us to make sense of the formal expressions in (1).

Recall that $\sigma = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ and $\rho = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right)$. These elements generate Γ, but of course not at all freely: e.g. one has $(\sigma^{-1}\rho)^2 = (\sigma^{-2}\rho^2)^6 = 1$. On the other hand, the subsemigroup Q of Γ generated by ρ and σ is the *free* semigroup generated by these two elements, i.e. its elements are all words in σ and ρ and all such words distinct. In fact, the set Q is contained in the larger sub-semigroup P of Γ consisting of all matrices with non-negative entries, which is easily seen to be the semigroup generated by the two elements $\kappa = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and σ, subject to the unique relation $\kappa^2 = 1$. Since $\rho = \sigma\kappa$, every element of P is either a word in ρ and σ or else κ times such a word, so $\mathsf{P} = \mathsf{Q} \cup \kappa\mathsf{Q}$ (disjoint union). This says that $\mathsf{Q} \setminus \{1\} = \sigma\mathsf{P}$, the subset of P consisting of words in κ and σ which begin with a σ, or equivalently of matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ satisfying $c \geq a \geq 0$, $d \geq b \geq 0$. In turn, the subset of Q consisting of words in σ and ρ which end in a ρ is the subset of those elements satisfying the additional inequalities $b \geq a \geq 0$, $d \geq c \geq 0$, i.e. precisely our set Red. This shows again that the elements of Red are uniquely expressible as products of the matrices $\sigma^{n-1}\rho = \left(\begin{smallmatrix} 0 & 1 \\ 1 & n \end{smallmatrix}\right)$ with $n \geq 1$. We define $\ell(\gamma)$ for any $\gamma \in \mathsf{Q}$ as the number of ρ's in the representation of γ as a word in ρ and σ; this agrees with our previous definition on the subset $\mathsf{Red} = \mathsf{Q}\rho$.

Let Q_n be the subset of Q consisting of words in ρ and σ of length n. The recursive description $\mathsf{Q}_0 = \{1\}$ and $\mathsf{Q}_{n+1} = \mathsf{Q}_n\sigma \cup \mathsf{Q}_n\rho$ implies the identity $\sum_{\gamma\in\mathsf{Q}_n}[\gamma] = \big([\sigma]+[\rho]\big)^n$ in the group ring $\mathbb{Z}[\Gamma]$. More generally, if we introduce a variable v and define

$$\mathcal{K}_v = [\sigma] + v\,[\rho] = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right] + v\left[\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right] \in \mathbb{Z}[\Gamma][v]\,,$$

then we have $\mathcal{K}_v{}^n = \sum_{\gamma\in\mathsf{Q}_n} v^{\ell(\gamma)}\,[\gamma]$. On the other hand, $\mathsf{Q} = \bigcup_{n=0}^{\infty}\mathsf{Q}_n$, so to deal with all of Q (or Red) we must work with infinite sums of elements of Γ. In particular, let

$$\mathcal{L}_w \;=\; (1 - w\,\sigma)^{-1}\rho \;=\; \sum_{n=1}^{\infty} w^{n-1}\left[\begin{smallmatrix} 0 & 1 \\ 1 & n \end{smallmatrix}\right].$$

This reduces to our previous formal expression $\mathcal{L}$ at $w = 1$, but now makes sense as an element in the ring $\mathbb{Z}[\Gamma][[w]]$ of formal power series in one variable over the group ring $\mathbb{Z}[\Gamma]$, or as an element of $\mathbb{C}[\Gamma]$ if $w \in \mathbb{C}$, $|w| < 1$. Then we have the identities

$$\mathcal{K}_v{}^{n-1}\,[\rho] \;=\; \sum_{\substack{\gamma\in\mathsf{Red}\\ n(\gamma)=n}} v^{\ell(\gamma)-1}\,[\gamma] \qquad \text{and} \qquad \mathcal{L}_w{}^{\ell} = \sum_{\substack{\gamma\in\mathsf{Red}\\ \ell(\gamma)=\ell}} w^{n(\gamma)}\,[\gamma]\,,$$

where $n(\gamma)$ for $\gamma \in \mathsf{Q}$ denotes the length of γ as a word in σ and ρ. Combining, we get

$$\big(1 - w\,\mathcal{K}_v\big)^{-1}[\rho] \;=\; \mathcal{L}_w\,\big(1 - vw\,\mathcal{L}_w\big)^{-1} = \sum_{\gamma\in\mathsf{Red}} v^{\ell(\gamma)-1}\,w^{n(\gamma)-1}\,[\gamma]\,.$$

Integrating with respect to v gives the identity

$$-\log\bigl(1-vw\,\mathcal{L}_w\bigr) \;=\; \sum_{\gamma\in\mathsf{Red}} \frac{v^{\ell(\gamma)}}{\ell(\gamma)}\, w^{n(\gamma)}\,[\gamma]\,,$$

and the content of §1 can now be summarized by saying that we computed $-\log Z_\Gamma(s)$ as the trace of π_s of this sum on $\mathbf{V}$ in the limit $v = w = 1$.

C. The Selberg zeta function and the dynamics of the Gauss map. The proof of equation (6) given originally by Mayer is parallel in many ways to the one given above, but was expressed in terms of ideas coming from symbolic dynamics. Specifically, he used the connection between closed geodesics on $\mathcal{H}/\Gamma_1$ and periodic continued fractions to relate the Selberg zeta function to the dynamics of the "continued fraction map" (Gauss map) $F : [0,1) \to [0,1)$ which maps x to the fractional part of $1/x$ (and, say, to 0 if $x = 0$). We give a very brief outline of the argument.

To a "dynamical system" $F : X \to X$ and a weight function $h : X \to \mathbb{C}$ one associates for each integer $n \geq 1$ a *partition function*

$$Z_n(F,h) = \sum_{x\in X,\ F^n x = x} h(x)\,h(Fx)\,h(F^2x)\,\cdots\,h(F^{n-1}x)$$

(sum over n-periodic points). In our case, $X = [0,1)$, F is the continued fraction map, and we take for $h(x)$ the function $h_s(x) = x^{2s}$ where $s \in \mathbb{C}$ with $\Re(s) > \frac{1}{2}$ (to make the series defining Z_n converge). Using the technique of "transfer operators" and Grothendieck's theory of nuclear operators, Mayer shows that

$$Z_n(F,h_s) = \mathrm{Tr}(L_s^n) - (-1)^n\,\mathrm{Tr}(L_{s+1}^n) \qquad (\forall n \geq 0). \tag{10}$$

On the other hand, the definition of the Selberg zeta function can be written $Z(s) = \prod_{k=0}^{\infty} \zeta_{SR}(s+k)^{-1}$, where $\zeta_{SR}(s)$ (the letters "SR" stand for Smale-Ruelle) is defined as the product over all closed primitive geodesics in $\mathcal{H}/\Gamma_1$ of $(1 - e^{-\lambda s})$, λ being the length of the geodesic. The connection between closed geodesics and periodic continued fractions leads to the equation $\zeta_{SR}(s) = \exp\bigl(\sum_{n=1}^{\infty} \frac{1}{n} Z_{2n}(F,h_s)\bigr)$. (Here only even indices occur because the map $x \mapsto x^{-1} - m$ implicit in the definition of F corresponds to a matrix of determinant -1, so that only even iterates of F correspond to the action of elements of Γ_1.) Together with (10) and the determinant formula $\exp\bigl(-\sum_{n=1}^{\infty} \frac{1}{n}\mathrm{Tr}(A^n)\bigr) = \det(1-A)$ this gives $\zeta_{SR}(s) = \dfrac{\det(1-L_{s+1}^2)}{\det(1-L_s^2)}$ and hence finally $Z(s) = \det(1-L_s^2)$.

A similar proof, of course, works also for equation (2), but now using all the $Z_n(F,h_s)$. This version of Mayer's theorem was developed by Efrat [1]. To connect this to the discussion in **A** above, we rewrite (9) slightly as

$$-\log Z(s) = \sum_{\gamma\in\mathsf{Red}} \frac{1+\det(\gamma)}{\ell(\gamma)}\,\chi_s(\gamma)\,,$$

or equivalently as the factorization $Z(s) = Z_+(s) \cdot Z_-(s)$ where $Z_+(s) = \sum_{\mathsf{Red}} \chi_s(\gamma)/\ell(\gamma)$ and $Z_-(s) = \sum_{\mathsf{Red}} \det(\gamma)\chi_s(\gamma)/\ell(\gamma)$. The first factor is $Z_\Gamma(s)$ by the calculation in Sections 1 and 2, so its zeros correspond to even Maass wave forms, while the zeros of the second factor $Z_-(s)$ correspond to the odd wave forms. See §5B for more on this.

Part II. Period functions of Maass wave forms

§4. Various descriptions of the period correspondence. We explained in the introduction how the identity (2) should lead one to expect some sort of correspondence between eigenfunctions of the Laplacian in $\mathcal{H}/\Gamma$ and holomorphic solutions of the three-term functional equation (3). In this section we give several descriptions of this "period correspondence," each of which puts into evidence certain of its properties. There does not seem to be any single description which exhibits all aspects of the correspondence simultaneously.

We first recall some basic facts about Maass wave forms and fix terminology. The Maass wave forms for the modular group $\Gamma_1 = \mathrm{SL}(2,\mathbb{Z})$ are the non-constant Γ_1-invariant eigenfunctions of the hyperbolic Laplacian $\Delta = -y^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)$ which are square-integrable on the modular surface $\mathcal{H}/\Gamma_1$. The space of these forms breaks up under the action of the involution $\iota : z \mapsto -\bar{z}$ into the spaces of even (invariant) and odd (anti-invariant) forms. In particular, the even forms are the eigenfunctions of Δ on $\mathcal{H}/\Gamma$, since Γ is generated by Γ_1 and ι. We will always use the letter u to denote a Maass form and the letter s for its *spectral parameter*, i.e. for the complex number s such that the eigenvalue of u under Δ is $s(1-s)$. (Note that the number $1-s$ is an equally good spectral parameter for u, but to describe the period correspondence $u \leftrightarrow \psi$ we must fix the choice of s since the functional equation (3) depends on s. However, this dependence is very simple because it is known that s always has real part $\frac{1}{2}$ and hence $1-s = \bar{s}$, so that the map $\psi(z) \mapsto \overline{\psi(\bar{z})}$ maps the space of solutions of (3) for one choice of s to the corresponding choice for the other.) The invariance of u under the translation map $T : z \mapsto z+1$ and the conjugation map ι implies that $u(x+iy)$ has a cosine expansion with respect to x, and the square-integrability of u and differential equation $\Delta u = s(1-s)\,u$ imply that this expansion has the form

$$u(x+iy) = \sqrt{y} \sum_{n=1}^{\infty} a_n\, K_{s-\frac{1}{2}}(2\pi n y)\, \cos(2\pi n x)\,, \tag{11}$$

where K_ν is a modified Bessel function.

A. Description of the period correspondence via integral transforms. A number of formulas for the period correspondence $u \leftrightarrow \psi$ were given in [2]. A particularly direct one is the integral formula

$$\psi(z) = z \int_0^\infty \frac{t^s\, u(it)}{(z^2+t^2)^{s+1}}\, dt \qquad (\Re(z) > 0)\,. \tag{12}$$

This was obtained after a number of intermediate steps. One of the most striking is that there is an an entire function $g(w)$ which is related to u by

$$g(\pm 2\pi i n) = \tfrac{1}{2}\,(2\pi n)^{-s+1/2}\, a_n \qquad (n = 1, 2, 3, \dots) \tag{13}$$

(i.e., g is a "holomorphic interpolation" of the Fourier coefficients of u) and to ψ by

$$g^{(k)}(0) = \frac{1}{\Gamma(2s+k)}\,\psi^{(k)}(1) \qquad (k = 0,\ 1,\ 2,\ \dots) \tag{14}$$

(so that the Taylor coefficients of g at 0 and ψ at 1 determine each other). The function g in turn is obtained from another intermediate function ϕ which is defined by

$$\phi(w) = w^{1-s} \int_0^\infty \sqrt{wt}\, J_{s-\frac{1}{2}}(wt)\, u(it)\, dt \tag{15}$$

(Hankel transform) and defines ψ by

$$\psi(z) = \int_0^\infty \phi(w)\, w^{2s-1}\, e^{-zw}\, dw \tag{16}$$

(Laplace transform). Substituting (15) into (16) gives (12), while substituting the Fourier expansion (11) into (15) and integrating term by term leads to the formula

$$\phi(w) = w \sum_{n=1}^{\infty} \frac{(2\pi n)^{-s+1/2}\, a_n}{w^2 + (2\pi n)^2}\,.$$

In particular, $\phi(w)$ has simple poles of residue $\frac{1}{2}(2\pi n)^{-s+1/2} a_n$ at $w = \pm 2\pi i n$ and no other poles, so the function $g(w) := (1 - e^{-w})\phi(w)$ is entire and satisfies (13), while on the other hand, once one has proved that $\psi(z)$ satisfies the three-term functional equation (3) one immediately gets

$$\int_0^\infty g(zw)\, w^{2s-1}\, e^{-w}\, dw = z^{-2s}\left[\psi(z^{-1}) - \psi(z^{-1}+1)\right] = \psi(z)\,,$$

and (14) follows easily. No single one of these formulas permits one to deduce in a direct way the properties of $\psi(z)$ (i.e., the analytic continuability to $\mathbb{C}' = \mathbb{C} \setminus (-\infty, 0]$ and the functional equation (3)) from the fact that u is a Maass form, and the proof of this in [2] is quite complex. On the other hand, they do give explicit ways to get from u to ψ and back: the forward direction is given by (12), while (13) and (14) determine the Fourier coefficients of u as special values of the power series $g(w) = \sum_k \psi^{(k)}(1) w^k / k!\,\Gamma(2s+k)$.

We refer to [2] and [3] for a more detailed discussion of these ideas and of other related approaches, including one based on a summation formula of Ferrar and another in terms of the Helgason automorphic boundary form of u, which are also important aspects of the story and provide useful perspectives.

B. Description in terms of Fourier expansions. The integral representation (12) makes visible the analyticity of $\psi(z)$ in a neighborhood of the positive real axis, but does not make it clear why ψ satisfies the three-term functional equation. In [3] a different description of ψ was given in which the functional equation becomes obvious and the key point is the continuability of ψ across the positive real axis. The starting point is the following simple algebraic fact.

Lemma. *If $\psi : \mathbb{C} \setminus \mathbb{R} \to \mathbb{C}$ is any function satisfying the three-term functional equation* (3) *then the function $f : \mathcal{H} \to \mathbb{C}$ defined by*

$$f(z) = \psi(z) + e^{-2\pi i s}\, \psi(-z) \tag{17}$$

is 1-periodic (i.e. T-invariant). Conversely, if $f : \mathcal{H} \to \mathbb{C}$ is any 1-periodic function, then the function $\psi : \mathbb{C} \setminus \mathbb{R} \to \mathbb{C}$ defined by

$$\psi(z) = \begin{cases} f(z) - z^{-2s} f(-1/z) & \text{if } \Im(z) > 0 \\ z^{-2s} f(1/z) - f(-z) & \text{if } \Im(z) < 0 \end{cases} \tag{18}$$

satisfies the functional equation (3). *Moreover, if $s \notin \mathbb{Z}$ then the correspondences* (17) *and* (18) *between 1-periodic functions in $\mathcal{H}$ and solutions of* (3) *in $\mathbb{C} \setminus \mathbb{R}$ are inverse maps to each other up to a non-zero scalar factor $1 - e^{-2\pi i s}$.*

Then we have the following very elegant description of the period correspondence.

Theorem 3. *Let u be an even Maass wave form with spectral parameter s and Fourier expansion given by* (11), *and $f : \mathcal{H} \to \mathbb{C}$ the 1-periodic holomorphic function defined by*

$$f(z) = \sum_{n=1}^{\infty} n^{s-\frac{1}{2}} a_n e^{2\pi i n z} \qquad (z \in \mathcal{H}). \tag{19}$$

Then the function ψ defined by (18) *extends holomorphically from $\mathbb{C} \setminus \mathbb{R}$ to $\mathbb{C}'$ and is bounded in the right half-plane. Conversely, if s is a complex number with $\Re(s) > 0$, $\psi : \mathbb{C}' \to \mathbb{C}$ a holomorphic solution of* (3) *which is bounded in the right half-plane, $f : \mathcal{H} \to \mathbb{C}$ the 1-periodic holomorphic function defined by* (17), *and $\{a_n\}$ the coefficients defined by the Fourier expansion* (19), *then the function $u : \mathcal{H} \to \mathbb{C}$ defined by the Fourier series* (11) *is an even Maass wave form with spectral parameter s.*

The proof of this theorem, given in [3], relies essentially on the properties of L-series. It is well-known that the L-series $L(\rho) = \sum_n a_n / n^\rho$ of a Maass wave form has a holomorphic extension to all complex values of ρ and satisfies a functional equation under $\rho \mapsto 1 - \rho$, and conversely that these properties of the coefficients a_n imply the Γ-invariance of the function u defined by (11). The L-series can be represented as the Mellin transform of the restrictions to the imaginary axis of either u or f (with different gamma-factors). We can now use the inverse Mellin transform to write the function ψ defined by (18) in the uppper and lower half-planes as integral transforms of $L(\rho)$, and the functional equation of L turns out to be just what is needed in order that these two formulas agree and define a holomorphic function in all of $\mathbb{C}'$. Conversely, if u is defined by (11) and f by (19) for some coefficients a_n (satisfying a growth condition), and if ψ is the function defined by (18), then the Mellin transforms of the restrictions of ψ to the positive and the negative imaginary axes are both linear combinations of $L(\rho)$ and $L(1-\rho)$. Now if ψ extends holomorphically across $\mathbb{R}_+$ and satisfies the growth condition, we can rotate the two paths of integration to $\mathbb{R}_+$, and the equality of these two linear combinations then gives the functional equation of the L-series.

This argument makes clear which properties of u correspond to which properties of ψ: if $\{a_n\}$ is any collection of coefficients (of not too rapid growth), then the function u defined by (11) is a T-invariant eigenfunction of the Laplacian with eigenvalue $s(1-s)$, while the function ψ defined by (19) and (18) is a holomorphic solution of the functional equation (3) in the upper and lower half-planes; this gives a bijection between translation-invariant

even eigenfunctions of Δ and functions ψ on $\mathbb{C} \smallsetminus \mathbb{R}$ satisfying (3), and under this bijection the eigenfunctions which are invariant under $z \mapsto -1/z$ correspond to the functions ψ which extend holomorphically across the positive real axis.

C. Unfolding from the positive real axis. In this subsection we state a result from [3] to the effect that the restriction map from the space of holomorphic solutions of (3) in $\mathbb{C}'$ to the space of analytic solutions of (3) on $\mathbb{R}_+$, which is obviously injective, is in fact bijective under suitable growth conditions. This complements the results of the two preceding subsections: in **A** we described how to get from u to $\psi|\mathbb{R}_+$ via an integral transform and how to get the Fourier coefficients of u from the Taylor expansion of ψ at $1 \in \mathbb{R}_+$, and in **B** we explained how to get the values of ψ off the real axis from u and vice versa.

Theorem 4. *Let s be a complex number with $\Re(s) > 0$. Then any bounded real-analytic solution of the functional equation* (3) *on the positive real axis extends to a holomorphic solution of* (3) *in the whole cut plane which is bounded in the right half-plane.*

The proof of this theorem is by a kind of "bootstrapping": by repeated applications of the functional equation (3) one successively extends ψ to larger and larger neighborhoods of $\mathbb{R}_+ \subset \mathbb{C}'$, while preserving the growth conditions. In fact, the growth conditions can be relaxed, e.g. if $\Re(s) = \frac{1}{2}$ then the assumption $\psi(x) = o(1/x)$ as $x \to 0$ already implies that ψ continues to a holomorphic function in $\mathbb{C}'$ and is bounded in $\Re(z) > 0$, which together with Theorem 3 implies that s is the spectral parameter associated to a Maass wave form. This is especially surprising because it turns out that *any* smooth solution of the functional equation on $\mathbb{R}_+$ is $O(1/x)$ as $x \to 0$ and that these solutions form an uncountable-dimensional vector space for any s, whereas the Maass forms exist only for special values of s and then form a finite-dimensional space.

§5. Complements. In the final section of the paper we give various examples of solutions of the functional equation (3), especially the polynomial solutions for negative integral values of s which give the link to the classical theory of periods of modular forms, and also indicate the changes that must be made when Γ is replaced by its subgroup Γ_1.

A. Examples and equivalent forms of the three-term functional equation. If we relax the growth conditions on the function ψ, then there are many more solutions of the functional equation (3). For example, an infinite class of solutions for any s is given by $\psi(z) = f(z) + z^{-2s} f(1/z)$ for any odd and 1-periodic entire function f. There are also more interesting examples which nearly satisfy the growth conditions of Theorem 3 and which correspond to the zeros of the Selberg zeta function other than the spectral parameters coming from Maass wave forms. These zeros occur at $s = 1$ and at the zeros of $\zeta(2s)$, where ζ is the Riemann zeta function (cf. [7], pp. 48-49). The solution of (3) for $s = 1$ is given by $\psi(z) = 1/z$. The solutions corresponding to the trivial zeros of $\zeta(2s)$ at $s = -1, -2, \ldots$ will be discussed in the next section. The solutions corresponding to the non-trivial zeros arise as follows. For $\Re(s) > 1$ define

$$\psi_s(z) = \zeta(2s)\bigl(1 + z^{-2s}\bigr) + 2 \sum_{m,n \geq 1} (mz + n)^{-2s},$$

a kind of "half-Eisenstein-series." The series converges absolutely and it is easy to check that it satisfies the functional equation (3). On the other hand, the shifted function $h_s(z) = \psi_s(z+1)$ is *not* a fixed point of the Mayer operator L_s; instead, as one checks in a straightforward way, one has $(L_s h_s)(z) = h_s(z) - \zeta(2s)$. It follows that the (easily obtained) analytic continuation of h_s gives a fixed point of (the analytic continuation of) L_s at the zeros of $\zeta(2s)$.

We also mention two equivalent forms of the period functional equation, as a sample of the algebraic character of the theory. The first is the equation

$$\psi(z) = (z+1)^{-2s}\left[\psi\left(\frac{z}{z+1}\right) + \psi\left(\frac{1}{z+1}\right)\right].$$

Written in the language of the group algebra $\mathbb{Z}[\Gamma]$, this says that $\pi_s(\mathcal{K})\psi = \psi$, where $\mathcal{K}$ is the element $\mathcal{K}_1 = [\sigma] + [\rho]$ of §3**B** and is related to the Mayer element $\mathcal{L}$ by the equation $\mathcal{L}(1-\mathcal{L})^{-1} = (1-\mathcal{K})^{-1}[\rho]$. The second says that ψ is fixed by the operator $\pi_s\left(\sum_{n\geq 0}[\rho^n\sigma]\right)$. Written out, this is the infinitely-many-term functional equation

$$\psi(z) = \sum_{n=1}^{\infty} \frac{1}{(F_n\, z + F_{n+1})^{2s}}\, \psi\left(\frac{F_{n-2}\, z + F_{n-1}}{F_n\, z + F_{n+1}}\right)$$

where $\{F_n\}$ are the Fibonacci numbers. Note that this series, unlike the one defining the Mayer operator L_s, is rapidly convergent if $\Re(s) > 0$ and $\Re(z) > -(1+\sqrt{5})/2$.

B. Even and odd Maass wave forms. We now consider the modular group Γ_1 instead of Γ. As mentioned at the beginning of §4, the Maass wave forms for Γ_1 break up into two kinds, the even ones (which are invariant under the map $u(z) \mapsto u(-\bar{z})$ and hence under all of Γ) and the odd ones (for which $u(z) = -u(-\bar{z})$). The spectral parameters corresponding to both kinds of Maass forms are zeros of the Selberg zeta function $Z(s)$ of Γ_1, with the ones corresponding to even forms being zeros of $Z_\Gamma(s)$. On the other hand, as we saw in Part I, $Z_\Gamma(s)$ is the determinant of the operator $1 - L_s$, while $Z(s)$ is the determinant of $1 - L_s^2 = (1-L_s)(1+L_s)$. The odd Maass forms should therefore correspond to the solutions in $\mathbf{V}$ of $L_s h = -h$ and hence, after the same shift $\psi(z) = h(z-1)$ as in the even case, to the solutions of the *odd three-term functional equation*

$$\psi(z) = \psi(1+z) - z^{-2s}\,\psi\left(1+\frac{1}{z}\right), \tag{20}$$

instead of the even functional equation (3). This is in fact true and, as one would expect, the description and properties of this "odd period correspondence" are very similar to those in the even case. The Fourier cosine expansion (11) is naturally replaced by the corresponding sine series. The integral transform (12), which must obviously be modified since $u(iy)$ is now identically zero, is replaced by

$$\psi(z) = \int_0^\infty \frac{t^s\, u_x(it)}{(z^2+t^2)^s}\, dt \qquad (\Re(z) > 0),$$

where $u_x = \dfrac{\partial u}{\partial x}$ $(z = x + iy)$. The algebraic correspondence described in the Lemma in §4**B** is true with appropriate sign changes (change the sign of the second term in (17) and of both terms in the second line of (18)), and Theorem 3 then holds *mutatis mutandum*.

Examples of non-Maass solutions of the odd functional equation are the function $1-z^{-2s}$ (or more generally $f(z)-z^{-2s}f(1/z)$ with f even and 1-periodic) for all s and $\psi(z)=\log z$ for $s=0$. The example $\psi_s(z)$ discussed in Subsection **A** has no odd analogue. (The analogous fact about Selberg zeta functions is that all the zeros of $Z_-(s)=Z(s)/Z_\Gamma(s)$ correspond to the odd spectral parameters, whereas the zeros of $Z_\Gamma(s)$ correspond both to even Maass forms and to zeros of the Riemann zeta function.) The two alternate forms of the even functional equation given at the end of**A** have the obvious odd analogues (replace $\mathcal{K}_{s,1}$ by $\mathcal{K}_{s,-1}$ and $\sum_{n\geq 0}\psi|_{\rho^n\sigma}$ by $\sum_{n\geq 0}(-1)^n\psi|_{\rho^n\sigma}$)).

Finally, one can give a uniform description of the period functions associated to Maass forms, without separating into the even and odd cases. These functions should correspond to the fixed points of L_s^2 on V, and this leads (after the usual shift $\psi(z)=h(z-1)$) to the "master functional equation"

$$\psi(z)=\psi(z+1)+(z+1)^{-2s}\,\psi\Big(\frac{z}{z+1}\Big)\,. \tag{21}$$

We will call a solution of (21) a *period function*. Since the involution $\psi(z)\mapsto z^{-2s}\psi(1/z)$ preserves this equation, every period function decomposes uniquely into an even (invariant) and odd (anti-invariant) part, and one checks easily that the even and odd period functions are precisely the solutions of (3) or (20), respectively. The description of the period correspondence given in §4**B** is now modified as follows. Any 1-periodic eigenfunction of Δ with eigenvalue $s(1-s)$ has a Fourier expansion of the form $u(x+iy)=\sqrt{y}\sum_{n\neq 0}a_n\,K_{s-\frac12}(2\pi|n|y)e^{2\pi inx}$. We then define a 1-periodic holomorphic function f on $\mathbb{C}\setminus\mathbb{R}$ by two different Fourier series, using the a_n with $n>0$ in the upper half-plane and the a_n with $n<0$ in the lower half-plane. In each half-plane there is a 1:1 correspondence between the space of 1-periodic functions and the space of solutions of (21) given by the (up to a scalar factor, inverse) transformations

$$f(z)\mapsto\psi(z):=f(z)-z^{-2s}f(-1/z)\,,\qquad \psi(z)\mapsto f(z):=\psi(z)+z^{-2s}\psi(-1/z)\,.$$

Then, just as before, the invariance of u under $z\mapsto -1/z$ is equivalent (under suitable growth conditions) to the analytic continuability of $\psi(z)$ across the positive real axis.

C. Integral values of s and classical period theory. Let s be a negative integer, which we write in the form $1-k$ with $k\geq 2$. The factor z^{-2s} in the master functional equation (21) (or in its even or odd versions (3) or (20)) now becomes a monomial and we can look for polynomial solutions ψ, which we will then call *period polynomials*. The degree of such a polynomial must be $\leq 2k-2$, so the problem of finding all solutions for a given k is just a matter of finite linear algebra. For $k=2$, 3, 4 and 5 we find that the only polynomial solution is $z^{2k-2}-1$ (which is an odd polynomial but an even period function), but for $k=6$ there are three linearly independent solutions $z^{10}-1$, $z^8-3z^6+3z^4-z^2$, and $4z^9-25z^7+42z^5-25z^3+4z$. This has to do with the fact that for $k=6$ the space S_{2k} of cusp forms of weight $2k$ on the modular group has a non-trivial element for the first time, namely the discriminant function

$$\Delta(z)=e^{2\pi iz}\prod_{n=1}^{\infty}(1-e^{2\pi inz})=\sum_{n=1}^{\infty}\tau(n)\,e^{2\pi inz}\,.$$

Associated to this cusp form is its *Eichler integral* $\tilde{\Delta}(z) = \sum_n n^{-11}\tau(n)e^{2\pi inz}$, which is not quite modular (of weight $2-2k=-10$) but instead satisfies $(cz+d)^{10}\tilde{\Delta}\left(\frac{az+b}{cz+d}\right) = \tilde{\Delta}(z) + P_\gamma(z)$ for any $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_1$ with P_γ a polynomial of degree ≤ 10, and the 3-dimensional space of period polynomials is generated by the odd and even parts of the polynomial P_γ for $\gamma = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, together with the polynomial $z^{10}-1$. In general, if one associates to any cusp form $f(z) = \sum A_n e^{2\pi inz} \in S_{2k}$ its Eichler integral $\tilde{f}(z) = \sum n^{-2k+1}A_n e^{2\pi inz}$, then the difference $\tilde{f}(z) - z^{2k-2}\tilde{f}(-1/z)$ is a polynomial $P = P_f$ of degree $\leq 2k-2$ which satisfies the *period conditions*

$$P(z) + z^{2k-2}P\left(\frac{-1}{z}\right) = P(z) + z^{2k-2}P\left(1-\frac{1}{z}\right) + (z-1)^{2k-2}P\left(\frac{1}{1-z}\right) = 0\,,$$

and the period theory of Eichler, Shimura and Manin tells us that this space has dimension $2\dim S_{2k-2}+1$ and is spanned by $z^{2k-2}-1$ and by the even and odd parts of the polynomials P_f. But an elementary calculation shows that polynomials satisfying the period conditions are precisely the polynomial solutions of (21) with $s = 1-k$ (and further that this space breaks up into the direct sum of its subspaces of odd and even polynomials and that these are precisely the polynomial solutions of (3) and of (20) respectively). This fits in very well with our picture since it is known that $s = 1-k$ is a zero of $Z_\Gamma(s)$ of multiplicity $\delta_k := \dim S_{2k}$ and a zero of $Z(s)$ of multiplicity $2\delta_k + 1$. What's more, one can get directly from cusp forms of weight $2k$ to nearly Γ_1-invariant eigenfunctions of the Laplace operator with eigenvalue $k(1-k)$. For instance, the eigenfunction defined by (11) with $s=-5$ and $a_n = \tau(n)/n^{11/2}$ is not only invariant under the translation T and the reflection ι, but is nearly invariant under $z \mapsto -1/z$, the difference $u(-1/z) - u(z)$ being a polynomial in x, y and $1/y$ with coefficients which are closely related to those of the odd period polynomial $4z^9 - 25z^7 + 42z^5 - 25z^3 + 4z$ above.

References

[1] I. Efrat, *Dynamics of the continued fraction map and the spectral theory of* $SL(2,\mathbb{Z})$, Invent. Math. **114** (1993), 207–218.

[2] J.B. Lewis, *Spaces of Holomorphic Functions equivalent to the even Maass Cusps Forms*, Invent. Math., to appear.

[3] J.B. Lewis and D. Zagier, *Period functions for Maass wave forms*, in preparation.

[4] D. Mayer, *The Thermodynamic formalism Approach to Selberg's Zeta function for* $PSL(2,\mathbb{Z})$, Bull. AMS **25** (1991), 55–60.

[5] D. Mayer, *Continued fractions and related transformations*, in "Ergodic Theory, Symbolic Dynamics and Hyperbolic Spaces," T. Bedford, M. Keane, C. Series (Eds.), Oxford University Press, Oxford 1991, pp. 175–222.

[6] A. Terras, "Harmonic Analysis on Symmetric Spaces and Applications," Vol. I, Springer, New York 1985.

[7] A.B. Venkov, "Spectral Theory of Automorphic Functions", Kluwer, Dordrecht 1990.

[8] D. Zagier, "Zetafunktionen und Quadratische Körper", Springer-Verlag, Berlin 1981.

STATISTICAL PROPERTIES OF RANDOM MATRICES AND THE REPLICA METHOD

G. PARISI
Dipartimento di Fisica, Università La Sapienza *and INFN Sezione di Roma I*
Piazzale Aldo Moro, Rome 00185

Abstract

I present here some results on the statistical behaviour of large random matrices in an ensemble where the probability distribution is not a function of the eigenvalues only. The perturbative expansion can be cast in a closed form and the limits of validity of this expansion are carefully analyzed. A comparison is done with a similar model with quenched disorder, where the solution can be found by using the replica method. Finally I will apply these results to a model which should describe the liquid-glass transition in high dimensions.

1 Introduction

I cannot work on random matrices without thinking to Claude Itzykson. I started to study this subject with him together with Edouard Brézin and Jean-Bernard Zuber[1]. I enjoyed very much to work in this group. I had collaborated before with each of them separately, but it was the first time I was working with all of them together. It was a very interesting and pleasant experience. I feel still nostalgia for the long afternoons spent observing my friends doing long and difficult computations at the blackboard practically without doing mistakes.

After that work our roads partially separated. My friends remained interested in the field [2,3], while I started to be interested in spin glasses. Claude and Jean-Bernard started to study the case with two interacting matrices and produced after much struggle the wonderful paper on the unitary integrals [2]. At that time I had frequent discussions with Claude; he was explaining me the difficulties they met with their problem and I was telling him those I had with spin glasses. We tried to make some progresses together.

This friendly exchange was quite fruitful[a]. Let me quote two examples.

[a] I use this opportunity to acknowledge an other debt I have with Claude. I was strongly influenced by his old paper with Abarbanel on the eikonal [4]. I read that paper when I was starting to work in physics. I learned from it the power of the functional method and the relation among classes of diagrams and the solution of the Schroedinger equation in a random Gaussian field. That paper was quite present in my mind when ten years later we started to work on what we called *quenched* gauge theories [5].

The interpretation of the breaking of replica symmetry in terms of many pure states was also triggered by Claude's observation that

$$\sum_b Q^s_{a,b} = \frac{\sum_{k_i} \langle \prod_{i=1,s} \sigma_{k_i} \rangle^2}{N^s}. \tag{1}$$

A that time this relation was quite mysterious but it turned out to be a crucial step for later developments [6,7].

During one of my visits to Saclay I found a simple derivation of a result of Bessis [8] on orthogonal polynomials. I showed it to Claude, who was very interested and started immediately to derive some of the formulae needed to cast the result in a more compact way. I was too busy with spin glasses so I didn't write it. The results were finally incorporated in the paper on the unitary integrals [2] and fortunately in this way they were saved from the oblivion.

Two years ago, when with Marinari and Ritort we started to apply the replica method to non random systems [9], it was a real pleasure for me to read again the paper of Claude and Jean-Bernard [2], which contained crucial results for our aims. At that time I was thinking that one day I will come to Saclay to explain our results to Claude and that I will hear his comments and suggestions, but sadly that day will never come.

The problem I will study here is the natural extension of that work of fifteen years ago [1] and it can be formulated as follows [10]. We have an $N \times N$ symmetric matrix M and we associate to it the following Hamiltonian:

$$H(M) = N\mathrm{Tr}(h(N^{-1/2}M)) + \frac{1}{N}\sum_{i,k} f(M_{i,k}). \tag{2}$$

When N goes to infinity, the Hamiltonian becomes a quantity of order N. We will consider the expectation values of intensive observables $A(M)$, which have a well defined limit when $N \to \infty$, e.g. they have the same form as the Hamiltonian apart from a factor N.

We are interested in the computation of the equilibrium quantities

$$\langle A \rangle_N \equiv \frac{\int dM \exp(-\beta H(M)) A(M)}{\int dM \exp(-\beta H(M))} \tag{3}$$

in the limit where $N \to \infty$.

We can also define dynamical expectation values: we introduce the Langevin equation

$$\frac{dM_{i,k}}{dt} = -\frac{\partial H}{\partial M_{i,k}} + \eta(t)_{i,k}, \tag{4}$$

where η is a random Gaussian distributed noise, uncorrelated for different pairs of indices, such that

$$\overline{\eta(t_1)_{i,k}\eta(t_2)_{i,k}} = \frac{2}{\beta}\delta(t_1 - t_2). \tag{5}$$

Here the bar denotes the average over different realizations of the noise.

The dynamical expectation values are defined as

$$\langle A\rangle_D \equiv \lim_{t\to\infty}\lim_{N\to\infty}\overline{A(M(t))}. \tag{6}$$

The order of the limits matters. Indeed a well known theorem states that

$$\langle A\rangle \equiv \lim_{N\to\infty}\lim_{t\to\infty}\overline{A(M(t))}. \tag{7}$$

If the two limits do not commute when A is the energy density, metastable states are present. While metastable states with strictly infinite mean life cannot be present in systems with a short range interaction, the study of infinite range models where metastable states do exist, may be quite useful for understanding real systems in which there are metastable states with finite, but extremely large, mean life.

There are many motivations for studying this problem.

The problem is interesting *per se*. It is a non-trivial generalization of the original problem of random matrices [1] with the extra difficulty that the probability distribution of the matrices is no more invariant under the rotation group $O(N)$. This invariance was a crucial tool in the previous investigations and it is not present here.

In these models at low temperature the averages in the dynamic approach may not coincide with those of the equilibrium approach. Systems for which the two limits (infinite volume and time respectively) cannot be exchanged have a strong interest [11] and their properties may be investigated using the replica method [12].

While there are no real doubts that in the dynamic approach the limit ($N \to \infty$) of intensive quantities is unique, it is possible (and in my opinion it is quite likely) that one can finds two (or more) different sequences N_r (i.e. $N_r \to \infty$ when $r \to \infty$) such that the limit of the equilibrium quantities at low temperatures along the sequence depends on the sequence. In similar cases the low temperature behaviour of equilibrium properties becomes linked with the solution of arithmetic problems [9,13].

The techniques we will use to study the model are based on the replica method. The possibility of applying the replica method to non random systems

is very interesting, especially because we hope that this extension may be useful in studying real glass.

Finally there is a model which is very interesting from the physical point: N interacting particles constrained to move on a sphere in αN dimension [14]. This model is the simplest non trivial model for interacting particles which should have a liquid-glass transition and (hopefully) can be solved.

This paper (which is based on the results of[9,10,14]) is organized as follows: in section 2 we present a diagrammatical analysis of the perturbative expansion which can be resummed in a compact way. In section 3 we present a model where the Hamiltonian is random (quenched disorder). Its properties coincide with those of our original model in the high temperature phase; the model can be solved analytically using the replica method. In section 4 we discuss the relations of the random model with our original model and we present and disprove some conjectures. In section 5 we present a physically motivated model [14] of interacting particles which can be written in the form given by eq. (2). Finally in section 6 we present some tentative conclusions.

2 The perturbative expansion

In this section we consider the perturbative expansion for the following model

$$H(M) = \frac{1}{2}\mathrm{Tr}(M^2) + \frac{g_E \mathrm{Tr}(M^4)}{N} + g_L \sum_{i,k} M_{i,k}^4. \tag{8}$$

This Hamiltonian corresponds to the choice

$$\begin{aligned} h(x) &= (\frac{1}{2} + C)x^2 + g_E x^4, \\ f(x) &= -Cx^2 + g_L x^4. \end{aligned} \tag{9}$$

The value of C is arbitrary, indeed

$$\mathrm{Tr}(M^2) = \sum_{i,k} M_{i,k}^2. \tag{10}$$

The same term in the previous equation can be represented in both forms. In the rest of the paper we will use $C = 0$; in this section we will make the choice $\beta = 1$.

Our aim is to investigate at the diagrammatical level if there are some simplifications when $N \to \infty$. Indeed in this limit where $f = 0$ (i.e. $g_L = 0$) only planar diagrams survive. Here the situation is more complicated.

A careful diagrammatical analysis shows the following[10] results.

Let us consider the contributions to a connected Green function and let us exclude diagrams containing self energy corrections. In this case there are no mixed terms and the result is the sum of two functions, depending only on g_E and g_L respectively.

If we consider the self energy corrections in the internal lines, mixed terms may appear. Indeed the two point function has only one possible form, irrespectively from the type of vertices which contribute to it (in other words, there is only one quadratic invariant).

If we denote by G the value of the propagator (I do not indicate the dependance on the indices), we find that any Green function (Γ) can be written as

$$\Gamma = \Gamma_E(g_E, G) + \Gamma_L(g_L, G). \tag{11}$$

The same conclusion is valid for the self energy diagrams, which can be written as

$$\Sigma = \Sigma_E(g_E, G) + \Sigma_L(g_L, G). \tag{12}$$

The propagator is thus given by

$$G = \frac{1}{1 + \Sigma_E(g_E, G) + \Sigma_L(g_L, G)}. \tag{13}$$

The solution of this equation gives the value of the propagator G and from it we can reconstruct the other properties of the model.

These diagrammatical findings can be summarized in the following elegant way which generalizes to matrices the representation found in[9] for vectors.

We introduce two $N \times N$ matrices B and z. The total Hamiltonian for these two matrices is

$$H(B, z) = \mathrm{Tr}\left(Nh(BN^{-1/2}) + \frac{R_E}{2}B^2\right) + \sum_{i,k}\left(f(z_{i,k}) + \frac{R_L}{2}z_{i,k}^2\right), \tag{14}$$

where

$$R_L = 1 + \Sigma_L(g_L, G), \qquad R_E = \Sigma_E(g_L, G) \tag{15}$$

The previous conditions become now

$$G = \langle B_{i,k}^2 \rangle = \langle z_{i,k}^2 \rangle = \frac{1}{R_L + R_E}. \tag{16}$$

The expectation value of the energy in the original model can be finally written as

$$E = \langle \mathrm{Tr}\left(Nh(BN^{-1/2})\right) + \sum_{i,k} f(z_{i,k}) \rangle \tag{17}$$

More generally

$$\langle \mathrm{Tr}\left(N d_E(MN^{-1/2})\right) + \sum_{i,k} d_L(M_{i,k})\rangle =$$
$$\langle \mathrm{Tr}\left(N d_E(BN^{-1/2})\right) + \sum_{i,k} d_L(z_{i,k})\rangle \qquad (18)$$

Under this form the result is quite compact and it can be generalized to other interactions which have higher powers of M.

Both the integrations on B and z can be done explicitly: The B integration can be done using the standard techniques, because the integrand is now rotational invariant. The z integration factorizes in the product of N^2 independent integrals. The values of R_L and of R_E can be found by solving the two equations (16) in two unknowns. All computations can be done explicitly.

In the nutshell the final formula corresponds to a Hartree-Fock type approximation or, better, to two parallel spherical approximations: the local interaction (f) in the eigenvalue representation (B) has the only effects of modifying the quadratic term and viceversa. This formula generalizes well known results:

- If $f = 0$, the probability distribution of a given matrix element of the matrix M is Gaussian.
- If $h = 0$, the probability distribution of the eigenvalues is the same as in the Gaussian model.

However we have not been able (with reasons) to obtain this result in a compact way. We may speculate that the result is not valid beyond perturbation theory. Indeed, if such a general result would be correct, there should be a non-diagrammatical proof.

Generally speaking a perturbative result may break for two different reasons:

- There are non analytic terms, e.g. of the form $\exp(-C/(g_E g_L))$.
- There is a first order phase transition [b] at a non zero value of $g_E g_L$.

We shall see later that the second possibility is exactly what happens. However, before finding the limitations of this perturbative result, it is convenient to consider an other case in which the same result is obtained.

[b] In this context we say that a phase transition is of first order if no precursor signs are present, i.e. its presence cannot be predicted by finding susceptibilities which diverges when approaching the transition.

3 A Random equivalent model

It is interesting to investigate if there are other models for which the previous formulae are exact and can be proved in a compact way. The model can be found among those with random quenched disorder. We consider the following Hamiltonian:

$$H_O(z) \equiv H(z, B(z)) = N\mathrm{Tr}\left(h(BN^{-1/2})\right) + \sum_{i,k} f(z_{i,k}), \tag{19}$$

where z is an $N \times N$ matrix, $B(z)$ is a short notation for

$$B_{i,k} = \sum_{j,l=1,N} O_{i,k;j,l} z_{j,l}, \tag{20}$$

and O is an $N^2 \times N^2$ orthogonal matrix, where each pair of indices (i.e. i, k and j, l) plays the role of one index.

The Hamiltonian depends on O via the constraint in eq. (20). When the matrix O is the identity, we recover the previous model.

Here we consider the case in which O is a random orthogonal symmetric matrix. Our aim is to compute the following quantity in the large N limit

$$-N\beta F(\beta) = \int d\mu(O) \ln\left(\int dz \exp(-\beta H_O(z))\right). \tag{21}$$

To this end it is convenient to consider the quantity

$$-nN\beta F^{(n)} = \ln\left(\int d\mu(O) \left(\int dz \exp(-\beta H_O(z)\right)^n\right). \tag{22}$$

It is trivial to check that

$$\lim_{n\to 0} F^{(n)}(\beta) = F(\beta). \tag{23}$$

The replica method consists in computing $F^{(n)}$ for integer n and eventually continuing analytically the result to $n = 0$.

The first step consists in writing

$$\int dz \exp(-\beta H_O(z)) = \int dz dB d\lambda$$

$$\exp\left(-\beta H(z, B) + i \sum_{i,k} \lambda_{i,k} (B_{i,k} - \sum_{j,l=1,N} O_{i,k;j,l} z_{j,l})\right). \tag{24}$$

Now, a symmetric orthogonal matrix can be written as

$$O = VDV^* \tag{25}$$

where D is a diagonal matrix (in the generic case about half of the diagonal elements are equal to 1, the other to -1) and V is a generic orthogonal matrix. Using the formulae of reference [2] the integral over V can be done (if n remains finite when $N \to \infty$). In this way the constraint is integrated over and we remain with two separate integrals [c].

After a long computation, one finds that $F^{(n)}$ is given by the stationary point of $F^{(n)}(R_L, R_E)$. Here the R's are $n \times n$ matrices and

$$\exp\left(-nNF^{(n)}(R_L, R_E)\right) = \int dz dB \exp\left(-\beta H_{eff}(z, B, R_L, R_E)\right),$$

$$\beta H_{eff}(z, B, R_L, R_E) = \sum_{a=1,n} \beta H(z^a, B^a) + \tag{26}$$

$$\frac{1}{2} \sum_{a,b=1,n} \left(R_L^{a,b} \mathrm{Tr}(z^a z^b) + \mathrm{Tr} \ln(R_L + R_E) + R_E^{a,b} \mathrm{Tr}(B^a B^b) \right)$$

In the perturbative region where at least one of the two functions f and h is small, the matrices R's are diagonal and we recover the formulae of the previous chapter by looking to the stationary point of the free energy.

At low temperatures many new phenomena may appear. One of the most interesting is replica symmetry breaking. In this situation the saddle point is no more symmetric under the action of the replica group. The computations in the broken replica region are technically difficult, but they seem to be feasible and work in this direction is in progress [16].

It is interesting to recall that the original model is one of those we are considering here, i.e. it corresponds to $O = 1$.

4 A bold conjecture

4.1 A naive conjecture

The simplest conjecture would be to assume that the original formulae are always valid. In order to test this conjecture, we could study the model in

[c] The Itzykson-Zuber formula applies in the case of integral over unitary matrices for all N. Here we only need the leading term when N goes to infinity and we can this apply their final expression to orthogonal matrices; in both case only planar diagrams survive. The two results differ only to a crucial factor 2, which can be checked at the first non trivial order.

some limiting case. It is convenient to consider a very simple non trivial case, i.e. matrices M constrained to have matrix elements of modulus 1;

$$M_{i,k} = \pm 1 \tag{27}$$

This can be realized in the previous setting by using a function f, such that

$$\exp(-f(z)) = \delta(z^2 - 1) \tag{28}$$

We can now set $g_E = 1$ and write the partition function of the model as

$$\sum_{M_{i,k}=\pm 1} \exp(-\frac{\beta}{N}\mathrm{Tr}(M^4)). \tag{29}$$

According to the previous considerations, we have to study an auxiliary model with partition function

$$\int dB \exp\left(-R\mathrm{Tr}(B^2) - \frac{\beta}{N}\mathrm{Tr}(B^4)\right) \tag{30}$$

where R has been chosen in such a way that $\langle B_{i,k}^2 \rangle = 1$. In other words the spherical and the Ising model coincide in the high temperature region.

The solution of this problem can be found directly using the formulae of [1] for small positive β and their extension to the case of non-connected support of the eigenvalue distribution for large β.

Is this solution correct?

Certainly not for negative β. As noticed by Zinn-Justin, in this case we recover some form of Z_2 lattice gauge theory. For negative β the unconfining solution (i.e. $M_{i,k} \approx 1$) gives an energy proportional to N^2. This is not a serious contradiction, as far also the corresponding model eq. (30) does not exist for negative β.

The most serious troubles appear when β becomes large and positive. Here we can compute the free energy of the model $F(\beta)$ and from it we can recover other thermodynamical quantities. It is possible to compute the entropy.

One finds (as expected in a spherical model) that the entropy density $S(\beta)$ behaves at large β as

$$S(\beta) \approx -\ln(\beta) + const \tag{31}$$

The entropy becomes negative at sufficiently high β. This is impossible; the spherical and the Ising model cannot coincide in the low temperature region. Consequently, the solution of this *soluble model* must be wrong, in the same way as in the Sherrington-Kirkpatrick model [6].

Something better must be devised. Numerical simulations of the model show a perfect agreement for not too large values of β. The deviations appears suddenly at a value of β at which the specific heat is nearly discontinuous, strongly suggesting the presence of a phase transition.

4.2 The simplest reasonable conjecture

A better conjecture, which does not go against the positivity of the entropy, consists in assuming that the model with $O = 1$ coincides with the one with random O also in the low temperature phase. It is quite evident that in the case of random O the entropy cannot be negative and this observation is enough to remove the entropy crisis.

This conjecture may be extended also to the dynamic averages, which can now be computed by using an appropriate formulation of the replica method [12].

In this context, the analytic continuation of the free energy from the high temperature region has the meaning of the *annealed* value of the free energy, i.e. $F^{(1)}(\beta)$.

In principle, we can solve the random model by using the analytic formulae of the previous section. This is rather hard from the technical point of view. Detailed computations are in slow progress.

A trivial bound,

$$E(\beta) \geq 1, \tag{32}$$

comes from the inequality

$$N\mathrm{Tr}(M^4) \geq (\mathrm{Tr}(M^2))^2. \tag{33}$$

The same inequality tells us that $E = 1$ only if $\mathrm{Tr}(M^2)$.

However, some information can be obtained in a simple way: the annealed free energy is a lower bound to the quenched free energy and the entropy is non negative. In this way one finds that

$$E(\beta) \geq 1.06 \tag{34}$$

An approximation which is often quite good for these systems [15], consists in assuming that the entropy is given by

$$\max(S_a(\beta), 0), \tag{35}$$

where $S_a(\beta)$ is the annealed free energy. In the framework of this approximation the previous bound is exact.

4.3 A counterexample

Can we test if the previous conjecture consequently is valid for all large N? Alas, yes and the answer is negative, at least for some N.

Indeed, if $N = 2^k$ for some integer k, it is possible to find a matrix M whose elements are all equal to ± 1 and which is the square root of the identity. This proves that (at least for this sequence of N) the energy $E(\beta)$ is equal to 1 at low temperature. The same result may be proved for many other values of N, but it is unclear what happens for the generic N [17].

In a related model [9,13] it seems that quite different results are obtained depending on the arithmetic properties of N and that there is a way of taking the infinite N limit (i.e. a sequence of appropriate values of N going to infinity) such that the model becomes equal to the random model.

The possibility of more than one thermodynamic limit for the same model (depending on the sequence) is quite interesting, but it is clear that there exist at least one way in which the conjecture we have proposed is not valid.

4.4 A more refined conjecture

The physical origin of the difference in the equilibrium properties of the random and of the original model is related to the existence of "small" regions in the phase space of low energy. On the contrary, it is quite likely that the dynamical averages do coincide with those of the random model. These configurations $(C_{i,k})$ have such a low energy (they correspond in some sense to a crystal) that they cannot be reached in the natural evolution of a system arriving from the high temperature region due to the presence of infinitely high barriers.

At low temperature the equilibrium properties of this model have not been carefully studied. Here we follow the natural choice of generalizing the conjectures which have been done in a similar context on a simpler model which can be studied in a more effective way [9,13].

We suppose that the dynamic expectations values are the same as those in the random model (no contradiction is present). On the contrary the equilibrium expectations are definitely different from those of the random model. We suppose the system has a real first order transition (with latent heat) at some temperature T_C. At temperatures higher than T_C the two models are equivalent. At temperatures lower than T_C the phase equivalent to the random model still survive as a metastable phase.

We may think to stabilize this phase by adding a new term in the Hamiltonian, which has zero effect on the properties of the system in the *random* phase, but it strongly increases the energy of the *crystal* phase. An example

of such a term could be

$$Nr\theta(\sum_{i,k}((M_{i,k}-C_{i,k})^2-q)) \tag{36}$$

With an appropriate choice of C,r and q, it may be possible to kill the transition to the crystal phase without affecting the free energy in the other phase. We can thus modify the model adding a *small* perturbation to the Hamiltonian in such a way that its free energy coincides with the free energy of the random model at all temperatures.

5 Hard spheres on a sphere

In this section we show the existence of a physically interesting model which can be put in the form eq. (2). In this way we prove there is (at least) one non-trivial application of the considerations presented here.

Let us consider the following Hamiltonian

$$H=\sum_{i,k=N}V(|x_i-x_k|^2)=\sum_{i,k=N}W(x_i\cdot x_k), \tag{37}$$

where the x's are N D-dimensional vectors which are confined on the surface of a sphere [d] of radius R. The quantity V is the interparticle potential as function of the distance squared and

$$W(S)=V(2(R^2-S)) \tag{38}$$

is the potential as function of the scalar product.

The usual partition function of interacting particles in a $D-1$ dimensional flat space is obtained by sending $R\to\infty$, $N\to\infty$ at fixed density $\rho=NR^{-(D-1)}$. We may hope that for large dimensions the model will be soluble. Here we will study the case in which N, D and R go to infinity together (and also V depends on D).

Generally speaking we can write the partition function under the following form

$$\int d\mu(S)\exp\left(-\beta\sum_{i,k=N}W(S_{i,k})\right), \tag{39}$$

where S is an $N\times N$ matrix and

$$d\mu(S)=\prod_i(dx_i)\prod_{i,k}\delta(S_{i,k}-x_i\cdot x_k). \tag{40}$$

[d] We could also put the particle in cubic box, but we choose a sphere for technical reasons.

We can now substitute the hard constraint $|x_i| = R$ with a term in the probability distribution of the x proportional to $\exp(-\omega|x_i|^2/2)$, where ω is chosen in such a way that $\langle |x_i|^2 \rangle = R^2$.

The final model is slightly different from the previous one

$$Z = \int \prod_i (dx_i) \exp\left(-\beta H(x) - \omega/2 \sum_{i=1,N} |x_i|^2 \right). \tag{41}$$

In sufficiently high dimensions the two models (37 and 41) should be the same, the measure on x should become concentrated on a sphere of fixed radius which depends on ω.

Now the measure $\mu(S)$ becomes

$$\mu(S) = \int \prod_i (dx_i) \prod_{i,k} \delta(S_{i,k} - x_i \cdot x_k) \exp\left(-\omega/2 \sum_{i=1,N} |x_i|^2 \right), \tag{42}$$

which is invariant under $O(N)$ rotations as can be seen by an explicit computation.

The model is thus reduced to the original form eq.(2), we have just to obtain the correct scaling for large N. In other words we have to choose the dependance of the various terms on N is such a way to obtain the needed result.

It is possible to prove by an explicit computation that the *right* scaling is obtained when $N \to \infty$ and $D \to \infty$ together at fixed $\alpha \equiv N/D$. The potential V should also scale in an appropriate way in this limit.

After doing the appropriate computations one arrives to a model which can be explicitly solved in the low density case and that behaves in a quite similar model to the $M_{i,k} = \pm 1$. The details of the computation can be found on the original paper [14].

6 Conclusions

We can summarize these findings in the following way.

- The matrix model is soluble in the high temperature phase.
- In the high temperature phase the matrix model is equivalent to a random model which displays a replica breaking transition to a glassy phase.
- The random model describes the equilibrium and the dynamical properties of the model both in the high temperature phase and in the glassy phase.

- For some values of N there is a transition to a crystal phase at low temperature. The properties of the crystal phase cannot easily investigated: they depend on the arithmetic properties of N.

- It may be possible to destroy the crystal phase with a *small* perturbation which does not affect the properties in the other phase.

Some of these results are only conjectural, some have a more solid basis and some have been numerically verified in simpler models [9]. At the present moment we have a coherent picture of the behaviour of the model. It would be extremely interesting to find out if we can collect more evidence for its correctness.

References

1. E. Brézin, C. Itzykson, G. Parisi and J.B. Zuber, *Comm. Math. Phys.* **59**, 35 (1978).
2. C. Itzykson and J.B. Zuber, *J. Math. Phys.* **21**, 411 (1980).
3. E. Brézin and D.J. Gross, *Phys. Lett.* **97B**, 120 (1980).
4. H. Abarbanel and C. Itzykson, *Phys. Rev. Lett.* **23**, 53 (1969).
5. E. Marinari, G. Parisi, C. Rebbi, *Phys. Rev. Lett.* **47**, 1795 (1981).
6. M. Mézard, G. Parisi and M.A. Virasoro in *Spin glass theory and beyond*, World Scientific (Singapore 1987).
7. G. Parisi in *Field Theory, Disorder and Simulations*, World Scientific, (Singapore 1992).
8. D. Bessis, *Comm. Math. Phys.* **69**, 147 (1979).
9. E. Marinari, G. Parisi and F. Ritort, *J. Phys.* A (Math.Gen.) **27**, 7615 (1994); *J. Phys.* A (Math.Gen.) **27**, 7647 (1994).
10. L. Cugliandolo, J. Kurchan, G. Parisi and F. Ritort, *Phys. Rev. Lett.* **74**, 1012 (1995).
11. L. Cugliandolo, J. Kurchan, *Phys. Rev. Lett.* **71**, 173 (1993) and references therein.
12. S. Franz and G. Parisi, *J. Phys.* I (France) **5**, 1401 (1995).
13. I. Borsari, S. Graffi and F. Unguendoli, *J. Phys.* A (Math.Gen.), to appear and *Deterministic spin models with a glassy phase transition*, cond-mat 9605133.
14. L. Cugliandolo, J. Kurchan, E. Monasson and G. Parisi, *Math. Gen.* **29**, 1347 (1996).
15. B. Derrida, *Phys. Rev.* B **24**, 2613 (1981).
16. G. Parisi and F. Russo, work in progress.

17. M.R. Schroeder in *Number Theory in Science and Communication*, Springer-Verlag (Berlin) 1984.

RENORMALISATION GROUP APPROACH TO REACTION-DIFFUSION PROBLEMS

JOHN CARDY
All Souls College and
Department of Physics, Theoretical Physics, 1 Keble Road,
Oxford OX1 3NP, England

We describe how the methods of quantum field theory and the renormalisation group may be applied to classical stochastic particle systems which appear in non-equilibrium statistical mechanics. An emphasis is placed on the similarities and differences between these methods and more conventional applications of quantum field theory. Some simple applications are discussed.

1 Introduction

Reaction-diffusion problems are simple examples of non-equilibrium statistical systems. While it has long been recognised that the methods of quantum field theory extend far beyond their original domain of application in elementary particle physics, to, for example, many body quantum systems in condensed matter physics, and to the theory of both the statics and dynamics of critical behaviour in equilibrium statistical mechanics, it is less well appreciated that they also provide a powerful tool for analysing classical statistical systems far from equilibrium. This application is certainly not new,[1,2,3] but it is only relatively recently that the full formalism of the renormalisation group has been brought to bear,[4,5,6] in a very similar manner to that employed in equilibrium critical behaviour.

Reaction-diffusion problems are examples of classical stochastic particle systems. These particles are in general labelled by their species $A, B, \ldots$, which may be thought of as corresponding to different chemical reactants. Their dynamics consists of two elements: first, the particles diffuse according to some kind of Brownian motion, with diffusion constants $D_A, D_B, \ldots$. In a simulation, this might be described by random walks on a lattice. Second, they undergo reactions, for example $A + B \to C$, at prescribed rates λ_{AB}^C, whenever they meet. It is important that these processes are *diffusion-limited*, that is, the particles have to diffuse around before they find each other and react. In a chemical system this would mean that the reactants should not be stirred. In practice this may be realised by allowing the reactants to diffuse in a gel, or on a substrate. It is important to realise that these reactions are not usually reversible, and indeed the most interesting cases occur when they are completely irreversible.

The potential applications of these ideas to systems in chemistry, biology and physics are limited only by the imagination of the reader, and this is not the place to discuss them. Rather, we shall focus on some general features as illustrated by some rather simple examples. In general, from some random initial state, these systems will evolve in time towards a steady state described by some stationary probability distribution. Since the dynamics does not satisfy detailed balance, this is not in general a Gibbs measure. In some very simple cases, the steady state is trivial. For example, in the simplest system of all, with a single species of particle A undergoing the pure annihilation reaction $A + A \rightarrow$ inert (the inert particles not influencing the reaction further), the steady state has no particles, and hence no fluctuations since it is a classical vacuum. But it turns out that the *approach* to this steady is critical in the sense that it exhibits universal scaling behaviour, critical exponents, and so on. This may be understood, and the universal features computed, within the quantum field theoretic renormalisation group approach to be described.

A second class of critical phenomena corresponds to a non-equilibrium phase transition in the steady state, as some parameter of the dynamics is varied. For example, if to the annihilation reaction described above the branching process $A \rightarrow (m+1)A$ at rate σ is added, it turns out that under some circumstances there can be a transition at some finite value of σ to a non-trivial stationary state with a finite density of particles.[3,7,8]. The universality class of this transition turns out to depend on the parity of m. As with equilibrium critical behaviour, the symmetries of a system are seen to play an important role in determining its universality class.

The layout of this paper is as follows. In the next section we describe the general formalism and illustrate it with the annihilation reaction $A + A \rightarrow$ inert. We comment on the similarities and the differences with ordinary many-body quantum theory, and on the connections between this approach to non-equilibrium behaviour and others via the Fokker-Planck equation and the Langevin equation. In Sec. 3 we discuss the renormalisation group approach to the $A + A \rightarrow$ inert reaction in more detail, and finally in Sec. 4 we consider the more difficult problem of when the branching process is added.

2 Formalism

The dynamics of such a stochastic particle system is described by a master equation governing the time evolution of the probabilities $p(\alpha; t)$ that the system be in a given microstate α. For a system of particles on a lattice, for example, the αs might label the occupation number basis $(n_1, n_2, \ldots)$, corresponding to

having n_j particles at site j. The master equation takes the form

$$dp(\alpha;t)/dt = \sum_{\beta} R_{\beta\to\alpha}\, p(\beta;t) - \sum_{\beta} R_{\alpha\to\beta}\, p(\alpha;t). \tag{1}$$

Here $R_{\alpha\to\beta}$ is the rate for transitions from state α into β; for a reaction diffusion-problem these are determined by the diffusion constants and the reaction rates.

Such classical particle problems have two features in common with relativistic many-body problems which renders a 'second-quantised' formalism particularly useful. First, particle numbers are not, in general, conserved: they are created and destroyed by the dynamics. Second, and more important, the master equation and the Schrödinger equation share the properties of being linear and first order in time. It is therefore not surprising that such a formalism is similarly successful for these stochastic problems.

The first step is to construct a Fock space from annihilation and creation operators satisfying the usual commutation relationships $[a_i, a^\dagger{}_j] = \delta_{ij}$, and define the state vector

$$|\Psi(t)\rangle \equiv \sum_{\alpha} p(n_1, n_2, \ldots; t)\, a^\dagger{}_1^{n_1} a^\dagger{}_2^{n_2} \ldots |0\rangle. \tag{2}$$

Note that this is not normalised in the conventional manner, and what plays the role of a probability amplitude in quantum mechanics now is a probability. We have also chosen a bosonic representation, corresponding to the case when multiple occupancy of the sites is allowed. In simulations it is often more convenient to restrict the values of the n_j to 0 or 1, in which case a representation in terms of Pauli operators is more appropriate. This leads to quantum spin models rather than immediately to quantum field theories, which, in some one-dimensional cases, turn out to be integrable. In fact there is a some very elegant mathematics in this branch of the subject.[9] For example the quantum group symmetry of certain spin chains appears very naturally from this perspective. However, from the point of view of understanding the renormalisation group approach and generalising to noninteger dimensions, the bosonic formulation is more useful. In any case, as long as we are studying problems where the average particle density is low, the probability of multiple occupation should be small and there should be no difference between the physical results of the two approaches.

The statement is now that the master equation (1) is completely equivalent to a Schrödinger-like equation for the state vector

$$d|\Psi(t)\rangle/dt = -H|\Psi(t)\rangle, \tag{3}$$

where the 'hamiltonian' H is simply expressed in terms of the as and the $a^\dagger$s. For example, for the reaction-diffusion problem $A + A \rightarrow$ inert, one finds that

$$H = D \sum_{(i,j)} (a^\dagger{}_i - a^\dagger{}_j)(a_i - a_j) - \lambda \sum_i (a_i^2 - a^\dagger{}_i^2 a_i^2). \tag{4}$$

The simple hopping form of the first term (where the sum is over nearest neighbour pairs (i,j)), which corresponds to pure diffusion, is not surprising, but the second term may require some explanation. To understand its form, consider the simpler problem of the reaction at a single site. If $p(n;t)$ is probability of finding n particles at this site, the master equation is simply

$$dp(n;t)/dt = \lambda(n+2)(n+1)p(n+2;t) - \lambda n(n-1)p(n;t), \tag{5}$$

where the factors of $(n+2)(n+1)$ reflect the number of ways of choosing the pair of reacting particles. Defining $|\Psi\rangle = \sum_n p(n;t) a^{\dagger n} |0\rangle$ as above, its equation of motion is

$$\begin{aligned} d|\Psi\rangle/dt &= \lambda \sum_n \Big((n+2)(n+1)p(n+2) - n(n-1)p(n) \Big) a^{\dagger n} |0\rangle & (6) \\ &= \lambda \sum_n \left(a^2 p(n+2) a^{\dagger n+2} - a^{\dagger 2} a^2 p(n) a^{\dagger n} \right) |0\rangle & (7) \\ &= \lambda (a^2 - a^{\dagger 2} a^2) \sum_n p(n) a^{\dagger n} |0\rangle. & (8) \end{aligned}$$

The second term in the reaction part of (4) therefore corresponds to the second term in the master equation (1), and is required by the conservation of probability.

From the lattice hamiltonian (4) we may, if interested in long wavelength properties, proceed to the formal continuum limit

$$H = \int \left[D(\nabla a^\dagger)(\nabla a) - \lambda (a^2 - a^{\dagger 2} a^2) \right] d^d x, \tag{9}$$

and thence to a representation as a path integral over fields $a(x,t)$ and $a^*(x,t)$ with a weight $\exp(-S[a,a^*])$, with an action

$$S \equiv \int \left[a^* \partial_t a + D(\nabla a^*)(\nabla a) - \lambda (a^2 - a^{*2} a^2) \right] dt\, d^d x. \tag{10}$$

2.1 Differences from quantum mechanics

There are two immediately apparent differences from ordinary quantum field theory: first, there is no factor of i in the Schrödinger equation (3) – but this is familiar from euclidean formulations of conventional quantum theories; second, the hamiltonian is not hermitian. In many cases it will turn out that, nevertheless, its eigenvalues are real. (Complex eigenvalues correspond to oscillating states which are known to occur in some chemical reactions.) However, the most important difference is one of interpretation: expectation values of observables $\mathcal{O}$ are not given by $\langle\Psi|\mathcal{O}|\Psi\rangle$, since this would be bilinear, rather than linear, in the probabilities $p(\alpha;t)$. Instead, for an observable which is diagonal in the occupation number basis, its expectation value is of course

$$\overline{\mathcal{O}} = \sum_{\{n_j\}} \mathcal{O}(\{n_j\})p(\{n_j\};t), \tag{11}$$

and it is straightforward to show that this may be expressed as

$$\overline{\mathcal{O}} = \langle 0|e^{\sum_j a_j}\mathcal{O}\, e^{-Ht}|\Psi(0)\rangle, \tag{12}$$

since the state $\langle 0|e^{\sum_j a_j}$ is a left eigenstate of all the $a^\dagger{}_j$, with unit eigenvalue.

Conservation of probability then requires that $\langle 0|e^{\sum_j a_j}H = 0$. This is equivalent to the requirement that H should formally vanish when every $a^\dagger{}_j$ is set to unity. The appearance of the state $\langle 0|e^{\sum_j a_j}$ may complicate some of the subsequent calculations, since the interaction part of the hamiltonian is not normal ordered with respect to it, and therefore the usual formalism of time-dependent perturbation theory and Wick's theorem do not immediately apply. This problem may be avoided by first commuting the factor of $e^{\sum_j a_j}$ through the operators $\mathcal{O}$ and H in (12). This has the effect of shifting $a^\dagger \to 1 + a^\dagger$, since $e^a a^\dagger = (1 + a^\dagger)e^a$. The factor $e^{\sum_j a_j}$ acting on the initial state $|\Psi(0)\rangle$ is usually something simple, and the operators are now normal ordered.

Note that such a shift is convenient only if we are interested in what (in the language of particle physics) may be termed 'inclusive' probabilities, for example the expectation value of the local density $\bar{n}_j = \overline{a^\dagger{}_j a_j}$. After the shift, we see that in fact $\bar{n}_j = \langle a_j\rangle$, where $\langle\cdot\rangle$ denotes the usual QFT expectation value. For so-called exclusive quantities, for example the probability $\overline{\delta_{n,1}\prod_{j\neq i}\delta_{n_j 0}}$ that there is only one particle in the system, at site i, the factor $e^{\sum_j a_j}$ simply reduces to a_j in the correlation function, and no shift is necessary.

2.2 Relation to other formalisms

Of course, there are several other important ways of formulating stochastic processes, through either the Langevin equation or its related Fokker-Planck equation. It is interesting to see how these emerge in the present formalism for the simple example under consideration. If we make the shift $a^* = 1 + \bar{a}$ in the path integral, we find an action

$$S[a, \bar{a}] = \int \left[\bar{a}\partial_t a + D(\nabla\bar{a})(\nabla a) + 2\lambda\bar{a}a^2 + \lambda\bar{a}^2a^2\right]dtd^dx. \tag{13}$$

The non-linear terms in $\bar{a}$ may be disentangled in terms of a Gaussian transformation

$$e^{-\lambda\bar{a}^2a^2} = \int e^{-\eta\bar{a}}P(\eta)d\eta, \tag{14}$$

where $P(\eta)$ is a suitable Gaussian distribution. The path integral over $\bar{a}$ may now be performed, yielding a functional delta function equivalent to the equation

$$\partial_t a = D\nabla^2 a - 2\lambda a^2 + \eta(x,t). \tag{15}$$

Neglecting the last term, this is just the so-called rate equation which one might write down as a first approximation to the equation of motion for the density (note that we have argued above that the expectation value of a is the density.) The rate equation approximation assumes that the annihilation rate, which is proportional to the probability of finding two particles at the same point, is simply given by the square of the density. This approximation clearly neglects the correlations between the particles, and is on the same footing as the mean field approximation in equilibrium critical behaviour. In this sense (15) looks very much like a Langevin equation, with a noise term η. However, such equations are usually derived from the master equation through some kind of approximate coarse-graining, and the exact form of the noise term is often unclear, especially when the dynamics does not constrain this through detailed balance. By contrast, the correlations of the noise here are completely explicit

$$\langle\eta(x,t)\eta(x',t')\rangle = -\lambda a^2\delta(x-x')\delta(t-t'). \tag{16}$$

That the noise should depend on a is expected, since there can be no noise when the density is zero. But the minus sign is surprising. It implies that the noise η is pure imaginary, so that the solution of (15) is complex!

This curious result may be traced to the fact that, although the 'quantum mechanical' average $\langle a\rangle$ is the mean density $\bar{n}$, this is not true of higher moments. For example, the mean square density $\overline{n^2}$ is given by the average of

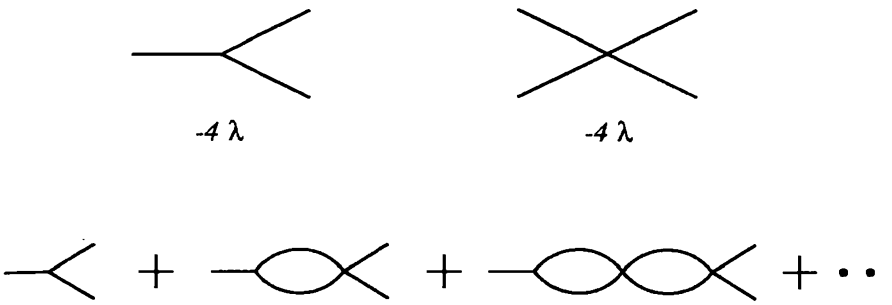


Figure 1: Vertices and renormalisation for $A + A \rightarrow$ inert.

$(a^\dagger a)^2 = {a^\dagger}^2 a^2 + a^\dagger a$. The operators $a^\dagger$ give unity acting to the left on the state $\langle 0|e^a$, so that in fact $\overline{n^2} = \langle a^2 \rangle + \langle a \rangle$. In general, one may show that, if a has a Gaussian distribution, as it would in the case of pure diffusion, then the density n would have a pure Poisson distribution. This is to be expected, since a simple random walk will have Poissonian statistics. The effect of the reactions is to modify this. In a sense the 'noise' η in (15) represents only that part of the physical noise which originates in the discreteness of the reaction process. Since, however, this cannot truly be disentangled from the diffusion noise, there is no need for its correlations to be positive.

The hamiltonian approach we have described above should be distinguished from another based on the Fokker-Planck equation. The latter begins from a coarse-grained Langevin description of the problem, and describes the time evolution of the probability distribution of the solution of this equation. Like the master equation, it is linear and first order in time, so may be cast in a hamiltonian formalism (in this case more usually called the liouvillean.) But the Fokker-Planck equation describes the diffusion in phase space and cannot easily accommodate processes where particles are continually being created and destroyed. It is therefore less useful for these types of problem.

3 Renormalisation group analysis

The field theory described by the action (13) is extremely simple. The bare propagator $(-i\omega + Dq^2)^{-1}$ is simply the Green function for the diffusion equation, and it may be represented by a directed line moving forward in time (conventionally from right to left.) The vertices are shown in Fig. 1. Immediately we see that, since the number of particles in a given intermediate state cannot increase as we move from right to left through a diagram, there can be no

loop corrections to the propagator, hence no wave function renormalisation for the fields a or $\bar{a}$. The only non-trivial renormalisation is that of the coupling constant λ, which may be seen through the loop corrections either to the vertex $\Gamma^{2,1}$ (shown in Fig. 1) or to $\Gamma^{2,2}$. (The fact that these renormalise in the same way is a consequence of probability conservation, which relates the untruncated Green functions $\langle 0|a(t,x)a^{\dagger^2}(0,0)|0\rangle$ and $\langle 0|a^2(t,x)a^{\dagger^2}(0,0)|0\rangle$. It is interesting to note that probability conservation is not expressed through a Noether symmetry in this formalism.) The diagrams in Fig. 1 correspond to a simple inclusion-exclusion argument for the probability of finding two particles at the same point given that they have not reacted in the past. They may be simply summed, with the result that the one-loop renormalisation group beta function is exact:

$$\beta(g_R) = -\epsilon g_R + b g_R^2, \tag{17}$$

where g_R is the dimensionless renormalised coupling, b is a positive constant, and $\epsilon = 2 - d$.

For $d > 2$, then, g_R is irrelevant in the infrared, and the rate equation (15) with no noise term is asymptotically valid, while for $d < 2$ it flows to a non-trivial fixed point $g^* = O(\epsilon)$. The consequences of this may be explored by writing down and solving the Callan-Symanzik equation in the usual way. Consider, for example, the mean density $n(t)$, which depends in principle on the initial density n_0 and the rate λ, expressed through g_R and the normalisation scale t_0. (We work in units where the diffusion constant $D = 1$.) Then

$$n(t, n_0, g_R, t_0) = (t/t_0)^{-d/2}\, n(t_0, n_0(t/t_0)^{d/2}, \tilde{g}_R(t/t_0), t_0) \tag{18}$$

where the running coupling $\tilde{g}_R \to g^*$ as $t/t_0 \to \infty$. The simple exponent in the prefactor on the right hand side reflects the lack of any anomalous dimensions for the density. The right hand side may now be evaluated as a power series in g^* (the lowest order being simply the mean field result), which is converted into a power series in ϵ. Term by term, one may show that the result is in fact independent of the initial density n_0, so that in fact the whole expression is a universal function of ϵ only. The result is $n(t) \sim A/t^{d/2}$ where[4]

$$A = \frac{1}{4\pi\epsilon} + \frac{2\ln 8\pi - 5}{16\pi} + O(\epsilon). \tag{19}$$

Universal forms for correlation functions may be derived in a similar manner. The result is that the whole probability distribution for fluctuations becomes universal in the late time regime.

4 Branching and annihilating random walks

A more interesting type of critical behaviour emerges if to the annihilation reaction $A+A \rightarrow$ inert we add the branching process $A \rightarrow (m+1)A$, with rate σ. Here m is a positive integer. In the rate equation approximation, the mean density now satisfies

$$dn/dt = \sigma m n - \lambda n^2, \tag{20}$$

which would suggest that, for any $\sigma > 0$, the steady state has a non-zero density of particles (it is 'active' in the language of catalysis.) In fact, simulations suggest that this is only true in sufficiently high dimensions $d > 2$. For $d \leq 2$ the fluctuations need to be taken into account, and this may be done using the formalism described above.

In one dimension, with $m = 2$, this model may be interpreted as a dynamic Ising model, where the particles play the role of domain walls.[10] In that case the processes of diffusion and annihilation are generated by single spin flips (these are assumed to occur at zero temperature so that pair creation of domain walls is suppressed), and the branching process $A \rightarrow 3A$ is associated with spin exchange (assumed to occur at infinite effective temperature). Since this model has two 'temperatures', it does not satisfy detailed balance and the stationary state is not Gibbsian. For that reason the model may undergo a nontrivial ordering transition, even in one dimension.

The additional term in the hamiltonian has the form

$$H_b = \sigma \int [a^{\dagger m+1} a - a^\dagger a] d^d x. \tag{21}$$

Note that there is a difference depending on the parity of m. If it is even, then the number of particles is locally conserved modulo 2 by both the annihilation and the branching process, while for m odd the latter violates this. For m even this is manifested in the formal symmetry of the hamiltonian under $(a, a^\dagger) \rightarrow (-a, -a^\dagger)$. Note that if we make the shift $a^\dagger \rightarrow 1 + \tilde{a}$ in order to develop the perturbative expansion for 'inclusive' processes, this symmetry becomes hidden. If we further make an expansion of the hamiltonian in powers of $\tilde{a}$, and drop higher orders on the grounds of irrelevance by power counting, the symmetry is completely lost. This is evidently a dangerous thing to do, since it is well known that symmetries play a very important role in influencing universality classes of critical behaviour. Fortunately the renormalisation group behaviour of the theory should be independent of which type of correlation functions we choose to study, and therefore may be computed in the unshifted theory where the symmetry is manifest.

4.1 Case of even m

The first question to be addressed[11] is whether the branching rate σ is relevant at the pure annihilation fixed point, that is, in the beta function $\beta_\sigma = -y\sigma + O(\sigma^2)$, is $y > 0$? If so, then as soon as $\sigma \neq 0$ it will flow away in the infrared into what is presumably the fixed point controlling the active phase. For $d > 2$, the pure annihilation reaction is controlled by the gaussian fixed point, so simple power counting suffices. This yields $y = 2$, consistent with the simulation results. For $d < 2$, the annihilation fixed point is accessible within the $\epsilon = 2 - d$ expansion, and so y may be computed only perturbatively. The result is that[11]

$$y = 2 - \tfrac{1}{2}m(m+1)\epsilon + O(\epsilon^2). \tag{22}$$

Notice that the $O(\epsilon)$ corrections are large in $d = 1$, but no conclusion may be drawn from this since the higher terms have been neglected.

Fortunately it is possible to compute y exactly in $d = 1$.[11] There are several ways of doing this, but the simplest is to realise that in this limit it becomes a kind of a free fermion problem. This is because going to the annihilation fixed point $g_R \to g^*$ corresponds to taking the limit of the bare coupling $\lambda \to \infty$. In that case the term $\lambda a^{\dagger 2} a^2$ in (9) corresponds to an infinite hard core repulsion, so that the particles behave in one dimension like free fermions, at least in those periods of time evolution during which the other terms in the hamiltonian do not play a role. (For this reason the problem is not completely equivalent to free fermions.) In that limit, it does not make sense to create the new particles at the same lattice site. The best one can do is to distribute them between m neighbouring sites, so that the corresponding term in the lattice hamiltonian has the form

$$\sigma \sum_j \prod_{i=j-\frac{m}{2}}^{j+\frac{m}{2}} c^\dagger{}_i c_j, \tag{23}$$

where $c^\dagger{}_i$ and c_j are now anticommuting operators. In the continuum limit, we may make a Taylor expansion of each $c^\dagger{}_i$ about $i = j$, in powers of the lattice spacing b. The lowest surviving term has the form

$$\tilde{\sigma} \int c^\dagger (\partial c^\dagger)(\partial^2 c^\dagger) \ldots (\partial^m c^\dagger)\, c\, dx, \tag{24}$$

where $\tilde{\sigma} = \sigma b^{m(m+1)/2}$ is now the effective expansion parameter in the continuum limit. This extra factor modifies the dimensional analysis, which then implies that

$$y = 2 - \tfrac{1}{2}m(m+1). \tag{25}$$

Thus the $O(\epsilon)$ result in (22) appears to be exact for $d = 1$. We have no simple explanation of this, as we have also computed explicitly the $O(\epsilon^2)$ terms and find them to be non-zero.

However, this does imply that the branching is irrelevant at the pure annihilation fixed point for $d = 1$, and hence the infrared or late time behaviour for sufficiently small σ should be that of the pure annihilation process, with finitely renormalised parameters (for example, the diffusion constant.) It may also be shown that, even if the original branching process does not allow for $m = 2$ processes, these will inevitably be generated under renormalisation, and, since this coupling is the most relevant, it controls the late time behaviour. Physically both these effects may be understood as follows. In pure annihilation, the surviving particles become strongly anticorrelated in space. This is because each sweeps out a region around itself: for $d < 2$, every test particle placed within that region has probability one of eventually annihilating with it. (For $d > 2$ the test particle may escape.) When a small branching rate is turned on, the single particles occasionally branch into bunches of $3, 5, \ldots$ particles but these stay close together, and almost always annihilate with their siblings before visiting other bunches. The effect is therefore of diffusing bunches, which behave in many ways like single particles with a reduced diffusion constant. Clearly even if branchings only with $m > 2$ are allowed, the pair annihilation process will generate an effective $m = 2$ branching rate.

For larger values of the branching rate σ there should be a transition to the active state, which should correspond to some nontrivial fixed point of the renormalisation group. But it seems to be very difficult to analyse this within any perturbative renormalisation group scheme. This is because the problem has two critical dimensionalities: $d = 2$ associated with the nontrivial nature of the annihilation, and $d \approx \frac{4}{3}$ where the value of y changes sign, and therefore no systematic ϵ-expansion is possible. So far we have not been able to find another small parameter, and the best we can do is a truncated loop expansion in fixed number of dimensions. This leads to the expected fixed point, but the estimated values for the critical exponents are far from those measured in simulations.

4.2 Case of odd m

Although the above analysis might suggest that for $m = 1$ the branching rate is relevant even when $d = 1$, so that there is no nontrivial transition, this is not the case, since now there is no conservation law modulo 2, and the process $A \to 0$ is immediately generated under renormalisation. This has eigenvalue 2 and corresponds to the generation of a mass gap in the theory. In fact

one may show that for small branching rates the mean density should decay exponentially to zero. This conclusion is valid even when $d = 2$: although in this case the annihilation rate which generates the new term is irrelevant, it is only logarithmically vanishing, and meanwhile the rate for the process $A \to 0$ is growing under the renormalisation group. Once again, for sufficiently large σ, there should be a transition to the active state. In this case it is rather easier to analyse. On including the effective term $\Delta(a - a^\dagger a)$ in the hamiltonian, corresponding to $A \to 0$, and making the shift $a^\dagger \to 1 + \bar{a}$, one finds an interaction hamiltonian of the form

$$H_{int} = \int [\delta \bar{a} a + \mu_1 \bar{a} a^2 - \mu_2 \bar{a}^2 a + \cdots] d^d x, \tag{26}$$

where μ_1 and μ_2 are positive constants, and δ may change sign (as it does at the mean field transition.) The omitted terms are of higher order, and their neglect is, this time, justified, since there is no symmetry relating them to the lower order terms. This is a well-known theory[12] which describes the universality class known as directed percolation (DP), although it was first studied by particle physicists in the context of reggeon field theory. Generically, any dynamical phase transition from an inactive state, with no noise, to an active one, is in the DP universality class, and this has been verified for a number of models. The branching and annihilating random walks for m even and $d = 1$ are therefore an interesting exception to this general rule. They evade it because they possess an additional conservation law. This is of course quite a familiar idea from equilibrium critical behaviour.

5 Conclusions

These simple examples I hope illustrate the point that quantum field theory still has many unexplored applications, which are not limited to quantum systems nor to equilibrium critical behaviour. Perhaps we are not yet at the stage when the mathematical beauty of such applications is apparent, but the richness of the subject is such that I believe that this may well emerge in the years to come.

Acknowledgements

This work was carried out under partial support of EPSRC Grant GR/J78044. I gratefully acknowledge my collaborators B. P. Lee, M. Howard, and U. Tauber who have all contributed to the research described above.

References

1. M Doi, J. Phys. A **9**, 1465 (1976)
2. L Peliti, J. Physique **46**, 1469 (1985)
3. P Grassberger, F Krause and T von der Twer, J. Phys. A **17**, L105 (1984)
4. B P Lee, J. Phys. A **27**, 2633 (1994)
5. B P Lee and J L Cardy, J. Stat. Phys. **80**, 971 (1995)
6. M Howard and J L Cardy, J. Phys. A **28**, 3599 (1995)
7. H Takayasu and A Yu Tretyakov, Phys. Rev. Lett. **68**, 3060 (1992)
8. I Jensen, Phys. Rev. E **50**, 3623 (1994)
9. F C Alcaraz, M Droz, M Henkel and V Rittenberg, Ann. Phys. **230**, 250 (1994)
10. N Menyhárd, J. Phys. A **27**, 6139 (1994)
11. J L Cardy and U Tauber, in preparation.
12. J L Cardy and R L Sugar, J. Phys. A **13**, L423 (1980)

ZERO TEMPERATURE GLAUBER DYNAMICS OF THE $1d$ POTTS MODEL

B. DERRIDA
Laboratoire de Physique Statistique, ENS,
24 rue Lhomond, 75005 Paris, France

For the zero temperature Glauber dynamics of the one dimensional Potts model, the analogy with a voter model can be used to write the exact expressions of the correlation functions between an arbitrary number of spins. This allowes one to obtain the exact distribution of domain sizes in the long time limit and to calculate exactly the exponent characteristic of the decay of the density of persistent spins.

This paper is dedicated to Claude Itzykson who was for so many years a friend as well as a colleague. Along with many others in Saclay, I have benefited greatly from his enthusiasm, his generosity and his experience. Beyond the outstanding achievements he left behind, Claude will also remain for us an example of a scientist driven by the purest motivations.

1 Introduction

The $1d$ Ising or Potts model evolving according to Glauber dynamics at zero temperature is one of the simplest systems which exhibits coarsening [1]. Although some quantities such as the pair correlations can be calculated easily using the analogy with voter models [2,3,4,5] and non-intersecting random walks, several properties characteristic of the long time regime of this coarsening phenomenon (like for example the distribution of domain sizes) have remained for a long time more difficult to obtain.

The goal of the present lecture is to review a few exact results obtained recently [6,7,8] in collaboration with Vincent Hakim, Vincent Pasquier and Reuven Zeitak. These results include both the exact expression of the distribution of domain sizes in the long time limit and of the exponent θ characterising the decay of the density of persistent spins (i.e. of the spins which never flip up to time t). They are all derived from exact expressions of correlation functions for an arbitrary large number of spins.

2 Distribution of domain sizes

The system considered throughout this work is a q-state Potts model evolving according to a continuous time zero temperature Glauber dynamics. Initially each spin $\sigma_i(0)$ at site i of an inifinite one dimensional lattice is assigned one of the q possible values at random. According to Glauber dynamics at zero

temperature, each time a spin of a $1d$ chain is updated, it takes the common value of its two neighbors if these two neighbors are in the same state and it chooses at random the value of one of its two neighbors, if these neighbors are in different states. In other words, during every infinitesimal time interval Δt, each spin in the system is updated according to

$$\sigma_i(t+\Delta t) = \begin{cases} \sigma_i(t) & \text{with probability } 1-2\Delta t \\ \sigma_{i-1}(t) & \text{with probability } \Delta t \\ \sigma_{i+1}(t) & \text{with probability } \Delta t \end{cases} \tag{1}$$

Clearly, the dynamics tends to align neighboring spins and to eliminate domain walls giving rise to a coarsening phenomenon with fewer and fewer domains of increasing sizes. From (1) it is easy to see that the probability $A_{x,y}(t)$ that two sites at positions x and y are in the same state at time t satisfies the following evolution equation for $x < y$

$$\frac{dA_{x,y}(t)}{dt} = A_{x+1,y}(t) + A_{x-1,y}(t) + A_{x,y+1}(t) + A_{x,y-1}(t) - 4\,A_{x,y}(t) \tag{2}$$

and $A_{x,x}(t) = 1$. This equation together with the initial condition

$$A_{x,y}(0) = \frac{1}{q} + \frac{q-1}{q}\delta_{x,y}$$

determines the two point correlation function $A_{x,y}(t)$ at any later time and the solution is for $x \le y$

$$A_{x,y}(t) = 1 - \frac{q-1}{q}\frac{1}{2\pi}\int_{-\pi}^{\pi} d\theta\, \frac{\sin\theta \sin(y-x)\theta}{1-\cos\theta}\, e^{-4(1-\cos\theta)t}\,. \tag{3}$$

This can be rewritten as

$$A_{x,y}(t) = 1 - \frac{q-1}{q} c_{x,y}(t) \tag{4}$$

where $c_{x,y}(t)$ is the probability that two walkers starting at positions x and y do not meet up to time t (during every infinitesimal time interval Δt, each walker hops to its right with probability Δt, to its left with probability Δt and does move with probability $1-2\Delta t$). These $c_{x,y}(t)$ satisfy the same evolution equation as the $A_{x,y}(t)$

$$\frac{dc_{x,y}(t)}{dt} = c_{x+1,y}(t) + c_{x-1,y}(t) + c_{x,y+1}(t) + c_{x,y-1}(t) - 4\,c_{x,y}(t) \tag{5}$$

with the obvious boundary conditions

$$c_{x,x}(t) = 0 \quad ; \quad c_{x,y}(0) = 1 - \delta_{x,y} \ .$$

The solution is given for $x \leq y$ by

$$c_{x,y} \equiv c_{x,y}(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\theta \ \frac{\sin\theta \sin(y-x)\theta}{1-\cos\theta} \ e^{-4(1-\cos\theta)t} \tag{6}$$

The key to the calculation of the distribution of domain sizes is that (4) can be generalised to give arbitrary equal time correlation functions in terms of the matrix $c_{x,y}$. For example, if $x_1 \leq x_2 \leq x_3 \leq x_4$, the probability $A_{x_1,x_2,x_3}(t)$ that $\sigma_{x_1}(t) = \sigma_{x_2}(t) = \sigma_{x_3}(t)$, and the probability $A_{x_1,x_2,x_3,x_4}(t)$ that $\sigma_{x_1}(t) = \sigma_{x_2}(t) = \sigma_{x_3}(t) = \sigma_{x_4}(t)$, are given by [8]

$$A_{x_1,x_2,x_3}(t) = 1 - \frac{q-1}{q^2}[c_{x_1,x_2} + c_{x_2,x_3}] - \left(\frac{q-1}{q}\right)^2 c_{x_1,x_3} \tag{7}$$

and

$$A_{x_1,x_2,x_3,x_4}(t) = 1 - \frac{q-1}{q^2}[c_{x_1,x_2} + c_{x_2,x_3} + c_{x_3,x_4}] - \left(\frac{q-1}{q}\right)^2 c_{x_1,x_4} + \frac{(q-1)^2}{q^3} c^{(2)}_{x_1,x_2,x_3,x_4} \tag{8}$$

where $c^{(2)}_{x_1,x_2,x_3,x_4}$ is the probability that there is no intersection up to time t between four random walkers starting at positions $x_1 < x_2 < x_3 < x_4$. Because one can represent non-intersecting random walkers by free fermions, $c^{(2)}_{x_1,x_2,x_3,x_4}$ is a Pfaffian and has the following expression in terms of the matrix $c_{x,y}$

$$c^{(2)}_{x_1,x_2,x_3,x_4} = c_{x_1,x_2} c_{x_3,x_4} + c_{x_1,x_4} c_{x_2,x_3} - c_{x_1,x_3} c_{x_2,x_4} \ . \tag{9}$$

More generally, for $x_1 \leq x_2 \leq \ldots \leq x_n$, the probability $A_{x_1,x_2 \ldots x_n}(t)$ that $\sigma_{x_1}(t) = \sigma_{x_2}(t) = \ldots = \sigma_{x_n}(t)$ is given by [8]

$$\begin{aligned} A_{x_1,x_2 \ldots x_n}(t) = 1 &- \mu \sum_{i=1}^{n-1} c_{x_i,x_{i+1}} + \mu^2 \sum_{i<j} c^{(2)}_{x_i,x_{i+1},x_j,x_{j+1}} \\ &- \mu^3 \sum_{i<j<k} c^{(3)}_{x_i,x_{i+1},x_j,x_{j+1},x_k,x_{k+1}} + \cdots \\ &- \lambda \left\{ \mu c_{x_1,x_n} - \mu^2 \sum_i c^{(2)}_{x_1,x_i,x_{i+1},x_n} + \mu^3 \sum_{i<j} c^{(3)}_{x_1,x_i,x_{i+1},x_j,x_{j+1},x_n} - \cdots \right\} \end{aligned} \tag{10}$$

where

$$\lambda = q - 1 \tag{11}$$

$$\mu = \frac{q-1}{q^2} \tag{12}$$

and $c^{(k)}_{x_1,x_2,\ldots x_{2k}}$ is the probability that there is no intersection up to time t among $2k$ walkers starting at positions $x_1 \leq x_2 \leq \ldots \leq x_{2k}$. This probability $c^{(k)}_{x_1,x_2,\ldots x_{2k}}$ is a Pfaffian given by

$$c^{(k)}_{x_1,x_2,\ldots,x_{2k-1},x_{2k}} = \frac{1}{2^k\ k!} \sum_{\sigma} \epsilon(\sigma)\ c_{x_{\sigma(1)},x_{\sigma(2)}} \cdots c_{x_{\sigma(2k-1)},x_{\sigma(2k)}} \tag{13}$$

where the sum runs over all the permutations σ of the indices $\{1,2,\cdots,2k\}$, $\epsilon(\sigma)$ is the signature of the permutation σ and the matrix c_{x_i,x_j} is given by (6) for $x_i < x_j$ and is antisymmetrized

$$c_{x_i,x_j} = -\ c_{x_j,x_i}\ < 0 \quad \text{when} \quad x_i > x_j\ .$$

One can easily check the validity of (7,8,9,10,13) by writing the evolution equations similar to (2,5) that $A_{x_1,x_2,\ldots,x_n}(t)$ or $c^{(k)}_{x_1,x_2,\ldots,x_{2k-1},x_{2k}}$ should satisfy and by verifying that they are satisfied with the right boundary conditions. One can also derive these expressions from the fact that non-intersecting random walkers in $1d$ can be described as free fermions[7,8].

Remark: other correlation functions than the probability that n spins are in the same state can be calculated from the knowledge of the matrix c_{x_i,x_j}. In fact one can express arbitrary correlation functions between n spins in term of the pair correlations only. For example in the Ising case ($q = 2$), if one defines Ising spins with $S_i = +1$ when $\sigma_i = 1$ and $S_i = -1$ when $\sigma_i = 2$, one has

$$\langle S_x S_y \rangle = 1 - c_{x,y}$$

and one can show that all the higher correlation functions can be expressed in terms of the two point function $\langle S_x S_y \rangle$: for $x_1 < x_2 < x_3 < x_4$,

$$\langle S_{x_1} S_{x_2} S_{x_3} S_{x_4} \rangle = \langle S_{x_1} S_{x_2} \rangle \langle S_{x_3} S_{x_4} \rangle + \langle S_{x_1} S_{x_4} \rangle \langle S_{x_2} S_{x_3} \rangle - \langle S_{x_1} S_{x_3} \rangle \langle S_{x_2} S_{x_4} \rangle$$

and more generally for $x_1 < x_2 < \ldots < x_{2n}$, the $2n$-point function $\langle S_{x_1} S_{x_2} \ldots S_{x_{2n}} \rangle$ is a Pfaffian (13).

In the long time limit, the expression (6) becomes a scaling form of the reduced variable $(y-x)/\sqrt{t}$

$$c_{x,y}(t) = F\left(\frac{y-x}{\sqrt{t}}\right) \tag{14}$$

where

$$F(z) = \frac{2}{\sqrt{\pi}} \int_0^{z/\sqrt{8}} e^{-u^2} du \tag{15}$$

and consequently all the correlation functions become scaling functions of the distances divided by $\sqrt{t}$.

In particular, for large t, the probability $A_{x,y}(t)$ that $\sigma_x(t) = \sigma_y(t)$ becomes

$$A_{x,y}(t) = 1 - \frac{q-1}{q} \frac{2}{\sqrt{\pi}} \int_0^{|y-x|/\sqrt{8}\sqrt{t}} e^{-u^2} du \tag{16}$$

showing that there is a characteristic length in the system increasing like $t^{1/2}$. The density of domain walls (which is also the inverse of the average size $\langle l \rangle$ of the domains) is of course $1 - A_{x,x+1}(t)$ and from (16), one finds that for large t

$$\langle l \rangle = \frac{1}{1 - A_{x,x+1}(t)} \simeq \frac{q}{q-1} \sqrt{2\pi}\sqrt{t} \tag{17}$$

The whole distribution of domain sizes is more difficult to obtain as it requires the knowledge of correlations between an arbitrary number of sites.

Expression (10) is exact and can be used to calculate the probability ψ_n that n consecutive spins are in the same state (i.e. that n consecutive spins are in the same domain). Then one can obtain the distribution $p(l)$ of domain sizes via

$$\psi_n = \sum_{l \geq n} p(l) \frac{l-n+1}{\langle l \rangle}$$

or equivalently

$$p(l) = \langle l \rangle \left[\psi_l - 2\psi_{l+1} + \psi_{l+2} \right]$$

In the long time limit, the ψ_n become scaling functions of the ratio $n/\sqrt{t}$. Thus the distribution $p(l)$ of domain sizes scales also in the same way and takes the form

$$p(l) = \frac{1}{\langle l \rangle} g_q \left(\frac{l}{\langle l \rangle} \right) \tag{18}$$

where the function $g_q(x)$ will be normalised such that

$$\int_0^\infty g_q(x) x dx = \int_0^\infty g_q(x) dx = 1$$

For $q \to \infty$, one can see from (10) that $\psi_n = 1 - c_{1,n}$ and this allowes one to obtain the following exact expression of $g_\infty(x)$

$$g_\infty(x) = \frac{\pi}{2} x \, e^{-x^2\pi/4} . \tag{19}$$

For other values of q, as n increeases, there are more and more terms in (10) which contribute to the ψ_n and $g_q(x)$ becomes an infinite series in powers of μ. However, when (14) and (15) are used into (10), one can obtain the expansion of ψ_n in powers on $n/\sqrt{t}$ and this leads[8] to the small x expansion of $g_q(x)$

$$\begin{aligned} g_q(x) &= \frac{\pi}{2}\frac{q}{q-1}x - \frac{\pi^2}{8}\frac{q^3}{(q-1)^3}x^3 + \frac{\pi^2}{24}\frac{q^3}{(q-1)^4}x^4 + \frac{\pi^3}{64}\frac{q^5}{(q-1)^5}x^5 \\ &- \frac{\pi^3}{120}\frac{q^5}{(q-1)^6}x^6 - \frac{\pi^4}{768}\frac{q^7}{(q-1)^7}x^7 + \frac{\pi^4}{26880}\frac{q^6(3+23q)}{(q-1)^8}x^8 \\ &+ \frac{\pi^5}{12288}\frac{q^9}{(q-1)^9}x^9 + O(x^{10}) \ . \end{aligned} \tag{20}$$

One can also write (10) as a determinant[7,8] and the large x behavior of $g_q(x)$ can be obtained from the theory of Toeplitz determinants. One finds for large x that

$$g_q(x) \simeq \exp[-A(q)x + B(q)] \tag{21}$$

where the coefficients $A(q)$ and $B(q)$ are given in terms of $\mu = \frac{q-1}{q^2}$

- For $q \leq 2$

$$A(q) = \frac{1}{4}\ \frac{q}{q-1}\sum_{n=1}^{\infty}\frac{(4\mu)^n}{n^{3/2}} \tag{22}$$

$$B(q) - 2\log A(q) = \log q - \frac{q-1}{q^2} + \frac{1}{4\pi}\sum_{n=2}^{\infty}\frac{(4\mu)^n}{n}\left\{-\pi + \sum_{p=1}^{n-1}\frac{1}{\sqrt{p(n-p)}}\right\} \ . \tag{23}$$

- For $q \geq 2$

$$A(q) = \frac{q\sqrt{\pi}}{q-1}\sqrt{-\log 4\mu} + \frac{1}{4}\frac{q}{q-1}\sum_{n=1}^{\infty}\frac{(4\mu)^n}{n^{3/2}} \tag{24}$$

$$\begin{aligned} B(q) - 2\log A(q) &= \log q - \frac{q-1}{q^2} + \frac{1}{4\pi}\sum_{n=2}^{\infty}\frac{(4\mu)^n}{n}\left\{-\pi + \sum_{p=1}^{n-1}\frac{1}{\sqrt{p(n-p)}}\right\} \\ &- \log 4 - \log(-\log 4\mu) - 2\sum_{n=1}^{\infty}\frac{1}{n\sqrt{\pi}}\int_{\sqrt{-n\log 4\mu}}^{\infty} dv\ e^{-v^2} \end{aligned} \tag{25}$$

In the limit $q \to \infty$, expression (24) gives $A(q) \to \infty$ and this confirms the decay (19) faster than exponential. For $q = 2$, one finds that $A(2) \simeq 1.3062..$ and $B(2) \simeq .517..$.

The two expressions (22,24) and (23,25) of $A(q)$ of $B(q)$ valid for $q < 2$ and $q > 2$ look different. However, they are analytic continuations of each other and therefore they represent the same function of q. This can be seen, in the case of $A(q)$, by writing for $q < 2$ the expression (22) as

$$A(q) = -\frac{1}{4\sqrt{\pi}} \frac{q}{q-1} \int_{-\infty}^{\infty} dk \log(1 - 4\mu e^{-k^2}) \ .$$

As $q \to 2$, the parameter μ tends to $1/4$ and two complex zeros of the integrand $\pm ik_0(q) = \pm i\sqrt{-\log 4\mu}$ approach the real axis. The analytic continuation (24) corresponds to an exchange of these two zeros (or equivalently to a deformation of the integration contour over k to prevent these two zeros from crossing the contour as q crosses 2). The same kind of reasoning can be used to show that (23) and (25) are two different expressions of the same function [8].

Remark: a priori, in the Potts model, the number of states q is an integer and so the case $q < 2$ is of little interest. Here however, it is easy to check, that if one considers the zero temperature dynamics of an Ising chain where the spins are initially uncorrelated but with a non-zero magnetization m, the distribution of the sizes of domains of + spins is exactly the same as for the q-state Potts model when

$$q = \frac{2}{1+m} \ .$$

For $m \neq 0$, there is no symmetry and the distributions of sizes of + domains and of − domains are different. They are given by $g_q(x)$ for $q = 2/(1+m)$ and $q = 2/(1-m)$ respectively.

Remark: It is interesting to note that the only way that the matrix $c_{x,y}$ enters in the expressions of (22,23,24,25) is through (14) with a function F given by (15). If instead of (15), a different function F was used in (14), the large x behavior of the normalised distribution of sizes $g_q(x)$ would remain of the form (21) with new expressions of $A(q)$ and $B(q)$. For example the expression (22) of $A(q)$ would become for $q \leq 2$

$$A(q) = -\frac{1}{4\pi F'(0)} \frac{q}{q-1} \int_{-\infty}^{\infty} dk \log\left(1 - 2\mu \int_{-\infty}^{\infty} dx e^{ikx} F'(|x|)\right) \ .$$

3 The number of persistent spins on an infinite line

The long time limit of a coarsening phenomenon, when it leads to a scaling regime, is in many respects similar to a critical point describing a second order

phase transition. This scaling regime is characterized by universal exponents (here (17) the typical size of domains grows with time like $t^{1/2}$) as well as universal scaling functions such as $g_q(x)$. It turns out that this scaling regime is characterised by other exponents than 1/2: it was found first numerically [9] that the fraction $r(t)$ of spins which never flip up to time t decays like

$$r(t) \sim t^{-\theta(q)} \tag{26}$$

where the exponent $\theta(q)$ depends on q. The exact value of the exponent $\theta(q)$ can be determined by a calculation rather similar to the one which gave the distribution of domain sizes (in fact the worke [8] on the distribution of domain sizes followed the calculation of the exponent $\theta(q)$ [6,7]). The result was that $\theta(q)$ is given by

$$\theta(q) = -\frac{1}{8} + \frac{2}{\pi^2}\left[\cos^{-1}\left(\frac{2-q}{\sqrt{2}\,q}\right)\right]^2 \tag{27}$$

which gives $\theta(2) \simeq .375$, $\theta(3) \simeq .5379508...$, $\theta(5) \simeq .6928365...$, $\theta(10) \simeq .8310356...$ It is remarkable that for all integer values of q except $q = 2$, $\theta(q)$ is irrational. The expression (27) has been obtained by two different approaches: the first one [7], summarized in this section, is based on a calculation done directly on the infinite line; the second one [6] uses exact properties of finite lattices and finite size scaling as described in the next section.

To calculate the probability $r(t)$ that a spin never flips up to time t, one can consider n different times $\tau_1 < \tau_2 < ... < \tau_n$ and try to calculate the probability that a given spin $\sigma_i(t)$ takes the same value at these n different times. Then taking the limit $(n \to \infty)$ of a set of times $\tau_1, \tau_2...\tau_n$ dense between 0 and t gives in principle $r(t)$.

To obtain $r(t)$ we used this idea for a semi-infinite system for which the spin σ_0 at the origin is updated according to

$$\sigma_0(t+\Delta t) = \begin{cases} \sigma_0(t) & \text{with probability } 1-\Delta t \\ \sigma_1(t) & \text{with probability } \Delta t \end{cases} \tag{28}$$

If $\phi(\tau_1, \tau_2..\tau_n)$ is the probability that $\sigma_0(\tau_1) = \sigma_0(\tau_2) = ...\sigma_0(\tau_n)$, the probability $r(t)$ that a spin of an infinite chain never flips up to time t is given by

$$r(t) = \left[\lim_{n\to\infty} \phi(\tau_1, \tau_2..\tau_n)\right]^2 \tag{29}$$

where as $n \to \infty$, the times $\tau_1, \tau_2...\tau_n$ become dense between 0 and t. To understand (29), one should notice that if a spin σ_i of an infinite chain never flips up to time t, the dynamics of the spins at its left are completely decorrelated

from the spins at its right. Thus the two sides of the spin σ_i can be considered as two independent semi-infinite chains.

As for equal time correlation functions (see section 2), the probability $\phi(\tau_1, \tau_2..\tau_n)$ that
$\sigma_0(\tau_1) = \sigma_0(\tau_2) = ...\sigma_0(\tau_n)$ can be expressed in terms of the matrix c_{τ_i,τ_j} where c_{τ_i,τ_j} is the probability that two random walkers starting at the origin of a semi-infinite chain at times $t - \tau_j < t - \tau_i$ do not meet up to time t. For $\tau_1 < \tau_2 < ... < \tau_n$,

$$\phi(\tau_1, \tau_2) = 1 - \frac{q-1}{q} c_{\tau_1,\tau_2}$$

$$\phi(\tau_1, \tau_2, \tau_3) = 1 - \frac{q-1}{q^2}[c_{\tau_1,\tau_2} + c_{\tau_2,\tau_3}] - \left(\frac{q-1}{q}\right)^2 c_{\tau_1,\tau_3}$$

$$\phi(\tau_1, \tau_2, \tau_3, \tau_4) = 1 - \frac{q-1}{q^2}[c_{\tau_1,\tau_2} + c_{\tau_2,\tau_3} + c_{\tau_3,\tau_4}] - \left(\frac{q-1}{q}\right)^2 c_{\tau_1,\tau_4} + \frac{(q-1)^2}{q^3} c^{(2)}_{\tau_1,\tau_2,\tau_3,\tau_4}$$

and more generally,

$$\phi(\tau_1, \tau_2, ...\tau_n) = 1 - \mu \sum_{i=1}^{n-1} c_{\tau_i,\tau_{i+1}} + \mu^2 \sum_{i<j} c^{(2)}_{\tau_i,\tau_{i+1},\tau_j,\tau_{j+1}}$$

$$- \mu^3 \sum_{i<j<k} c^{(3)}_{\tau_i,\tau_{i+1},\tau_j,\tau_{j+1},\tau_k,\tau_{k+1}} + \cdots$$

$$-\lambda \left\{ \mu c_{\tau_1,\tau_n} - \mu^2 \sum_i c^{(2)}_{\tau_1,\tau_i,\tau_{i+1},\tau_n} + \mu^3 \sum_{i<j} c^{(3)}_{\tau_1,\tau_i,\tau_{i+1},\tau_j,\tau_{j+1},\tau_n} - \cdots \right\} \quad (30)$$

These expressions for $\phi(\tau_1, \tau_2, ...\tau_n)$ are exactly the same as those (4,7,8,10) of $A_{x_1,x_2,...x_{2k}}$ and as in section 2, the $c^{(k)}_{\tau_1,\tau_2,..\tau_{2k}}$ are Pfaffians (9,13). The only difference is in the expression of the matrix $c_{\tau,\tau'}$ which is still antisymmetric and which is given here for $\tau < \tau'$ by

$$c_{\tau,\tau'} = -c_{\tau',\tau} = \sum_{0 \le x \le y} p(x,\tau)p(y,\tau') - p(x,\tau')p(y,\tau)$$

where $p(x,t)$ is the probability of finding, on site x at time t, a random walker which starts at the origin of a semi-infinite chain at $t = 0$

$$p(x,t) = \frac{1}{2\pi} \int_0^{2\pi} d\theta \; [\, \cos(x\theta) + \cos((x+1)\theta) \,] \; \exp[-2(1-\cos\theta)t]$$

Remark: instead of replacing the problem of a peristent spin on an infinite

line by the same problem on two independent semi-infinite lines, one could try to write directly for the infinite line an expression similar to (30). This expression would however be more complicated: the simplicity of the semi-infinite line comes from the fact that if there are n non-intersecting walkers starting at the origin at times $t - \tau_n < t - \tau_{n-1} < ...t - \tau_1$, the positions of these walkers remain for ever in the same order $x_1 < x_2 < ..x_n$. One can then easily represent these walkers by free fermions. One can also express the probabilities of all the meeting events between coalescing random walkers starting at times $t - \tau_n < t - \tau_{n-1} < ...t - \tau_1$ in terms of the matrix $c_{\tau,\tau'}$ [7]. If we had worked directly on an infinite chain, the positions $x_1, x_2, ...x_n$ of walkers starting at a given site at times $t - \tau_n < t - \tau_{n-1} < ...t - \tau_1$ could appear in many different orders and this would make the calculation of the probabilities of meeting events between coalescing random walkers much more difficult than for a semi-infinite line.

The last step to calculate the exponent $\theta(q)$ in (26) is to estimate the large n and large t behavior of $\phi(\tau_1, \tau_2, ...\tau_n)$ (when the times $\tau_1, \tau_2, ..\tau_n$ become dense) (29,30). For large τ and τ', $c_{\tau,\tau'}$ becomes a scaling function of the ratio τ/τ'

$$c_{\tau,\tau'} = -c_{\tau',\tau} \simeq f\left(\frac{\tau}{\tau'}\right) \tag{31}$$

where

$$f(z) = \frac{4}{\pi} \tan^{-1} \sqrt{\frac{1}{z}} \quad - \quad 1 \; . \tag{32}$$

As for the distribution of domain sizes (see section 2), one can rewrite (30) as a determinant and use the theory of of Toeplitz determinants to show that when the $c_{\tau,\tau'}$ have the form (31), the large t and n behavior of $\phi(\tau_1, \tau_2, ...\tau_n)$ is given by

$$\phi(\tau_1, \tau_2, ...\tau_n) \sim t^{-\theta(q)}$$

with

$$\theta(q) = -\frac{1}{2\pi} \int_{-\infty}^{\infty} dk \; \log\left[1 + 2\mu \int_{-\infty}^{\infty} e^{ikv} \; e^{v} \; f'(e^{v}) dv\right]$$

and when $f(z)$ is given by (32), this leads to (27).

The exact expression (27) is in fact in good agreement with the numerical estimates which had been previously obtained either by MonteCarlo simulations [9,10] or by finite size scaling [11].

4 Persistent spins on a finite lattice

Another way of calculating the exponent $\theta(q)$ is to consider the zero-temperature Glauber dynamics on a finite lattice of L sites[6] with random initial conditions and to calculate the probability $\rho_L(q)$ that any given spin never flips between $t = 0$ and $t = \infty$. The size of the domains increases as $t^{1/2}$ for an infinite system and one expects that the dynamics on the finite system will stop at a time $t \sim L^2$. Thus one expects

$$\rho_L(q) \sim L^{-2\theta(q)} \tag{33}$$

The calculation of $\rho_L(q)$ is rather easy for small systems. For example for a system of two sites with periodic boundary conditions, there is a probability $1/q$ that initially the two spins are identical, in which case the two spins will never flip, and a probability $(q-1)/q$ that the two spins are initially different, in which case one of the two spins moves once and then the dynamics stops. Therefore

$$\rho_2(q) = \frac{1}{q} + \frac{1}{2}\frac{q-1}{q} = \frac{q+1}{2q} \tag{34}$$

The same kind of reasoning can be used to obtain for lattices with periodic boundary conditions

$$\begin{aligned}
\rho_1(q) &= 1 \\
\rho_2(q) &= \frac{q+1}{2q} \\
\rho_3(q) &= \frac{2q^2+4q+1}{7q^2} \\
\rho_4(q) &= \frac{8q^3+23q^2+12q+1}{44q^3} \\
\rho_5(q) &= \frac{62q^4+224q^3+176q^2+32q+1}{495q^4} \\
\rho_6(q) &= \frac{912q^5+3864q^4+3983q^3+1135q^2+81q+1}{9976q^5} \\
\rho_7(q) &= \frac{25086q^6+119732q^5+150210q^4+58176q^3+6756q^2+200q+1}{360161}
\end{aligned}$$

and so on.

Initially[11] this exact calculation of $\rho_L(q)$ was done on lattices up to $L = 14$ and these expressions analyzed by a finite size scaling method based on (33) gave the following estimates for the exponent $\theta(q)$:

$$\theta(2) = .3750 \pm .0001 \quad ; \quad \theta(3) = .5379 \pm .0002 \quad ; \quad \theta(5) = .6928 \pm .0003$$

to be compared with the exact values given by (27)

$$\theta(2) = .3750 \quad ; \quad \theta(3) = .5379508... \quad ; \quad \theta(5) = .6928365...$$

Later [6], in collaboration with Vincent Hakim and Vincent Pasquier, we could derive the general expression of $\rho_L(q)$ for arbitrary L and q. Again $\rho_L(q)$ can be written in the same form as (10) or (30)

$$\rho_L(q) = 1 - \mu \sum_{i=1}^{L-1} c_{i,i+1} + \mu^2 \sum_{i<j} c^{(2)}_{i,i+1,j,j+1} - \mu^3 \sum_{i<j<k} c^{(3)}_{i,i+1,j,j+1,k,k+1} + \cdots$$
$$- \lambda \left\{ \mu c_{1,L} - \mu^2 \sum_{i} c^{(2)}_{1,i,i+1,L} + \mu^3 \sum_{i<j} c^{(3)}_{1,i,i+1,j,j+1,L} - \cdots \right\} \quad (35)$$

where the matrix $c_{i,j}$ is a $L \times L$ antisymmetric matrix satisfying for $i < j$ the following equation

$$c_{i+1,j} + c_{i-1,j} + c_{i,j+1} + c_{i,j-1} - 4c_{i,j} = 0 \quad (36)$$

with the boundary conditions that for all $1 \leq i \leq L$

$$c_{i,i} = 0 \quad ; \quad c_{0,i} = c_{i,L} = 1$$

This matrix $c_{i,j}$ can be reinterpreted as the probability that two random walkers starting at positions i and j with $1 \leq i \leq j \leq L$ will not meet before one of them hits one of the two boundaries 0 or $L+1$. This gives
for $L = 2$

$$c_{1,2} = 1/2$$

for $L = 3$

$$c_{1,2} = c_{2,3} = 3/7 \quad ; \quad c_{1,3} = 5/7$$

for $L = 4$

$$c_{1,2} = c_{3,4} = 18/44 \quad ; \quad c_{1,3} = c_{2,4} = 28/44 \quad ; \quad c_{2,3} = 14/44 \quad ; \quad c_{1,4} = 36/44$$

and more generally for $i < j$ and arbitrary L

$$c_{i,j} = 1 - \sum_{k \ even} \sum_{k' odd} \frac{8 \sin k\alpha \ \sin k'\alpha \ (\sin ki\alpha \ \sin k'j\alpha \ - \ \sin kj\alpha \ \sin k'i\alpha)}{(L+1)^2 \ (\cos k'\alpha \ - \ \cos k\alpha) \ (2 - \cos k\alpha - \cos k'\alpha)}$$

with $\alpha = \pi/(L+1)$, $2 \leq k \leq L$, $1 \leq k' \leq L$

Here also, the $\rho_L(q)$ can be written as a determinant [6] and the large L behavior of this determinant leads to the same expression of $\theta(q)$ as in (27).

5 Conclusion

For the growth of ferromagnetic domains induced by the zero temperature Glauber dynamics of the $1d$ Potts model, one can calculate exactly several properties characteristic of the long time regime beyond the pair correlation function: the distribution of domains sizes and the exponent $\theta(q)$ characteristic of the decay of the persistent spins. These results are obtained from the knowledge of the exact expressions of correlation functions for an arbitrarily large number of spins (10,30,35). It is remarkable that different quantities are given by three identical expressions (10,30,35) with only a different matrix c for each case. A difficult, though essential, step (which can be found in [7,8]) in the derivation of the results summarised here is the calculation of the asymptotics of (10,30,35), once they are written as determinants.

One interest in the exponent $\theta(q)$ is that it seems to be one of the simplest non trivial exponents one can measure in statistical mechanics. One can write very simple programs to measure $\theta(q)$ accurately. Of course, one can try to predict $\theta(q)$ in higher dimension [10,12], for various lattices and ask natural questions as the degree of universality of this exponent or the existence of an upper critical dimension. These questions remain up to now unsolved and so far only approximate schemes to determine $\theta(q)$ have been developped [13].

With the restricted definition of $r(t)$ (the fraction of spins which never flip up to time t), it is clear that as soon as the temperature is non zero, thermal noise makes $r(t)$ decay exponentially. Still, in $d > 1$, if a system is quenched into a coexistence phase, one can observe the growth of ferromagnetic domains. When the size of the domains becomes larger than the equilibrium correlation length, one can tell which region of space is in a given phase. Then it becomes meaningful to define $r(t)$ as the fraction of space which remains in the same phase up to time t. One difficulty in measuring $r(t)$ in a simulation at non zero temperature is to distinguish the growing domains from the thermal fluctuations. There is a way [14] of overcoming this difficulty by comparing two systems (system A with a random initial condition where coarsening takes place and system B with a fully ordered initial configuration which serves as a reference). By using the same noise for the two systems, one can define $r(t)$ as the fraction of spins which remain identical up to time t in systems A and B. With this definition of $r(t)$ one can eliminate flips due thermal fluctuations (as they occur in the same way in the two systems A and B) and measure flips due to the motion of the domain walls while the system coarsens. This possibility of measuring θ at finite temperature should give useful informations on the degree of universality of θ. In particular, one can try to see whether the presence of anisotropy due to the lattice [15] may affect θ.

The exponent θ which can be thought as the exponent characteristic of a first passage problem can be defined and measured in several other contexts: gaussian processes [16,17], directed percolation [18], Ginsburg Landau equations [19], soap froth [20], critical dynamics [21], driven diffusive systems [22] and even in experiments [23,24]. For one dimensional systems, the problem of the persistent spins can be thought as a two species reaction diffusion model [25,26,27,28]: the domain walls diffuse and coalesce or annihilate whereas the persistent spins are motionless particles which are annihilated by the domain walls. The calculation described in section 3 can be extended, at least in a pertubative way, to treat the case of reaction diffusion systems where both species move [29].

Acknowledgments

This lecture is based on several pieces of works on which I had the pleasure to collaborate with A. J. Bray, C. Godrèche, V. Hakim, P.M.C. de Oliveira, D. Stauffer, V. Pasquier and R. Zeitak.

References

1. A.J. Bray, Adv. Phys. **43**, 357 (1994)
2. T.M. Liggett, *Interacting Particle Systems*, NY: Springer Verlag (1985)
3. Z. Rácz, Phys. Rev. Lett. **55**, 1707 (1985)
4. A.J. Bray, J. Phys. A **23**, L67 (1990)
5. J.G. Amar and F. Family, Phys. Rev. A **41**, 3258 (1990)
6. B. Derrida, V. Hakim and V. Pasquier, Phys. Rev. Lett. **75**, 751 (1995)
7. B. Derrida, V. Hakim and V. Pasquier, preprint 1996, J. Stat. Phys. in press
8. B. Derrida and R. Zeitak, preprint 1996, Phys. Rev. E in press
9. B. Derrida, A. J. Bray and C. Godrèche, J. Phys. A **27**, L357 (1994)
10. D. Stauffer, J. Phys. A **27**, 5029 (1994)
11. B. Derrida, J. Phys. A **28**, 1481 (1995)
12. B. Derrida, P.M.C. de Oliveira and D. Stauffer, Physica A **224** , 604 (1996)
13. S. N. Majumdar and C. Sire Phys. Rev. Lett. **77** 1420 (1996)
14. B. Derrida, work in progress
15. A.D. Rutenberg, Phys. Rev. E **54**, 2181 (1996)
16. S.N. Majumdar, C. Sire, A.J. Bray and S.J. Cornell Phys. Rev. Lett. **77** (1996) in press
17. B. Derrida, V. Hakim and R. Zeitak, Phys. Rev. Lett. **77** (1996) in press

18. H. M. Koduvely and H. Hinrichsen, Poster at the Trieste Conference, August 1996
19. A. J. Bray, B. Derrida and C. Godrèche, Europhys. Lett. **27**, 175 (1994)
20. B. Levitan and E. Domany, preprint 1996
21. S.N. Majumdar, C. Sire, A.J. Bray and S.J. Cornell, preprint 96
22. S.J. Cornell and A.J. Bray, preprint 96
23. M. Marcosmartin, D. Beysens, J. P. Bouchaud, C. Godrèche and I. Yekutieli, Physica A **214**, 396 (1995)
24. J. Stavans, private communication
25. P.L. Krapivsky, E. Ben-Naim and S. Redner, Phys. Rev. E **50** , 2474 (1994)
26. P.L. Krapivsky, S. Redner and F. Leyvraz, Phys. Rev. E **51**, 3977 (1995)
27. E. Ben-Naim, L. Frachebourg and P.L. Krapivsky, Phys. Rev. E **53**, 3078 (1996) and preprint 96
28. J. L. Cardy, J. Phys. A **28**, L19 (1995)
29. C. Monthus, preprint 1996, to appear in Phys. Rev. E

THE VERLINDE FORMULA FOR $\mathbf{PGL}_p$

A. BEAUVILLE [1]

DMI - École Normale Supérieure (URA 759 du CNRS)
45 rue d'Ulm, F-75230 PARIS *Cedex 05*

To the memory of
Claude ITZYKSON

Introduction

The Verlinde formula expresses the number of linearly independent conformal blocks in any rational conformal field theory. I am concerned here with a quite particular case, the Wess-Zumino-Witten model associated to a complex semi-simple group [2] G. In this case the space of conformal blocks can be interpreted as the space of holomorphic sections of a line bundle on a particular projective variety, the moduli space M_G of holomorphic G-bundles on the given Riemann surface. The fact that the dimension of this space of sections can be explicitly computed is of great interest for mathematicians, and a number of rigorous proofs of that formula (usually called by mathematicians, somewhat incorrectly, the "Verlinde formula") have been recently given (see e.g. [F], [B-L], [L-S]).

These proofs deal only with simply-connected groups. In this paper we treat the case of the projective group $\mathbf{PGL}_r$ when r is prime.

Our approach is to relate to the case of $\mathbf{SL}_r$, using standard algebro-geometric methods. The components $M^d_{\mathbf{PGL}_r}$ $(0 \leq d < r)$ of the moduli space $M_{\mathbf{PGL}_r}$ can be identified with the quotients M^d_r/J_r, where M^d_r is the moduli space of vector bundles on X of rank r and fixed determinant of degree d, and J_r the finite group of holomorphic line bundles α on X such that $\alpha^{\otimes r}$ is trivial. The space we are looking for is the space of J_r-invariant global sections of a line bundle $\mathcal{L}$ on M^d_r; its dimension can be expressed in terms of the character of the representation of J_r on $H^0(M^d_r, \mathcal{L})$. This is given by the Lefschetz trace formula, with a subtlety for $d = 0$, since M^0_r is not smooth. The key point (already used in [N-R]) which makes the computation quite easy is that the fixed point set of any non-zero element of

[1] Partially supported by the European HCM project "Algebraic Geometry in Europe" (AGE).

[2] This group is the complexification of the compact semi-simple group considered by physicists.

J_r is an abelian variety – this is where the assumption on the group is essential. Extending the method to other cases would require a Chern classes computation on the moduli space M_H for some semi-simple subgroups H of G; this may be feasible, but goes far beyond the scope of the present paper. Note that the case of $M^1_{\mathbf{PGL}_2}$ has been previously worked out in [P] (with an unfortunate misprint in the formula).

In the last section we check that our formulas agree with the predictions of Conformal Field Theory, as they appear for instance in [S-Y]. Note that our results are slightly more precise (in this particular case): we get a formula for $\dim H^0(M^d_{\mathbf{PGL}_r}, \mathcal{L})$ for every d, while CFT only predicts the sum of these dimensions (see Remark 4.3).

1. The moduli space $M_{\mathbf{PGL}_r}$

(1.1) Throughout the paper we denote by X a compact (connected) Riemann surface, of genus $g \geq 2$; we fix a point p of X. Principal $\mathbf{PGL}_r$-bundles on X correspond in a one-to-one way to projective bundles of rank $r-1$ on X, i.e. bundles of the form $\mathbf{P}(E)$, where E is a rank r vector bundle on X; we say that $\mathbf{P}(E)$ is semi-stable if the vector bundle E is semi-stable. The semi-stable projective bundles of rank $r-1$ on X are parameterized by a projective variety, the moduli space $M_{\mathbf{PGL}_r}$.

Two vector bundles E, F give rise to isomorphic projective bundles if and only if F is isomorphic to $E \otimes \alpha$ for some line bundle α on X. Thus a projective bundle can always be written as $\mathbf{P}(E)$ with $\det E = \mathcal{O}_X(dp)$, $0 \leq d < r$; the vector bundle E is then determined up to tensor product by a line bundle α with $\alpha^r = \mathcal{O}_X$. In particular, the moduli space $M_{\mathbf{PGL}_r}$ has r connected components $M^d_{\mathbf{PGL}_r}$ $(0 \leq d < r)$. Let us denote by M^d_r the moduli space of semi-stable vector bundles on X of rank r and determinant $\mathcal{O}_X(dp)$, and by J_r the kernel of the multiplication by r in the Jacobian JX of X; it is a finite group, canonically isomorphic to $H^1(X, \mathbf{Z}/(r))$. The group J_r acts on M^d_r, by the rule $(\alpha, E) \mapsto E \otimes \alpha$; it follows from the above remarks that the component $M^d_{\mathbf{PGL}_r}$ is isomorphic to the quotient M^d_r/J_r.

(1.2) We will need a precise description of the line bundles on $M_{\mathbf{PGL}_r}$. Let me first recall how line bundles on M^d_r can be constructed [D-N]: a simple way is

to mimic the classical definition of the theta divisor on the Jacobian of X (i.e. in the rank 1 case). Put $\delta = (r,d)$; let A be a vector bundle on X of rank r/δ and degree $(r(g-1)-d)/\delta$. These conditions imply $\chi(E \otimes A) = 0$ for all E in M_r^d; if A is general enough, it follows that the condition $H^0(X, E \otimes A) \neq 0$ defines a (Cartier) divisor Θ_A in M_r^d. The corresponding line bundle $\mathcal{L}_d := \mathcal{O}(\Theta_A)$ does not depend on the choice of A, and generates the Picard group $\mathrm{Pic}(M_r^d)$.

(1.3) The quotient map $q : M_r^d \to M_{\mathbf{PGL}_r}^d$ induces a homomorphism $q^* : \mathrm{Pic}(M_{\mathbf{PGL}_r}^d) \to \mathrm{Pic}(M_r^d)$, which is easily seen to be injective. Its image is determined in [B-L-S]: it is generated by $\mathcal{L}_d^\delta$ if r is odd, by $\mathcal{L}_d^{2\delta}$ if r is even.

(1.4) Let $\mathcal{L}'$ be a line bundle on $M_{\mathbf{PGL}_r}^d$. The line bundle $\mathcal{L} := q^*\mathcal{L}'$ on M_r^d admits a natural action of J_r, compatible with the action of J_r on M_r^d (this is often called a J_r-linearization of $\mathcal{L}$). This action is characterized by the property that every element α of J_r acts trivially on the fibre of $\mathcal{L}$ at a point of M_r^d fixed by α. In the sequel we will always consider line bundles on M_r^d of the form $q^*\mathcal{L}'$, and endow them with the above J_r-linearization.

This linearization defines a representation of J_r on the space of global sections; essentially by definition, the global sections of $\mathcal{L}'$ correspond to the J_r-invariant sections of $\mathcal{L}$. Therefore our task will be to compute the dimension of the space of invariant sections; as indicated in the introduction, we will do that by computing, for any $\alpha \in J_r$ of order r, the trace of α acting on $H^0(M_r^d, \mathcal{L})$.

2. The action of J_r on $H^0(M_r^d, \mathcal{L}_d^k)$

We start with the case when r and d are coprime, which is easier to deal with because the moduli space is smooth.

Proposition 2.1 .– *Assume r and d are coprime. Let k be an integer; if r is even we assume that k is even. Let α be an element of order r in JX. Then the trace of α acting on $H^0(M_r^d, \mathcal{L}_d^k)$ is $(k+1)^{(r-1)(g-1)}$.*

Proof: The Lefschetz trace formula reads [A-S]

$$\mathrm{Tr}(\alpha \mid H^0(M_r^d, \mathcal{L}_d^k)) = \int_P \mathrm{Todd}(T_P)\ \lambda(N_{P/M_r^d}, \alpha)^{-1}\ \widetilde{\mathrm{ch}}(\mathcal{L}_{d\,|P}^k, \alpha)\ .$$

Here P is the fixed subvariety of α; whenever F is a vector bundle on P and φ a diagonalizable endomorphism of F, so that F is the direct sum of its eigen-sub-

bundles F_λ for $\lambda \in \mathbf{C}$, we put

$$\widetilde{ch}(F,\varphi) = \sum \lambda \, ch(F_\lambda) \quad ; \quad \lambda(F,\varphi) = \prod_\lambda \sum_{p\geq 0} (-\lambda)^p \, ch(\Lambda^p F_\lambda^*) \, .$$

We have a number of informations on the right hand side thanks to [N-R]:

(2.1 *a*) Let $\pi : \widetilde{X} \to X$ be the étale r-sheeted covering associated to α; put $\xi = \alpha^{r(r-1)/2} \in JX$. The map $L \mapsto \pi_*(L)$ identifies any component of the fibre of the norm map $Nm : J^d\widetilde{X} \to J^dX$ over $\xi(dp)$ with P. In particular, P is isomorphic to an abelian variety, hence the term $Todd(T_P)$ is trivial.

(2.1 *b*) Let $\theta \in H^2(P,\mathbf{Z})$ be the restriction to P of the class of the principal polarization of $J^d\widetilde{X}$. The term $\lambda(N_{P/M_r^d},\alpha)$ is equal to $r^{r(g-1)}e^{-r\theta}$.

(2.1 *c*) The dimension of P is $N = (r-1)(g-1)$, and the equality $\int_P \frac{\theta^N}{N!} = r^{g-1}$ holds.

With our convention the action of α on $\mathcal{L}_{d|P}^k$ is trivial. The class $c_1(\mathcal{L}_{d|P})$ is equal to $r\theta$: the pull back to P of the theta divisor Θ_A (1.2) is the divisor of line bundles L in P with $H^0(L \otimes \pi^*A) \neq 0$; to compute its cohomology class we may replace π^*A by any vector bundle with the same rank and degree, in particular by a direct sum of r line bundles of degree $r(g-1)-d$, which gives the required formula.

Putting things together, we find

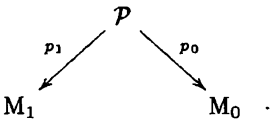
$$Tr(\alpha \mid H^0(M_r^d, \mathcal{L}_d^k)) = \int_P r^{-r(g-1)} e^{r\theta} e^{kr\theta} = (k+1)^{(r-1)(g-1)} \, . \quad \blacksquare$$

We now consider the degree 0 case:

Proposition 2.2 .– *Let k be a multiple of r, and of $2r$ if r is even; let α be an element of order r in JX. Then the trace of α acting on $H^0(M_r^0, \mathcal{L}_0^k)$ is $(\frac{k}{r}+1)^{(r-1)(g-1)}$.*

Proof: We cannot apply directly the Lefschetz trace formula since it is manageable only for smooth projective varieties; instead we use another well-known tool, the Hecke correspondence (this idea appears for instance in [B-S]). For simplicity we write M_d instead of M_r^d. There exists a Poincaré bundle $\mathcal{E}$ on $X \times M_1$, i.e. a vector bundle whose restriction to $X \times \{E\}$, for each point E of M_1, is isomorphic to E. Such a bundle is determined up to tensor product by a line bundle coming from

M_1 ; we will see later how to normalize it. We denote by $\mathcal{E}_p$ the restriction of $\mathcal{E}$ to $\{p\} \times M_1$, and by $\mathcal{P}$ the projective bundle $\mathbf{P}(\mathcal{E}_p^*)$ on M_1 . A point of $\mathcal{P}$ is a pair (E, φ) where E is a vector bundle in M_1 and $\varphi : E \to \mathbf{C}_p$ a non-zero homomorphism, defined up to a scalar; the kernel of φ is then a vector bundle $F \in M_1$, and we can view equivalently a point of $\mathcal{P}$ as a pair of vector bundles (F, E) with $F \in M_0$, $E \in M_1$ and $F \subset E$. The projections p_d on M_d $(d = 0, 1)$ give rise to the "Hecke diagram"



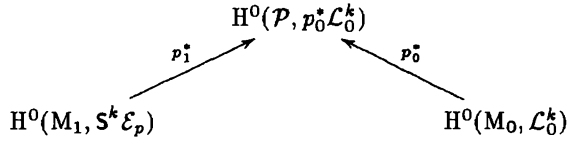
Lemma 2.3 .– *The Poincaré bundle $\mathcal{E}$ can be normalized (in a unique way) so that $\det \mathcal{E}_p = \mathcal{L}_1$; then $\mathcal{O}_{\mathcal{P}}(1) \cong p_0^* \mathcal{L}_0$.*

Proof: Let $E \in M_1$. The fibre $p_1^{-1}(E)$ is the projective space of non-zero linear forms $\ell : E_p \to \mathbf{C}$, up to a scalar. The restriction of $p_0^* \mathcal{L}_0$ to this projective space is $\mathcal{O}(1)$ (choose a line bundle L of degree $g - 1$ on X ; if E is general enough, $H^0(X, E \otimes L)$ is spanned by a section s with $s(p) \neq 0$, and the condition that the bundle F corresponding to ℓ belongs to Θ_L is the vanishing of $\ell(s(p))$). Therefore $p_0^* \mathcal{L}_0$ is of the form $\mathcal{O}_{\mathcal{P}}(1) \otimes p_1^* \mathcal{N}$ for some line bundle $\mathcal{N}$ on M_1 . Replacing $\mathcal{E}$ by $\mathcal{E} \otimes \mathcal{N}$ we ensure $\mathcal{O}_{\mathcal{P}}(1) \cong p_0^* \mathcal{L}_0$.

An easy computation gives $K_{\mathcal{P}} = p_1^* \mathcal{L}_1^{-1} \otimes p_0^* \mathcal{L}_0^{-r}$ ([B-L-S], Lemma 10.3). On the other hand, since $\mathcal{P} = \mathbf{P}(\mathcal{E}_p^*)$, we have $K_{\mathcal{P}} = p_1^*(K_{M_1} \otimes \det \mathcal{E}_p) \otimes \mathcal{O}_{\mathcal{P}}(-r)$; using $K_{M_1} = \mathcal{L}_1^{-2}$ [D-N], we get $\det \mathcal{E}_p = \mathcal{L}_1$. ∎

We normalize $\mathcal{E}$ as in the lemma; this gives for each $k \geq 0$ a canonical isomorphism $p_{1*} p_0^* \mathcal{L}_0^k \cong S^k \mathcal{E}_p$. Let α be an element of order r of JX . It acts on the various moduli spaces in sight; with a slight abuse of language, I will still denote by α the corresponding automorphism. There exists an isomorphism $\alpha^* \mathcal{E} \xrightarrow{\sim} \mathcal{E} \otimes \alpha$, unique up to a scalar ([N-R], lemma 4.7); the induced isomorphism $u : \alpha^* \mathcal{E}_p \xrightarrow{\sim} \mathcal{E}_p$ induces the action of α on $\mathcal{P}$. Imposing $u^r = \mathrm{Id}$ determines u up to a r-th root of unity, hence determines completely $S^k u$ when k is a multiple of r . Since the Hecke

diagram is equivariant with respect to α, it gives rise to a diagram of isomorphisms



which is compatible with the action of α; in particular, the trace we are looking for is equal to the trace of α on $H^0(M_1, S^k\mathcal{E}_p)$.

We are now in the situation of Prop. 2.1, and the Lefschetz trace formula gives:

$$\mathrm{Tr}(\alpha \mid H^0(M_1, S^k\mathcal{E}_p)) = \int_P \mathrm{Todd}(T_P)\ \lambda(N_{P/M_1}, \alpha)^{-1}\ \widetilde{\mathrm{ch}}(S^k\mathcal{E}_{p\,|P}, \alpha)\ .$$

The only term we need to compute is $\widetilde{\mathrm{ch}}(S^k\mathcal{E}_{p\,|P}, \alpha)$. Let $\mathcal{N}$ be the restriction to $\widetilde{X} \times P$ of a Poincaré line bundle on $\widetilde{X} \times J^1\widetilde{X}$; let us still denote by $\pi : \widetilde{X} \times P \to X \times P$ the map $\pi \times \mathrm{Id}_P$. The vector bundles $\pi_*(\mathcal{N})$ and $\mathcal{E}_{|X\times P}$ have the same restriction to $X \times \{\gamma\}$ for all $\gamma \in P$, hence after tensoring $\mathcal{N}$ by a line bundle on P we may assume they are isomorphic ([R], lemma 2.5). Restricting to $\{p\} \times P$ we get $\mathcal{E}_{p\,|P} = \bigoplus_{\pi(q)=p} \mathcal{N}_q$, with $\mathcal{N}_q = \mathcal{N}_{|\{q\}\times P}$.

We claim that the $\mathcal{N}_q$'s are the eigen-sub-bundles of $\mathcal{E}_{p\,|P}$ relative to α. By (2.1 *a*), a pair $(E, F) \in \mathcal{P}$ is fixed by α if and only if $E = \pi_*L$, $F = \pi_*L'$, with $\mathrm{Nm}(L) = \xi(p)$, $\mathrm{Nm}(L') = \xi$; because of the inclusion $F \subset E$ we may take L' of the form $L(-q)$, for some point $q \in \pi^{-1}(p)$. In other words, the fixed locus of α acting on $\mathcal{P}$ is the disjoint union of the sections $(\sigma_q)_{q\in\pi^{-1}(p)}$ of the fibration $p_1^{-1}(P) \to P$ characterized by $\sigma_q(\pi_*L) = (\pi_*L, \pi_*(L(-q)))$. Viewing $\mathcal{P}$ as $\mathbf{P}(\mathcal{E}^*_{p\,|P})$, the section σ_q corresponds to the exact sequence

$$0 \to \pi_*(\mathcal{N}(-q))_{|\{p\}\times P} \longrightarrow \pi_*(\mathcal{N})_{|\{p\}\times P} \cong \mathcal{E}_{|\{p\}\times P} \longrightarrow \mathcal{N}_q \to 0\ .$$

Therefore on each fibre $\mathbf{P}(E_p)$, for $E \in P$, the automorphism α has exactly r fixed points, corresponding to the r sub-spaces $\mathcal{N}_{(q,E)}$ for $q \in \pi^{-1}(p)$; this proves our claim.

The line bundles $\mathcal{N}_q$ for $q \in \widetilde{X}$ are algebraically equivalent, and therefore have the same Chern class. We thus have $c_1(\mathcal{E}_{p\,|P}) = r\,c_1(\mathcal{N}_q)$. On the other hand we know that $\det \mathcal{E}_p = \mathcal{L}_1$ (lemma 2.3), and that $c_1(\mathcal{L}_{1\,|P}) = r\theta$ (proof of Prop. 2.1). By comparison we get $c_1(\mathcal{N}_q) = \theta$. Putting things together we obtain

$$\widetilde{\mathrm{ch}}(S^k\mathcal{E}_{p\,|P}, \alpha) = \int_P \mathrm{Tr}\, S^k D_r\ e^{k\theta} r^{-r(g-1)} e^{r\theta}$$

where D_r is the diagonal r-by-r matrix with entries the r distinct r-th roots of unity.

Lemma 2.4 .– *The trace of* $S^k D_r$ *is* 1 *if* r *divides* k *and* 0 *otherwise.*

Consider the formal series $s(T) := \sum_{i\geq 0} T^i \operatorname{Tr} S^i u$ and $\lambda(T) := \sum_{i\geq 0} T^i \operatorname{Tr} \Lambda^i u$. The formula $s(T)\lambda(-T) = 1$ is well-known (see e.g. [Bo], § 9, formula (11)). But

$$\lambda(-T) = \sum_{i=0}^{r} (-T)^i \operatorname{Tr} \Lambda^i u = \prod_{\zeta^r=1} (1 - \zeta T) = 1 - T^r \ ,$$

hence the lemma. Using (2.1 c) the Proposition follows. ∎

3. Formulas

In this section I will apply the above results to compute the dimension of the space of sections of the line bundle $\mathcal{L}_d^k$ on the moduli space $M^d_{\mathbf{PGL}_r}$. Let me first recall the corresponding Verlinde formula for the moduli spaces M_r^d. Let $\delta = (r, d)$; we write $\mathcal{L}_d = \mathcal{D}^{r/\delta}$, with the convention that we only consider powers of $\mathcal{D}$ which are multiple of r/δ (the line bundle $\mathcal{D}$ actually makes sense on the *moduli stack* $\mathcal{M}_r^d$, and generates its Picard group). We denote by μ_r the center of $\mathbf{SL}_r$, i.e. the group of scalar matrices ζI_r with $\zeta^r = 1$.

Proposition 3.1 .– *Let* T_k *be the set of diagonal matrices* $t = \operatorname{diag}(t_1, \ldots, t_r)$ *in* $\mathbf{SL}_r(\mathbf{C})$ *with* $t_i \neq t_j$ *for* $i \neq j$, *and* $t^{k+r} \in \mu_r$; *for* $t \in T_k$, *let* $\delta(t) = \prod_{i<j}(t_i - t_j)$. *Then*

$$\dim H^0(M_r^d, \mathcal{D}^k) = r^{g-1}(k+r)^{(r-1)(g-1)} \sum_{t\in T_k/\mathfrak{S}_r} \frac{((-1)^{r-1}t^{k+r})^{-d}}{|\delta(t)|^{2g-2}}$$

Proof: According to [B-L], Thm. 9.1, the space $H^0(M_r^d, \mathcal{D}^k)$ for $0 < d < r$ is canonically isomorphic to the space of conformal blocks in genus g with the representation $V_{k\varpi_{r-d}}$ of $\mathbf{SL}_r$ with highest weight $k\varpi_{r-d}$ inserted at one point. The Verlinde formula gives therefore (see [B], Cor. 9.8[1]):

$$\dim H^0(M_r^d, \mathcal{D}^k) = r^{g-1}(k+r)^{(r-1)(g-1)} \sum_{t\in T_k/\mathfrak{S}_r} \frac{\operatorname{Tr}_{V_{k\varpi_{r-d}}}(t)}{|\delta(t)|^{2g-2}} \ ;$$

[1] There is a misprint in the first equality of that corollary, where one should read T_t^{reg}/W instead of T_t^{reg} ; the second equality (and the proof!) are correct.

this is still valid for $d = 0$ with the convention $\varpi_r = 0$.

The character of the representation $V_{k\varpi_{r-d}}$ is given by the Schur formula (see e.g. [F-H], Thm. 6.3):

$$\mathrm{Tr}_{V_{k\varpi_{r-d}}}(t) = \frac{1}{\delta(t)} \begin{vmatrix} t_1^{k+r-1} & t_2^{k+r-1} & \cdots & t_r^{k+r-1} \\ t_1^{k+r-2} & t_2^{k+r-2} & & t_r^{k+r-2} \\ \vdots & \vdots & & \vdots \\ t_1^{k+d} & t_2^{k+d} & & t_r^{k+d} \\ t_1^{d-1} & t_2^{d-1} & & t_r^{d-1} \\ \vdots & \vdots & & \vdots \\ 1 & 1 & & 1 \end{vmatrix} .$$

Writing $t^{k+r} = \zeta I_r \in \mu_r$, the big determinant reduces to $\zeta^{r-d}(-1)^{d(r-d)} \det(t_j^{d-i})$, and finally, since $\prod t_i = 1$, to $((-1)^{r-1}\zeta)^{-d}\delta(t)$, which gives the required formula. ∎

Corollary 3.2 .– *Let* T'_k *be the set of matrices* $t = \mathrm{diag}(t_1, \ldots, t_r)$ *in* $\mathbf{SL}_r(\mathbf{C})$ *with* $t_i \neq t_j$ *if* $i \neq j$, *and* $t^{k+r} = (-1)^{r-1} I_r$. *Then*

$$\sum_{d=0}^{r-1} \dim \mathrm{H}^0(\mathrm{M}_r^d, \mathcal{D}^k) = r^g (k+r)^{(r-1)(g-1)} \sum_{t \in \mathrm{T}'_k/\mathfrak{S}_r} \frac{1}{|\delta(t)|^{2g-2}} \cdot \quad \blacksquare$$

We now consider the moduli space $\mathrm{M}_{\mathbf{PGL}_r}$. We know that the line bundle $\mathcal{D}^k$ on M_r^d descends to $\mathrm{M}_{\mathbf{PGL}_r}^d = \mathrm{M}_r^d/\mathrm{J}_r$ exactly when k is a multiple of r if r is odd, or of $2r$ if r is even (1.3). When this is the case we obtain a line bundle on $\mathrm{M}_{\mathbf{PGL}_r}^d$, that we will still denote by $\mathcal{D}^k$; its global sections correspond to the J_r-invariant sections of $\mathrm{H}^0(\mathrm{M}_r^d, \mathcal{D}^k)$.

We will assume that r is *prime*, so that every non-zero element α of J_r has order r. Then Prop. 2.1 and 2.2 lead immediately to a formula for the dimension of the J_r-invariant subspace of $\mathrm{H}^0(\mathrm{M}_r^d, \mathcal{D}^k)$ as the average of the numbers $\mathrm{Tr}(\alpha)$ for α in J_r. Using Prop. 3.1 we conclude:

Proposition 3.3 .– *Assume that* r *is prime. Let* k *be a multiple of* r; *if* $r = 2$ *assume* $4 \mid k$. *Then*

$$\begin{aligned} \dim \mathrm{H}^0(\mathrm{M}_{\mathbf{PGL}_r}^d, \mathcal{D}^k) &= r^{-2g} \dim \mathrm{H}^0(\mathrm{M}_r^d, \mathcal{D}^k) \; + \; (1 - r^{-2g})(\frac{k}{r} + 1)^{(r-1)(g-1)} \\ &= r^{-2g} (\frac{k}{r} + 1)^{(r-1)(g-1)} \left(r^{r(g-1)} \sum_{t \in \mathrm{T}_k/\mathfrak{S}_r} \frac{((-1)^{r-1} t^{k+r})^{-d}}{|\delta(t)|^{2g-2}} + r^{2g} - 1 \right) . \end{aligned}$$

Summing over d and plugging in Cor. 3.2 gives the following rather complicated formula:

Corollary 3.4 .–

$$\dim \mathrm{H}^0(\mathrm{M}_{\mathbf{PGL}_r}, \mathcal{D}^k) = r^{1-2g}\,(\frac{k}{r}+1)^{(r-1)(g-1)}\,\Big(r^{r(g-1)}\sum_{t\in \mathrm{T}'_k/\mathfrak{S}_r}\frac{1}{|\delta(t)|^{2g-2}} + r^{2g}-1\Big)\ .$$

As an example, if k is an integer divisible by 4, we get

$$(3.5)\quad \dim \mathrm{H}^0(\mathrm{M}_{\mathbf{PGL}_2}, \mathcal{D}^k) = 2^{1-2g}\,(\frac{k}{2}+1)^{g-1}\Big(\sum_{\substack{l \text{ odd}\\ 0<l<k+2}}\frac{1}{(\sin\frac{l\pi}{k+2})^{2g-2}} + 2^{2g}-1\Big)\ .$$

4. Relations with Conformal Field Theory

(4.1) According to Conformal Field Theory, the space $\mathrm{H}^0(\mathrm{M}_{\mathbf{PGL}_r}, \mathcal{D}^k)$ should be canonically isomorphic to the space of conformal blocks for a certain Conformal Field Theory, the WZW model associated to the projective group. This implies in particular that its dimension should be equal to $\sum_j |\mathrm{S}_{0j}|^{2-2g}$, where (S_{ij}) is a unitary symmetric matrix. For instance in the case of the WZW model associated to $\mathbf{SL}_2$, we have

$$\mathrm{S}_{0j} = \frac{\sin\frac{(j+1)\pi}{k+2}}{\sqrt{\frac{k}{2}+1}} \qquad \text{, with} \quad 0 \le j \le k\ ,$$

where the index j can be thought as running through the set of irreducible representations $\mathrm{S}^1, \ldots, \mathrm{S}^k$ of $\mathbf{SL}_2$ (or equivalently $\mathbf{SU}_2$), with $\mathrm{S}^j := \mathrm{S}^j(\mathbf{C}^2)$.

We deduce from (3.5) an analogous expression for $\mathbf{PGL}_2$: we restrict ourselves to even indices and write

$$\mathrm{S}'_{0j} = 2\,\mathrm{S}_{0j} \qquad \text{for} \quad j \text{ even } < k/2 \quad ; \qquad \mathrm{S}'_{0,\frac{k}{2}(1)} = \mathrm{S}'_{0,\frac{k}{2}(2)} = \mathrm{S}_{0\frac{k}{2}}$$

In other words, we consider only those representations of $\mathbf{SL}_2$ which factor through $\mathbf{PGL}_2$ and we identify the representation S^{2j} with S^{k-2j}, doubling the coefficient S_{0j} when these two representations are distinct, and counting twice the representation which is fixed by the involution (this process is well-known, see e.g. [M-S]).

(4.2) The case of $\mathbf{SL}_r$ is completely analogous; we only need a few more terminology from representation theory (we follow the notation of [B]). The primary

fields are indexed by the set P_k of dominant weights λ with $\lambda(H_\theta) \leq k$, where H_θ is the matrix $\mathrm{diag}(1, 0, \ldots, 0, -1)$. For $\lambda \in P_k$, we put $t_\lambda = \exp 2\pi i \dfrac{\lambda+\rho}{k+r}$ (we identify the Cartan algebra of diagonal matrices with its dual using the standard bilinear form); the map $\lambda \mapsto t_\lambda$ induces a bijection of P_k onto $T_k/\mathfrak{S}_r$ ([B], lemma 9.3 *c*)). In view of Prop. 3.1, the coefficient $S_{0\lambda}$ for $\lambda \in P_k$ is given by

$$S_{0\lambda} = \frac{\delta(t_\lambda)}{\sqrt{r}(k+r)^{(r-1)/2}} .$$

Passing to $\mathbf{PGL}_r$, we first restrict the indices to the subset P'_k of elements $\lambda \in P_k$ such that t_λ belongs to T'_k; this means that λ belongs to the root lattice, i.e. that the representation V_λ factors through $\mathbf{PGL}_r$. The center μ_r acts on T_k by multiplication; this action preserves T'_k, and commutes with the action of $\mathfrak{S}_r$. The corresponding action on P_k is deduced, via the bijection $\lambda \mapsto \frac{\lambda+\rho}{k+r}$, from the standard action of μ_r on the fundamental alcove A with vertices $\{0, \varpi_1, \ldots, \varpi_{r-1}\}$. [1]

We identify two elements of P'_k if they are in the same orbit with respect to this action. The action has a unique fixed point, the weight $\frac{k}{r}\rho$, which corresponds to the diagonal matrix D_r (2.4); we associate to this weight r indices $\nu^{(1)}, \ldots, \nu^{(r)}$, and put

$$S'_{0\lambda} = r\, S_{0\lambda} \quad \text{for } \lambda \in P'_k/\mu_r ,\ \lambda \neq \frac{k}{r}\rho ; \qquad S'_{0,\nu^{(i)}} = S_{0,\frac{k}{r}\rho} \quad \text{for } i = 1, \ldots, r .$$

From Cor. 3.4 follows easily the formula $\dim H^0(M_{\mathbf{PGL}_r}, \mathcal{D}^k) = \sum |S'_{0\lambda}|^{2-2g}$, where λ runs over $P'_k/\mu_r \cup \{\nu^{(1)}, \ldots, \nu^{(r)}\}$.

Remark 4.3.– It is not clear to me what is the physical meaning of the space $H^0(M^d_{\mathbf{PGL}_r}, \mathcal{D}^k)$, in particular if its dimension can be predicted in terms of the S-matrix. It is interesting to observe that the number $N(g)$ given by Prop. 3.3, which is equal to $\dim H^0(M^d_{\mathbf{PGL}_r}, \mathcal{D}^k)$ for $g \geq 2$, *is not necessarily an integer* for $g = 1$: for $d = 0$ we find $N(1) = 1 + \dfrac{(k+1)^{r-1} - 1}{r^2}$, which is not an integer unless $r^2 \mid k$.

[1] The element $\exp \varpi_1$ of the center gives the rotation of A which maps 0 to ϖ_1, ϖ_1 to ϖ_2, ..., and ϖ_{r-1} to 0.

REFERENCES

[A-S] M.F. ATIYAH, I.M. SINGER: *The index of elliptic operators III.* Ann. of Math. **87**, 546-604 (1968).

[B] A. BEAUVILLE: *Conformal blocks, Fusion rings and the Verlinde formula.* Proc. of the Hirzebruch 65 Conf. on Algebraic Geometry, Israel Math. Conf. Proc. **9**, 75-96 (1996).

[B-L] A. BEAUVILLE, Y. LASZLO: *Conformal blocks and generalized theta functions.* Comm. Math. Phys. **164**, 385-419 (1994).

[B-L-S] A. BEAUVILLE, Y. LASZLO, Ch. SORGER: *The Picard group of the moduli of G-bundles on a curve.* Preprint alg-geom/9608002.

[B-S] A. BERTRAM, A. SZENES: *Hilbert polynomials of moduli spaces of rank 2 vector bundles II.* Topology **32**, 599-609 (1993).

[Bo] N. BOURBAKI: *Algèbre*, Chap. X (Algèbre homologique). Masson, Paris (1980).

[D-N] J.M. DREZET, M.S. NARASIMHAN: *Groupe de Picard des variétés de modules de fibrés semi-stables sur les courbes algébriques.* Invent. math. **97**, 53-94 (1989).

[F] G. FALTINGS: *A proof for the Verlinde formula.* J. Algebraic Geometry **3**, 347-374 (1994).

[F-H] W. FULTON, J. HARRIS: *Representation theory.* GTM **129**, Springer-Verlag, New York Berlin Heidelberg (1991).

[L-S] Y. LASZLO, Ch. SORGER: *The line bundles on the moduli of parabolic G-bundles over curves and their sections.* Annales de l'ENS, to appear; preprint alg-geom/9507002.

[M-S] G. MOORE, N. SEIBERG: *Taming the conformal zoo.* Phys. Letters B **220**, 422-430 (1989).

[N-R] M.S. NARASIMHAN, S. RAMANAN: *Generalized Prym varieties as fixed points.* J. of the Indian Math. Soc. **39**, 1-19 (1975).

[P] T. PANTEV: *Comparison of generalized theta functions.* Duke Math. J. **76**, 509-539 (1994).

[R] S. RAMANAN: *The moduli spaces of vector bundles over an algebraic curve.* Math. Ann. **200**, 69-84 (1973).

[S-Y] A.N. SCHELLEKENS, S. YANKIELOWICZ: *Field identification fixed points in the coset construction.* Nucl. Phys. **B 334**, 67 (1990).

GALOIS ACTIONS FOR GENUS ONE RATIONAL CONFORMAL FIELD THEORIES

M. BAUER
CEA/Saclay, Service de Physique Théorique
F-91191 Gif-sur-Yvette Cedex, France

We describe a conjectural Galois action on representations of the modular group arising in rational conformal field theories on the torus : the Galois action is by automorphisms and we propose an explicit formula for the Galois action on S and T, the generators of $SL_2(\mathbb{Z})$. We prove our conjecture for Wess–Zumino–Witten models on simply connected groups.

A la mémoire de Claude Itzykson

1 Introduction.

1.1 The purpose of these notes is to describe a conjectural universal Galois action on the representations of the modular group of the torus $SL_2(\mathbb{Z})$ that arise in rational conformal field theories.

1.2 The plan of these notes is the following.

Basic definitions are given in section 2.

Section 3 starts with a reminder on a theorem due to Coste and Gannon[1], and continues with a preliminary discussion of our conjecture on Galois actions. Using properties of the restricted characters as modular functions, we show how, in special cases, the conjecture fits into the framework of modular function fields for principal congruence subgroups. Then we discuss some rationality properties, relating the conjecture with the theorem of Coste and Gannon.

In the rest of the paper, we forget about modular functions, and study a particular class of representations of $SL_2(\mathbb{Z})$ from the point of view of algebraic number theory. We apply the results to an interesting class of rational conformal field theories.

Quadratic structures and the representations of the modular group that they underlie are introduced in section 4. This construction is related to the metaplectic representation in mathematics.

Section 5 contains the proof that representations of the modular group associated to quadratic structures satisfy the conjectures on Galois actions made in section 3.

A discrete version of quantum mechanics is used in section 6 to analyse representations of the modular group associated to quadratic structures.

Section 7 is of more general nature. We prove that representations of the modular group which satisfy the conjecture on Galois actions form a category stable under direct sums, tensor products, sub- and quotient representations.

Physical applications are given in section 8. We show how to go from representations associated to quadratic structures to representations arising in Wess-Zumino-Witten models on simply-connected groups. We also show how to deal with certain cosets and orbifolds, but we certainly cannot claim that this exhausts all possibilities.

Appendices A and B give a short reminder on Galois theory. Appendix C establishes some arithmetic properties of quadratic structures and of the associated representations of the modular group.

1.3 It is more than appropriate here to stress the influence that Claude Itzykson had not only on these notes, but on the whole subject. He prophesied very early that Galois theory had to play a role to study $SL_2(\mathbb{Z})$ actions in conformal field theory, and especially modular invariant partition functions. His handwritten notes on the famous proof of the ADE classification [2] already contain observations and computations in this direction. Two years later, while computing the (non necessarily positive) modular invariants for $SU(N)$ Wess-Zumino-Witten models, he foresaw again a manifestation of Galois theory when the commutant with integral coefficients turned out to be as large as the commutant with complex coefficients. In fact our conjecture, to be discussed below, gives a natural explanation for this. After that, his enthousiasm for Galois actions found another playground : the theory of "dessins d'enfants" [3,4], a theme that is surprisingly closely related to the present discussion [5].

1.4 I had the immense luck and pleasure to discuss with him day after day during those years, first as a student and then as a collaborator[a]. I learned a lot from him, from a technical, a conceptual and last but not least a human point of view. Claude was one of the few scientists I know with a deep and broad understanding of physics and mathematics : he was an "honnête homme" [b] as we use to say in French, a language he cherished and used so elegantly.

I would also like to stress how wonderful he was as a thesis advisor. A few months after I began, we became used to chat every day, Claude patiently and gently spending time to correct my mistakes, to help me understand points I missed, to suggest me other possible ways to attack a problem. We always used

[a] He gave me that "title" at some point.

[b] Although there is no question about Claude's honesty, the use of "honnête homme" here is to stress how cultured he was, with interests not only in science, but in any noble human activity.

a blackboard and made detailed computations. But he also kept me informed of what was going on in other areas of physics. He shared his insights with me. He was right most of the time and it was hard for me to convince him that he was wrong , in the very rare cases when it happened[c]. It usually took more than an hour of sharp discussion. But when he was convinced, suddenly he was very happy, and I could see that he was proud of me, just like a father can be proud of his child.

I have the feeling sometimes that he still there, behind me, looking over my shoulder with a smile on his face, just the way he used to do when he entered my office and I was not aware of his presence. It is a wonderful sensation...as long as it lasts.

1.5 It is a real pleasure to thank Daniel Altschuler for discussions, Arnaud Beauville for help with metaplectic representations, Antoine Coste for explanations of his work with Terry Gannon [1] and remarks on the modular group, Pascal Degiovanni for discussions and for friendly communication of his computations on orbifolds of holomorphic conformal field theories, Gareth Jones for sharing with me his knowledge of outer automorphism groups of finite quotients of the modular group, Terry Gannon for bringing [6] (and especially the relevance of chapter 6. for our conjectures) to my knowledge, Philippe Ruelle for discussions, early checks of the conjecture and for a careful reading of the manuscript, and finally Jean-Bernard Zuber for help with fixed point resolutions. Without their kind encouragements, those notes would not be the same. Last but not least, I thank Jean-Michel Drouffe and Jean-Bernard Zuber for giving me this opportunity to express my deep gratitude to Claude.

2 Definitions.

2.1 The modular group $SL_2(\mathbb{Z})$ plays a central role in what follows. It is the group of 2 by 2 matrices with integral entries and determinant 1. It is generated by the matrices

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \text{ and } T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

The relations are $S^2T = TS^2$, $S^4 = 1$, $(ST)^3 = S^2$. We set

$$C = S^2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

[c] I also tried to convince him that he was wrong in several instances when he was right. But Claude didn't mind. He even delayed the publication of one of his articles once because of some unfounded criticism of mine.

We shall also make frequent use of the principal congruence subgroups of $SL_2(\mathbb{Z})$. If n is a positive integer, the principal congruence subgroup of level n, denoted by Γ_n, is the (invariant) subgroup of $SL_2(\mathbb{Z})$ formed by matrices equal to the identity modulo n. Its index in $SL_2(\mathbb{Z})$ is finite. Note that Γ_1 is $SL_2(\mathbb{Z})$ itself. The quotient $\Gamma_1/\Gamma_n \equiv \hat{\Gamma}_n$ is isomorphic to $SL_2(\mathbb{Z}_n)$. By an abuse of notation we use the same letters S and T for the generators of $SL_2(\mathbb{Z})$ and their class in the quotient $\hat{\Gamma}_n$.

2.2 Any rational conformal field theory comes with a certain number of data.

First, there is a chiral algebra (with central charge c) and its associated primary fields ϕ_i, $i \in I$, of conformal weight h_i, indexed by a finite set I. We assume that ϕ_0 is the identity operator (so that $h_0 = 0$).

Second there is a representation of the modular group $SL_2(\mathbb{Z})$ by matrices with rows and columns indexed by I. This representation is fixed once one knows the matrices $\mathcal{S}$ and $\mathcal{T}$ representing S and T. The matrix $\mathcal{T}$ is diagonal and $\mathcal{T}_{ii} = e^{2i\pi(h_i - c/24)}$.

2.3 To the primary field ϕ_i is associated a representation R_i of the chiral algebra. If the (so called restricted) character

$$\chi_i(\tau) \equiv \mathrm{Tr}_{R_i} e^{2i\pi\tau(L_0 - c/24)}$$

is a well defined holomorphic function in the upper half plane $\mathfrak{H}$ for any $i \in I$, then $\chi_i(-1/\tau) = \sum_j \mathcal{S}_{ji}\chi_j(\tau)$ and (by construction) $\chi_i(\tau+1) = \sum_j \mathcal{T}_{ji}\chi_j(\tau)$.

When the restricted characters are linearly independent (this covers a number of interesting cases, for instance the $A_1^{(1)}$ and Virasoro minimal models, but is relatively rare) the above two equations define $\mathcal{S}$ and $\mathcal{T}$, which build a representation of $PSL_2(\mathbb{Z})$ ($\mathcal{C}$ is trivial). Otherwise, one has to use a more refined (and, as far as the author understands, not totally straightforward to implement) prescription.

The fact that modular transformations of restricted characters do not define $\mathcal{S}$ in general will reappear in several places in these notes and makes it difficult to find a systematic approach to our problem.

3 Facts and conjectures.

3.1 A lot is known [1,7] about the action of Galois transformations on the matrix elements of the matrix $\mathcal{S}$ (representing the transformation $\tau \to -1/\tau$ on characters). This is because $\mathcal{S}$ diagonalises a set of commuting matrices with integral entries, the matrices of fusion rules. This is the celebrated Verlinde

formula, which remains conjectural when stated in full generality, but for which there is such an overwhelming evidence that we shall take it as true.

Theorem 1 (Coste and Gannon, after De Boer and Goere) [1,7]

The field K obtained by extending $\mathbb{Q}$ with the matrix elements of $\mathcal{S}$ is an Abelian extension of $\mathbb{Q}$. There is a unique representation of $\mathrm{Gal}\,\mathrm{K}/\mathbb{Q}$ *as a group of permutations of the set I indexing the rows (or columns) of $\mathcal{S}$* [d] *and a unique map ε from* $\mathrm{Gal}\,\mathrm{K}/\mathbb{Q} \times \mathrm{I}$ *to $\{-1, 1\}$ such that*

$$\sigma(\mathcal{S}_{ij}) = \varepsilon_\sigma(i)\mathcal{S}_{\sigma(i)j} \text{ for } \sigma \in Gal\, K/\mathbb{Q}$$

3.2 Our approach is a tentative to deal with the full modular group (or equivalently with S and T) at the same time. It turns out that a nice conjectural picture emerges. This picture has not produced any spectacular application yet.

Our conjecture is :

Conjecture 2 *Let $\mathcal{S}$ and $\mathcal{T}$ be the matrices implementing the transformations $\tau \to -1/\tau$ and $\tau \to \tau + 1$ on the characters of a rational conformal field theory. Let n be the order of $\mathcal{T}$. Then the field of definition of the representation of the modular group is $\mathbb{Q}[e^{2i\pi/n}]$. If σ_a is the action of the element a of the Galois group $\mathbb{Z}_n^*$ on the representation and b is its inverse, then*

$$\sigma_a(\mathcal{T}) = \mathcal{T}^a \qquad \sigma_a(\mathcal{S}) = \mathcal{S}\mathcal{T}^b\mathcal{S}\mathcal{T}^a\mathcal{S}\mathcal{T}^b\mathcal{S}^{-1}.$$

We have no complete proof of this (if so, it would be a theorem) but there is some evidence that it is completely general. We shall give two completely different arguments and also check the consistency of this conjecture with the result of Coste and Gannon.

3.3 The first type of argument uses the fact that the restricted characters are modular functions. It breaks down when they are not linearly independent.

Lemma 3 *Let f_i, $i \in I$ be a finite number of linearly independent holomorphic functions on $\mathfrak{H}$ which carry a linear representation of $PSL_2(\mathbb{Z})$ whose kernel Γ has finite index in $PSL_2(\mathbb{Z})$. Let $\hat{\Gamma} = PSL_2(\mathbb{Z})/\Gamma$ be the quotient, which acts faithfully. Assume furthermore that the functions f_i are meromorphic at infinity (so the functions f_i are modular functions for Γ), and that their expansion at infinity has rational coefficients. Then the matrix elements (in the basis f_i) of the representation of the modular group generate a Galois extension of $\mathbb{Q}$ and Galois transformations act by automorphisms of $\hat{\Gamma}$.*

[d] And we use the same notation for both.

There is an immediate corollary.

Corollary 4 *If the restricted characters of a rational conformal field theory are well defined and linearly independent, and the modular group acts by a finite quotient $\hat{\Gamma}$, then the matrix elements of the representation of the modular group generate a finite Galois extension of $\mathbb{Q}$ and Galois transformations act by automorphisms of $\hat{\Gamma}$.*

This is much weaker than the conjecture (which states that the Galois action is by automorphisms in any rational conformal field theory and gives the explicit Galois action) but already slightly non-trivial. A physicist derivation can be found in [5], and we just give the idea of the proof :

The functions f_i have rational coefficients, so that the polynomial relations among them and with the modular invariant j are defined over $\mathbb{Q}$, i.e. they can be expanded as linear combinations of $\mathbb{Q}$-polynomial relations [e]. These relations define a Riemann surface (with punctures) which is the quotient of the upper half plane $\mathfrak{H}$ by Γ. All the automorphism of this Riemann surface fixing j are by construction given by elements of $\hat{\Gamma}$ acting linearly on the f_i's. Those linear transformations act on the polynomial relations. Start with a $\mathbb{Q}$-polynomial relation, act by an element of $\hat{\Gamma}$ and re-expand on $\mathbb{Q}$-polynomials. Then act on both sides with a Galois transformation. On one side the Galois transformation changes the matrix elements of $\hat{\Gamma}$. On the other side it changes the coefficients of the expansion but does not touch $\mathbb{Q}$-polynomial relations. This means that the Galois transform of an element of $\hat{\Gamma}$ is again an automorphism, fixing j. So it has to be an element of the finite group $\hat{\Gamma}$ again. This shows that by adding the matrix elements of the representation of $\hat{\Gamma}$ to $\mathbb{Q}$ gives a finite Galois extension, and that the Galois group of this extension acts by automorphims on $\hat{\Gamma}$.

The main feature of this argument for a physicist is that it does not depend on the Verlinde formula. Its conclusions are very modest when compared to theorem 1.

3.4 We can get more if we make more restrictive hypotheses. The results of this paragraph are motivated by a discussion with Terry Gannon [f], and Philippe Ruelle.

---

[e] This is just because to check a polynomial relation among the f_i and j one can expand in the parameter at infinity and check the vanishing term by term : this constraints the coefficients of the polynomial by linear equations defined over $\mathbb{Q}$. The same kind of argument will reappear in the next paragraph

[f] Who first realised that standard properties of modular function fields (see [6] chapter 6.) are equivalent to special cases of our conjecture.

Clearly the statement that

$$\sigma_a(T) = T^a \qquad \sigma_a(\mathcal{S}) = \mathcal{S}T^b\mathcal{S}T^a\mathcal{S}T^b\mathcal{S}^{-1}$$

induces automorphisms of $\hat{\Gamma}$ can be used to get elements of Γ, the kernel of the representation of $SL_2(\mathbb{Z})$ given by the rational conformal theory under study.

For instance, the above automorphisms make sense for $\hat{\Gamma}_n \cong SL_2(\mathbb{Z}_n)$: for $a \in \mathbb{Z}_n^*$ with inverse b the automorphism of $SL_2(\mathbb{Z}_n)$ changing T to T^a and S to $ST^bST^aST^bS^{-1}$ is

$$\begin{pmatrix} p & q \\ r & s \end{pmatrix} \in SL_2(\mathbb{Z}_n) \rightarrow \begin{pmatrix} p & aq \\ br & s \end{pmatrix}.$$

This automorphism is inner if and only if a is a square in $\mathbb{Z}_n^*$.

Now observe that if we view $SL_2(\mathbb{Z}_n)$ as a subgroup of $GL_2(\mathbb{Z}_n)$ the above automorphisms extend an become inner :

$$\begin{pmatrix} p & q \\ r & s \end{pmatrix} \in GL_2(\mathbb{Z}_n) \rightarrow \begin{pmatrix} p & aq \\ br & s \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & b \end{pmatrix} \begin{pmatrix} p & q \\ r & s \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & a \end{pmatrix}). \tag{3.1}$$

This remains true if one divides by the invariant subgroup $\{I, -I\}$.

The appearance of this group is very natural in the theory of modular functions. We refer to[6] chapter 6. for background. We start with

Lemma 5 *The quotient $\mathfrak{H}/\Gamma_n$ is a Riemann surface with punctures. There is a canonical way to fill the punctures and get a compact Riemann surface $\overline{\mathfrak{H}/\Gamma_n}$. Its field of meromorphic functions, $F_{n,\mathbb{C}}$, is a Galois extension of $\mathbb{C}(j)$ [9] with Galois group $SL_2(\mathbb{Z}_n)$.*

There exist nice generators for the extension $F_{n,\mathbb{C}}/F_{1,\mathbb{C}}$, the so-called Fricke functions $f_{r,s}(\tau)$, $r, s \in \mathbb{Z}_n$, $(r,s) \neq (0,0)$. If $\mathcal{P}(z,\tau)$ denotes the standard Weierstrass function associated to the lattice $\mathbb{C}$ with generators 1 and τ, $f_{r,s}(\tau)$ is $\mathcal{P}(z,\tau)$ at $z = \frac{r\tau+s}{n}$ (a division point of order n), times a known function of τ (but independent of n, r and s) needed to make $f_{r,s}(\tau)$ a homogeneous function of degree 0 on the set of lattices in $\mathbb{C}$. The Fricke functions are invariant under $\tau \rightarrow \tau + n$ so they can be expanded in powers of $e^{2i\pi\tau/n}$. We call this the q-expansion at infinity, by analogy with characters of conformal field theories. One can prove

Lemma 6 *The Fricke functions are invariant under Γ_n and meromorphic at the punctures, $F_{n,\mathbb{C}} = F_{1,\mathbb{C}}(f_{r,s})_{r,s}$. The Galois group $SL_2(\mathbb{Z}_n)$ permutes the*

[9] The function $j(\tau)$ is the standard modular invariant, a generator of $F_{1,\mathbb{C}}$, the function field of $\overline{\mathfrak{H}/\Gamma_1}$.

Fricke functions : it acts on the right on the row vector (r, s). The coefficients in the q-expansion of the Fricke functions at the punctures are in $\mathbb{Z}[e^{2i\pi/n}]$.

This last property makes it possible to get finer information. Define $F_{1,\mathbb{Q}} \equiv \mathbb{Q}(j)$ and $F_{n,\mathbb{Q}} \equiv F_{1,\mathbb{Q}}(f_{r,s})_{r,s}$. This is the so-called modular function field of level n.

Lemma 7 *The modular function field of level n is a Galois extension of $F_{1,\mathbb{Q}}$. The Galois group is $GL_2(\mathbb{Z}_n)/\pm I$. It permutes the Fricke functions : it acts on the right on the row vector (r, s). The intersection of the modular function field of level n with $\mathbb{C}$ is $\mathbb{Q}[e^{2i\pi/n}]$. The Galois group acts on n^{th}-roots of unity by the determinant homomorphism $GL_2(\mathbb{Z}_n)/\pm I \stackrel{\det}{\to} \mathbb{Z}_n^*$.*

We have now the necessary information to get a more precise version of lemma 3 and corollary 4.

Lemma 8 *Suppose that a familly of functions f_i satisfying the hypotheses of lemma 3, is such that furthermore there is a positive integer n for which Γ contains $\Gamma_n/\pm I$. Then $\hat{\Gamma}$ is a quotient of $SL_2(\mathbb{Z}_n)$. The matrix elements (in the basis f_i) of the representation of the modular group generate an Abelian extension of $\mathbb{Q}$ and Galois transformations act by automorphisms of $\hat{\Gamma}$ as predicted by our conjecture.*

with the immediate corollary

Corollary 9 *For a conformal field theory whose restricted satisfy the hypotheses of corollary 4, with the further assumption that there is a positive integer n such that Γ contains $\Gamma_n/\pm I$, conjecture 2 is true.*

The proof starts with two remarks.

By hypothesis, the functions $f_i, i \in I$ belong to $F_{n,\mathbb{C}}$. We use the fact that their q-expansion at infinity has rational coefficients to show that the do in fact belong to $F_{n,\mathbb{Q}}$. Try to write f_i as a quotient of two polynomials P_i and Q_i with variables j and the Fricke functions of level n. The equality $P_i = f_i Q_i$ is satisfied if and only if it becomes an identity when q-expansions at infinity are substituted on both sides. So the coefficients of P_i and Q_i have to satisfy linear equations defined over $\mathbb{Q}[e^{2i\pi/n}]$. We know that they have a solution in $\mathbb{C}$ so they have a solution in $\mathbb{Q}[e^{2i\pi/n}]$, which means that f_i belongs to $F_{n,\mathbb{Q}}$. Let us observe that as the functions f_i have no poles away from the punctures, one can take Q_i to be 1.

Now, we claim that the matrix elements of the representation of $SL_2(\mathbb{Z})$ in the basis f_i are in $\mathbb{Q}[e^{2i\pi/n}]$. The argument is analogous : write $f_i = P_i$ as above. A modular transformation on the left side gives a linear combination of the $f_j = P_j$. On the right side, a modular transformation permutes the Fricke functions. Then, we expand both sides around infinity : the matrix coefficients of the representation are constrained by linear equations defined

over $\mathbb{Q}[e^{2i\pi/n}]$. They have a unique solution in $\mathbb{C}$ because the functions f_i are linearly independent. So this solution has to be in $\mathbb{Q}[e^{2i\pi/n}]$.

After these preliminaries, let us act on the functions f_i with an arbitrary element M of the Galois group $GL_2(\mathbb{Z}_n)/\pm I$. We can write

$$M = M_l \begin{pmatrix} 1 & 0 \\ 0 & \det M \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \det M \end{pmatrix} M_r.$$

The matrices M_l and M_r are in $SL_2(\mathbb{Z}_n)$. Acting on f_i with the first decomposition, only M_l acts because the other factor is a Galois transformation on the coefficients of the q expansion of f_i at infinity, and this involves only rationals. Acting on f_i with the second decomposition, M_r expands f_i in terms of the f_j. Then the second factor acts only on the coefficients of this expansion (because as above it acts trivially on f_j) as a standard Galois transformation. By linear independence of the functions f_i, this proves that the Galois transform of the matrix representing M_r on the f_i by $\det M$ is the matrix representing M_l on the f_i, giving exactly the transformation (3.1). This finishes the proof.

Under the hypotheses of 4, it has been observed long ago in concrete examples that if n is the order of T in the conformal field theory, Γ_n and Γ are closely related. However, Γ_n does not always act trivially. Sometimes, some characters carry a non-trivial one dimensional representation of Γ_n. However, there may be a nice generalisation of the above arguments to cover such cases.

It should be interesting to study the following family of groups. For any positive integer n, let G_n be the groups given by generators and relations as follows. The generators are S, T, and M_a for $a \in \mathbb{Z}_n^*$. Now come the relations. First the subgroup of G_n generated by S and T is a quotient of $SL_2(\mathbb{Z})$, and $T^n = 1$. Second, S^2 is central. Third, the map $a \to M_a$ from $\mathbb{Z}_n^*$ to G_n is a one to one group homomorphism. Fourth, for $a \in \mathbb{Z}_n^*$, with inverse b, $SM_a = M_a S T^b S T^a S T^b S^{-1}$ and $TM_a = M_a T^a$.

By construction, the map

$$S \to \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \qquad T \to \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad M_a \to \begin{pmatrix} 1 & 0 \\ 0 & a \end{pmatrix}$$

extends to a group homomorphism from G_n to $GL_2(\mathbb{Z}_n)$, and this homomorphism is onto.

A simpler presentation of G_n would be desirable. Simplicity suggest that G_n might be a finite central extension of $GL_2(\mathbb{Z}_n)$.

It would also be nice to know if G_n appears naturally as a Galois group of an extension of $\mathbb{Q}(j)$ or a more intricate object related to modular forms may be.

Anyway, the relationship between Γ_n and Γ will enter our discussion again in section 6 on finite quantum mechanics. The plausibility of this relationship is enhanced by theorem 1. In fact, the elements of $\hat{\Gamma}_n$ invariant under the automorphisms (3.1) are the diagonal matrices, and those are exactly of the form $\sigma_a(S)S^{-1}$. In a representation of $\hat{\Gamma}_n$ having the same properties as the ones we attribute to $\hat{\Gamma}$, the representative of $\sigma_a(S)S^{-1}$, a Galois invariant, should have rational entries. Theorem 1 says that in the conformal field theory context $\sigma(\mathcal{S})\mathcal{S}^{-1}$ is just a signed permutation matrix, so has (very special) rational entries. This analogy suggests that the matrices $M(\sigma)_{ij} = \varepsilon_\sigma(i)\delta_{\sigma(i),j}$ are the only rational matrices in the representation of $\hat{\Gamma}$.

3.5 A modular invariant partition functions is associated to a matrix $\mathcal{N}$ with (non-negative) integral coefficients commuting with $\mathcal{S}$ and $\mathcal{T}$. Such a matrix $\mathcal{N}$ is fixed by Galois transformations and commutes with any Galois transform of $\mathcal{S}$ (resp. $\mathcal{T}$) if it commutes with $\mathcal{S}$ (resp. $\mathcal{T}$). Thus our conjecture does not say anything new for modular invariant partition functions. This is disappointing.

On the other hand, we observe that if the conjecture is correct, the question of commuting with the action of the modular group can always be expressed by linear equations with rational coefficients. If U is an element of $SL_2(\mathbb{Z})$, $\mathcal{U}$ its representative and $a \in \mathbb{Z}_n^*$ (where n is the order of $\mathcal{T}$) an element of the Galois group, we know that $\sigma_a(\mathcal{U})$ also represents an element of $SL_2(\mathbb{Z})$. So a matrix H commuting with the action of $SL_2(\mathbb{Z})$ commutes with

$$\mathcal{U}(q) \equiv \sum_{a \in \mathbb{Z}_n^*} e^{2i\pi qa/n} \sigma_a(\mathcal{U})$$

for any $q \in \mathbb{Z}_n$. But obviously $\mathcal{U}(q)$ is Galois invariant, so has rational entries. Conversely, as

$$\mathcal{U} = \frac{1}{n} \sum_{q \in \mathbb{Z}_n} e^{-2i\pi q/n} \mathcal{U}(q)$$

it is clear that the commutation with the action of $SL_2(\mathbb{Z})$ is equivalent to the commutation with rational matrices. This is the explanation for the observation made by Claude Itzykson and mentioned in the introduction : why is the commutant over the rationals as large as the commutant over the complex numbers ?

Let us note that we do not use the explicit formulæ of the Galois action : we use that the extension is Abelian (also known to be true from theorem 1) and that the Galois action is by automorphisms (which we know how to prove under the hypotheses of corollary 4).

3.6 The second argument is our main point and occupies the rest of the paper. It involves the description of a long known explicit family of representations (related to the so-called metaplectic representations) of the modular group for which the conjecture is true. Then we argue that this family is large enough to establish the main conjecture for rational conformal field theories associated to the Wess–Zumino–Witten models for simply connected semi-simple groups. Our line of attack works for certain coset models and orbifolds as well as well, but the situation is much more complicated. We have been forced to this because of the problem of degeneracy of restricted characters.

It is part of the standard lore that all rational conformal field theories are obtained from Wess–Zumino–Witten models by clever combinations of coset and orbifold constructions. So our results are already encouraging, but a uniform proof instead of our case by case approach would be highly desirable. A first step in this direction has been accomplished by Degiovanni. He has proved our conjecture for orbifolds of holomorphic rational conformal field theories [8] and hopefully general orbifolds are tractable as well.

This is the evidence we have for the validity of the main conjecture.

4 A class of representations of $SL_2(\mathbb{Z})$.

4.1 This section is devoted to the construction of a family of representations of $SL_2(\mathbb{Z})$ that we use in the sequel.

Let M be a finite Abelian group, and m be theexponent of M, that is, the smallest positive integer such that

$$mx = 0 \; \forall x \in M.$$

Let ζ be a primitive $2m^{th}$ root of unity and set $\xi = \zeta^2$. Let B be a map from M to $\mathbb{Z}_{2m}$ and $(\ |\ )$ a map from $M \times M$ to $\mathbb{Z}_m$ such that

1. B is a quadratic form i.e. $B(ax) = a^2 B(x) \; \forall x \in M, \forall a \in \mathbb{Z}$.
2. $(\ |\ )$ is symmetric and bilinear i.e. $(x|y) = (y|x)$ and $(x|y+z) = (x|y) + (x|z) \; \forall x, y, z \in M$.
3. The functions $\{\chi_x\}_{x \in M}$ from M to $\mathbb{C}$ defined by $\chi_x(y) = \xi^{(x|y)}$ are the characters of M.
4. $B(x+y) - B(x) - B(y) = 2(x|y) \; \forall \, x, y \in M$

Definition 10 *The data $\{M, B, (\ |\ ), \zeta\}$ will be called a quadratic structure in the sequel.*

4.2 To any such quadratic structure, one can associate a representation of the modular group $SL_2(\mathbb{Z})$ as follows. We start with a finite dimensional Hilbert space H with an orthonormal basis $\{e_x\}_{x\in M}$ indexed by the elements of M. We define two endomorphisms of H, $\mathcal{S}$ and $\mathcal{T}$ represented by unitary symmetric matrices

$$\mathcal{S}_{x,y} = \frac{1}{|M|^{1/2}}\xi^{-(x|y)} \qquad \mathcal{T}_{x,y} \equiv \alpha^{-1}\tilde{\mathcal{T}}_{x,y} = \alpha^{-1}\zeta^{B(x)}\delta_{x,y} \tag{4.1}$$

where α is a fixed cube root of

$$\beta = \frac{1}{|M|^{1/2}}\sum_{x\in M}\zeta^{B(x)}. \tag{4.2}$$

It is not obvious that $\beta \neq 0$ (which is needed to define $\mathcal{T}$). Once this is proved, we build a representation of the modular group with generators $\mathcal{S}$ and $\mathcal{T}$. Note that $\mathcal{S}$ is a finite Fourier transform.

Lemma 11 *The number β is a root of unity. The map $S \to \mathcal{S}$ and $T \to \mathcal{T}$ extends to a representation of $SL_2(\mathbb{Z})$.*

For this we have to check the relations.

First a simple computation shows that $\left(\mathcal{S}^2\right)_{x,y} = \delta_{x,-y}$. So $\mathcal{S}^4 = 1$, and as $B(-x) = B(x)$, $\mathcal{S}^2$ commutes with $\mathcal{T}$. Note that $\mathcal{S}$ is symmetric and unitary. Now repeated application of $B(x+y) - B(x) - B(y) = 2(x|y)$ shows that

$$\left(\mathcal{S}\tilde{\mathcal{T}}\right)^3 = \frac{1}{|M|^{1/2}}\sum_{x\in M}\zeta^{B(x)}\mathcal{S}^2 = \beta\mathcal{S}^2.$$

Taking the determinant of both sides, we see that β is a root of unity because $\mathcal{S}$ and $\tilde{\mathcal{T}}$ are of finite order. As $\mathcal{T}$ is normalised by $\alpha^3\beta = 1$ the map $S \to \mathcal{S}$, $T \to \mathcal{T}$ extends to a representation of $SL_2(\mathbb{Z})$.

The order of $\mathcal{S}$ divides 4 and the order of $\tilde{\mathcal{T}}$ divides $2m$ so we have shown that the order of β divides $4m|M|$. But this bound is very poor. Later, we shall prove that $\beta^8 = \alpha^{24} = 1$. But to do that, we need further arithmetic properties of the representation of $SL_2(\mathbb{Z})$ associated to a quadratic structure. Anyway, given β, there are three possible values for α [h] and we fix one of them.

4.3 This construction has a property analogous to the so-called arithmetic functions : if m can be decomposed as $m = pq$ were p and q are relatively prime integers, there is a corresponding decomposition of the quadratic form,

[h] In the context of rational conformal field theorie, this amounts to the possibility of tensoring with the Wess–Zumino–Witten model E_8 at level 1.

the scalar product and the representation of $SL_2(\mathbb{Z})$. We describe this factorisation explicitly in appendix C, and use it in the next paragraph to show that $\beta^8 = 1$.

4.4 We already know that β is a root of unity, but the proof that its order divides 8 is a little bit complicated.

To build the argument, we use the splitting properties explicated in appendix C. They make clear that the property $\beta^8 = 1$ will be true for any m if it is true whenever m is odd or a power of 2. So we only prove those two special cases.

Lemma 12 *If m is odd, $\beta^4 = 1$.*

If m is odd, multiplication by 2 is an automorphism of M so

$$\beta = \frac{1}{|M|^{1/2}} \sum_{x \in M} \zeta^{B(2x)} = \frac{1}{|M|^{1/2}} \sum_{x \in M} \xi^{2(x|x)}.$$

This is nothing but the trace of the matrix $\mathcal{S}$ associated to the new quadratic structure $(M, B, (\ |\ ), -\zeta^2)$[i]. But $\mathcal{S}$ is of order 4 so its trace is in $\mathbb{Z}[i]$. Thus $\beta^4 = 1$ if m is odd.

Lemma 13 *If m is a power of 2, $\beta^8 = 1$.*

If m is a power of 2, then so is $|M|$[j]. Hence, the order of β, which divides $8m|M|$, is a power of 2 as well. Also $\sqrt{2} = e^{2i\pi/8} + e^{-2i\pi/8}$. Now Galois theory comes into the game. Let a be an odd integer. Raising the roots of unity involved in the equation defining β to the a^{th} power is a Galois transformation. Assume that a is the square of an odd integer. Then $a = 1 \bmod 8$ so the action on $\sqrt{2}$ is trivial. Moreover, in the sum $\sum_{x \in M} \zeta^{aB(x)}$ a can be re-absorbed in the definition of x, so the sum is invariant. Hence $\beta^a = \beta$ whenever a is the square of an odd integer. This implies that $\beta^8 = 1$ if m is a power of 2 (take for instance $a = 9$).

Using arithmetic factorisation we get the following

Corollary 14 *For any quadratic structure, $\beta^8 = \alpha^{24} = 1$.*

---

[i] Because m is odd, $-\zeta^2$ is a primitive $2m^{th}$ root of unity, then $(-\zeta^2)^2 = \xi^2$.

[j] This comes from the structure theorem for finite Abelian groups, due to Kroenecker. But an elementary argument goes as follows : we show that any prime r dividing $|M|$ divides m or equivalently that $|M|$ divides a certain power of m. Set $M_0 \equiv M$. Suppose we have constructed M_i for $i = 0, \cdots, r$ and x_i for $i = 1, \cdots, r$. If M_r is the trivial group, we stop. If not we choose a non-zero element $x_{i+1} \in M_r$ and define M_{r+1} as the quotient of M_r by the subgroup generated by x_{r+1}. As M_{i+1} is a proper quotient of M_i this process terminates after a finite number of steps, say s. As m annihilates $M = M_0$, m annihilates M_i for $i = 1, \cdots, s$. Thus the order of x_i divides m for $i = 1, \cdots, s$. So $|M_i|$ divides $m|M_{i+1}|$ for $i = 0, \cdots, s-1$ and finally $|M|$ divides m^s.

5 Galois transformations

In this section, we show that the conjecture 2 is true for representations associated to quadratic structure. More precisely we prove

Lemma 15 *For any quadratic structure* $\{M, B, (\ |\ ), \zeta\}$ *and for any choice of* α, *the matrix elements of the associated representation of the modular group belong to* $\mathbb{Q}[e^{2i\pi/n}]$ *where* n *is the order of* $\mathcal{T}$. *If* σ_a *is the Galois transformation associated to* $a \in \mathbb{Z}_n^*$ *with inverse* b *then*

$$\sigma_a(\mathcal{T}) = \mathcal{T}^a \qquad \sigma_a(\mathcal{S}) = \mathcal{S}\mathcal{T}^b\mathcal{S}\mathcal{T}^a\mathcal{S}\mathcal{T}^b\mathcal{S}^{-1}.$$

Let n be the order of $\mathcal{T}$ and n' a common multiple of 24 and $2m$. The matrix elements of $\mathcal{S}$ and $\mathcal{T}$ can be expressed in terms of α, which is such that $\alpha^{24} = 1$ and ζ, a primitive $2m^{th}$ root of unity. For instance, formula (4.2) expresses $|M|^{1/2}$ in terms of α and ζ. All the matrix elements of the representation are obtained from products of $\mathcal{S}$ and $\mathcal{T}$, so that the extension K of $\mathbb{Q}$ generated by the matrix elements of the representation is contained in $\mathbb{Q}[e^{2i\pi/n'}]$. In particular, any Galois transformation of the matrix elements can be represented (usually in more than one way) by an element a of $\mathbb{Z}_{n'}^*$, acting on the roots of unity involved by raising to the a^{th} power. We denote by σ_a the Galois transformation corresponding to a. Let b be the inverse of a in $\mathbb{Z}_{n'}^*$.

We start with the computation of $\mathcal{S}\mathcal{T}^b\mathcal{S}\mathcal{T}^a\mathcal{S}\mathcal{T}^b\mathcal{S}^{-1}$. The case when $a = b = 1$ has already been considered in the previous section. By repeated use of $B(x+y) - B(x) - B(y) = 2(x|y)$ one finds that

$$\left(\mathcal{S}\mathcal{T}^b\mathcal{S}\mathcal{T}^a\mathcal{S}\mathcal{T}^b\right)_{xy} = \frac{\alpha^{-a-2b}}{|M|^{3/2}} \sum_{u,v} \xi^{-(u|x+by)} \zeta^{bB(u+y-av)}.$$

We take v and $w = u + y + av$ as independent variables and set

$$\beta_b = \frac{1}{|M|^{1/2}} \sum_{x \in M} \zeta^{bB(x)}.$$

Then the right-hand side of the above equality is just

$$\alpha^{a+2b} \beta_b \delta_{x,-by}.$$

Hence

$$\left(\mathcal{S}\mathcal{T}^b\mathcal{S}\mathcal{T}^a\mathcal{S}\mathcal{T}^b\mathcal{S}^{-1}\right)_{xy} = \alpha^{-a-2b} \left(\sum_{z \in M} \zeta^{bB(z)} \right) \frac{1}{|M|} \xi^{-a(x|y)}.$$

It is easy to apply σ_b to the right-hand side. For instance

$$\sigma_b(\alpha^{-a-2b}) = \alpha^{-1-2b^2}.$$

Also

$$\sigma_b(\sum_{z\in M} \zeta^{bB(z)}) = \sum_{z\in M} \zeta^{b^2B(z)} = |M|^{1/2}\alpha^3$$

where in the last equality we have used the quadratic property of B. Finally

$$\sigma_b(\xi^{-a(x|y)}) = \xi^{-(x|y)}.$$

Putting all this together, we see that

$$\sigma_b\left(\mathcal{S}\mathcal{T}^b\mathcal{S}\mathcal{T}^a\mathcal{S}\mathcal{T}^b\mathcal{S}^{-1}\right)_{xy} = \alpha^{2(1-b^2)}\mathcal{S}_{xy}.$$

Now b is prime to 24, so $b^2-1 = 0 \bmod 24$[k] and $\alpha^{2(1-b^2)} = 1$. This proves that

$$\sigma_a\left(\mathcal{S}_{xy}\right) = \left(\mathcal{S}\mathcal{T}^b\mathcal{S}\mathcal{T}^a\mathcal{S}\mathcal{T}^b\mathcal{S}^{-1}\right)_{xy}.$$

The right-hand side is unchanged if a and b are changed by a multiple of n, the order of $\mathcal{T}$. Moreover it is plain that $\sigma_a(\mathcal{T}_{xy}) = (\mathcal{T}^a)_{xy}$. This means that the matrix elements of $\mathcal{S}$ are in $\mathbb{Q}[e^{2i\pi/n}]$ (if not, the above Galois transformations could not be exhaustive) and that the representations of $SL_2(\mathbb{Z})$ (corresponding to the possible values of α) associated to a quadratic structure satisfy the conjecture.

6 Finite quantum mechanics

6.1 The purpose of this section is to emphasise the close relationship between representations of $SL_2(\mathbb{Z})$ associated to quadratic structures and principal congruence subgroups. This gives an alternative way to interpret Galois actions.

6.2 We start from a quadratic structure $\{M, B, (\ |\ ), \zeta\}$, and the Hilbert space H with basis $\{e_x\}_{x\in M}$ indexed by the elements of M. The notation are as in paragraph 4.1 and 4.2. We define a discrete analog of the canonical commutation relations of quantum mechanics : the momenta $\{\mathcal{P}^x\}_{x\in M}$ and positions $\{\mathcal{Q}^x\}_{x\in M}$ are two families of unitary operators on H representing M. Their action on H is given by

$$\mathcal{P}^x e_y = e_{x+y} \qquad \mathcal{Q}^x e_y = \xi^{-(x|y)} e_y \tag{6.1}$$

[k] Claude Itzykson enjoyed this elementary fact and used it in several instances, see e.g. [5].

and they satisfy

$$\mathcal{P}_x \mathcal{Q}_y = \xi^{(x|y)} \mathcal{Q}_y \mathcal{P}_x. \tag{6.2}$$

The operators $\{\mathcal{P}^x \mathcal{Q}^y\}_{x,y \in M}$ form an orthonormal basis of End H for the trace. First Tr $\mathcal{P}^x \mathcal{Q}^y$ is easily seen to be 0 unless $x = y = 0$ and then the trace is dim $H = |M|$. Then the cyclic property of the trace gives

$$\text{Tr } \mathcal{Q}^{-y} \mathcal{P}^{-x} \mathcal{P}^{x'} \mathcal{Q}^{y'} = |M| \delta^{x,x'} \delta^{y,y'}.$$

6.3 It is obvious that $\mathcal{T}$ commutes with position operators and a simple computation shows that $\mathcal{T}^{-1} \mathcal{P}^x \mathcal{T} = \zeta^{-B(x)} \mathcal{P}^x \mathcal{Q}^x$. As in ordinary quantum mechanics, the finite Fourier transform $\mathcal{S}$ exchanges the momentum and position representations so $\mathcal{S}^{-1} \mathcal{Q}^x \mathcal{S} = \mathcal{P}^x$ and $\mathcal{S}^{-1} \mathcal{P}^x \mathcal{S} = \mathcal{Q}^{-x}$. Putting things together

$$\mathcal{T}^{-1} \mathcal{P}^x \mathcal{Q}^y \mathcal{T} = \zeta^{-B(x)} \mathcal{P}^x \mathcal{Q}^{x+y} \qquad \mathcal{S}^{-1} \mathcal{P}^x \mathcal{Q}^y \mathcal{S} = \xi^{(x|y)} \mathcal{P}^y \mathcal{Q}^{-x}. \tag{6.3}$$

6.4 On the other hand, there is a canonical right action of $SL_2(\mathbb{Z})$ on $M \times M$ given by $(x, y) \to (x, y)U$ for $(x, y) \in M \times M$ and $u \in SL_2(\mathbb{Z})$. We transfer this action on End (End H) by acting on the exponents of $\mathcal{P}^x \mathcal{Q}^y$. We denote this action by $(\mathcal{P}^x \mathcal{Q}^y)_U$: for

$$U = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}) \tag{6.4}$$

we have

$$(\mathcal{P}^x \mathcal{Q}^y)_U = \mathcal{P}^{ax+cy} \mathcal{Q}^{bx+dy}. \tag{6.5}$$

we observe the striking analogy with the action of the modular group on Fricke functions, recalled in paragraph 3.4.

6.5 The interesting fact is that this action coïncides up to phases with the adjoint action associated to the quadratic structure. So if $U \in SL_2(\mathbb{Z})$ is represented by $\mathcal{U}$ on the quadratic structure then

$$\mathcal{U}^{-1} \mathcal{P}^x \mathcal{Q}^y \mathcal{U} \propto (\mathcal{P}^x \mathcal{Q}^y)_U \tag{6.6}$$

It is sufficient to check this on the generators S and T and this is just the content of equation (6.3).

6.6 The proportionality factor $Coc(U,(x,y))$ is a cocycle, which means that

$$Coc(U,(x,y))Coc(V,(x,y)U) = Coc(UV,(x,y)).$$

If $\mathcal{U}$ belongs to Aut H, we denote by $\ell(\mathcal{U})$ the element of End (End H) which is conjugation by $\mathcal{U}$. We restrict now to the case when $\mathcal{U}$ is the action of some $U \in SL_2(\mathbb{Z})$ on H. Then $\ell(\mathcal{U})$ gives a right action of $SL_2(\mathbb{Z})$ on End H, and we have

$$\ell(\mathcal{U})(\mathcal{P}^x\mathcal{Q}^y) = Coc(U,(x,y))(\mathcal{P}^x\mathcal{Q}^y)_U \tag{6.7}$$

There is a trick to compute the cocycle. It is to do as if $\zeta^{(x|y)}$ were a well-defined quantity (only its square is). This is because $[x,y] \equiv \zeta^{-(x|y)}\mathcal{P}^x\mathcal{Q}^y$ transforms formally without any phase under the adjoint action of $\mathcal{S}$ and $\mathcal{T}$ [1]. This leads to guess that

$$\mathcal{U}^{-1}\mathcal{P}^x\mathcal{Q}^y\mathcal{U} = \zeta^{-abB(x)-cdB(y)}\xi^{-\frac{ad+bc-1}{2}(x|y)}\mathcal{P}^{ax+cy}\mathcal{Q}^{bx+dy}. \tag{6.8}$$

One can check that this phase works for $\mathcal{S}$ and $\mathcal{T}$, and has the cocycle property, so the above formula is correct.

6.7 The above formula is rather remarkable, because in general it is very hard to reconstruct the action of a generic element of $SL_2(\mathbb{Z})$ starting only from the action of S and T. In particular, we see that if U belongs to Γ_m, the principal congruence subgroup of level m, $\mathcal{U}$ commutes or anti-commutes with P^xQ^y. If U belongs to Γ_{2m}, $\mathcal{U}$ is in the center of End H so it is a scalar matrix. There is a reciprocal. If U is represented by a scalar matrix, then first U belongs to Γ_m and second $\zeta^{-abB(x)-cdB(y)}\xi^{-\frac{ad+bc-1}{2}(x|y)} = 1$ for any $x,y \in M$. The first condition implies the second if and only if the quadratic form B is even (that is $B(x) \in 2\mathbb{Z}_{2m}$ for any $x \in M$). If m is odd, B is always even. If m is even and B odd, let Γ'_m be the invariant subgroup of $SL_2(\mathbb{Z})$ made of matrices in Γ_m with off-diagonal entries equal to 0 modulo $2m$ (and not only mod m). We conclude that U is represented by a scalar matrix if and only if either B is even and $U \in \Gamma_m$ or m is even, B is odd and $u \in \Gamma'_m$. Thus $\hat{\Gamma}$, the group acting effectively (i.e. the quotient of $SL_2(\mathbb{Z})$ by the kernel of the representation associated to the quadratic structure), is a central extension of $\hat{\Gamma}_m$ or $\hat{\Gamma}'_m$ depending on M and B, a connection announced in paragraph 3.4. The relationship could be made more precise, but there is no need for this in the sequel.

[1] This means that the cocycle is formally trivial, in fact it is trivial up to signs.

6.8 We are now in position to do Galois theory using finite quantum mechanics. The matrix elements of momenta are integers. The matrix elements of positions are 0 off diagonal and roots of unity on the diagonal. So Galois transformations don't act on momenta and act on positions by raising to a power. We act on both sides of equation (6.8) with σ_r, raising all roots of unity involved to the power r. The outcome is

$$\sigma_r(\mathcal{U}^{-1})\mathcal{P}^x\mathcal{Q}^{ry}\sigma_r(\mathcal{U}) = \zeta^{-rabB(x)-rcdB(y)}\xi^{-r\frac{ad+bc-1}{2}(x|y)}\mathcal{P}^{ax+cy}\mathcal{Q}^{rbx+rdy}. \tag{6.9}$$

Now, we recall that the adjoint action $\ell(\mathcal{U})$ descends to $\hat{\Gamma}_{2m}$, and write $\hat{U}$ for the projection of U. Let σ_s be the inverse of σ_r and set $z = ry$ (so $y = sz$) to get

$$\ell(\sigma_r(\mathcal{U}))(\mathcal{P}^x\mathcal{Q}^z) = Coc(\sigma_r(\hat{U}),(x,y))(\mathcal{P}^x\mathcal{Q}^y)_{\sigma_r(\hat{U})} \tag{6.10}$$

where $\sigma_r(\hat{U})$ is defined by

$$\hat{U} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \to \sigma_r(\hat{U}) = \begin{pmatrix} a & rb \\ sc & d \end{pmatrix} \tag{6.11}$$

as announced in paragraph 3.4.

Of course, this is slightly weaker than the results from the previous section, because finite quantum mechanics works up to phases.

7 Generalisations

7.1 Up to now, we have only been dealing with representations of $SL_2(\mathbb{Z})$ associated to quadratic structures. The purpose of this section is to work in the category $\mathfrak{R}$ of representations of $SL_2(\mathbb{Z})$ that have the conjectured behaviour under Galois transformations and see that it is closed under a certain number of constructions. First $\mathfrak{R}$ contains the one-dimensional representations of $SL_2(\mathbb{Z})$ as well as the ones associated to quadratic structures. Second $\mathfrak{R}$ is stable under direct sums and and tensor products, under duals, and under passing to sub- or quotient representations.

A certain familiarity with appendix B is useful to read this section.

7.2 Before we give a precise definition of $\mathfrak{R}$, let us make a few remarks. Up to now, we have given explicit bases in the representations of $SL_2(\mathbb{Z})$ we have been dealing with. But as far as our conjecture is concerned, nothing would have changed under a change of basis with coefficients in $\mathbb{Q}$. This is a sign of the flexibility of our computations and we shall use it in the sequel. But it is also a drawback because the situation is more rigid in conformal field

theory : there is a prefered basis anyway, given by the characters and in which the representation is unitary. For instance in theorem 1 the fact that Galois transformations permute rows of $\mathcal{S}$ is of course not true after a base change in general. This explains why we don't insist on unitarity, symmetry and all that in the sequel.

7.3 A member of $\mathfrak{R}$ is made of a $\mathbb{Q}$-vector space V, a cyclotomic extension of $K = \mathbb{Q}[e^{2i\pi/n}]$ (where n is a positive integer) and a representation R of $SL_2(\mathbb{Z})$ on $V^K \equiv V \otimes_{\mathbb{Q}} K$ [m] such that $\mathrm{Gal}\, K/\mathbb{Q}$ acts on $\mathcal{T}$ and $\mathcal{S}$ (representing T and S on V^K) as predicted in conjecture 2 i.e. as follows : n is the order of $\mathcal{T}$ and (denoting by σ_a the element of $\mathrm{Gal}\, K/\mathbb{Q}$ sending $e^{2i\pi/n}$ to $e^{2i\pi a/n}$ for $a \in \mathbb{Z}_n^*$ with inverse b)

$$\sigma_a(\mathcal{T}) = \mathcal{T}^a \qquad \sigma_a(\mathcal{S}) = \mathcal{S}\mathcal{T}^b\mathcal{S}\mathcal{T}^a\mathcal{S}\mathcal{T}^b\mathcal{S}^{-1}. \tag{7.1}$$

The following elementary observation turns out to be useful because in general the order of the representative of T is affected by the constructions leaving $\mathfrak{R}$ stable. Suppose we have a representation of $SL_2(\mathbb{Z})$ on $V^{K'}$ where $K' = \mathbb{Q}[e^{2i\pi/n'}]$ and that the formulæ (7.1) hold for any integer a prime to n'. Let n be the order of $\mathcal{T}$. As σ_a is periodic of period n' whereas powers of $\mathcal{T}$ are periodic of period n, the only possibility is that n' is a multiple of n. Moreover, if some matrix element (in any basis made of elements of V) of $\mathcal{T}$ and $\mathcal{S}$ were not in $K = \mathbb{Q}[e^{2i\pi/n}]$, formulae (7.1) could not reproduce all the Galois transformations. Thus, $SL_2(\mathbb{Z})$ acts on V^K to give an element of $\mathfrak{R}$. So one can freely relax the condition $n = n'$ to the condition that n' is a multiple of n in the definition of $\mathfrak{R}$. That's what we do in the sequel.

7.4 We start with

Lemma 16 *One-dimensional representations of* $SL_2(\mathbb{Z})$ *are in* $\mathfrak{R}$.

This is established by a simple direct calculation. If V is one-dimensional End V^K is canonically isomorphic to K so there is no choice of basis to make. We start from numbers t and s (representing T and S). They commute so they have to satisfy $s = t^{-3}$ and $s^4 = t^{-12} = 1$. So t is a root of unity and s belongs to $\mathbb{Q}[t]$. We take $n' = 12$. A given Galois transformation on t can be represented by an integer prime to 12, say a. Then a is its own inverse. We have to check that $s^a = st^a st^a st^a s^{-1}$. The right hand side is $s^2 t^{3a} = s^{2-a}$, which is s^a because a is odd and $s^4 = 1$.

It is clear from the results of section 5 that

[m] Remember that Galois transformations in $\mathrm{Gal}\, K/\mathbb{Q}$ act on the tensor product by acting on K, leaving V pointwise fixed.

Lemma 17 *Representations of $SL_2(\mathbb{Z})$ associated to quadratic structures belong to $\mathfrak{R}$.*

7.5 We turn to direct sums and tensor products.

Lemma 18 *Let (V_i, n_i', R_i), $i = 1, 2$ be two members of $\mathfrak{R}$. Let n' be a common multiple of n_1' and n_2'. Then the direct sum $(V1 \oplus V_2, n', R_1 \oplus R_2)$ and the tensor product $(V1 \otimes V_2, n', R_1 \otimes R_2)$ are again in $\mathfrak{R}$.*

As $K = \mathbb{Q}[e^{2i\pi/n'}]$ is just the compositum of $\mathbb{Q}[e^{2i\pi/n_1'}]$ and $\mathbb{Q}[e^{2i\pi/n_2'}]$, the direct sum and tensor product representations are realised in $(V_1 \oplus V_2)^{K'}$ and $(V_1 \otimes V_2)^{K'}$ respectively. A given Galois transformation can be implemented on both representations by an integer a prime to n'. Formulæ (7.1) for the direct sum and tensor product representation are then direct consequences of the membership of the original representations to $\mathfrak{R}$. Note that the order n of T is the least common multiple of n_1 and n_2 in the direct sum. But in general, for a tensor product the order n of T simply divides the least common multiple of n_1 and n_2.

It is clear that Galois transformations commute with transposition and inversion of matrices so it is clear that

Lemma 19 *The category $\mathfrak{R}$ is stable under duals.*

7.6 The stability of $\mathfrak{R}$ for sub-representations and quotient representation is again essentially seen by manipulating definitions. We have

Lemma 20 *If (V, n', R) is a member of $\mathfrak{R}$, $K \equiv \mathbb{Q}[e^{2i\pi/n'}]$ and V_0 is a subspace of V such that V_0^K is an invariant subspace for R then $(V_0, n', R_{|V_0^K})$ and $(V/V_0, n', R/R_{|V_0^K})$ are again members of $\mathfrak{R}$.*

We let $\mathrm{End}_0\, V$ be the space of endomorphisms of V sending V_0 into V_0. As recalled in paragraph B.5, the action of $\mathrm{Gal}\, K/k$ commutes with the projections $\mathrm{End}_0\, V^K \to \mathrm{End}\, V_0^K$ and $\mathrm{End}_0\, V^K \to \mathrm{End}\, (V/V_0)^K$. So the elements of $\mathrm{Gal}\, K/k$ act as predicted by formulæ (7.1) on the sub- and quotient representation. Again this implies that, although the order of the representative of T may be different in the sub- or quotient representation and in the original representation, we end up with members of $\mathfrak{R}$.

This finishes the proof that $\mathfrak{R}$ is stable under sub- and quotient representations.

8 Applications to conformal field theories

8.1 In this last section, we illustrate the above constructions on concrete conformal field theories. We start with $\hat{\mathfrak{su}}(n)_1$ and make some remarks on

models for which the fusion rules are the algebra of representations of a finite Abelian group. Then we deal with Wess–Zumino–Witten models for simply-connected groups. We finish with examples of the coset constructions and remarks on fixed point resolutions.

8.2 Quadratic structures apply directly to the $\hat{\mathfrak{su}}(N)_1$ model. Basic references are [9,10]. This theory contains N characters χ_i, $i = 0 \cdots N-1$. The restricted characters χ_i and χ_{N-i} are equal for $i = 1, \cdots, N-1$ because of complex conjugation. The central charge is $c = N-1$ and the conformal weights are

$$h_i = \frac{i(N-i)}{2N}.$$

The defines the action of T :

$$\mathcal{T}_{jk} = \delta_{jk} e^{2i\pi(\frac{i(N-j)}{2N} - \frac{N-1}{24})}.$$

The action of S is given by a finite Fourier transform :

$$\mathcal{S}_{jk} = \frac{1}{N^{1/2}} e^{2i\pi jk/N}.$$

To make contact with quadratic structures, we identify the elements of $\{0, \cdots, N-1\}$ with their representatives in $\mathbb{Z}_N$ which we take to be M. So $m = N$. We choose $B(i) \equiv i(N-i) \equiv (N-1)i^2 \bmod 2N$ [n] and $(i|j) \equiv -ij \bmod N$. Taking $\zeta = e^{i\pi/N}$ and $\alpha = e^{i\pi(N-1)/12}$ we associate to this quadratic structure the representation of the modular group for $\hat{\mathfrak{su}}(n)_1$. So $\hat{\mathfrak{su}}(n)_1$ theories satisfy conjecture 2.

8.3 Taking tensor products for different values of N, we see that any finite Abelian group M can appear in a conformal field theory, but with a very special choice of quadratic form and scalar product. However, it is unlikely that any quadratic structure can appear in conformal field theories [o]. On the other hand, representations of $SL_2(\mathbb{Z})$ associated to quadratic structures share a lot of properties with those appearing in rational conformal field theories. They are unitary , $\mathcal{S}$ is symmetric, the matrix elements $\mathcal{S}_{0x}$ are real and positive so one can take the 0 element of M to play the role of the identity operator, and

[n] Because B has values in $\mathbb{Z}_{2N}$ and not simply in $\mathbb{Z}_N$, one has to check that B is indeed well defined and quadratic. This is a trivial computation in this case.

[o] A study of this question requires a classification of quadratic structures first. We shall not embark on this problem here, although it looks tractable, and is likely to be part of standard mathematics.

the fusion rules are just addition in M so they define an associative algebra with integral non-negative entries. Thus it seems that even if conjecture 2 is true for all rational conformal field theories, the behaviour under Galois transformations is not enough to characterise the representations of the modular group coming from conformal field theories.

8.4 In this paragraph, we prove that if $\mathring{\mathfrak{g}}$ is a complex simple Lie algebra, the representation of $SL_2(\mathbb{Z})$ carried by the characters of the integrable representations of the associated affine algebra $\mathfrak{g}$ at level k (a non-negative integer) is in $\mathfrak{R}$. The proof for the twisted case would follow the same pattern with slight additional complications. We refer to [9], especially Chapter 13, for background.

The starting point is a lattice L [p] in a real vector space $\mathring{\mathfrak{h}}$ of dimension ℓ, endowed with a positive definite bilinear form $(\ |\ )$ making L an integral even lattice. This means that $(a|b) \in \mathbb{Z}$ and $(a|a) \in 2\mathbb{Z}$ for any $a, b \in L$. Let $L^* = \{x \in \mathring{\mathfrak{h}}, (a|x) \in \mathbb{Z}\ \forall a \in L\}$ be the dual of L with respect to $(\ |\ )$. Then L is a sub-lattice of L^*. Choose a positive integer k. Define $M_k = L^*/kL$, a finite quotient of L^*. Let m_1 be the least positive integer such that $m_1 L^* \subset L$ and set $m_k = km_1$. We have

Lemma 21 *The integer m_k is the exponent of M_k, i.e. the least positive integer m such that $mM_k = 0$.*

The fact that $m_k M_k = 0$ is clear. For the reciprocal, observe that if m is an integer, $mM_k = 0$ is equivalent to $mL^* \subset kL$. Choose a basis for L and take the dual basis as a basis for L^*. The matrix of $(\ |\ )$ in the basis of L and in the dual basis of L^* are inverse of each other. Let A be the matrix of $(\ |\ )$ in the basis of L. By definition A has integral entries. Then $mM_k = 0$ if and only if $mA^{-1} = kB$ for a certain matrix B with integral entries. Taking the determinant on both sides, we get that a power of the rational m/k is an integer. So m/k itself is an integer and k divides m. So $(m/k)L^* \subset L$ and m is a multiple of m_k.

Now we define a quadratic structure $\{M_k, B_k, (\ |\ )_k), e^{i\pi/m_k}\}$ as follows. For $x, y \in L^*$, consider $(x|y)_k$, the residue class modulo m_k of $m_1(x|y)$, and $B_k(x)$, the residue class modulo $2m_k$ of $m_1(x|x)$. For $a \in kL$, $m_1(a|y)$ belongs to $m_k\mathbb{Z}$ and $m_1[(a|a) + 2(a|x)]$ belongs to $2m_k\mathbb{Z}$ (here, we use that L is even). So $(\ |\ )$ and B_k both descend to M_k, and we keep the same notation for those projections. It is plain that the polarisation of B_k is $(\ |\ )_k$, so the only thing that remains is to show that $(\ |\ )_k$ is non-degenerate. Equivalently, we have to

[p] Denoted by M in [9], but we want to avoid confusion with the finite group of a quadratic structure.

prove that if $x \in L^*$ is such that $(x|y)$ is a multiple of k for any $y \in L^*$ then $x \in kL$. But this is obvious. Thus we have shown that

Lemma 22 *For any positive integer k, $\{M_k, B_k, (\ |\ )_k), e^{i\pi/m_k}\}$ (defined above) is a quadratic structure.*

Definition 23 *The quadratic structure $\{M_k, B_k, (\ |\ )_k), e^{i\pi/m_k}\}$ is denoted by $\{L, (\ |\ )\}_k$ in the sequel.*

As usual (and again [9] is a nice self-contained reference) one can associate a family of functions $\{\Theta_x^{(k)}\}_{x \in L^*}$ to L and k : they are the so-called classical theta functions of level k. One shows that $\Theta_x^{(k)}$ and $\Theta_y^{(k)}$ coïncide if $x - y \in kL$ and that these are the only linear relations (over the complex numbers) among the classical theta functions of level k. So we have a family of linearly independent functions $\Theta_x^{(k)}$ indexed by elements of M_k. Of course, $\Theta_x^{(k)}$ depends on certain variables. What we need to know here is that one of those is a parameter $\tau \in \mathfrak{H}$. Let v be a collective notation for the other variables. Second, there is a function $v'(v, \tau)$ such that the following formulæ hold : for any $x \in M_k$

$$\Theta_x^{(k)}(-1/\tau, v') = (-i\tau)^{\ell/2} \sum_{y \in M_k} \mathcal{S}_{y,x}^{(k)} \Theta_y^{(k)}(\tau, v)$$

and

$$\Theta_x^{(k)}(\tau + 1, v) = \tilde{T}_{x,x}^{(k)} \Theta_x^{(k)}(\tau, v)$$

where $\mathcal{S}^{(k)}$ and $\tilde{T}^{(k)}$ are the matrices associated to the quadratic structure

$$\{L, (\ |\ )\}_k \equiv \{M_k, B_k, (\ |\ )_k), e^{i\pi/m_k}\}$$

as defined in 4.2. This is the well-known fact that modular transformations of theta functions are related to the finite Fourier transform.

We can now come to the heart of the matter. Suppose L is the lattice corresponding to the affine algebra $\mathfrak{g}$ associated to the complex simple Lie algebra $\overset{\circ}{\mathfrak{g}}$. Then $\overset{\circ}{\mathfrak{h}}$ is a Cartan subalgebra of $\overset{\circ}{\mathfrak{g}}$ and $\overset{\circ}{W}$, the (finite) Weyl group of $\overset{\circ}{\mathfrak{g}}$ acts on L and L^* as a group of isometries. Moreover, W^k, the affine Weyl group, acts on $\overset{\circ}{\mathfrak{h}}$ as the semi-direct product of $\overset{\circ}{W}$ and kL. By a choice of simple roots, one determines an open simplex S_+^k whose closure is a fundamental domain for the action of W^k on $\overset{\circ}{\mathfrak{h}}$. For $x \in L$, one defines anti-invariant classical theta functions

$$A_x^{(k)} = \sum_{w \in \overset{\circ}{W}} \varepsilon(w) \Theta_{w(x)}^{(k)}.$$

Their linear span is denoted by Th_k^-. For any $w \in W^{(k)}$ one has $A^{(k)}_{w(x)} = \varepsilon(w)A^{(k)}_x$ and these are the only linear relations among the anti-invariant classical theta functions. In particular, $A^{(k)}_x = 0$ if x is fixed by an element of W^k because by Chevalleys lemma, the stabilizer of x, if non-trivial, contains reflections. The first result is

Lemma 24 (Kac) [9]

Let $h^\vee$ be the dual Coxeter number of $\mathring{\mathfrak{g}}$. Let k be a positive integer. If $k < h^\vee$ then S^k_+ contains no point of L^*. If $k = h^\vee$, S^k_+ contains a single point of L^* denoted by ρ. The space Th_k^- is trivial for $k < h^\vee$ and 1-dimensional for $k = h^\vee$.

This makes it natural to translate k by $h^\vee$. From now on, k is a non-negative integer. We define $P^k_{++} = S^{k+h^\vee}_+ \cap L^*$ and $P^k_+ = P^k_{++} - \rho$.The following theorem is crucial.

Theorem 25 (Kac) [9]

The set P^k_+ is the set of highest weights of integrable representations of $\mathfrak{g}$ at level k. The anti-invariant classical theta functions $A^{(k+h^\vee)}_{\Lambda+\rho}, \Lambda \in P^k_+$ form a basis of $Th^-_{k+h^\vee}$ and the character $\chi^{(k)}_\Lambda$ of the highest weight integrable representation of $\mathfrak{g}$ at level k of highest weight Λ is

$$\chi^{(k)}_\Lambda = A^{(k+h^\vee)}_{\Lambda+\rho}/A^{(h^\vee)}_\rho.$$

As $\mathring{W}$ maps L into L, its action descends to $M_{k+h^\vee}$, and this action preserves $(\ |\ )_{k+h^\vee}$ and $B_{k+h^\vee}$. For instance, $\mathcal{S}^{(k+h^\vee)}_{y,w(x)} = \mathcal{S}^{(k+h^\vee)}_{w^{-1}(y),x}$. Moreover, it is harmless to identify P^k_{++} and P^k_+ with their images in M_k. From this one concludes that for $x \in P^k_{++}$

$$A^{(k+h^\vee)}_x(-1/\tau, v') = (-i\tau)^{\ell/2} \sum_{w\in\mathring{W}} \sum_{y\in P^k_{++}} \varepsilon(w)\mathcal{S}^{(k+h^\vee)}_{w(y),x} A^{(k+h^\vee)}_y(\tau, v)$$

and

$$A^{(k+h^\vee)}_x(\tau+1, v') = \tilde{\mathcal{T}}^{(k+h^\vee)}_{x,x} A^{(k+h^\vee)}_y(\tau, v).$$

In particular for $k = 0$ one can simplify the above formulæ :

$$A^{(h^\vee)}_\rho(-1/\tau, v') = (-i\tau)^{\ell/2}(-i)^{|\mathring{\Delta}_+|} A^{(h^\vee)}_\rho(\tau, v)$$

and

$$A^{(h^\vee)}_\rho(\tau+1, v) = e^{i\pi\,\dim\mathring{\mathfrak{g}}/12} A^{(h^\vee)}_\rho(\tau, v)$$

where $|\overset{\circ}{\Delta}_+|$ is the number of positive roots of $\overset{\circ}{\mathfrak{g}}$ and we have used the "strange formula" $12(\rho|\rho) = h^\vee \dim \overset{\circ}{\mathfrak{g}}$. Finally, for $\Lambda \in P_+^k$

$$\chi_\Lambda^{(k)}(-1/\tau, v') = i^{|\overset{\circ}{\Delta}_+|} \sum_{w \in \overset{\circ}{W}} \sum_{\Lambda' \in P_+^k} \varepsilon(w) \mathcal{S}^{(k+h^\vee)}_{w(\Lambda'+\rho),\Lambda+\rho} \chi_{\Lambda'}^{(k)}(\tau, v)$$

and

$$\chi_\Lambda^{(k)}(\tau+1, v) = \frac{\tilde{\mathcal{T}}^{(k+h^\vee)}_{\Lambda+\rho,\Lambda+\rho}}{\tilde{\mathcal{T}}^{(h^\vee)}_{\rho,\rho}} \chi_\Lambda^{(k)}(\tau, v).$$

This is more than enough to prove the result we were after :

Lemma 26 *The representation of $SL_2(\mathbb{Z})$ carried by the characters of the integrable highest weight representations of $\mathfrak{g}$ at level k is obtained from a representation of $SL_2(\mathbb{Z})$ associated to the quadratic structure $\{L, (\ |\ )\}_k$ first by going to the sub-representation of anti-invariant classical theta functions (defined over $\mathbb{Q}$) and then by tensoring with a 1-dimensional representation. In particular, this representation is in $\mathfrak{R}$*

8.5 We want to make a few remarks about the coset construction. The most common version is about cosets of Wess–Zumino–Witten models, but Witten proposed a vast abstract generalisation of the notion in [11]. Very briefly, if the chiral algebra $\mathcal{A}$ of a conformal field theory (with stress tensor $T_\mathcal{A}$ of central charge $c_\mathcal{A}$) contains a chiral subalgebra $\mathcal{B}$ that contains a stress tensor $T_\mathcal{B}$ (of central charge $c_\mathcal{B}$) such that $T_\mathcal{A}$ restricted to the primary fields of $\mathcal{B}$ and $T_\mathcal{B}$ coïncide, then the set of fields in $\mathcal{A}$ that have no short distance singularities with $\mathcal{B}$ is again a chiral algebra denoted by $\mathcal{A}/\mathcal{B}$ (with stress tensor $T_{\mathcal{A}/\mathcal{B}} = T_\mathcal{A} - T_\mathcal{B}$ of central charge $c_{\mathcal{A}/\mathcal{B}} = c_\mathcal{A} - c_\mathcal{B}$). Clearly, $\mathcal{A}/\mathcal{B}$ intertwines between $\mathcal{A}$ and $\mathcal{B}$.

This means that if $\mathcal{B}$ is the chiral algebra of another rational conformal field theory, and if $\mathcal{H}_I$ is an irreducible representation of $\mathcal{A}$, one can decompose $\mathcal{H}_I = \sum_i \mathcal{H}_I^i \otimes \mathcal{H}_i$ where the sum runs over the finite set of irreducible representations of $\mathcal{B}$ and $\mathcal{H}_I^i$ is a representation of $\mathcal{A}/\mathcal{B}$.

The most favorable situation is when $\mathcal{H}_I^i$ describes the set of irreducible representations of $\mathcal{A}/\mathcal{B}$ when I and i describe the sets of irreducible representations of $\mathcal{A}$ and $\mathcal{B}$ respectively. Let us call this the naïve situation. Then the representation of the modular group for $\mathcal{A}/\mathcal{B}$ is the tensor product of the representation for $\mathcal{A}$ and the dual representation for $\mathcal{B}$. So we see that

Lemma 27 *In the naïve situation, the representation of $SL_2(\mathbb{Z})$ associated to the coset algebra $\mathcal{A}/\mathcal{B}$ is in $\mathfrak{R}$ whenever the representations of $SL_2(\mathbb{Z})$ associated to $\mathcal{A}$ and $\mathcal{B}$ are in $\mathfrak{R}$.*

On the other hand, even if the naïve situation is generic (which is by no means clear), many (most ?!) interesting examples are not naïve. The spaces $\mathcal{H}_I^i$ need not be irreducible or distinct. The case when they are irreducible and exhaust the possible representations of $\mathcal{A}$ but with redundancy is not too bad. We call it the semi-naïve case. Essentially one has to identify some representations, and this means going to a quotient representation of the naïve representation. The identifications do not introduce any new irrationality, so again we have

Lemma 28 *In the semi-naïve situation, the representation of $SL_2(\mathbb{Z})$ associated to the coset algebra $\mathcal{A}/\mathcal{B}$ is in $\mathfrak{R}$ whenever the representations of $SL_2(\mathbb{Z})$ associated to $\mathcal{A}$ and $\mathcal{B}$ are in $\mathfrak{R}$.*

Let us note that getting the Virasoro minimal models as cosets of $SU(2)$ falls in this class. In fact a direct argument from the explicit form of $\mathcal{S}$ and $\mathcal{T}$ is not difficult : again, it is a simple folding of a finite Fourier transform. More generally, because the cyclic symmetry of the Dynkin diagrams of type $\hat{\mathfrak{su}}(n)$ for $n \geq 2$, their is always a cyclic group $\mathbb{Z}_n$ of automorphisms acting on the cosets $\frac{\hat{\mathfrak{su}}(n)_k \times \hat{\mathfrak{su}}(n)_l}{\hat{\mathfrak{su}}(n)_{k+l}}$. For this case the label I is a pair of integrable highest weights, one for level l and one for level l, whereas the label i is for an integrable highest weight at level $k+l$. When k and l have no common factors with n, this symmetry leads simply to identifications. But in the presence of common factors, the $\mathbb{Z}_n$ symmetry acts with fixed points. This leads us to the more complicated situation.

The case when the $\mathcal{H}_I^i$ are not all irreducible, but every representation of $\mathcal{A}/\mathcal{B}$ appears in the decomposition, is very interesting and very hard. Again the fact that the $\mathcal{H}_I^i$ are not all distinct is not essential. The fact that the $\mathcal{H}_I^i$ is not irreducible implies that the naïve representation is not enough because the true $\mathcal{S}$ has to be invertible (even unitary) so it has to distinguish between the different pieces $\mathcal{H}_I^i$ may contain. This is a problem very close to what is known as the problem of resolution of fixed points mentioned above. This phenomenon appears in many other contexts, for instance orbifolds and simple currents. There has been substantial recent progress to deal with fixed points [13], and this has also led to a proposal for the $\mathcal{S}$ matrix of most cosets of Wess–Zumino–Witten models. We have not yet tried to check our conjecture for those cases, because the proposal is not totally straightforward to manipulate. So we shall only make a few remarks in the next and last paragraph.

8.6 We start from a rational conformal field theory with chiral algebra $\mathcal{A}$, and irreducible representations V_a, indexed by $a \in A$ with restricted characters χ_a. The unitarity of the representation of $SL_2(\mathbb{Z})$ carried by the characters

implies that

$$\sum_{a\in A} |\chi_a|^2$$

is modular invariant.

Sometimes one can construct other sesquilinear modular invariants. For instance there is a partition $A = \cup_{b\in B} A_b$ such that

$$\sum_{b\in B} |\sum_{a\in A_b} \chi_a|^2$$

is again modular invariant. In such cases it is tempting to define a new chiral algebra $\mathcal{B}$ extending $\mathcal{A}$ and such that its irreducible representations are indexed by B with restricted characters $\chi_b^{(B)} \equiv \sum_{a\in A_b} \chi_a$. Indeed, this can always be done [12]. The modular transformations of the new characters are easily obtained from the transformations of the original ones, and they are automatically unitary.

On the other hand, more intriguing situations may appear. For instance, it may happen that for some positive integers m_b

$$Z^{new} = \sum_{b\in B} m_b |\sum_{a\in A_b} \chi_a|^2$$

is modular invariant. Very often, the value of $m_b|A_b|$ doesn't depend on b, so it is tempting to see the partition $A = \cup_{b\in B} A_b$ as a set of orbits, with multiplicities appearing due to fixed points, a justification for the name given to those phenomena. Again, this leads to hope that there is some kind of new extended symmetry, but this time, modular transformations are unitary on $m_b^{1/2} \sum_{a\in A_b} \chi_a$, not on $\sum_{a\in A_b} \chi_a$. However, unless m_b is a perfect square for any b, the first expression cannot be a candidate for a restricted character. The only way out of this dilemma is that in the new theory, the same restricted characters may appear more than once, which means that one has to split m_b in pieces which can only be distinguished with complete characters [q]. But then, real new work is needed to get the modular transformations of the new characters : $\mathcal{T}$ is not a problem, but only sum rules are known for the new $\mathcal{S}$ matrix. Moreover, nothing guarantees a priori that there is a unique way of splitting. In fact, the answer to this questions requires extra information, for instance, the consistency of fusion rules derived from $\mathcal{S}$ with the Verlinde

[q] The situation is analogous to a situation met in algebraic geometry : at singular points, some branches meet, and a resolution of singularities is needed to separate those branches. Hence the name fixed point resolution.

formula. This makes any check of our conjecture difficult for the new theory, even if we know that it is correct for the original one.

On the other hand, we have seen above that when the restricted characters are linearly independent, our conjectures can be attacked using the standard theory of modular forms. Even if Wess–Zumino–Witten models give examples of two- or threefold degeneracy of the restricted chatacters, it is clear that fixed point resolutions offer a much richer structure of degeneracy to confront our ideas with. So progress in this direction is crucial.

Anyway, the outcome of the computations in concrete examples often reveals the following structure.

1. The set B can be partitioned as $B = \cup_{c \in C} B_c$ in such a way that the new characters are indexed by b and by another index whose range depends on the class of b. So for any $c \in C$ and any $b \in B_c$ we have a set $\chi_{b,i} \propto \sum_{a \in A_b} \chi_a$, $i \in I_c$.

2. The new $\mathcal{S}$ matrix can be expressed in terms of the old $\mathcal{S}$ matrix (restricted to the sub-representation responsible for the modular invariance of Z^{new}) and on a family of matrices $\mathcal{S}^c$, $c \in C$ acting on a vector space with a basis indexed by $b \in B_c$. The complete formula involves projectors in the space of i indices and looks complicated, but all the irrational numbers are in $\mathcal{S}^{old}$ and in the various $\mathcal{S}^c$ matrices, so $\mathcal{S}^{new}$ looks simply like a direct sum from our point of view. This gives rise to a representation of $SL_2(\mathbb{Z})$ if T^c (the restriction of T^{old} to B_c) and $\mathcal{S}^c$ build a representation of $SL_2(\mathbb{Z})$.

3. The representation build by $\mathcal{S}^c$ and T^c is the tensor product of a one dimensional representation of $SL_2(\mathbb{Z})$ and a representation coming from a rational conformal field theory.

To resume, in certain cases, resolution of fixed points leads to a direct sum representation of $SL_2(\mathbb{Z})$. So if one can show that the components satisfy our conjecture, then the full new theory satisfies it as well. The third observation above is encouraging because the components come from simpler conformal field theories : again this leads to hope that our conjecture can be proved starting from simple examples. Far from satisfactory as it is, this is nevertheless an appropriate point to close our discussion.

Appendices

A Galois theory on a post-stamp.

A1 In this section, we mention very briefly the basics of Galois theory and the results we use in the rest of the paper. For a pedagogical and elementary account, we recommend [14] ; the more general and abstract presentation of [15] is also valuable.

As far as Galois theory is concerned in this paper, we shall only deal with zero characteristic[r].

A2 Let k be a field contained in a larger field K. We say that K is an extension of k. We denote this situation by K/k although no quotient is involved here.

If K/k is an extension, K is a vector space over k, whose dimension $[K : k]$ is called the degree of the extension. If $[K : k] < \infty$ we say that the extension K/k is finite. Then the elements of K are algebraic over k because the only a finite number of powers of an element $x \in K$ can be linearly independent so that x has to be a zero of some polynomial with coefficients in k. We shall mostly concentrate on finite extensions in the sequel.

A3 According to Steinitz's theorem, any field k has an algebraic closure k^c, unique up to isomorphism. This means first that k^c is an algebraic extension of k (k^c contains k, and any element of k^c is a zero of a polynomial with coefficients in k) and second that any non constant polynomial with coefficients in k^c splits as a product of linear factors. In other words, k^c is an algebraic extension of k that admits no non-trivial algebraic extensions.

In general, k^c is not a finite extension of k. But k^c contains a subfield isomorphic to any given finite (or algebraic) extension. From two algebraic extensions K_1 and K_2 of k given as subfields of k^c, we define the compositum K of K_1 and K_2 as the smallest subfield of k^c containing K_1 and K_2.

A4 We say that K/k is a Galois extension if every polynomial in $k[X]$ with a zero in K splits in K [s], the" one root in, all roots in" property. For example, k^c is a Galois extension of k. Thus every algebraic extension is contained in a Galois extension. In fact, any finite algebraic extension is contained in a finite Galois extension.

[r] So that a Galois extension and a normal extension are the same.

[s] I.e. is a product of linear factors in K.

A5 Let Aut K/k be the subgroup of the automorphism group of K that fixes every element of k.

The set of elements of K fixed by Aut K/k is a field k' with $k \subset k' \subset K$. If K is a Galois extension of k then $k = k'$. To rephrase, if K/k is a Galois extension, to check that $x \in K$ is an element of k it is sufficient to check that x is fixed by K/k. There is a reciprocal for finite extensions : if K/k is finite and $k = k'$ then K/k is Galois.

The automorphism group Aut K/k of a Galois extension is also denoted by Gal K/k.

Galois extension have other characterisations that show their usefulness. For instance, if K is obtained from k by adding all the roots of a family of polynomials with coefficients in k, then K/k is a Galois extension. Moreover, this gives all Galois extensions.

A6 The fundamental result of Galois theory is the Galois correspondence : if L/k is a finite Galois extension, there is a one to one inclusion reversing correspondence between intermediate fields and subgroups of Gal L/k. To an intermediate field K ($k \subset K \subset L$) is associated Aut L/K and to a subgroup of Gal L/k is associated its fixed field. Those two maps are inverse of each other. Moreover, if K is an intermediate field, K/k is Galois if and only if Aut L/K is an invariant subgroup of Gal L/k. In this case, Gal K/k is isomorphic to the quotient of Gal L/k by Aut L/K.

Let us mention that any algebraic (resp. any finite) extension K/k is contained in a (resp. a finite) Galois extension L/k, so that the Galois correspondence gives another criterion for Galois extensions : it is easy to check that K/k is Galois if and only if any element of Gal L/k maps K into itself. If this is the case, there is a (restriction) homomorphism from Gal L/k to Gal K/k, and this homomorphism is onto.

A7 Especially relevant for us are Abelian extensions. They are the Galois extensions such that Gal K/k is Abelian. By the Galois correspondence, any extension contained in a finite Abelian extension is again an Abelian extension.

The field $\mathbb{Q}[e^{2i\pi/n}]$ obtained by adding $e^{2i\pi/n}$ [t] to $\mathbb{Q}$ is a Abelian extension of $\mathbb{Q}$. The corresponding Galois group is isomorphic to $\mathbb{Z}_n^*$.

Let K be an extension of $\mathbb{Q}$ such that $\mathbb{Q} \subset K \subset \mathbb{Q}[e^{2i\pi/n}]$ for a certain n. Because the restriction homomorphism is onto, any element of Gal K/k can be represented (usually in more than one way) as an element of $\mathbb{Z}_n^*$. This is

[t] Or any other primitive n^{th}-root of unity, because primitive n^{th}-roots of unity are powers of each other.

a property that proves very useful : a (deep) result by Kroenecker and Weber asserts that any Abelian extension of $\mathbb{Q}$ is contained in $\mathbb{Q}[e^{2i\pi/n}]$ for a certain n.

A8 Later we shall almost only encounter Galois extensions and very often even Abelian extensions. So it is worth giving an example of an extension that is not Galois. Let x be the real cube root of 2. Let K be $\mathbb{Q}$ vector space with basis $1, x, x^2$, seen as a subset of $\mathbb{R}$. Using Bezout's theorem, one can check easily that K is a subfield of $\mathbb{R}$. It contains exactly one cube root of 2 (the other two are complex), so $K/\mathbb{Q}$ is not a Galois extension.

B Galois actions on representations

B1 In this appendix, we recall some useful facts about Galois actions in linear algebra. In the main text, we use explicit bases most of the time, but here we prefer a more canonical approach.

B2 It is possible to extend parts of Galois theory to vector spaces. If V is a (finite dimensional in all applications) vector space over k and K is an extension of k we define V^K to be the tensor product $V \otimes_k K$. There is a natural inclusion $V \to V^k$ sending $v \in V$ to $v \otimes 1$.

B3 If $k \subset K \subset L$ is a family of field extensions, there is no ambiguity in writing V^L because $V \otimes_k L$ and $V^K \otimes_K L$ are canonically isomorphic. Moreover, for analogous reasons, if V_0 is a subspace of V and W is another k-vector space, one can write either $(V/V_0)^K$ or V^K/V_0^K for quotients, $(V^*)^K$ or $(V^K)^*$ for duals, $(V \oplus W)^K$ or $V^K \oplus W^K$ for sums, $((\mathrm{Hom}_k(V,W))^K$ or $\mathrm{Hom}_K(V^K, W^K)$ for homomorphisms and $(V \otimes_k W)^K$ or $V^K \otimes_K W^K$ for tensor products.

B4 An automorphism $\sigma \in \mathrm{Aut}\, K/k$ acts canonically on V^K by acting on the second factor : for $v \in V$ and $\lambda \in K$,

$$\sigma(v \otimes \lambda) \equiv v \otimes \sigma(\lambda).$$

B5 If V_0 is a subspace of V we denote by $\mathrm{End}_0 V$ the set of linear maps from V to V sending V_0 to V_0. There is of course a restriction map from $\mathrm{End}_0 V$ to $\mathrm{End}\, V_0$ and it is onto because vector subspaces are summands. There is also a natural projection map from $\mathrm{End}_0 V$ to $\mathrm{End}\, V/V_0$. Those operations commute

with Galois transformations. Explicitly for $\sigma \in \operatorname{Aut} K/k$ the diagrams

$$\begin{array}{ccc} \operatorname{End}_0 V^K & \xrightarrow{\sigma} & \operatorname{End}_0 V^K \\ \downarrow & & \downarrow \\ \operatorname{End} V_0^K & \xrightarrow{\sigma} & \operatorname{End} V_0^K \end{array} \qquad \begin{array}{ccc} \operatorname{End}_0 V^K & \xrightarrow{\sigma} & \operatorname{End}_0 V^K \\ \downarrow & & \downarrow \\ \operatorname{End} (V/V_0)^K & \xrightarrow{\sigma} & \operatorname{End} (V/V_0)^K \end{array}$$

commute. This is obvious if we take a basis for V_0, complete it to get a basis of V and look at matrix elements : an element of $\operatorname{End}_0 V^K$ has a block structure $\begin{pmatrix} A & B \\ 0 & D \end{pmatrix}$. This induces A on $\operatorname{End} V_0^K$ and D on $\operatorname{End} (V/V_0)^K$.

B6 Let K be an extension of k and V' be a K-vector space. A subset I of $\operatorname{End} V'$ is said to be defined over k if there is a k-vector space V and an isomorphism $V^K \cong V'$ such that I is in the image of $\operatorname{End} V$ under the composition

$$\operatorname{End} V \to \operatorname{End} V^K \to \operatorname{End} V'.$$

This means exactly that there is a basis for V' such that the matrix elements of the members of I in this basis are all in k.

In particular, if G is a group with a representation on V', it makes sense to ask whether this representation is defined over k or not. Of course if the answer is yes, the character takes its values in k. The reciprocal is in general wrong, although the two problems are closely related [16].

C Arithmetic factorisation

C1 We keep the notations of section 4. The construction of quadratic structures and representations of $SL_2(\mathbb{Z})$ starting from a finite Abelian group M share a common property with the so-called arithmetic functions : if m can be decomposed as $m = pq$ were p and q are relatively prime integers, there is a corresponding decomposition of the quadratic form, the scalar product and the representation of $SL_2(\mathbb{Z})$.

Suppose that $m = pq$ with p and q as above. At least one of p and q is odd, so we can assume without loss of generality that q is odd. We fix a pair of integers (u, v) solution of Bezout's equation $2up + vq = 1$ [u].

[u] If m is even the factor 2 is crucial. If m is odd, the factor 2 is irrelevant. Its presence makes it possible to deal with all values of m at once.

C2 For any $x \in M$, $2upx + vqx = x$. So x can be written as a sum $x = py + qz$ with $y = 2ux$ and $z = vx$. Moreover if $px = 0$ (resp. $qx = 0$), x can be written as qy (resp. pz) and if $px = qx = 0$, $x = 0$. The converse statements are obvious because $m = pq$ annihilates any element of M.

Define $M_p \equiv \{x \in M,\ px = 0\} = \{qx\}_{x \in M}$ and $M_q \equiv \{x \in M,\ qx = 0\} = \{px\}_{x \in M}$, which are obviously subgroups of M. We can rephrase the above properties by saying that $M = M_p \oplus M_q$. Let $x = x_p + x_q$ denote the decomposition of a generic element of M.

C3 This is an orthogonal decomposition because

$$(x_p|x_q) = 2up(x_p|x_q) + vq(x_p|x_q) = 2u(px_p|x_q) + v(x_p|qx_q) = 0.$$

Hence $(x_p|x_q) = 0$ and $B(x) = B(x_p) + B(x_q)$.

Moreover $2pB(x_p) = 0 = qB(x_q)$. The first equality is because by polarisation $2B(x_p) = 2(x_p|x_p)$ so $2pB(x_p) = 2(px_p|x_p) = 0$. For the second equality, we use that q is odd to define $z_q = \frac{q+1}{2} x_q$, which is such that $x_q = 2z_q$ so $B(x_q) = 4B(z_q)$ and then the argument is the same as for p.

C4 The Hilbert space H splits as a tensor product $H = H_p \otimes H_q$ with $e_x \equiv e_{x_p} \otimes e_{x_q}$. We look for quadratic structures $\{M_p, B_p, (\ |\ )_p, \zeta_p\}$ and $\{M_q, B_q, (\ |\ )_q, \zeta_q\}$ such that the representation splits. Explicitly, we want that for $x, y \in M$

$$\zeta^B(x) = \zeta_p^{B_p(x_p)} \zeta_q^{B_q(x_q)} \qquad \xi^{(x|y)} = \xi_p^{(x_p|y_p)_p} \xi_q^{(x_q|y_q)_q}.$$

We show that ζ_p and ζ_q can be chosen arbitrarily, and that once fixed, the quadratic structure on M_p and M_q satisfying the tensor product property exist and are unique. As a change in ζ_p or ζ_q can be re-absorbed in the normalisation of the quadratic forms and scalar products, we can assume without loss of generality that $\zeta_p = \zeta^q$ and $\zeta_q = \zeta^p$. Then the above conditions become

$$B(x) = qB_p(x_p) + pB_q(x_q) \qquad (x|y) = q(x_p|y_p)_p + p(x_q|y_q)_q. \tag{C.1}$$

C5 Take $x, y \in M_p$ (so $x_q = y_q = 0$) and multiply both equations by v. This gives

$$vB(x_p) = (1 - 2up)B_p(x_p) \qquad v(x_p|y_p) = (1 - 2up)(x_p|y_p)_p.$$

So the only possibility is that B_p (resp. $(\ |\ )_p$) is the restriction of vB (resp. $v(\ |\ )$) to M_p projected modulo $2p$ (resp. modulo p).

Now we take $x, y \in M_q$ and multiply the above conditions by $2u$. This gives

$$2uB(x_q) = (1 - vq)B_q(x_q) \qquad 2u(x_q|y_q) = (1 - vq)(x_q|y_q)_q.$$

Observe that because q is odd, $qB_q = 0$ in $\mathbb{Z}_{2q}$ (set $z_q = \frac{q+1}{2}x_q$ so $x_q = 2z_q$: $B_q(x_q) = 4B_q(z_q)$ is 0 modulo 2). So again the only possibility is that B_q (resp. $(\ |\ )_q$) is the restriction of $2uB$ (resp. $2u(\ |\ )$) to M_p projected modulo $2q$ (resp. modulo q). This proves the uniqueness of the solution if there is one and gives explicitly the only candidate.

C6 To show that this candidate passes the test, we observe that B_p and B_q are quadratic, that $(\ |\ )_p$ and $(\ |\ )_q$ are bilinear symmetric, and that the polarisation identities hold. To show non-degeneracy, it is enough to show that $\xi^{(x|y)} = \xi_p^{(x_p|y_p)_p}\xi_q^{(x_q|y_q)_q}$ [v]. This is equivalent to part of formulæ (C.1), and we have to prove this anyway. In $\mathbb{Z}_{pq}$

$$\begin{aligned} q(x_p|y_p)_p + p(x_q|y_q)_q &= vq(x_p|y_p) + 2up(x_q|y_q) = (x_p|y_p) + (x_q|y_q) \\ &= (x|y) - (x_p|y_q) - (x_q|y_p) = (x|y). \end{aligned}$$

Now, in $\mathbb{Z}_{2pq}$

$$\begin{aligned} qB_p(x_p) + pB_q(x_q) &= vqB(x_p) + 2upB(x_q) = B(x_p) + B(x_q) \\ &= B(x) - 2(x_p|x_q) = B(x). \end{aligned}$$

In those two arguments we use the orthogonality of M_p and M_q and its consequences (see paragraph C.3).

C7 This terminates the proof that the representations of $SL_2(\mathbb{Z})$ associated to quadratic structures have arithmetic factorisation, a useful fact to show that $\beta^8 = 1$ in paragraph 4.4.

References

1. A. Coste and T. Gannon, *Remarks on Galois symmetry in rational conformal field theories*, Phys. Lett. B323 (1994) 316–321.
2. A. Cappelli, C. Itzykson and J.-B. Zuber, *The A-D-E classification of minimal and $A_1^{(1)}$ conformal invariant theories*, Commun. Math. Phys. 113 (1987) 1–26.

[v] Because missing characters in M_p or M_q would imply missing characters for M.

3. M.Bauer and C. Itzykson. *Triangulations* in *The Grothendieck theory of dessins d'enfants*, L. Schneps ed., LMSLNS 200, Cambridge Univ. Press.
4. P. Di Francesco and C. Itzykson, *A generating function for fatgraphs*, Ann. Inst. Henry Poincaré 59 (1993) 117–139.
5. M. Bauer, A. Coste, C. Itzykson and P. Ruelle, *Comments on the links between su(3) modular invariants, simple factors in the Jacobian of Fermat curves and rational triangular billiards*, to appear in J. Geom. Phys. (**hep-th 9604104**).
6. S. Lang, Elliptic Functions, second edition, Springer–Verlag (1987).
7. J. de Boer, J. Goeree, *Markov traces and II_1 factors in CFT*, Commun. Math. Phys 139 (1991) 267–304.
8. P. Degiovanni, *private communication.*
9. V. G. Kac, Infinite Dimensional Lie Algebras, third edition, Cambridge University Press (1990).
10. C.Itzykson, *Level one Kac-Moody characters and modular invariance*, R.C.P. 25, vol. 39 (1988) 69–84.
11. E. Witten, *The central charge in three dimensions*, in Physics and Mathematics of Strings, World Scientific, 1990.
12. G. Moore and N. Seiberg, *Naturality in conformal field theory*, Nucl. Phys. B313 (1989) 16–40.
13. J. Fuchs, A. N. Schellekens and C. Schweigert, *A matrix S for all simple current extensions*, (**hep-th 9601078**).
14. I. Stewart, Galois Theory, second edition, Chapman and Hall (1992).
15. S. Lang, Algebra, third edition, Addison–Wesley Publishing Company (1994).
16. J.–P. Serre, Représentations Linéaires des Groupes Finis, Hermann (1967).

SOFTLY BROKEN $N = 2$ QCD

L. ÁLVAREZ-GAUMÉ

Theory Division, CERN, 1211 Geneva 23, Switzerland

M. MARIÑO

Theory Division, CERN, 1211 Geneva 23, Switzerland
and
Departamento de Física de Partículas, Universidade de Santiago de Compostela,
E-15706 Santiago de Compostela, Spain

We analyze the possible soft breaking of $N = 2$ supersymmetric Yang-Mills theory with and without matter flavour preserving the analiticity properties of the Seiberg-Witten solution. We present the formalism for an arbitrary gauge group and obtain an exact expression for the effective potential. We describe in detail the onset of the confinement description and the vacuum structure for the pure $SU(2)$ Yang-Mills case and also some general features in the $SU(N)$ case. A general mass formula is obtained, as well as explicit results for the mass spectrum in the $SU(2)$ case.

1 Introduction and Conclusions.

In two remarkable papers [1,2], Seiberg and Witten obtained exact information on the dynamics of $N = 2$ supersymmetric gauge theories in four dimensions with gauge group $SU(2)$ and $N_f \leq 4$ flavour multiplets. Their work was extended to other groups in [3,4,5,6]. One of the crucial advantages of using $N = 2$ supersymmetry is that the low-energy effective action in the Coulomb phase up to two derivatives (*i.e.* the Kähler potential, the superpotential and the gauge kinetic function in $N = 1$ superspace language) are determined in terms of a single holomorphic function called the prepotential [7]. In references [1,2], the exact prepotential was determined using some plausible assumptions and many consistency conditions. For $SU(2)$ the solution is neatly presented by associating to each case an elliptic curve together with a meromorphic differential of the second kind whose periods completely determine the prepotential. For other gauge groups [6] the solution is again presented in terms of the period integrals of a meromorphic differential on a Riemann surface whose genus is the rank of the group considered. It was also shown in [1,2] that by soft breaking $N = 2$ down to $N = 1$ (by adding a mass term for the adjoint $N = 1$ chiral

multiplet in the $N = 2$ vector multiplet) confinement follows due to monopole condensation [8].

For $N = 1$ theories exact results have also been obtained [9] using the holomorphy properties of the superpotential and the gauge kinetic function, culminating in Seiberg's non-abelian duality conjecture [10].

With all this new exact information it is also tempting to obtain exact information about ordinary QCD. The obvious problem encountered is supersymmetry breaking. A useful avenue to explore is soft supersymmetry breaking. The structure of soft supersymmetry breaking in $N = 1$ theories has been known for some time [11]. In [12,13] soft breaking terms are used to explore $N = 1$ supersymmetric QCD (SQCD) with gauge group $SU(N_c)$ and N_f flavours of quarks, and to extrapolate the exact results in [9] concerning the superpotential and the phase structure of these theories in the absence of supersymmetry. This leads to expected and unexpected predictions for non-supersymmetric theories which may eventually be accessible to lattice computations. In some cases however for instance when $N_f \geq N_c$) it is known in the supersymmetric case that the origin of moduli space is singular, and therefore some of the assumptions made about the Kähler potential for meson and baryon operators are probably too strong. Since the methods of [1,2] provide us with the effective action up to two derivatives, the kinetic and potential term for all low-energy fields are under control, and therefore in this paper we prefer to explore in which way we can softly break $N = 2$ SQCD directly to $N = 0$ while at the same time preserving the analyticity properties of the Seiberg-Witten solution. This is a very strong constraint and there is, essentially, only one way to accomplish this task: we make the dynamical scale Λ of the $N = 2$ theory a function of an $N = 2$ vector multiplet which is then frozen to become a spurion whose F and D-components break softly $N = 2$ down to $N = 0$. If we want to interpret physically the spurion, one can recall the string derivation of the Seiberg-Witten solution in [14,15] based on type II-heterotic duality. In the field theory limit in the heterotic side (in order to decouple string and gravity loops) the natural scaling is taken to be $Me^{iS} = \Lambda$, where M is the Planck mass, S is the dilaton (in the low-energy theory $S = \theta/2\pi + 4\pi i/g^2$, with g the gauge coupling constant and θ the CP-violating phase), and Λ the dynamical scale of the gauge theory which is kept fixed while $M \to \infty$ and $iS \to \infty$. Since the dilaton sits in a vector multiplet of $N = 2$ when the heterotic string is compactified on $K3 \times T_2$, this is precisely the field we want to make into a spurion, and procedure is compatible with the Seiberg-Witten monodromies. In this way we obtain a theory at $N = 0$ with a more restricted structure that those used in [12,13].

As soon as the soft breaking terms are turned on monopole condensation

appears, and we get a unique ground state (near the massless monopole point of [1,2]). Furthermore, in the Higgs region we can compute the effective potential, and we can verify that this potential drives the theory towards the region where condensation takes place. When the supersymmetry breaking parameter is increased, the minimum displaces to the right along the real u-axis. At the same time, the region in the u-plane in which the monopole condensate is energetically-favoured expands. Near the massless dyon point of [1,2], we find that dyon condensation is energetically favourable but, unlike monopole condensation, it is not sufficiently-strong an effect to lead to another minimum of the effective potential. Eventually, when the soft supersymmetry breaking parameter is made sufficiently large, the regions where monopole and dyon condensation are favoured begin to overlap. At this point, it is clear that our methods break down, and new physics is needed to describe the dynamics of these mutually-nonlocal degrees of freedom.

One advantage of this method of using the dilaton spurion to softly break supersymmetry from $N = 2$ to $N = 0$ is its universality. It works for any gauge group and any number of massive or massless quarks. We work out the general structure of soft breaking by the dilaton spurion in an arbitrary gauge group paying special attention to the monodromies and the properties of the spurion couplings, and we find the general features of the vacuum structure for the case of $SU(N)$. We also study the evolution of the mass eigenvalues in the case of the $SU(2)$ and also show in general that with this soft breaking procedure there is a general sum rule satisfied by the masses of all the multiplets.

The organization of this paper is as follows: In section one we present the general formalism for the breaking of supersymmetry due to a dilaton spurion for a general gauge group, and we study the symplectic transformations of the various quantities involved. The results agree with the general structure derived in [16] concerning the modification of the symplectic transformations of special geometry in the presence of background $N = 2$ vector superfields. In section three we study the effective potential and vacuum structure. In section four we particularize the formalism to the case of $SU(2)$ where the analysis can be made more explicit. In section five we analyze in some detail the case for $SU(N)$ without hypermultiplets. Finally in section six we present a general mass sum rule for the general case, and also obtain explicit results of the masses in the $SU(2)$ case. It is clear that for the moment we cannot take the supersymmetry decoupling limit due to the fact that as the supersymmetry breaking parameter increases we find that regions where mutually non-local operators acquire vacuum expectation values overlap. This raises the fascinating issue that in order to reach the real QCD limit we have to understand the dynamics of the Argyres-Douglas phases [17].

2 Breaking $N = 2$ with a dilaton spurion: general gauge group

In this section we present the generalization of the procedure introduced in [21] to $N = 2$ Yang-Mills theories with a general gauge group G of rank r and massless matter hypermultiplets.

The low energy theory description of the Coulomb phase [1] involves r abelian $N = 2$ vector superfields A^i, $i = 1, \cdots, r$ corresponding to the unbroken gauge group $U(1)^r$. The holomorphic prepotential $\mathcal{F}(A^i, \Lambda)$ depends on the r superfields A^i and the dynamically generated scale of the theory, Λ. The low energy effective lagrangian takes the form (in $N = 1$ notation) [1]:

$$\mathcal{L} = \frac{1}{4\pi}\mathrm{Im}\Big[\int d^4\theta \frac{\partial \mathcal{F}}{\partial A^i}\overline{A}^i + \frac{1}{2}\int d^2\theta \frac{\partial^2 \mathcal{F}}{\partial A^i \partial A^j} W_\alpha^i W^{\alpha j}\Big], \tag{2.1}$$

We define the dual variables, as in the $SU(2)$ case, by

$$a_{D,i} \equiv \frac{\partial \mathcal{F}}{\partial a^i}. \tag{2.2}$$

The Kähler potential and effective couplings associated to (2.1) are:

$$K(a, \bar{a}) = \frac{1}{4\pi}\mathrm{Im} a_{D,i}\bar{a}^i,$$

$$\tau_{ij} = \frac{\partial^2 \mathcal{F}}{\partial a^i \partial a^j}, \tag{2.3}$$

and the metric of the moduli space is given accordingly by:

$$(ds)^2 = \mathrm{Im}\frac{\partial^2 \mathcal{F}}{\partial a^i \partial a^j} da^i d\bar{a}^j. \tag{2.4}$$

We introduce now a complex space $\mathbf{C}^{2r}$ with elements of the form

$$v = \begin{pmatrix} a_{D,i} \\ a^i \end{pmatrix}. \tag{2.5}$$

The metric (2.4) can then be written as

$$\begin{aligned} (ds)^2 &= -\frac{i}{2}\sum_i (da_{D,i} d\bar{a}^i - d\bar{a}_{D,i} da^i) \\ &= -\frac{i}{2}\begin{pmatrix} da_{D,i} & da^i \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} d\bar{a}_{D,i} \\ d\bar{a}^i \end{pmatrix}, \end{aligned} \tag{2.6}$$

which shows that the transformations of v preserving the form of the metric are matrices $\Gamma \in Sp(2r, \mathbf{Z})$. They verify $\Gamma^{\mathrm{T}}\Omega\Gamma = \Omega$, where Ω is the $2r \times 2r$ matrix appearing in (2.6), and can be written as:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{2.7}$$

where the $r \times r$ matrices A, B, C, D satisfy:

$$A^{\mathrm{T}}D - C^{\mathrm{T}}B = \mathbf{1}_r, \quad A^{\mathrm{T}}C = C^{\mathrm{T}}A, \quad B^{\mathrm{T}}D = D^{\mathrm{T}}B. \tag{2.8}$$

The vector v transforms then as:

$$\begin{pmatrix} a_D \\ a \end{pmatrix} \to \Gamma \begin{pmatrix} a_D \\ a \end{pmatrix} = \begin{pmatrix} Aa_D + Ba \\ Ca_D + Da \end{pmatrix}. \tag{2.9}$$

From this we can obtain the modular transformation properties of the prepotential $\mathcal{F}(a^i)$ (see [19]). Since

$$\begin{aligned} \frac{\partial \mathcal{F}_\Gamma}{\partial a^k} &= \frac{\partial a^i_\Gamma}{\partial a^k}\frac{\partial \mathcal{F}_\Gamma}{\partial a^i_\Gamma} = \left(C^{ip}\tau_{pk} + D^i_k\right)\left(A^j_i a_{D,j} + B_{ij}a^j\right) \\ &= (D^{\mathrm{T}}B)_{kj}a^j + (D^{\mathrm{T}}A)^j_k \frac{\partial \mathcal{F}}{\partial a^j} + (C^{\mathrm{T}}B)^p_j \frac{\partial a_{D,p}}{\partial a^k}a^j \\ &+ (C^{\mathrm{T}}A)^{pj}\frac{\partial a_{D,p}}{\partial a^k}a_{D,j}, \end{aligned} \tag{2.10}$$

using the properties (2.8) of the symplectic matrices we can integrate (2.10) to obtain:

$$\begin{aligned} \mathcal{F}_\Gamma &= \mathcal{F} + \frac{1}{2}a^k(D^{\mathrm{T}}B)_{kj}a^j + \frac{1}{2}a_{D,k}(C^{\mathrm{T}}A)^{pj}a_{D,j} \\ &+ a^k(B^{\mathrm{T}}C)^j_k a_{D,j}. \end{aligned} \tag{2.11}$$

Starting with (2.11) we can prove that the quantity $\mathcal{F} - 1/2\sum_i a^i a_{D,i}$ is a monodromy invariant, and evaluating it asymptotically, one obtains the relation [18,19,20]:

$$\mathcal{F} - \frac{1}{2}\sum_i a^i a_{Di} = -4\pi i b_1 u, \tag{2.12}$$

where b_1 is the coefficient of the one-loop beta function (for $SU(N_c)$ with N_f hypermultiplets in the fundamental representation, $b_1 = (2N_c - N_f)/16\pi^2$)

and $u = \langle \mathrm{Tr}\phi^2 \rangle$. With the normalization for the electric charge used in [2] and [5], the r.h.s. of (2.12) is $-2\pi i b_1 u$.

As in the $SU(2)$ case, presented in[21], we break $N = 2$ supersymmetry down to $N = 0$ by making the dynamical scale Λ a function of a background vector superfield S, $\Lambda = e^{iS}$. This must be done in such a way that s, $s_D = \partial\mathcal{F}/\partial s$ be monodromy invariant. To see this, we will derive a series of relations analogous to the ones in the $SU(2)$ case[21], starting with the following expression for the prepotential in terms of local coordinates:

$$\mathcal{F} = \sum_{ij} a^i a^j f_{ij}(a^l/\Lambda), \tag{2.13}$$

where we take $f_{ij} = f_{ji}$. We define now a $(r+1) \times (r+1)$ matrix of couplings including the dilaton spurion $a^0 = s$:

$$\tau_{\alpha\beta} = \frac{\partial^2 \mathcal{F}}{\partial a^\alpha a^\beta}. \tag{2.14}$$

Greek indices α, β go from 0 to r, and latin indices i, j from 1 to r. We obtain:

$$a_{D,k} = 2\sum_i a^i f_{ik} + \frac{1}{\Lambda}\sum_{ij} a^i a^j f_{ij,k},$$

$$\tau_{ij} = 2f_{ij} + \frac{2}{\Lambda}\sum_k a^k (f_{ik,j} + f_{jk,i}) + \frac{1}{\Lambda^2} a^k a^l f_{kl,ij},$$

$$\tau_{0i} = -\frac{i}{\Lambda}\sum_{jk} a^j a^k (2f_{ij,k} + f_{jk,i}) - \frac{i}{\Lambda^2}\sum_{jkl} a^j a^k a^l f_{jk,li},$$

$$\tau_{00} = -\frac{1}{\Lambda}\sum_{ijk} a^i a^j a^k f_{ij,k} - \frac{1}{\Lambda^2}\sum_{ijkl} a^i a^j a^k a^l f_{ij,kl}, \tag{2.15}$$

and the dual spurion field is given by:

$$s_D = \frac{\partial \mathcal{F}}{\partial s} = -\frac{i}{\Lambda}\sum_{ijk} a^i a^j f_{ij,k} \tag{2.16}$$

The equations (2.15) and (2.16) give the useful relations:

$$\tau_{0i} = i(a_{D,i} - \sum_j a^j \tau_{ji}), \quad \frac{\partial \tau_{0i}}{\partial a^k} = -i\sum_j a^j \frac{\partial \tau_{ij}}{\partial a^k},$$

$$\frac{\partial \tau_{00}}{\partial a^k} = i\tau_{0k} - \sum_{ij} a^i a^j \frac{\partial \tau_{ij}}{\partial a^k}. \tag{2.17}$$

Using now (2.12) one can prove that s_D is a monodromy invariant,

$$\frac{\partial \mathcal{F}}{\partial s} = i\Big(2\mathcal{F} - \sum_i a^i a_{D,i}\Big) = 8\pi b_1 u \tag{2.18}$$

and from (2.17) and (2.18) we get

$$\tau_{0i} = 8\pi b_1 \frac{\partial u}{\partial a^i},$$

$$\tau_{00} = 8\pi i b_1 \Big(2u - \sum_i a^i \frac{\partial u}{\partial a^i}\Big) \tag{2.19}$$

Now we will present the transformation rules of the gauge couplings τ_{ij} under a monodromyy matrix Γ in $Sp(2r, \mathbf{Z})$. In terms of the local coordinates $a^i_\Gamma = C^{ip} a_{D,p}(a^j, s) + D^i_q a^q$ we have the couplings

$$\tau^\Gamma_{\alpha\beta} = \frac{\partial^2 \mathcal{F}}{\partial a^\alpha_\Gamma \partial a^\beta_\Gamma}. \tag{2.20}$$

The change of coordinates is given by the matrix:

$$\begin{pmatrix} \dfrac{\partial a^i_\Gamma}{\partial a^j} & \dfrac{\partial a^i_\Gamma}{\partial s} \\ \dfrac{\partial s}{\partial a^j} & \dfrac{\partial s}{\partial s} \end{pmatrix} = \begin{pmatrix} C^{ip}\tau_{pj} + D^i_j & C^{ip}\tau_{0p} \\ 0 & 1 \end{pmatrix}, \tag{2.21}$$

with inverse

$$\begin{pmatrix} \dfrac{\partial a^i}{\partial a^j_\Gamma} & \dfrac{\partial a^i}{\partial s} \\ \dfrac{\partial s}{\partial a^j_\Gamma} & \dfrac{\partial s}{\partial s} \end{pmatrix} = \begin{pmatrix} \Big((C\tau + D)^{-1}\Big)^i_j & -\Big((C\tau + D)^{-1}\Big)^i_k C^{kp}\tau_{p0} \\ 0 & 1 \end{pmatrix}. \tag{2.22}$$

Therefore we have:

$$\Big(\frac{\partial}{\partial a^j_\Gamma}\Big)_{\Gamma-\text{basis}} = \Big((C\tau + D)^{-1}\Big)^i_j \frac{\partial}{\partial a^i},$$

$$\left(\frac{\partial}{\partial s}\right)_{\Gamma-\text{basis}} = \frac{\partial}{\partial s} - \left[(C\tau+D)^{-1}C\tau\right]^i_0 \frac{\partial}{\partial a^i}; \tag{2.23}$$

which lead to the transformation rules for the couplings:

$$\begin{aligned} \tau^\Gamma_{ij} &= \Big(A\tau+B\Big)\Big(C\tau+D\Big)^{-1}_{ij}, \qquad \tau^\Gamma_{0i} = \tau_{0j}\Big((C\tau+D)^{-1}\Big)^j_i, \\ \tau^\Gamma_{00} &= \tau_{00} - \tau_{0i}\Big[(C\tau+D)^{-1}C\tau\Big]^i_0. \end{aligned} \tag{2.24}$$

3 Effective potential and vacuum structure

In this section we will obtain, starting from the formalism developed in the previous section, the effective potential in the Coulomb phase of the softly broken $N = 2$ theory, for a general group of rank r.

To break $N = 2$ down to $N = 0$ we freeze the spurion superfield to a constant. The lowest component is fixed by the scale Λ, and we only turn on the auxiliary F^0 (*i.e.* we take $D^0 = 0$). We must include in the effective lagrangian $r + 1$ vector multiplets, where r is the rank of the gauge group:

$$A^\alpha = (A^0, A^I), \qquad I = 1, \cdots, r. \tag{3.1}$$

There are submanifolds in the moduli space where extra states become massless and we must include them in the effective lagrangian. They are BPS states corresponding to monopoles or dyons, so we introduce n_H hypermultiplets near these submanifolds in the low energy description:

$$(M_i, \widetilde{M}_i), \qquad i = 1, \cdots, n_H \tag{3.2}$$

We suppose that these BPS states are mutually local, hence we can find a symplectic transformation such that they have $U(1)^r$ charges $(q^I_i, -q^I_i)$ with respect to the I-th $U(1)$ (we follow the $N = 1$ notation). The full $N = 2$ effective lagrangian contains two terms:

$$\mathcal{L} = \mathcal{L}_{\text{VM}} + \mathcal{L}_{\text{HM}}, \tag{3.3}$$

where $\mathcal{L}_{\text{VM}}$ is given in (2.1), and

$$\begin{aligned} \mathcal{L}_{\text{HM}} &= \sum_i \int d^4\theta \big(M_i^* \mathrm{e}^{2q_i^I V^{(I)}} M_i + \widetilde{M}_i^* \mathrm{e}^{-2q_i^I V^{(I)}} \widetilde{M}_i\big) \\ &+ \sum_{I,i} \Big(\int d^2\theta \sqrt{2} A^I q_i^I M_i \widetilde{M}_i + \text{h.c.}\Big) \end{aligned} \tag{3.4}$$

The terms in (3.3) contributing to the effective potential are

$$\begin{aligned} V &= b_{IJ}F^I\overline{F}^J + b_{0I}\big(F^0\overline{F}^I + \overline{F}^0F^I\big) + b_{00}|F^0|^2 \\ &+ \frac{1}{2}b_{IJ}D^ID^J + D^Iq_i^I(|m_i|^2 - |\widetilde{m}_i|^2) + |F_{m_i}|^2 + |F_{\widetilde{m}_i}|^2 \\ &+ \sqrt{2}\Big(F^Iq_i^Im_i\widetilde{m}_i + a^Iq_i^Im_iF_{\widetilde{m}_i} + a^Iq_i^I\widetilde{m}_iF_{m_i} + \text{h.c.}\Big), \end{aligned} \tag{3.5}$$

where all repeated indices are summed and $b_{\alpha\beta} = \text{Im}\tau_{\alpha\beta}/4\pi$. We eliminate the auxiliary fields and obtain:

$$\begin{aligned} D^I &= -(b^{-1})^{IJ}q_i^J(|m_i|^2 - |\widetilde{m}_i|^2), \\ F^I &= -(b^{-1})^{IJ}b_{0J}F^0 - \sqrt{2}(b^{-1})^{IJ}q_i^J\overline{m}_i\overline{\widetilde{m}}_i, \\ F_{m_i} &= -\sqrt{2}\overline{a}^Iq_i^I\overline{\widetilde{m}}_i, \qquad F_{\overline{m}_i} = -\sqrt{2}\overline{a}^Iq_i^I\overline{m}_i. \end{aligned} \tag{3.6}$$

We denote $(q_i,q_j) = \sum_{IJ}q_i^I(b^{-1})^{IJ}q_j^I$, $(q_i,b_0) = \sum_{IJ}q_i^I(b^{-1})^{IJ}b_{0J}$, $a\cdot q_i = \sum_I a^Iq_i^I$. Substituting in (3.5) we obtain:

$$\begin{aligned} V &= \frac{1}{2}\sum_{ij}(q_i,q_j)(|m_i|^2 - |\widetilde{m}_i|^2)(|m_j|^2 - |\widetilde{m}_j|^2) + 2\sum_{ij}(q_i,q_j)m_i\widetilde{m}_i\overline{m}_j\overline{\widetilde{m}}_j \\ &+ 2\sum_i|a\cdot q_i|^2(|m_i|^2 + |\widetilde{m}_i|^2) + \sqrt{2}\sum_i(q_i,b_0)\Big(F^0m_i\widetilde{m}_i + \overline{F}^0\overline{m}_i\overline{\widetilde{m}}_i\Big) \\ &- |F^0|^2\frac{\det b_{\alpha\beta}}{\det b_{IJ}}, \end{aligned} \tag{3.7}$$

where $\det b_{\alpha\beta}/\det b_{IJ} = b_{00} - b_{0I}(b^{-1})^{IJ}b_{0J}$ is the cosmological term. This term in the potential is a monodromy invariant. To prove this it is sufficient to prove invariance under the generators of the symplectic group $Sp(2r,\mathbf{Z})$:

$$\begin{pmatrix} A & 0 \\ 0 & (A^{\mathrm{T}})^{-1} \end{pmatrix}, \qquad A \in Gl(r,\mathbf{Z}),$$

$$T_\theta = \begin{pmatrix} \mathbf{1} & \theta \\ 0 & \mathbf{1} \end{pmatrix}, \qquad \theta_{ij} \in \mathbf{Z}, \qquad \Omega = \begin{pmatrix} 0 & \mathbf{1} \\ -\mathbf{1} & 0 \end{pmatrix}. \tag{3.8}$$

Invariance under T_θ and the matrix involving only A is obvious, and for Ω one can check it easily.

The vacuum structure is determined by the minima of (3.7). As in [21], we first minimize with respect to m_i, $\overline{m}_i$:

$$\begin{aligned} \frac{\partial V}{\partial\overline{m}_i} &= \sum_j(q_i,q_j)(|m_j|^2 - |\widetilde{m}_j|^2)m_i + 2|a\cdot q_i|^2m_i \\ &+ 2\sum_j(q_i,q_j)m_j\widetilde{m}_j\overline{\widetilde{m}}_i + \sqrt{2}\overline{F}^0(q_i,b_0)\overline{\widetilde{m}}_i = 0, \end{aligned} \tag{3.9}$$

$$\frac{\partial V}{\partial \overline{\widetilde{m}}_i} = \sum_j (q_i,q_j)(-|m_j|^2+|\widetilde{m}_j|^2)\widetilde{m}_i + 2|a\cdot q_i|^2\widetilde{m}_i$$
$$+ \; 2\sum_j (q_i,q_j)m_j\widetilde{m}_j\overline{m}_i + \sqrt{2}F^0(q_i,b_0)\overline{m}_i = 0. \qquad (3.10)$$

Multiplying (3.9) by $\overline{m}_i$, (3.10) by $\overline{\widetilde{m}}_i$ and substracting, we get

$$\sum_j (q_i,q_j)(|m_j|^2-|\widetilde{m}_j|^2)(|m_i|^2+|\widetilde{m}_i|^2) + 2|a\cdot q_i|^2(|m_i|^2-|\widetilde{m}_i|^2) = 0. \quad (3.11)$$

Suppose now that, for some indices $i \in I$, $|m_i|^2+|\widetilde{m}_i|^2 > 0$. Multiplying (3.11) by $|m_i|^2-|\widetilde{m}_i|^2$ and summing over i we obtain

$$\sum_{ij}(q_i,q_j)(|m_i|^2-|\widetilde{m}_i|^2)(|m_j|^2-|\widetilde{m}_j|^2) = -\sum_{i\in I}\frac{2|a\cdot q_i|^2}{|m_i|^2+|\widetilde{m}_i|^2}(|m_i|^2-|\widetilde{m}_i|^2)^2. \qquad (3.12)$$

The matrix $(b^{-1})^{IJ}$ is positive definite, and if the charge vectors q_i^I are linearly independent it follows that the matrix (q_i,q_j) is positive definite too. Then the l.h.s. of (3.12) is ≥ 0 while the r.h.s. is ≤ 0. The only way for this equation to be consistent is if

$$|m_i| = |\widetilde{m}_i|, \qquad i = 1,\cdots,n_H. \qquad (3.13)$$

In this case we can write the equation (3.9), after absorbing the phase of $F^0 = f_0 e^{i\gamma}$ in $\widetilde{m}_i$, as:

$$2|a\cdot q_i|^2 m_i + 2\sum_j (q_i,q_j)m_j\widetilde{m}_j\overline{\widetilde{m}}_i + \sqrt{2}f_0(q_i,b_0)\overline{\widetilde{m}}_i = 0. \qquad (3.14)$$

Multiplying by $\overline{m}_i$ and summing over i, we obtain

$$2\sum_i |a\cdot q_i|^2|m_i|^2 + \sqrt{2}f_0\sum_i (q_i,b_0)\overline{m}_i\overline{\widetilde{m}}_i = -2\sum_{ij}(q_i,q_j)m_j\overline{m}_i\widetilde{m}_j\overline{\widetilde{m}}_i, \quad (3.15)$$

hence $\sqrt{2}f_0\sum_i (q_i,b_0)\overline{m}_i\overline{\widetilde{m}}_i$ is real. We can insert in (3.7) and get the following expression for the effective potential:

$$V = -f_0^2\frac{\det b_{\alpha\beta}}{\det b_{IJ}} - 2\sum_{ij}(q_i,q_j)m_j\overline{m}_i\widetilde{m}_j\overline{\widetilde{m}}_i. \qquad (3.16)$$

If (3.13) holds, we can fix the gauge in the $U(1)^r$ factors and write

$$m_i = \rho_i, \qquad \widetilde{m}_i = \rho_i e^{i\phi_i} \qquad (3.17)$$

and (3.14) reads:

$$\rho_i^2\Big(|a\cdot q_i|^2+\sum_j(q_i,q_j)\rho_j^2 e^{i(\phi_j-\phi_i)}+\frac{f_0(q_i,b_0)}{\sqrt{2}}e^{-i\phi_i}\Big)=0. \tag{3.18}$$

Apart form the trivial solution $\rho_i=0$, we have:

$$|a\cdot q_i|^2+\sum_j(q_i,q_j)\rho_j^2 e^{i(\phi_j-\phi_i)}+\frac{f_0(q_i,b_0)}{\sqrt{2}}e^{-i\phi_i}=0 \tag{3.19}$$

and we can have a monopole (or dyon) VEV in some regions of the moduli space. Notice that for groups of rank $r>1$ there is a coupling between the different $U(1)$ factors and one needs a numerical study of the equation above once the values of the charges q_i^I are known. In addition, the moduli space is in that case very complicated and explicit solutions for the prepotential and gauge couplings of the $N=2$ theory are difficult to find. However we still can have some qualitative information in many cases under some mild assumptions, as we will see.

4 Vacuum structure of the $SU(2)$ Yang-Mills theory

4.1 The Seiberg-Witten Solution

In [1] Seiberg and Witten obtained the structure of the quantum moduli space of the $N=2$ $SU(2)$ Yang-Mills theory and also the exact solution for the prepotential $\mathcal{F}$ including all the non-perturbative corrections. Some of the properties of this solution are:

i) The moduli space $\mathcal{M}_u$ is parametrized by $u=\langle \mathrm{Tr}\phi^2\rangle$ and can be understood as the complex u-plane. The $SU(2)$ symmetry is never restored, and the theory stays in the Coulomb phase throughout the moduli space.

ii) $\mathcal{M}_u$ has a symmetry $u\to -u$ (the non-anomalous subset of the $U(1)_R$ group), and at the points $u=\Lambda^2$, $-\Lambda^2$ singularities in the holomorphic prepotential $\mathcal{F}$ develop. Physically they correspond respectively to a massless monopole and dyon with charges $(q_e,q_m)=(0,1),(-1,1)$. Hence near $u=\Lambda^2$, $-\Lambda^2$ the correct effective action should include together with the photon vector multiplet monopole or dyon hypermultiplets.

iii) The vector ${}^t v=(a_D,a)$ defines a flat $SL_2(\mathbf{Z})$ vector bundle over the moduli space $\mathcal{M}_u$. Its properties are determined by the singularities and the monodromies around them. Since $\partial^2\mathcal{F}/\partial a^2$ or $\partial a_D/\partial a$ is the coupling constant, these data are obtained from the β-function in the three patches: large-u, the Higgs phase, the monopole and the dyon regions. From the BPS mass

formula[22,23] the mass of a BPS state of charge (q_e, q_m) (with q_e, q_m coprime for the charge to be stable) is:

$$M = \sqrt{2}|q_e a + q_m a_D|. \tag{4.1}$$

If at some point u_0 in $\mathcal{M}_u$, $M(u_0) = 0$, the monodromy around this point is given by [1,2,6]

$$\begin{pmatrix} a_D \\ a \end{pmatrix} \to M(q_e, q_m) \begin{pmatrix} a_D \\ a \end{pmatrix}, \tag{4.2}$$

$$M(q_e, q_m) = \begin{pmatrix} 1 + 2q_e q_m & 2q_e^2 \\ -2q_m^2 & 1 - 2q_e q_m \end{pmatrix}. \tag{4.3}$$

Also for large u, $\mathcal{F}$ is dominated by the perturbative one loop contribution, obtained from the one loop β-function:

$$\mathcal{F}_{1-\text{loop}}(a) = \frac{i}{2\pi} a^2 \ln \frac{a^2}{\Lambda} \tag{4.4}$$

Hence we also have monodromy at infinity. The three generators of the monodromy are therefore:

$$M_\infty = \begin{pmatrix} -1 & 2 \\ 0 & -1 \end{pmatrix}, \quad M_{\Lambda^2} = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}, \quad M_{-\Lambda^2} = \begin{pmatrix} -1 & 2 \\ -2 & 3 \end{pmatrix}; \tag{4.5}$$

and they satisfy:

$$M_\infty = M_{\Lambda^2} M_{-\Lambda^2}. \tag{4.6}$$

These matrices generate the subgroup $\Gamma_2 \subset SL_2\mathbf{Z}$ of 2×2 matrices congruent to the unit matrix modulo 2.

We learn from (4.1)-(4.3) that in the Higgs, monopole and dyon patches, the natural independent variables to use are respectively $a^{(h)} = a$, $a^{(m)} = -a_D$, $a^{(d)} = a_D - a$. Thus in each patch we have a different prepotential:

$$\mathcal{F}^{(h)}(a), \quad \mathcal{F}^{(m)}(a^{(m)}), \quad \mathcal{F}^{(d)}(a^{(d)}). \tag{4.7}$$

iv) The explicit form of $a(u)$, $a_D(u)$ is given in terms of the periods of a meromorphic differential of the second kind on a genus one surface described by the equation:

$$y^2 = (x^2 - \Lambda^4)(x - u), \tag{4.8}$$

describing the double covering of the plane branched at $\pm\Lambda^2$, u, ∞. We choose the cuts $\{-\Lambda^2, \Lambda^2\}$, $\{u, \infty\}$. The correctly normalized meromorphic 1-form is:

$$\lambda = \Lambda \frac{\sqrt{2}}{2\pi} \frac{dx\sqrt{x - u/\Lambda^2}}{\sqrt{x^2 - 1}}. \tag{4.9}$$

Then:

$$a(u) = \Lambda \frac{\sqrt{2}}{\pi} \int_{-1}^{1} \frac{dt\sqrt{u/\Lambda^2 - t}}{\sqrt{1-t^2}}; \tag{4.10}$$

$$a_D(u) = \Lambda \frac{\sqrt{2}}{\pi} \int_{1}^{u/\Lambda^2} \frac{dt\sqrt{u/\Lambda^2 - t}}{\sqrt{1-t^2}}. \tag{4.11}$$

Using the hypergeometric representation of the elliptic functions[24]:

$$K(k) = \frac{\pi}{2} F(1/2, 1/2, 1; k^2); \quad K'(k) = K(k');$$

$$E(k) = \frac{\pi}{2} F(-1/2, 1/2, 1; k^2); \quad E'(k) = E(k'), \quad k'^2 + k^2 = 1, \tag{4.12}$$

we obtain :

$$k^2 = \frac{2}{1 + u/\Lambda^2}, \quad k'^2 = \frac{u - \Lambda^2}{u + \Lambda^2}, \tag{4.13}$$

$$a(u) = \frac{4\Lambda}{\pi k} E(k), \quad a_D(u) = \frac{4\Lambda}{i\pi} \frac{E'(k) - K'(k)}{k}. \tag{4.14}$$

Using the elliptic function identities:

$$\frac{dE}{dk} = \frac{E - K}{k}, \quad \frac{dK}{dk} = \frac{1}{kk'^2}(E - k'^2 K), \tag{4.15}$$

$$\frac{dE'}{dk} = -\frac{k}{k'^2}(E' - K'), \quad \frac{dK'}{dk} = -\frac{1}{kk'^2}(E' - k^2 K'), \tag{4.16}$$

the coupling constant becomes:

$$\tau_{11} = \frac{\partial a_D}{\partial a} = \frac{da_D/dk}{da/dk} = \frac{iK'}{K}, \tag{4.17}$$

which is indeed the period matrix of the curve (4.8).

4.2 Vacuum structure of the softly broken $SU(2)$ theory

When we softly break the $N = 2$ $SU(2)$ Yang-Mills theory we obtain an effective potential including the couplings τ_{01} and τ_{00}. In the normalization of[1], and with $b_1 = 1/4\pi^2$, the spurion-induced couplings are

$$\tau_{01} = \frac{2}{\pi} \frac{\partial u}{\partial a}, \qquad \tau_{00} = \frac{2i}{\pi} \Big(2u - a \frac{\partial u}{\partial a} \Big). \tag{4.18}$$

The monodromy transformations of the couplings (2.24) have a simple expression in the $SU(2)$ case:

$$\begin{aligned} \tau_{11}^{\Gamma} &= \frac{\alpha\tau_{11}+\beta}{\gamma\tau_{11}+\delta}, \qquad \tau_{01}^{\Gamma} = \frac{\tau_{01}}{\gamma\tau_{11}+\delta}, \\ \tau_{00}^{\Gamma} &= \tau_{00} - \frac{\gamma\tau_{01}^2}{\gamma\tau_{11}+\delta}. \end{aligned} \tag{4.19}$$

From the exact Seiberg-Witten solution (4.10), (4.11) and the previous equations we can compute the couplings τ_{ij} in the Higgs and monopole region.

i) Higgs region:

$$a_D^{(h)} = \frac{4\Lambda}{i\pi}\frac{E'-K'}{k}, \quad a^{(h)} = \frac{4\Lambda}{\pi k}E(k),$$

$$\tau_{11}^{(h)} = \frac{iK'}{K}, \quad \tau_{01}^{(h)} = \frac{2\Lambda}{kK}, \quad \tau_{00}^{(h)} = -\frac{8i\Lambda^2}{\pi}\Big(\frac{E-K}{k^2K}+\frac{1}{2}\Big). \tag{4.20}$$

ii) Monopole region:

$$a_D^{(m)} = \frac{4\Lambda}{\pi k}E(k), \quad a^{(m)} = -\frac{4\Lambda}{i\pi}\frac{E'-K'}{k},$$

$$\tau_{11}^{(m)} = \frac{iK}{K'}, \quad \tau_{01}^{(m)} = \frac{2i\Lambda}{kK'}, \quad \tau_{00}^{(m)} = \frac{8i\Lambda^2}{\pi}\Big(\frac{E'}{k^2K'}-\frac{1}{2}\Big). \tag{4.21}$$

In the analysis of the effective potential (3.7) we must first minimize with respect to the monopole (or dyon) field. For $r=1$ the equation for the VEV (3.19) is

$$\rho^2 + b_{11}|a|^2 + \frac{b_{01}e^{-i\phi}f_0}{\sqrt{2}} = 0, \tag{4.22}$$

and the last term must be real so $e^{-i\phi} = \epsilon = \pm 1$. The charge is $q=1$ in the $SU(2)$ Yang-Mills theory, both in the monopole and in the dyon regions. Apart from the solution $\rho = 0$ we can have

$$\rho^2 = -b_{11}|a|^2 - \frac{b_{01}\epsilon f_0}{\sqrt{2}} > 0. \tag{4.23}$$

Note that $b_{11} = \frac{1}{4\pi}\mathrm{Im}\ \tau_{11}$ is always positive, and therefore (4.23) determines a region in the u-plane where the monopoles acquire a VEV. Depending on the sign of b_{01} we choose the sign of ϵ. In fact we can replace (4.23) by:

$$\rho^2 = -b_{11}|a|^2 + \frac{1}{\sqrt{2}}|b_{01}|f_0 > 0 \tag{4.24}$$

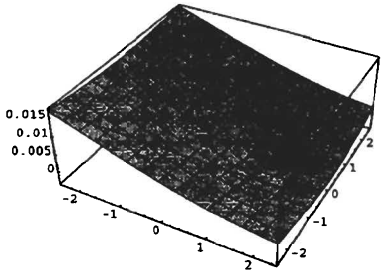


Figure 1: Effective potential, $V^{(h)}$, (4.26).

and f_0 is always measured in units of Λ. Thus for the numerical plots we set $\Lambda = 1$. From (3.16) we get the effective potential:

$$V = -\frac{2}{b_{11}}\rho^4 - \frac{\det b}{b_{11}} f_0^2 \tag{4.25}$$

This is good news. It implies that the region where the monopoles acquire a VEV is energetically favored, and we have a first order phase transition to confinement. Depending on the sign of b_{01}, m and $\widetilde{m}$ are either aligned or antialigned. The $SU(2)_R$ symmetry of $N = 2$ supersymmetry is broken by the explicit off-diagonal term $b_{01} m\widetilde{m}/b_{11}$ in (3.7) and by the VEV $\rho \neq 0$.

Where $\rho^2 \to 0$, the potential maps smoothly onto the potential for the Higgs region,

$$V^{(h)} = -\frac{\det b^{(h)}}{b_{11}^{(h)}} f_0^2, \tag{4.26}$$

where, we recall, $\det b/b_{11}$ is monodromy-invariant. In the monopole region, a nonzero monopole VEV is favoured, and the effective potential is given by (4.25) and written in terms of magnetic variables:

$$V^{(m)} = -\frac{2}{b_{11}^{(m)}}\rho^4 - \frac{\det b^{(m)}}{b_{11}^{(m)}} f_0^2 \tag{4.27}$$

where $b^{(h)}$, $b^{(m)}$ are given in (4.20), (4.21).

In the Higgs region, the effective potential is given by (4.26) and we plot it in fig. 1. It has no minimum outside the monopole region near $u = \Lambda^2$ (where, as we shall see, the energy can be further lowered by giving the monopoles a VEV). One sees that the shape of the potential makes the fields roll towards the

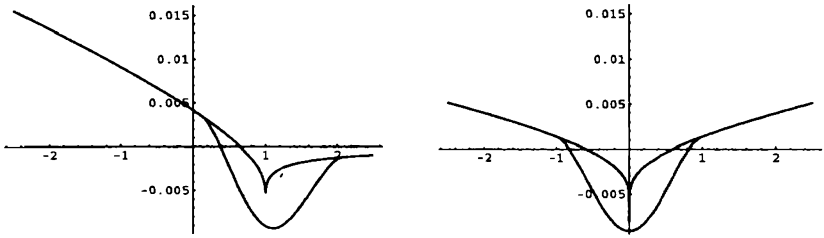


Figure 2: Effective potential, $V^{(h)}$, (4.26) (top) and, $V^{(m)}$, (4.27) (bottom) along the real axis (left) and for $u = \Lambda^2(1 + iy)$ (right). Both are plotted for $f_0 = 0.3\Lambda$.

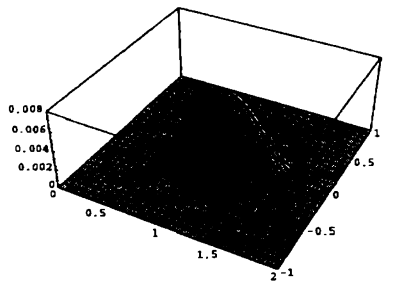


Figure 3: Monopole expectation value ρ^2 for $f_0 = 0.1\Lambda$ on the u-plane.

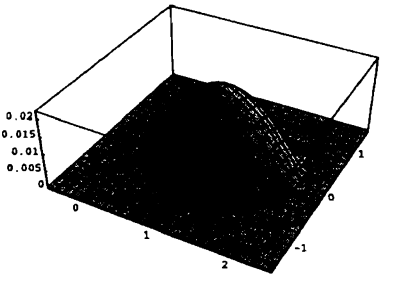


Figure 4: Monopole expectation value ρ^2 for $f_0 = 0.3\Lambda$ on the u-plane.

monopole region. In fig. 2, we plot slices of the potential $V^{(h)}$ along the real u-axis and parallel to the imaginary u-axis with $\mathrm{Re}(u) = \Lambda^2$. For comparison, we also plot $V^{(m)}$. Note that they agree in the Higgs region (where the monopole VEV vanishes), and that $V^{(m)}$ lowers the energy (and smooths out the cusp in $V^{(h)}$ at $u = \Lambda^2$) in the monopole region.

Next we look at the monopole region (4.24). a (*i.e.* $a^{(m)}$) is a good coordinate in this region vanishing at $u = \Lambda^2$. As soon as f_0 is turned on monopole condensation and confinement occur. In figs. 3,4 we plot ρ^2 in the u-plane for values of $f_0 = 0.1\Lambda$, 0.3Λ; and in figs. 5,6 the effective potential (4.27) for the same values of the supersymmetry breaking parameter f_0.

One can see that the minimum is stable and that the size of the monopole VEV is $\sim f_0$. There are two features worth noticing. The first is that the absolute minimum occurs along the real u-axis. This is seen numerically and also as a consequence of the reality properties of the elliptic functions. Second, as f_0 is increased, the region where (4.24) holds becomes wider. This is seen in

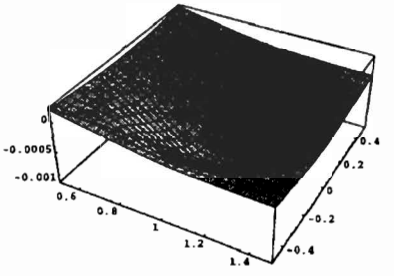


Figure 5: Effective potential (4.27) for $f_0 = 0.1\Lambda$.

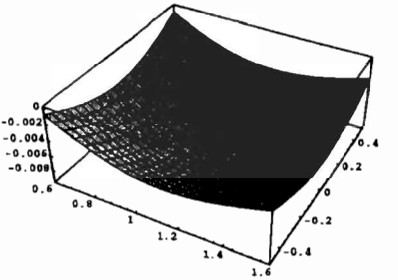


Figure 6: Effective potential (4.27) for $f_0 = 0.3\Lambda$.

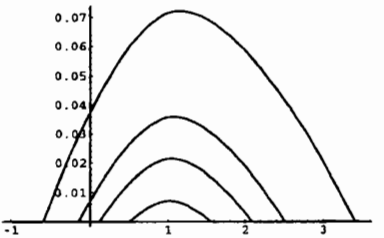


Figure 7: Plot of ρ^2 along the real u-axis, for $f_0/\Lambda =$ (from bottom to top) 0.1, 0.3, 0.5, 1.0.

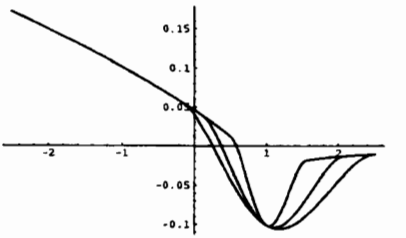


Figure 8: $V^{(m)}/f_0^2$ along the real u-axis for $f_0 = 0.1\Lambda$ (top), 0.5Λ (middle) and Λ (bottom).

fig. 7, where ρ^2 is plotted along the real u-axis as a function of f_0. Accordingly, the minimum of the effective potential moves to the right along the real u-axis, as one can see in fig. 8, where $V^{(m)}/f_0^2$ is plotted for three increasing values of f_0 (we have divided by f_0^2 to fit the three potentials on the same graph).

Finally, we turn to the dyon region. To understand what happens in the dyon region, we study the transformation rules of the τ_{ij} couplings under the residual $\mathbf{Z}_8 \subset U(1)_R$ symmetry whose generator acts on the u-plane as $u \mapsto -u$. The reason why we need to analyze in general the behavior under $\mathbf{Z}_8$ is because the representation we have chosen for the Seiberg-Witten solution in sections 2,3 is well adapted to study the monopole region. Naively applying them to the dyon region, we may encounter some discontinuities due to the position of the cuts. Outside the curve of marginal stability one can write the prepotential as [1]:

$$\mathcal{F} = \frac{i}{2\pi} a^2 \log \frac{a^2}{\Lambda^2} + a^2 \sum_{k \geq 1} c_k \left(\frac{\Lambda}{a}\right)^{4k}. \tag{4.28}$$

If $\omega = e^{2\pi i/8}$ is the generator of the $\mathbf{Z}_8$ symmetry, it is easy to show that the couplings τ_{ij} transform according to[b]:

$$\begin{aligned} a &\mapsto ia, \qquad a_D \mapsto i(a_D - a), \\ \tau_{11} &\mapsto \tau_{11} - 1, \qquad \tau_{01} \mapsto i\tau_{01}, \qquad \tau_{00} \mapsto -\tau_{00}. \end{aligned} \tag{4.29}$$

So the relation between the dyon and monopole variables is:

$$\begin{aligned} a^{(d)}(u) &= ia^{(m)}(-u), \qquad a_D^{(d)}(u) = i\left(a_D^{(m)}(-u) - a^{(m)}(-u)\right), \\ \tau_{11}^{(d)}(u) &= \tau_{11}^{(m)}(-u) - 1, \qquad \tau_{01}^{(d)}(u) = i\tau_{01}^{(m)}(-u), \quad \tau_{00}^{(d)}(u) = -\tau_{00}^{(m)}(-u), \end{aligned} \tag{4.30}$$

with $a_D^{(d)} = -a_D$. Using the expressions for the monopole couplings in (4.21), which are well-behaved near $u = \Lambda^2$, we obtain expressions for the dyon couplings which are well-behaved near $u = -\Lambda^2$. The analysis of (4.24) changes crucially once these rules are implemented. Near the monopole region $a^{(m)} \sim i(u - \Lambda^2)$, hence $\tau_{01}^{(m)} \sim i$ is purely imaginary. In (4.27) although

[b] There is one more aspect of the $\mathbf{Z}_8$ transformation rules worth noticing. If we implement these rules we find that the condensate moves to the dyon region, and one might be tempted to conclude that with this choice it is the dyon that condenses. This is not the case. Using the one-loop β-function, we know that $\Lambda^4 \sim \exp(-\frac{8\pi^2}{g^2} + i\theta)$. The action of $\mathbf{Z}_8$ amounts to the change $\Lambda \mapsto i\Lambda$ or what is the same, $\theta \mapsto \theta + 2\pi$. Using the relation found in [25], when we make this change the massless state at $u = -\Lambda^2$ (before supersymmetry breaking) has zero electric charge, while the state at $u = \Lambda^2$ acquires charge one. Thus we find again a monopole condensate, in a way consistent with the $\mathbf{Z}_2$-symmetry.

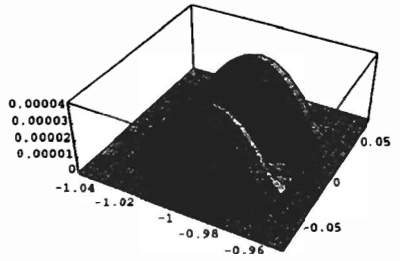


Figure 9: Dyon expectation value $\rho^2_{(d)}$ for $f_0 = 0.3\Lambda$ on the u-plane.

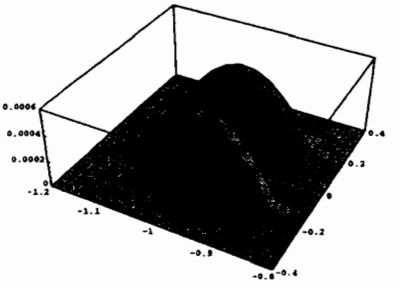


Figure 10: Dyon expectation value $\rho^2_{(d)}$ for $f_0 = \Lambda$ on the u-plane.

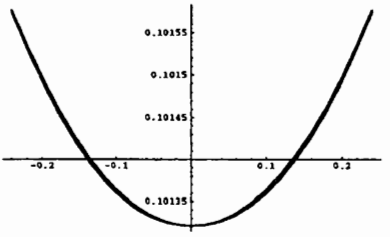


Figure 11: Plot of $V^{(h)}(u)$ (top) and $V^{(d)}(u)$ (bottom) versus $\mathrm{Im}(u)$ for $\mathrm{Re}(u) = -\Lambda^2$ and $f_0 = \Lambda$.

b_{11} diverges at $u = \Lambda^2$ the divergence is cancelled by the vanishing of $a^{(m)}$ at the same point. Since $\mathrm{Im}\tau^{(m)}_{01} > 0$ as soon as $f_0 \neq 0$ the monopoles condense. Using (4.30), however, we see that $a^{(d)} \sim (u+\Lambda^2)$ with a real coefficient. Thus $\mathrm{Im}\tau^{(d)}_{01} = 0$ at $u = -\Lambda^2$ and we conclude from (6.21) that the dyon condensate *vanishes* along the real u-axis. Nevertheless, a dyon condensate *is* energetically favoured in a pair of complex-conjugate regions in the u-plane centered about $u = -\Lambda^2$. We plot $\rho^2_{(d)}$, for two different values of f_0 in figs. 9,10.

Unlike the monopole VEV, the magnitude of the dyon VEV is *tiny* on the scale of $V^{(h)}$. It therefore makes an all-but-negligible contribution to the effective potential (fig. 11). In particular, $V^{(d)}$ does *not* have a minimum in the dyon region. The only minimum of the full effective potential is the one we previously found in the monopole region.

As we have already noted, the monopole region (in which $\rho^2_{(m)} \neq 0$) expands as f_0 is increased. Eventually, for $f_0 \sim 1.3\Lambda$, it reaches the dyon region (in which $\rho^2_{(d)} \neq 0$). At this point, it is clear that our whole approximation of

including *just* the monopole field (or *just* the dyon field) in the effective action breaks down.

What are the other limitations of our approximations? First, we have neglected certain soft supersymmetry breaking terms which arise when we derive the soft breaking terms from spontaneously broken $N = 2$ supergravity. These additional terms scale to zero in the rigid limit, that is, they are suppressed by powers of $\log \frac{\Lambda}{M_{\rm Pl}}$ or $\frac{\Lambda}{M_{\rm Pl}}$ and, for our purposes are negligible. We have also neglected higher-spinor-derivative corrections to the Seiberg-Witten effective action. These clearly cannot affect the vacuum structure in the supersymmetric limit. They also, *by definition* must be supersymmetric; otherwise they lead to explicitly hard supersymmetry breaking terms, which is an entirely different matter from the soft supersymmetry breaking we are considering. Nevertheless, once supersymmetry is broken, they can, in principle, lead to corrections to the scalar potential suppressed by higher powers of f_0^2/Λ^2. For the moderate values of f_0 that we are considering, these corrections are numerically rather small, and do not affect the qualitative features of the solutions we have found. *A priori*, if the higher spinor derivative terms in the Seiberg-Witten effective action were known, we could systematically improve our approximations by going to higher order in f_0^2/Λ^2.

However, the fundamental obstacle to pushing our approximation to larger values of the soft supersymmetry breaking parameters would remain. The mutual non-locality of the monopoles and dyons leads to our inability to calculate the effective potential where the monopole and dyon regions overlap. Since this is, at least initially, far from the monopole vacuum, we expect that the monopole vacuum persists, at least as metastable minimum, even beyond the critical value of f_0. But we do not know when (or if) a new, lower minimum develops once the monopole and dyon regions overlap. If a new vacuum does appear there, then we would have a first order phase transition to this new confining phase [c]. This raises the exciting possibility that the correct description of the QCD vacuum requires the introduction of mutually non-local monopoles and dyons. Phases of this nature have been shown to arise in the $N = 2$ moduli space for gauge group $SU(3)$ [17]. Perhaps the way to approach the true QCD vacuum in the correct phase is to start with one of these $N = 2$-superconformal field theories and turn on a relevant, soft supersymmetry-breaking perturbation.

[c] An explicit realization of this phase transition due to the overlapping of monopole and dyon regions occurs in the softly broken $SU(2)$ theory with one massless hypermultiplet [26]

5 Vacuum structure of the $SU(N)$ Yang-Mills theory

The moduli space of vacua of the $N = 2$ $SU(N)$ Yang-Mills can be parametrized in a gauge-invariant way by the elementary symmetric polynomials s_l, $l = 2, \cdots, N$ in the eigenvalues of $\langle\phi\rangle$, ϕ_i. The vacuum structure of the theory is associated to the hyperelliptic curve [3]:

$$y^2 = P(x)^2 - \Lambda^{2N},$$

$$P(x) = \frac{1}{2}\det(x - \langle\phi\rangle) = \frac{1}{2}\prod_i (x - \phi_i), \tag{5.1}$$

where Λ is the dynamical scale of the $SU(N)$ theory and $P(x)$ can be written in terms of the variables s_l as $P(x) = 1/2\sum_l (-1)^l s_l x^{N-l}$. Once the hyperelliptic curve is known, one can compute in principle the metric on the moduli space and the exact quantum prepotential, but explicit solutions are difficult to find (they have been obtained in [4] for the $SU(3)$ case). But as in the $SU(2)$ case one expects that the minima of the effective potential for the $SU(N)$ theory are near the $N = 1$ points (at least for a small supersymmetry breaking parameter). The physics of the $N = 1$ points in $SU(N)$ theories has a much simpler description because it involves only small regions of the moduli space, and has been studied in [5]. The $N = 1$ points correspond to points in the moduli space where $N - 1$ monopoles coupling to each $U(1)$ become massless simultaneously. From the point of view of the hyperelliptic curve this corresponds to a simultaneous degeneration of the $N - 1$ α-cycles, associated to monopoles. This means in turn that the polynomial $P(x)^2 - \Lambda^{2N}$ must have $N - 1$ double zeros and two single zeros. If we set $\Lambda = 1$, this can be achieved with the Chebyshev polynomials

$$P(x) = \cos\left(N \arccos\frac{x}{2}\right), \tag{5.2}$$

and the corresponding eigenvalues are $\phi_i = 2\cos\pi(i - \frac{1}{2})/N$. The other $N - 1$ points, corresponding to the simultaneous condensation of $N - 1$ mutually local dyons, are obtained with the action of the anomaly-free discrete subgroup $\mathbf{Z}_{4N} \subset U(1)_R$. One can perturb slightly the curve (5.2) to obtain the effective lagrangian (or equivalently, the prepotential) at lowest order. What is found is that, in terms of the dual monopole variables $a_{D,I}$, the $U(1)$ factors are decoupled and $\tau^D_{IJ} \sim \delta_{IJ}\tau_I$. Near the $N = 1$ point where $N - 1$ monopoles become massless one can then simplify the equation (3.19) for the monopole VEVs, because $q^I_i = \delta^I_i$, $(b^{-1})^{IJ} = \delta^{IJ}b_I^{-1}$. The equation reduces then to $r = N - 1$ $SU(2)$-like equations, and in particular the phase factors $e^{-i\phi_I}$ must

be real. We then set $e^{-i\phi_I} = \epsilon_I$, $\epsilon_I = \pm 1$. The VEVs are determined by:

$$\rho_I^2 = -b_I |a_{D,I}|^2 - \frac{f_0 b_{0I} \epsilon_I}{\sqrt{2}}, \qquad I = 1, \cdots, r. \tag{5.3}$$

The effective potential (3.16) reads:

$$V = -f_0^2 \Big(b_{00} - \sum_I \frac{b_{0I}^2}{b_I} \Big) - 2 \sum_I \frac{1}{b_I} \rho_I^4. \tag{5.4}$$

The quantities that control, at least qualitatively, the vacuum structure of the theory, are b_{0I} and b_{00}. If $b_{0I} \neq 0$ at the $N = 1$ points, we have a monopole VEV for ρ_I around this point. If $b_{0I} = 0$, we still can have a VEV, as it happens in the $SU(2)$ case in the dyon region, but we expect that it will be too tiny to produce a local minimum. When one has monopole condensation at one of these $N = 1$ points in all the $U(1)$ factors, the value of the potential at this point is given by

$$V = -f_0^2 b_{00}, \tag{5.5}$$

and if the local minimum is very near to the $N = 1$ point, we can compare the energy of the different $N = 1$ points according to (5.5) and determine the true vacuum of the theory. Hence, to have a qualitative picture of the vacuum structure, and if we suppose that the minima of the effective potential will be located near the $N = 1$ points, we only need to evaluate b_{0I}, b_{00} at these points. This can be done using the explicit solution in [5] and the expressions (2.19).

To obtain the correct normalization of the constant appearing in (2.18) we can evaluate $\sum_I a_{D,I} da/du - a da_{D,I}/du$ in the $N = 1$ points, obtaining the constant value $4\pi i b_1$. The value of the quadratic Casimir at the $N = 1$ point described by (5.2) is

$$u = \langle \mathrm{Tr} \phi^2 \rangle = 4 \sum_{i=1}^{N} \cos^2 \frac{\pi(i - 1/2)}{N} = 2N, \tag{5.6}$$

and the values at the other $N = 1$ points are given by the action of $\mathbf{Z}_N$ (u has charge 4 under $U(1)_R$): $u^{(k)} = 2\omega^{4k} N$, $\omega = e^{\pi i/2N}$ with $k = 0, \cdots, N-1$. To compute τ_{0I} we must also compute $\partial u / \partial a_{D,I}$. Using the results of [5], we have:

$$\frac{\partial u}{\partial a_{D,I}} = -4i \sin \frac{\pi I}{N}, \tag{5.7}$$

and using $b_1 = 2N/16\pi^2$, we obtain

$$\tau_{0I} = 4\pi b_1 \frac{\partial u}{\partial a_{DI}} = -\frac{2Ni}{\pi} \sin \frac{\pi I}{N}. \tag{5.8}$$

At the $N = 1$ point where $N - 1$ monopoles condense, $a_{D,I} = 0$, therefore

$$\tau_{00} = 8\pi i u = \frac{2i}{\pi} N^2. \tag{5.9}$$

(5.8) indicates that monopoles condense at this point in all the $U(1)$ factors, but with different VEVs. This is a consequence the spontaneous breaking of the S_N symmetry permuting the $U(1)$ factors [5].

To study the other $N = 1$ points we must implement the $\mathbf{Z}_N$ symmetry in the u-plane. The local coordinates $a_I^{(k)}$ vanishing at these points are given by a $Sp(2r, \mathbf{Z})$ transformation acting on the coordinates a_I, $a_{D,I}$ around the monopole point. The $\mathbf{Z}_N$ symmetry implies that

$$\frac{\partial u}{\partial a_I^{(k)}}(u^{(k)}) = \omega^{2k} \frac{\partial u}{\partial a_{D,I}}(u^{(0)}), \tag{5.10}$$

and then we get

$$b_{0I}^{(k)} = \frac{1}{4\pi} \mathrm{Im} \tau_{0I}^{(k)} = -\frac{N}{2\pi^2} \cos\frac{\pi k}{N} \sin\frac{\pi I}{N},$$

$$b_{00}^{(k)} = \frac{1}{4\pi} \mathrm{Im} \tau_{00}^{(k)} = \frac{1}{2}\left(\frac{N}{\pi}\right)^2 \cos\frac{2\pi k}{N}. \tag{5.11}$$

The first equation tells us that generically we will have dyon condensation at all the $N = 1$ points, and the second equation together with (5.5) implies that the condensate of $N - 1$ monopoles at $u = 2N$ is energetically favoured, and then it will be the true vacuum of the theory. Notice that the $\mathbf{Z}_N$ symmetry works in such a way that the size of the condensate, given by $|\cos\frac{\pi k}{N}|$, corresponds to an energy given by $-\cos\frac{2\pi k}{N}$: as one should expect, the bigger the condensate the smaller its energy. In fact, for N even the $N = 1$ point corresponding to $k = N/2$ has no condensation. In this case the energy is still given by (5.5), as the effective potential equals the cosmological term with $b_{0I} = 0$, and is the biggest one.

6 Mass formula in softly broken $N = 2$ theories

6.1 A general mass formula

In some cases the mass spectrum of a softly broken supersymmetric theory is such that the graded trace of the square of the mass matrix is zero as it happens in supersymmetric theories [27]. We will see in this section that this is also the case when we softly break $N = 2$ supersymmetry with a dilaton spurion.

We will then compute the trace of the squared mass matrix which arises from the effective lagrangian (3.3), once the supersymmetry breaking parameter is turned on. The fermionic content of the theory is as follows: we have fermions ψ^I, λ^I coming from the $N = 2$ vector multiplet A^I (in $N = 1$ language, ψ^I comes from the $N = 1$ chiral multiplet and λ^I from the $N = 1$ vector multiplet). We also have "monopolinos" ψ_{m_i}, $\psi_{\widetilde{m}_i}$ from the n_H matter hypermultiplets. To obtain the fermion mass matrix, we just look for fermion bilinears in (3.3). From the gauge kinetic part and the Kähler potential in $\mathcal{L}_{\rm VM}$ we obtain:

$$\frac{i}{16\pi}F^\alpha\partial_\alpha\tau_{IJ}\lambda^I\lambda^J + \frac{i}{16\pi}\overline{F}^\alpha\partial_\alpha\tau_{IJ}\psi^I\psi^J. \tag{6.12}$$

where $F^0 = f_0$ and the auxiliary fields F^I are given in (3.6). From the kinetic term and the superpotential in $\mathcal{L}_{\rm HM}$ we get:

$$\begin{aligned} & i\sqrt{2}\sum_i q_i\cdot\lambda\left(\overline{m}_i\psi_{m_i} - \overline{\widetilde{m}}_i\psi_{\widetilde{m}_i}\right) \\ - \quad & \sqrt{2}\sum_i\left(a\cdot q_i\psi_{m_i}\psi_{\widetilde{m}_i} + q_i\cdot\psi\psi_{\widetilde{m}_i}m_i + q_i\cdot\psi\psi_{m_i}\widetilde{m}_i\right) \end{aligned} \tag{6.13}$$

If we order the fermions as $(\lambda, \psi, \psi_{m_i}, \psi_{\widetilde{m}_i})$ and denote $\mu^{IJ} = iF^\alpha\partial_\alpha\tau_{IJ}/4\pi$, $\hat{\mu}^{IJ} = i\overline{F}^\alpha\partial_\alpha\tau_{IJ}/4\pi$, the "bare" fermionic mass matrix reads:

$$M_{1/2} = \begin{pmatrix} \mu/2 & 0 & i\sqrt{2}q_i^I\overline{m}_i & -i\sqrt{2}q_i^I\overline{\widetilde{m}}_i \\ 0 & \hat{\mu}/2 & -\sqrt{2}q_i^I\widetilde{m}_i & -\sqrt{2}q_i^I m_i \\ i\sqrt{2}q_i^I\overline{m}_i & -\sqrt{2}q_i^I\widetilde{m}_i & 0 & -\sqrt{2}a\cdot q_i \\ -i\sqrt{2}q_i^I\overline{\widetilde{m}}_i & -\sqrt{2}q_i^I m_i & -\sqrt{2}a\cdot q_i & 0 \end{pmatrix}, \tag{6.14}$$

but we must take into account the wave function renormalization for the fermions λ^I, ψ^I and consider

$$\mathcal{M}_{1/2} = ZM_{1/2}Z, \qquad Z = \begin{pmatrix} b^{-1/2} & 0 & 0 & 0 \\ 0 & b^{-1/2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{6.15}$$

The trace of the squared fermionic matrix can be easily computed:

$$\begin{aligned} \mathrm{Tr}\mathcal{M}_{1/2}\mathcal{M}_{1/2}^\dagger &= \frac{1}{4}\mathrm{Tr}[\mu b^{-1}\overline{\mu}b^{-1} + \hat{\mu}b^{-1}\overline{\hat{\mu}}b^{-1}] \\ &+ 4\sum_i|a\cdot q_i|^2 + 8\sum_i(q_i, q_i)(|m_i|^2 + |\widetilde{m}_i|^2). \end{aligned} \tag{6.16}$$

The scalars in the model are the monopole fields m_i, $\widetilde{m}_i$ and the lowest components of the $N = 1$ chiral superfields in the A^I, a^I. To compute the trace of the scalar mass matrix we need

$$\frac{\partial^2 V}{\partial m_i \partial \overline{m}_i} = \sum_l (q_i, q_l)(|m_l|^2 - |\widetilde{m}_l|^2) + (q_i, q_i)(|m_i|^2 + 2|\widetilde{m}_i|^2) + 2|a \cdot q_i|^2,$$

$$\frac{\partial^2 V}{\partial \widetilde{m}_i \partial \overline{\widetilde{m}}_i} = -\sum_l (q_i, q_l)(|m_l|^2 - |\widetilde{m}_l|^2) + (q_i, q_i)(2|m_i|^2 + |\widetilde{m}_i|^2) + 2|a \cdot q_i|^2,$$

$$\begin{aligned} \frac{\partial^2 V}{\partial a^I \partial \overline{a}_J} &= f_0^2 \frac{\partial^2 (b_0, b_0)}{\partial a^I \partial \overline{a}_J} + 2 \sum_{k,l} \frac{\partial^2 (q_k, q_l)}{\partial a^I \partial \overline{a}_J} m_k \widetilde{m}_k \overline{m}_l \overline{\widetilde{m}}_l \\ &+ 2 \sum_k q_k^I q_k^J (|m_k|^2 + |\widetilde{m}_k|^2) \\ &+ \sqrt{2} \sum_k \frac{\partial^2 (q_k, b_0)}{\partial a^I \partial \overline{a}_J} f_0 (m_k \widetilde{m}_k + \overline{m}_k \overline{\widetilde{m}}_k). \end{aligned} \tag{6.17}$$

In the last expression we used that, due to the holomorphy of the couplings $\tau_{\alpha\beta}$, $\partial^2_{IJ} b_{\alpha\beta} = 0$. If we assume that we are in the conditions of section 2, at the minimum we have $|m_i| = |\widetilde{m}_i|$, and the trace of the squared scalar matrix is

$$\mathrm{Tr}\mathcal{M}_0^2 = 6 \sum_i (q_i, q_i)(|m_i|^2 + |\widetilde{m}_i|^2) + 8 \sum_i |a \cdot q_i|^2 + 2(b^{-1})^{IJ} \frac{\partial^2 V}{\partial a^I \partial \overline{a}_J}, \tag{6.18}$$

where we have included the wave function renormalization for the scalars a^I. The mass of the dual photon is given by the monopole VEV through the magnetic Higgs mechanism:

$$\mathrm{Tr}\mathcal{M}_1^2 = 2 \sum_i (q_i, q_i)(|m_i|^2 + |\widetilde{m}_i|^2). \tag{6.19}$$

Taking into account all these contributions, the graded trace of the squared matrix is:

$$\begin{aligned} \sum_j (-1)^{2j}(2j+1)\mathrm{Tr}\mathcal{M}_j^2 &= -\frac{1}{2}\mathrm{Tr}[\mu b^{-1} \overline{\mu} b^{-1} + \hat{\mu} b^{-1} \overline{\hat{\mu}} b^{-1}] \\ &+ 2 f_0^2 \mathrm{Tr} b^{-1} \partial\overline{\partial}(b_0, b_0) + 4 \sum_{k,l} \mathrm{Tr} b^{-1} \partial\overline{\partial}(q_k, q_l) m_k \widetilde{m}_k \overline{m}_l \overline{\widetilde{m}}_l \\ &+ 2\sqrt{2} \sum_k \mathrm{Tr} b^{-1} \partial\overline{\partial}(q_k, b_0) f_0 (m_k \widetilde{m}_k + \overline{m}_k \overline{\widetilde{m}}_k). \end{aligned} \tag{6.20}$$

To see that this is zero, we write the bilinears in the monopole fields in terms of the auxiliary fields F^I, $\overline{F}^I$, using (3.6):

$$\sum_i q_i^I \overline{m}_i \tilde{\overline{m}}_i = -\frac{1}{\sqrt{2}}(b_{IJ}F^J + b_{0I}f_0). \tag{6.21}$$

Then we can group the terms in (6.20) depending on the number of F^I, $\overline{F}^I$, and check that they cancel separately. For instance, for the terms with two auxiliaries, we have from the first term in (6.20):

$$-2(F^I\overline{F}^J + \overline{F}^I F^J)\partial_I b_{MN}(b^{-1})^{NP}\partial_{\overline{J}} b_{PQ}(b^{-1})^{QM} \tag{6.22}$$

and from the third term

$$\begin{aligned} & 2F^I\overline{F}^J \partial_M b_{JN}(b^{-1})^{NP}\partial_{\overline{Q}} b_{PI}(b^{-1})^{QM} \\ + \quad & 2F^I\overline{F}^J \partial_M b_{PI}(b^{-1})^{NP}\partial_{\overline{Q}} b_{JN}(b^{-1})^{QM}. \end{aligned} \tag{6.23}$$

Taking into account the holomorphy of the couplings and the Kähler geometry, we have $\partial_M b_{PI} = \partial_I b_{PM}$, $\partial_{\overline{Q}} b_{JN} = \partial_{\overline{J}} b_{QN}$, so (6.22) and (6.23) add up to zero. With a little more algebra one can verify that the terms with one F^I (and their conjugates with $\overline{F}^I$) and without any auxiliaries add up to zero too. The result is then:

$$\sum_j (-1)^{2j}(2j+1)\mathrm{Tr}\mathcal{M}_j^2 = 0. \tag{6.24}$$

6.2 Mass spectrum in the $SU(2)$ case

In the $SU(2)$ case we can obtain much more information about the mass matrix and also determine its eigenvalues. First we consider the fermion mass matrix. Taking into account that at the minimum of the effective potential $m = \overline{m} = \rho$, $\tilde{m} = \epsilon m$, we can introduce the linear combination:

$$\eta_\pm = \frac{1}{\sqrt{2}}(\psi_m \pm \epsilon\psi_{\tilde{m}}). \tag{6.25}$$

With respect to the new fermion fields $(\lambda, \eta_+, \psi, \eta_-)$, the bare fermion mass matrix reads:

$$M_{1/2} = \begin{pmatrix} \frac{1}{2}\mu & -2\epsilon\rho & 0 & 0 \\ -2\epsilon\rho & -\sqrt{2}\epsilon a & 0 & 0 \\ 0 & 0 & \frac{1}{2}\mu & 2i\rho \\ 0 & 0 & 2i\rho & -\sqrt{2}\epsilon a \end{pmatrix}, \tag{6.26}$$

Notice that, in the $SU(2)$ case, the auxiliary field F is real and $\mu = \hat{\mu}$. $\mathcal{M}_{1/2}\mathcal{M}_{1/2}^{\dagger}$ can be easily diagonalized . From (6.26) it is easy to see that the squared fermion mass matrix is block-diagonal with the same 2×2 matrix in both entries:

$$\begin{pmatrix} b_{11}^{-2}\mu\overline{\mu}/4 + 4b_{11}^{-1}\rho^2 & -\epsilon b_{11}^{-3/2}\mu\rho + 2\sqrt{2}\overline{a}\rho \\ -\epsilon b_{11}^{-3/2}\overline{\mu}\rho + 2\sqrt{2}a\rho & 4b_{11}^{-1}\rho^2 + 2|a|^2 \end{pmatrix}. \tag{6.27}$$

Hence there are two different eigenvalues doubly degenerated. In terms of the determinant and trace of (6.27),

$$\begin{aligned} \alpha &= (m_1^F)^2 + (m_2^F)^2 = \frac{1}{4b_{11}^2}\mu\overline{\mu} + 2|a|^2 + \frac{8}{b_{11}}\rho^2 \\ \beta &= (m_1^F)^2(m_2^F)^2 = \frac{1}{b_{11}^2}|4\rho^2 + \frac{\epsilon}{\sqrt{2}}a\mu|^2, \end{aligned} \tag{6.28}$$

the eigenvalues are:

$$(m_{1,2}^F)^2 = \frac{\alpha}{2} \pm \frac{1}{2}\sqrt{\alpha^2 - 4\beta}. \tag{6.29}$$

The computation of the scalar mass matrix is more lengthy. First we must compute the second derivatives of the effective potential, evaluated at the minimum. To obtain more simple expressions, we can use the identities (2.19) to express all the derivatives of the couplings in terms only of $\partial b_{11}/\partial a$, $\partial^2 b_{11}/\partial a^2$. The results are:

$$\begin{aligned} \frac{\partial^2 V}{\partial m \partial \overline{m}} &= \frac{3}{b_{11}}\rho^2 + 2|a|^2, \quad \partial^2 V/\partial m^2 = \frac{\partial^2 V}{\partial \widetilde{m}^2} = \frac{1}{b_{11}}\rho^2, \\ \frac{\partial^2 V}{\partial m \partial \widetilde{m}} &= \frac{\epsilon}{b_{11}}\rho^2 + \frac{\sqrt{2}b_{01}}{b_{11}}f_0, \quad \frac{\partial^2 V}{\partial m \partial \overline{\widetilde{m}}} = \frac{\epsilon}{b_{11}}\rho^2 \\ \frac{\partial^2 V}{\partial m \partial a} &= 2\rho\Big[\overline{a} - \big(b_{11}\frac{\partial}{\partial a}\frac{1}{b_{11}}\big)\big(|a|^2 - \frac{i\epsilon}{\sqrt{2}}af_0\big)\Big] \\ \frac{\partial^2 V}{\partial a^2} &= -b_{11}^2\big(\frac{\partial}{\partial a}\frac{1}{b_{11}}\big)f_0(af_0 + 2\sqrt{2}i\epsilon|a|^2) \\ &\quad -b_{11}^2\big(\frac{\partial^2}{\partial a^2}\frac{1}{b_{11}}\big)(af_0 + \sqrt{2}i\epsilon|a|^2)^2, \\ \frac{\partial^2 V}{\partial \widetilde{m} \partial a} &= \epsilon\frac{\partial^2 V}{\partial m \partial a}, \quad \frac{\partial^2 V}{\partial \overline{m} \partial a} = \frac{\partial^2 V}{\partial m \partial a}, \quad \frac{\partial^2 V}{\partial \overline{\widetilde{m}} \partial a} = \frac{\partial^2 V}{\partial \widetilde{m} \partial a}, \\ \frac{\partial^2 V}{\partial \overline{a} \partial a} &= 4\rho^2 + \frac{1}{2b_{11}}\mu\overline{\mu}, \end{aligned} \tag{6.30}$$

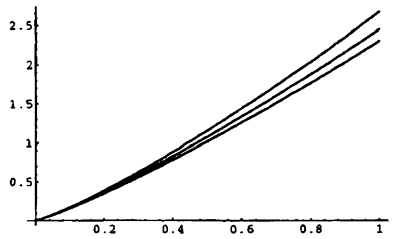


Figure 12: Fermion masses (6.29) (top and bottom) and photon mass (6.19) (middle) in softly broken $SU(2)$ Yang-Mills, as a function of $0 \leq f_0 \leq 1$.

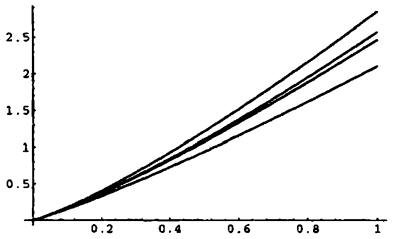


Figure 13: Masses of the scalars in softly broken $SU(2)$ Yang-Mills, as a function of $0 \leq f_0 \leq 1$.

and the rest of the derivatives are obtained through complex conjugation. In the last line we used the result of the previous section. To obtain the bosonic mass matrix we must take into account the wave-function renormalization of the a, $\overline{a}$ variables, as in (6.18). Its eigenvalues are as follows: we have a zero eigenvalue corresponding to the Goldstone boson of the spontaneously broken $U(1)$ symmetry. There is also an eigenvalue with degeneracy two given by:

$$2\Big(\frac{\partial^2 V}{\partial m \partial \overline{m}} - \frac{\partial^2 V}{\partial \widetilde{m}^2}\Big) = -\frac{2\sqrt{2}\epsilon}{b_{11}} f_0 b_{01}. \tag{6.31}$$

Notice that this is always positive if we have a non-zero VEV for ρ. The other three eigenvalues are best obtained numerically, as they are the solutions to a third-degree algebraic equation.

As an application of these general results, we can plot the mass spectrum as a function of the supersymmetry breaking parameter f_0 in the $SU(2)$ Yang-Mills case, where the minimum corresponds to the monopole region and $\epsilon = -1$. We have only to compute the derivatives of the magnetic coupling, with the result:

$$\frac{\partial \tau_{11}^{(m)}}{\partial a^{(m)}} = \frac{\pi^2}{8}\frac{k}{k'^2 K'^3}, \qquad \frac{\partial^2 \tau_{11}^{(m)}}{\partial a^{(m)2}} = -\frac{\pi i}{32}\frac{k^2}{k'^4 K'^4}\Big(k'^2 - k^2 + \frac{3E'}{K'}\Big). \tag{6.32}$$

These derivatives diverge at the monopole singularity $u = 1$, and we may think that this can give some kind of singular behaviour there. In fact this is not so. The position of the minimum, u_0, behaves almost linearly with respect to f_0, $u_0 - 1 \sim f_0$, and this guarantees that the behaviour very near to $u = 1$ (corresponding to a very small f_0) is perfectly smooth, as one can see in the figures. In fig. 12 we plot the fermion masses (6.29) (top and bottom)

and the photon mass given in (6.19) (middle). In fig. 13 we plot the masses of the scalars, where the second one from the top corresponds to the doubly degenerated eigenvalue (6.31).

Acknowledgments

One of us (L. A.-G.) would like to thank J.M. Drouffe and J.B. Zuber for the opportunity to present this work at the conference in honour of C. Itzykson "The Mathematical Beauty of Physics". We would also like to thank J. Distler and C. Kounnas for an enjoyable collaboration.

References

1. N. Seiberg and E. Witten, Nucl. Phys. **B426** (1994) 19, hep-th/9407087.
2. N. Seiberg and E. Witten, Nucl. Phys. **B431** (1994) 484, hep-th/9408099.
3. A. Klemm, W. Lerche. S. Theisen and S. Yankielowicz, Phys. Lett. **B344** (1995) 169, hep-th/9411048;
   P.C. Argyres and A.E. Faraggi, Phys. Rev. Lett. **74** (1995) 3931, hep-th/9411057.
4. A. Klemm, W. Lerche. and S. Theisen, Int. J. Mod. Phys. **A11** (1996) 1929, hep-th/9505150.
5. M. Douglas and S.H. Shenker, Nucl. Phys. **B447** (1995) 271, hep-th/9503163.
6. A. Brandhuber and K. Landsteiner, Phys. Lett. **B358** (1995) 73, hep-th/9507008;
   U.H. Danielsson and B. Sundborg, Phys. Lett. **B358** (1995) 273, hep-th/9504102;
   A. Hanany and Y. Oz, Nucl. Phys. **B452** (1995) 283, hep-th/9505075;
   P.C. Argyres, M.R. Plesser and A.D. Shapere, Phys. Rev. Lett. **75** (1995) 1699, hep-th/9505100.
7. B. de Wit and A. Van Prœyen, Nucl. Phys. **B245** (1984) 89;
   see also P. Fré and P. Soriani, "The $N = 2$ Wonderland", World Scientific, 1995, for a complete set of references.
8. G. 't Hooft, 1976, in "High Energy Physics", edited by A. Zichichi, Palermo, 1976;
   S. Mandelstam, Phys. Rep. **C23** (1976) 245.
9. I. Affleck, M. Dine and N. Seiberg, Nucl. Phys. **B241** (1984) 493; **B256** (1985) 557;
   D. Amati, G.C. Rossi, G. Veneziano, Nucl. Phys. **B249** (1985) 1; D.

Amati, K. Konishi, Y. Meurice, G.C. Rossi and G. Veneziano, Phys. Rep. **162** (1988) 169;
T.R. Taylor, G. Veneziano and S. Yankielowicz, Nucl. Phys. **B218** (1982); G. Veneziano and S. Yankielowicz, Phys. Lett. **113B** (1982) 231;
N. Seiberg, Phys. Lett. **B318** (1993) 469, hep-ph/9309335; Phys. Rev. **D49** (1994) 6857, hep-th/9402044;
K. Intriligator, R. Leigh and N. Seiberg, Phys. Rev. **D50** (1994) 1052, hep-th/9403198;
K. Intriligator, Phys. Lett. **B336** (1994) 409, hep-th/9407106;
K. Intriligator and N. Seiberg, Nucl. Phys. **B431** (1994) 551, hep-th/9408155.

10. N. Seiberg, Nucl. Phys. **B435** (1995) 129, hep-th/9411149;
P.C. Argyres, M.R. Plesser, N. Seiberg and E. Witten, Nucl. Phys. **B461** (1996) 71, hep-th/9511154.
11. L. Girardello and M.T. Grisaru, Nucl. Phys. **B194** (1982) 65.
12. O. Aharony, J. Sonnenschein, M.E. Peskin and S. Yankielowicz, Phys. Rev. **D52** (1995) 6157, hep-th/9507013.
13. N. Evans, S.D.H. Hsu and M. Schwetz, Phys. Lett. **B355** (1995) 475, hep-th/9503186;
N. Evans, S.D.H. Hsu, M. Schwetz, S.B. Selipsky, Nucl. Phys. **B456** (1995) 205, hep-th/9508002.
14. S. Kachru and C. Vafa, Nucl. Phys. **B450** (1995) 69, hep-th/9505105.
15. S. Kachru, A. Klemm, W. Lerche, P. Mayr, and C. Vafa, Nucl. Phys. **B459** (1996) 537, hep-th/9508155.
16. B. de Wit, hep-th/9602060.
17. P.C. Argyres and M. Douglas, Nucl. Phys. **B448** (1995) 166, hep-th/9505062.
18. M. Matone, Phys. Lett. **B357** (1995) 342, hep-th/9506102.
19. J. Sonnenschein, S. Theisen and S. Yankielowicz, Phys. Lett. **B367** (1996) 145, hep-th/9510129.
20. T. Eguchi and S.-K. Yang, Mod. Phys. Lett. **A11** (1996) 131, hep-th/9510183.
21. L. Álvarez-Gaumé, J. Distler, C. Kounnas and M. Mariño, hep-th/9604004.
22. M.K. Prasad and C.M. Sommerfield, Phys. Rev. Lett. **35** (1975) 760;
E.B. Bogomolny, Sov. J. Nucl. Phys. **24** (1976) 449.
23. E. Witten and D. Olive, Phys. Lett. **B78** (1978) 97.
24. I.S. Gradshteyn and I.M. Ryzhik, "Tables of series, products and integrals", Academic Press.

25. E. Witten, Phys. Lett. **B86** (1979) 283.
26. L. Álvarez-Gaumé and M. Mariño, to appear.
27. S. Ferrara, L. Girardello and F. Palumbo, Phys. Rev. **D20** (1979) 403.

POLYGONAL BILLIARDS AND APERIODIC TILINGS

J.M. LUCK
CEA/Saclay, Service de Physique Théorique,
F-91191 Gif-sur-Yvette Cedex, France

This is an overview of several topics related to polygons and triangles, including the spectrum of the Laplace operator in a polygonal domain (functional determinant, spectral zeta function, arithmetical degeneracies of integrable cases), and self-similar structures in discrete geometry, namely aperiodic chains and tilings, built from deterministic inflation rules (geometrical characteristics, relationship to quasicrystals, nature of diffraction spectrum).

Preamble

Very simple objects, such as the hydrogen atom, the harmonic oscillator or Platonic solids, have been the cornerstone of some of the most celebrated papers by Claude Itzykson. Simplicity thus appears as one of the facets of the mathematical beauty of Physics in Claude's work.

This contribution is devoted to yet another class of simple objects that also amazed and intrigued Claude quite much, namely polygons, and especially triangles. Section 1 deals with the spectrum of the Laplace operator in a polygonal domain of the plane. I have been one of those who shared Claude's enthusiasm on this subject. As far as the second part (Section 2) is concerned, i.e., aperiodic chains and tilings, I wish to recall what Claude told me once in the early days of quasicrystals about the origins of his interest for group theory. As a young boy he was fascinated by the crystals he saw in museums. Much later he got involved in group theory in order to unravel, among other things, the mathematical beauty... of crystals.

1 Polygonal billiards

This section is a review of various properties of the spectrum of the Laplace operator ∇^2 in a bounded domain $\mathcal{D}$ of the plane. The eigenvalue equation

$$(\nabla^2 + E_n)\phi_n(\mathbf{x}) = 0, \tag{1}$$

with, for definiteness, Dirichlet boundary conditions ($\phi_n(\mathbf{x}) = 0$ for $\mathbf{x} \in \partial\mathcal{D}$), has an infinite sequence of positive eigenvalues ($E_1 \leq E_2 \leq \cdots$). In the following $\mathcal{D}$ will most often be a polygon, and especially a triangle.

1.1 Polygonal billiards in Classical and Quantum Mechanics

Consider the free motion of a point particle in a domain $\mathcal{D}$, with reflecting boundary. The *billiard* so defined is a prototypical dynamical system. In Classical Mechanics, the particle moves with a constant velocity and bounces elastically on the boundary. The quantum-mechanical problem is defined by the stationary Schrödinger equation, which is of the form (1).

For a generic domain $\mathcal{D}$, the classical billiard is expected to behave as a typical *chaotic* system. The quantum-mechanical billiard is expected to exhibit *quantum chaos*, i.e., features of quantum systems which are chaotic in the classical limit, such as level repulsion.[1] The same holds for a polygonal billiard with generic angles, incommensurate to π.

A different dynamical behavior takes place in a rational polygonal billiard, whose inner angles are all commensurate to π. Let us write them as $\{\pi\alpha_i\}$, where

$$\alpha_i = \frac{p_i}{q_i} \tag{2}$$

is an irreducible fraction. P. Richens and M. Berry have qualified these billiard systems as *pseudointegrable*.[2] The classical motion takes place on invariant surfaces, which, however, have in general a topology different from that of a torus, which is characteristic of integrable systems.[3]

To be more specific, let us focus on rational triangles. The vertices of a triangle are labeled as $i = 0, 1, \infty$, the corresponding angles being denoted as $\pi\alpha_i$, as in Eq. (2), and $\alpha_0 + \alpha_1 + \alpha_\infty = 1$. If Q denotes the least common multiple of the three denominators q_i, there is a way of gluing $2Q$ copies of the triangular billiard together, so as to form a closed invariant surface. The Riemann surface constructed in this way is characterized by its genus

$$g = 1 + \frac{1}{2} \sum_{i=0,1,\infty} \frac{Q}{q_i}(p_i - 1). \tag{3}$$

The connection between these Riemann surfaces and algebraic curves has been investigated by E. Aurell and C. Itzykson.[4] A classification of the rational triangular billiards is thus obtained (see Table 1).

The genus $g = 1$ corresponds to a torus. The classical and the quantum-mechanical motion are then *integrable*. There are three cases of integrable triangular billiards, which tile the plane under the Coxeter group of reflections with respect to their edges. Their properties will be illustrated in Section 1.3, on the example of the equilateral triangle. C. Itzykson has generalized the above construction, obtaining a classification of the irreducible integrable polyhedral billiards in any space dimension, in correspondence with Lie algebras.[5]

Table 1: Classification of rational triangular billiards, according to the genus of the associated invariant surfaces. The Lie algebras in correspondence with the integrable cases (genus $g = 1$) are given in the last column (after Ref. 4).

genus	p_0/q_0	p_1/q_1	p_∞/q_∞	$2Q$	triangle	algebra
1	1/3	1/3	1/3	6	equilateral	A_2
	1/2	1/4	1/4	8	rectangular-isosceles	B_2
	1/2	1/3	1/6	12	drawer's square	G_2
2	2/5	2/5	1/5	10	Robinson's P	
	3/5	1/5	1/5	10	Robinson's Q	
	2/3	1/6	1/6	12		
	1/2	3/8	1/8	16		
	1/2	1/5	3/10	20		
	1/2	2/5	1/10	20		
etc.						

The genus $g = 2$ of the double torus is already generic. The first two of the six cases listed in Table 1 correspond to Robinson's P and Q triangles, which will play a role in Section 2. Figure 1 shows how to assemble ten Robinson's Q triangles, so as to form a double torus.

1.2 Spectral zeta functions and functional determinants

The eigenfunctions of the Laplace operator in a domain $\mathcal{D}$, obeying Eq. (1), enter the calculation of partition functions which show up in various areas of Quantum Field Theory, such as string theories or conformal field theories. Consider a free scalar field $\phi(\mathbf{x})$ living in $\mathcal{D}$, with Dirichlet boundary conditions. The associated *partition function* is defined as

$$Z_{\mathcal{D}} = \int [\mathrm{d}\phi(\mathbf{x})] \exp\left(-\frac{1}{2}\int_{\mathcal{D}} \mathrm{d}^2\mathbf{x}(\nabla\phi)^2\right). \tag{4}$$

It is formally equal to the inverse square-root of the *functional determinant* of (minus) the Laplace operator ∇^2, namely

$$Z_{\mathcal{D}} = \left[\det(-\nabla^2)\right]^{-1/2} = \left(\prod_{n\geq 1} E_n\right)^{-1/2} \tag{5}$$

The above infinite-product representation is divergent, so that a ultraviolet regularization is needed in order to make sense of Eq. (5). To be more specific,

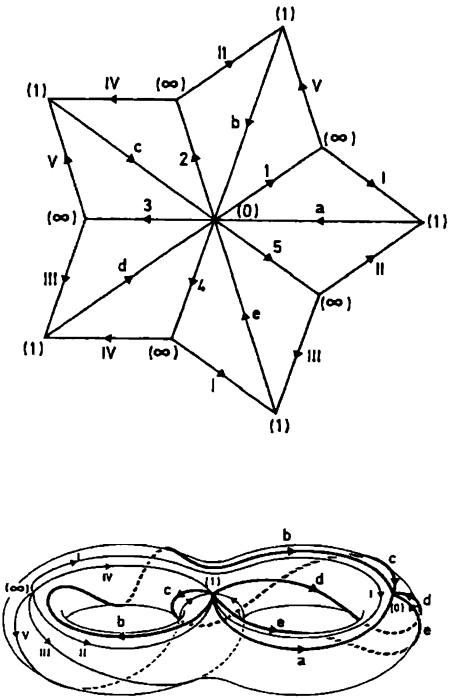


Figure 1: Construction of the Riemann surface associated with a triangular billiard: a double torus is obtained by gluing ten Robinson's Q triangles (the second case with genus $g = 2$ in Table 1) (after Ref. 4).

the counting function $N(E)$, defined as being the number of eigenvalues E_n of Eq. (1) less than E, admits the following asymptotic estimate for large E:

$$N(E) \approx \frac{1}{4\pi}\left(\mathcal{A}E - \mathcal{P}E^{1/2}\right), \tag{6}$$

with $\mathcal{A}$ and $\mathcal{P}$ being the area of the domain $\mathcal{D}$ and the perimeter of its boundary, respectively. The area term in Eq. (6) is referred to as the Weyl term. The counting function exhibits (possibly unbounded) oscillations with respect to its mean behavior (6), which have been explained by R. Balian and C. Bloch in terms of classical trajectories.[6]

A first and very commonly used way of regularizing Eq. (5) consists in

introducing the *spectral zeta function*[7,8]

$$\zeta_{\mathcal{D}}(s) = \mathrm{tr}\left[(-\nabla^2)^{-s}\right] = \sum_{n\geq 1} E_n^{-s}, \tag{7}$$

in analogy with the definition of the Riemann zeta function of Number Theory. The estimate (6) ensures that the sum in Eq. (7) converges for $\mathrm{Re}\, s > 1$. The spectral zeta function $\zeta_{\mathcal{D}}(s)$ usually has a meromorphic continuation to the whole s-plane, with poles at $s = 1$ and $s = 1/2$, corresponding to both terms in Eq. (6). It is analytic near $s = 0$, so that we have formally

$$Z_{\mathcal{D}} = \exp\left(-\frac{1}{2}\sum_{n\geq 1} \ln E_n\right) = \exp\left(\frac{1}{2}\zeta'_{\mathcal{D}}(0)\right). \tag{8}$$

This expression is considered to be the definition of $Z_{\mathcal{D}}$, within the zeta-function regularization scheme.

An alternative way of evaluating the partition function (4) consists in using discretization on a square lattice as a ultraviolet regulator. If a denotes the lattice spacing, the partition function $Z(a)$ behaves as follows in the continuum limit ($a \to 0$):

$$Z(a) \approx z \left(\frac{\mathcal{A}}{a^2}\right)^{\zeta_{\mathcal{D}}(0)} \exp\left(-f_0 \frac{\mathcal{A}}{a^2} - f_1 \frac{\mathcal{P}}{a}\right), \tag{9}$$

where

- the exponential contains the contributions of the area $\mathcal{A}$ and of the perimeter $\mathcal{P}$ of the domain. The corresponding specific free energies, f_0 and f_1, are non-universal;
- the power of the area leads to the interpretation of $\zeta_{\mathcal{D}}(0)$ as the anomalous dimension of vacuum in the domain $\mathcal{D}$;
- the absolute prefactor z is universal. It is considered as the lattice-regularized partition function.

An instructing example of a domain is the torus $\mathbf{T}$, for which both regularizations can be compared to each other. The zero mode $E_0 = 0$ of the Laplace operator on the torus, corresponding to the constant $\phi(\mathbf{x}) = 1$, is supposed not to be included in the infinite-product representation (5). The partition function (4) on a torus, viewed as a rectangle consisting of $M \times N$ sites with periodic boundary conditions, has been calculated by B. Duplantier and F.

David.[9] In the asymptotic regime (M and N large), their result is of the general form (9). The area free energy reads $f_0 = 2G/\pi$, with G being Catalan's constant, while f_1 cannot be defined, since the torus has no boundary. The anomalous dimension is $\zeta_{\rm T}(0) = -1/2$, while the lattice-regularized partition function reads

$$z = \frac{1}{t^{1/2} q^{1/12} \big(P(q)\big)^2}, \tag{10}$$

with $t = M/N$, $q = \exp(-2\pi t)$ is the modular parameter of the torus, and

$$P(q) = \prod_{n \geq 1} (1 - q^n) \tag{11}$$

is Euler's product. The expression (10) is a modular invariant, as it should. On the other hand, the partition function of a free field on the torus has been evaluated using the zeta-function regularization by C. Itzykson and J.B. Zuber.[10] As compared to the lattice calculation, this approach does not yield the non-universal contribution involving f_0. Its prediction for the zeta-function regularized partition function z is in full agreement with Eq. (10). This example suggests to make use of the zeta-function regularization, since the calculations involved are simpler in general.

Let us turn to the spectral zeta function of triangles. C. Itzykson, P. Moussa, and the author have derived a sequence of *sum rules* for the spectrum of the triangular billiard, i.e., integral expressions for $\zeta_{\mathcal{D}}(n)$, with $n \geq 2$ being an integer.[11] These sum rules yield, among other things, a sequence of lower bounds for the ground-state energy E_1. This property has been exploited in other contexts, such as billiards embracing an Aharonov-Bohm flux.[8] The starting point of our work is to observe that the Green's function of the Laplace operator reads

$$G(t, t') = \frac{1}{2\pi} \, {\rm Re} \ln \frac{z - \overline{z}'}{z - z'}. \tag{12}$$

In this expression, a complex co-ordinate $t \in \mathcal{D}$ is used to describe the triangle, the bar denotes complex conjugation, and z (respectively z') is the image of t (respectively t') by the Schwarz-Christoffel mapping, whose inverse maps the upper half-plane (${\rm Im}\, z > 0$) onto the inside of the triangle $\mathcal{D}$, according to

$$z \mapsto t = \int_0^z \frac{{\rm d}y}{p(y)}, \quad \text{with} \quad p(y) = y^{1-\alpha_0} (1-y)^{1-\alpha_1}, \tag{13}$$

so that $0 \mapsto t_0 = 0$, $1 \mapsto t_1 = \Gamma(\alpha_0)\Gamma(\alpha_1)/\Gamma(1-\alpha_\infty)$, and $\infty \mapsto t_\infty = \exp(i\pi\alpha_0)\Gamma(\alpha_0)\Gamma(\alpha_\infty)/\Gamma(1-\alpha_1)$. The inner angle at the vertex i is $\pi\alpha_i$, while

the area of $\mathcal{D}$, namely

$$\mathcal{A}_0 = \frac{\pi}{2} \prod_{i=0,1,\infty} \frac{\Gamma(\alpha_i)}{\Gamma(1-\alpha_i)}, \tag{14}$$

provides a natural scale, which will enter several formulas in the following.

The sum rules thus obtained[11] have the following n-fold complex integral expressions

$$\zeta_{\mathcal{D}}(n) = \left(\frac{\mathcal{A}}{\mathcal{A}_0}\right)^n \int_{\mathrm{Im}\, z_k > 0} \prod_{k=1}^{n} \left(\frac{\mathrm{d}^2 z_k}{2\pi |p(z_k)|^2} \, \mathrm{Re} \ln \frac{z_k - \overline{z}_{k+1}}{z_k - z_{k+1}} \right) \quad (n \geq 2), \tag{15}$$

with the convention $z_{n+1} = z_1$. An adaptation of this procedure leads to an explicit formula for the finite part of the spectral zeta function at its rightmost pole at $s = 1$, corresponding to the Weyl term in Eq. (6), i.e.,

$$\zeta_{\mathcal{D}}(s) \approx \frac{\mathcal{A}}{4\pi} \left(\frac{1}{1-s} + \ln \frac{\mathcal{A}}{4\mathcal{A}_0} + \gamma_E + \sum_{i=0,1,\infty} \left[\alpha_i \psi(\alpha_i) - (1-\alpha_i)\psi(1-\alpha_i) \right] \right), \tag{16}$$

with γ_E being Euler's constant, and ψ being the logarithmic derivative of Euler's gamma function.

The calculation of the zeta-function regularized partition function (8) for a triangular domain turned out to be more difficult. Extending earlier ideas of A. Polyakov on conformal invariance,[12] W. Weisberger calculated the zeta-function regularized functional determinant of the Laplace operator in a disk and in an annulus, with Dirichlet or Neumann boundary conditions.[13] This appealing method could, however, not be extended to polygons, because of their following peculiar feature. For any domain $\mathcal{D}$ with a smooth boundary, the anomalous dimension reads

$$\zeta_{\mathcal{D}}(0) = \frac{1}{6}. \tag{17}$$

It is therefore invariant under a smooth deformation of the domain. To the contrary, for a polygonal domain with inner angles $\{\pi\alpha_i\}$, we have the expression

$$\zeta_{\mathcal{D}}(0) = \frac{1}{24} \sum_i \left(\frac{1}{\alpha_i} - \alpha_i \right), \tag{18}$$

which depends continuously on the angles. This dependence induces logarithmic ultraviolet divergences in calculations based on conformal invariance, which are difficult to master.

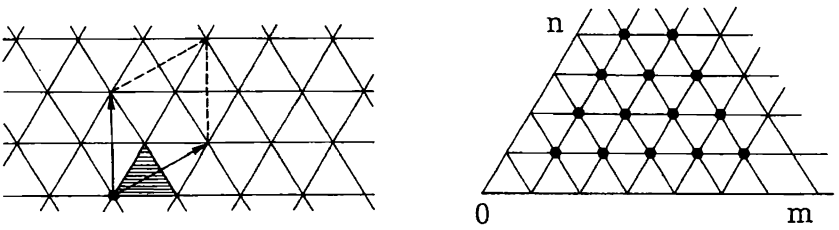


Figure 2: Integrability of the equilateral triangular billiard. (a) Classical Mechanics: six copies of the triangle build a torus, so that trajectories are lifted as straight lines. (b) Quantum Mechanics: eigenfunctions are indexed by points of the dual lattice of the torus (after Ref. 16).

Using more traditional heat-kernel techniques *à la* Sommerfeld, E. Aurell and P. Salomonson derived an explicit expression for $\zeta'_{\mathcal{D}}(0)$ in an arbitrary polygonal domain of the plane.[14] In the case of a triangle, their result reads

$$\zeta'_{\mathcal{D}}(0) = \zeta_{\mathcal{D}}(0) \ln \frac{\mathcal{A}}{\mathcal{A}_0} + \sum_{i=0,1,\infty} J(\alpha_i), \tag{19}$$

with

$$J(\alpha) = \frac{1}{12}\left(\frac{1}{\alpha} - \alpha\right)(\gamma_E - \ln 2) - \frac{1}{12}\left(\frac{1}{\alpha} + 3 + \alpha\right)\ln\alpha + \int_0^\infty \frac{\mathrm{d}y}{e^y - 1}\left[\frac{1}{2y}\left(\coth\frac{y}{2\alpha} - \alpha\coth\frac{y}{2}\right) - \frac{1}{12}\left(\frac{1}{\alpha} - \alpha\right)\right]. \tag{20}$$

1.3 Arithmetical degeneracies of integrable billiards

It has been recalled in section 1.1 that there are only three integrable triangular billiards. These triangles, corresponding to genus $g = 1$ in Table 1, are characterized by the property that the numerators of their angles read $p_i = 1$ in Eq. (2). Such a triangle tiles the plane under the Coxeter group of reflections with respect to its edges. The classical and quantum-mechanical motions are integrable. The spectrum of the Laplace operator usually exhibits arithmetical degeneracies.[15,16] All these properties will be illustrated on the example of the equilateral triangle.

Let $\triangle$ be the equilateral triangle of side a. Figure 2a shows that six copies of $\triangle$ build a torus, which tiles the plane under a lattice of translations.

Integrability at the classical level follows at once, since trajectories are lifted as straight lines in this plane. Integrability at the quantum-mechanical level is less obvious. The eigenfunctions of Eq. (1) in $\triangle$ seem to have been first found by Lamé, in the context of vibrations of elastic sheets,[17] and re-discovered many times since then. The eigenfunctions are superpositions of $2Q = 6$ plane waves, with symmetry-related wavevectors, according to the Lie algebra A_2 of the group $SU(3)$. The energy eigenvalues are indexed by two integers, (m, n), which parametrize the dual of the lattice of translations mentioned above, as shown in Figure 2b. They read explicitly

$$E_{\triangle}(m,n) = \left(\frac{4\pi}{3a}\right)^2 \mathcal{N}_{\triangle}(m,n), \quad \mathcal{N}_{\triangle}(m,n) = m^2+mn+n^2 \quad (m>0,\ n>0). \tag{21}$$

We shall consider in parallel the spectrum of the Laplace operator in the square $\square$ of side b. In this case it is an elementary exercise to find the eigenfunctions and eigenvalues. The latter are still indexed by two integers, (m, n), according to

$$E_{\square}(m,n) = \left(\frac{\pi}{b}\right)^2 \mathcal{N}_{\square}(m,n), \quad \mathcal{N}_{\square}(m,n) = m^2+n^2 \quad (m>0,\ n>0). \tag{22}$$

The structure of the above formulas is general: the energy levels of polyhedral integrable billiards are given, in suitably chosen units, by quadratic forms in integers.[5] The number of independent integer quantum numbers involved is equal to the dimension of the polyhedron, namely two in the present cases.

One of the most remarkable features of these integrable spectra is the occurrence of *arithmetical degeneracies*, which were noticed by Lord Rayleigh,[18] in the case of a rectangle with commensurate sides a and b, before being revived by M. Berry[15], and investigated in detail by C. Itzykson and the author.[16] For the equilateral triangle and the square, the first degeneracies of order two and four in the spectra (21) and (22) are

$$\begin{aligned}
7 &= \mathcal{N}_{\triangle}(1,2) = \mathcal{N}_{\triangle}(2,1),\\
5 &= \mathcal{N}_{\square}(1,2) = \mathcal{N}_{\square}(2,1),\\
91 &= \mathcal{N}_{\triangle}(1,9) = \mathcal{N}_{\triangle}(9,1) = \mathcal{N}_{\triangle}(5,6) = \mathcal{N}_{\triangle}(6,5),\\
65 &= \mathcal{N}_{\square}(1,8) = \mathcal{N}_{\square}(8,1) = \mathcal{N}_{\square}(4,7) = \mathcal{N}_{\square}(7,4).
\end{aligned} \tag{23}$$

The twofold degeneracies, corresponding to exchanging m and n, are a consequence of the presence of axes of symmetry. The occurrence of larger multiplicities is not physically intuitive, since it cannot be explained by any symmetry of the problem. There is a profound connection between these degeneracies and

Number Theory. Indeed, let us denote by $d_\triangle(\mathcal{N})$ the multiplicity of the integer $\mathcal{N}$ in the spectrum (21) of the equilateral triangle (plus one, if $\mathcal{N}$ is a perfect square), and by $d_\square(\mathcal{N})$ the multiplicity of the integer $\mathcal{N}$ in the spectrum (22) of the square (again plus one, if $\mathcal{N}$ is a perfect square). The latter quantity is nothing but the number of ways $\mathcal{N}$ can be written as a sum of two squares. This celebrated number-theoretical problem has received contributions from Diophantus, Fermat, and Jacobi.[19,20,22] The four representations of 65 given in the last line of Eq. (23) were known to Diophantus.[19]

The multiplicities $d_\triangle(\mathcal{N})$ and $d_\square(\mathcal{N})$ have the following general expressions. Factor the integer $\mathcal{N}$ over the primes as

$$\mathcal{N} = 3^c \prod_i p_i^{a_i} \prod_j q_j^{b_j} = 2^{c'} \prod_i {p'_i}^{a'_i} \prod_j {q'_j}^{b'_j}, \tag{24}$$

where the p_i (respectively, the q_j) are the primes equal to 1 (respectively, to -1) modulo 3, and the p'_i (respectively, the q'_j) are the primes equal to 1 (respectively, to -1) modulo 4. One has

$$\begin{aligned} d_\triangle(\mathcal{N}) &= \prod_i (1+a_i) \prod_j \left(\frac{1+(-1)^{b_j}}{2} \right), \\ d_\square(\mathcal{N}) &= \prod_i (1+a'_i) \prod_j \left(\frac{1+(-1)^{b'_j}}{2} \right). \end{aligned} \tag{25}$$

Many results have been derived from these remarkable expressions.[16]

First, the spectral zeta functions of the two integrable domains read

$$\zeta_\triangle(s) = \left(\frac{3a}{4\pi} \right)^{2s} \left[\zeta(s) L_\triangle(s) - \zeta(2s) \right], \quad \zeta_\square(s) = \left(\frac{b}{\pi} \right)^{2s} \left[\zeta(s) L_\square(s) - \zeta(2s) \right]. \tag{26}$$

In these formulas, $\zeta(s)$ is the Riemann zeta function, and $L_\triangle(s)$ and $L_\square(s)$ are the Dirichlet characters

$$\begin{aligned} L_\triangle(s) &= \sum_{n \geq 0} \left[(3n+1)^{-s} - (3n+2)^{-s} \right], \\ L_\square(s) &= \sum_{n \geq 0} \left[(4n+1)^{-s} - (4n+3)^{-s} \right]. \end{aligned} \tag{27}$$

The result (26) provides explicit checks of the general expressions (16), (18), (19). We have in particular

$$\zeta_\triangle(0) = \frac{1}{3}, \quad \zeta_\square(0) = \frac{1}{4}, \tag{28}$$

and

$$\zeta'_{\triangle}(0) = \ln \frac{\left(3^{7/4}\pi^2 a^2\right)^{1/3}}{\Gamma(1/3)}, \quad \zeta'_{\square}(0) = \ln \frac{\left(64\pi^3 b^2\right)^{1/4}}{\Gamma(1/4)}. \tag{29}$$

Second, the arithmetical degeneracies are responsible for the occurrence of anomalous fluctuations in the spectra of integrable polygonal billiards. Their statistical properties have been investigated by looking at the asymptotic behavior as $\mathcal{N} \to \infty$ of the average of various powers of the multiplicities $d_{\triangle}(\mathcal{N})$ and $d_{\square}(\mathcal{N})$. In the following, we denote by $\langle x(\mathcal{N}) \rangle$ the following *Cesàro* average of a function $x(\mathcal{N})$ defined over the integers

$$\langle x(\mathcal{N}) \rangle = \frac{1}{\mathcal{N}} \sum_{\mathcal{N}'=1}^{\mathcal{N}} x(\mathcal{N}'). \tag{30}$$

The various moments of the multiplicity functions $d_{\triangle}(\mathcal{N})$ and $d_{\square}(\mathcal{N})$ have been evaluated in closed form both domains:[16]

- the averages of the multiplicities themselves, i.e.,

$$\langle d_{\triangle} \rangle \approx \frac{\pi}{3\sqrt{3}}, \quad \langle d_{\square} \rangle \approx \frac{\pi}{4}, \tag{31}$$

  yield the mean density of states of the Laplace operator in each domain. These constants agree with the Weyl term in Eq. (6).

- the averages of the zero-th powers of the multiplicities, which give a measure of the support of the spectra, decay according to

$$\langle d^0_{\triangle} \rangle \approx A_{\triangle} (\ln \mathcal{N})^{-1/2}, \quad \langle d^0_{\square} \rangle \approx A_{\square} (\ln \mathcal{N})^{-1/2}, \tag{32}$$

  where

$$A_{\triangle} = \left(2\sqrt{3} \prod_j (1 - q_j^{-2}) \right)^{-1/2} = 0.638909,$$

$$A_{\square} = \left(2 \prod_j (1 - q_j'^{-2}) \right)^{-1/2} = 0.764224, \tag{33}$$

  with the notations introduced in Eq. (24). The spectrum of the Laplace operator in either domain is thus asymptotically concentrated on a small subset of the integers, whose density falls off very slowly, on a logarithmic scale.

- the averages of higher powers of the multiplicities diverge according to

$$\langle d_{\triangle}^{k} \rangle \approx B_{k,\triangle} (\ln \mathcal{N})^{\tau(k)}, \quad \langle d_{\square}^{k} \rangle \approx B_{k,\square} (\ln \mathcal{N})^{\tau(k)}, \tag{34}$$

with

$$\tau(k) = 2^{k-1} - 1. \tag{35}$$

The remarkable feature of this result is that the exponent $\tau(k)$ bears a non-trivial dependence on the index k, and grows faster than any power of k. This phenomenon is one of the manifestations of *multifractality*[23,24] [see section 2.7]. The spectra of integrable polygonal billiards in two dimensions are thus characterized by a multifractal distribution of degeneracies, having a purely number-theoretical origin.

2 Aperiodic chains and tilings

This section is devoted to self-similar geometrical objects in one and two dimensions, namely chains and tilings, built from deterministic inflation rules. The first motivation to these investigations is the experimental discovery of *quasicrystals*, recalled below.

A broader line of thought consists in viewing incommensurate structures and quasicrystals as the first two steps of a whole hierarchy of intermediate structures in condensed matter, with novel kinds of order, between the periodic state (crystals) and the randomly disordered one (glasses, amorphous materials). The chains and tilings presented in the following provide examples of such intermediate types of order; they also illustrate earlier ideas of S. Aubry on *weak periodicity*.[25]

2.1 Quasicrystals: experimental facts

In 1984, D. Shechtman, I. Blech, D. Gratias, and J. Cahn report on the unusual diffraction properties of a rapidly quenched sample of AlMn alloy.[26] Such materials have soon been called *quasicrystals* by D. Levine and P. Steinhardt.[27] The historical quasicrystalline *binary* alloys could only be produced by rapid quenches, i.e., far from equilibrium. More recently, several aluminum-based *ternary* alloys, such as AlMnPd, AlCuFe, AlMnSi, have been shown to possess thermodynamically stable quasicrystalline phases.

Figure 3 is an experimental pattern, obtained by electron diffraction, showing a two-dimensional section of the diffraction spectrum of quasicrystalline AlMnSi. As usual in solid-state physics, the only observable quantity in diffraction spectra is the Fourier intensity, i.e., the squared modulus of the Fourier

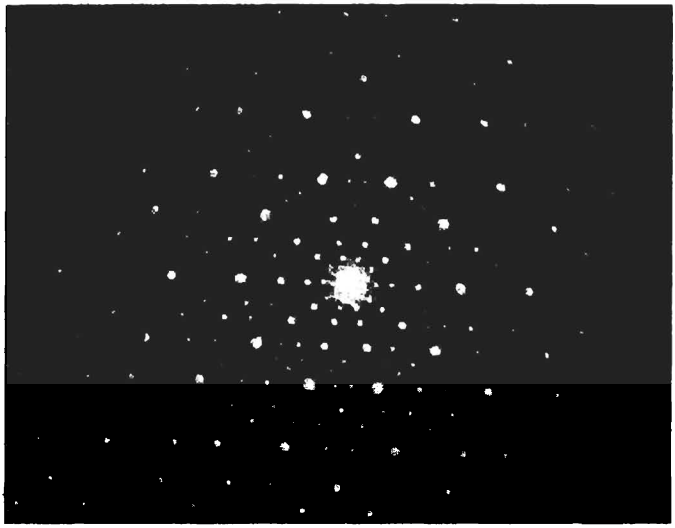


Figure 3: Electron diffraction pattern of the icosahedral phase of AlMnSi, in a plane perpendicular to a fivefold axis (courtesy of C.E.C.M.-C.N.R.S., Vitry-sur-Seine, France).

transform. In particular the symmetry of a structure is, by definition, the symmetry of its Fourier intensity. A quantitative description of diffraction spectra will be given in section 2.7.

The graph shown in Figure 3 illustrates the main characteristic features of quasicrystals:

- *Long-range order*

  The pattern consists of sharp spots, referred to as Bragg peaks, which are only broadened by instrumental resolution. Their presence is the signature of a perfect long-range order.

- *Crystallographic versus non-crystallographic symmetry*

  Another remarkable feature of Figure 3 is the presence of ten equivalent brightest spots, testifying fivefold symmetry. There are altogether six fivefold axes, so that the full three-dimensional diffraction pattern has the symmetry of the icosahedron, or equivalently of its dual the dodecahedron. Most quasicrystalline phases have icosahedral symmetry.

  This feature also shows up at the macroscopic scale, e.g. in growth shapes. Figure 4 shows almost perfectly dodecahedral growing grains of a quasicrystalline phase.

Table 2: Classification of crystallographic point symmetries in two dimensions, in terms of solutions to Eq. (36).

N	m	lattice
3	-1	hexagonal
4	0	square
6	1	triangular

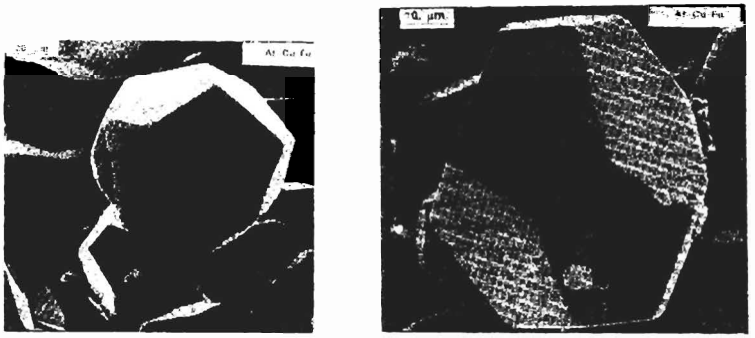


Figure 4: Dodecahedral growth shapes of icosahedral AlCuFe (courtesy of C.E.C.M.-C.N.R.S., Vitry-sur-Seine, France).

Icosahedral symmetry is forbidden by crystallography. This means that no structure can have both icosahedral point symmetry and a lattice of translations. The icosahedral phase is therefore necessarily an *aperiodic* structure. This statement is easier to prove in two dimensions. Consider a lattice $\mathcal{L}$ in the Euclidean plane, with N-fold point symmetry. This means that $\mathcal{L}$ is invariant under a rotation by an angle $\theta = 2\pi/N$. Let $\mathbf{e}$ be any of the shortest lattice vectors, and $\mathbf{e}_\pm$ the vectors obtained by rotating $\mathbf{e}$ by $\pm\theta$. Their sum $\mathbf{s} = \mathbf{e}_+ + \mathbf{e}_- = 2\cos\theta\,\mathbf{e}$ is a lattice vector parallel to $\mathbf{e}$, hence $\mathbf{s} = m\mathbf{e}$, with m being an integer. The two-dimensional crystallographic symmetries thus obey the Diophantine condition

$$2\cos\frac{2\pi}{N} = m \in \mathbf{Z}. \tag{36}$$

The non-trivial solutions of this equation are listed in Table 2. It is worth noticing that the smallest non-trivial point symmetry which is absent from this list is precisely $N = 5$, characteristic of icosahedral symmetry.

- *Self-similarity*

  The diffraction pattern in Figure 3 also exhibits self-similarity. The bright Bragg peaks can be joined together in several ways, so as to form e.g. regular pentagons of various sizes. The scaling factor of the structure, namely the ratio between the side lengths of two successive sizes of pentagons, is the irrational number

$$\tau = 2\cos\frac{\pi}{5} = \frac{1+\sqrt{5}}{2}. \tag{37}$$

  This number, which obeys the equation $\tau^2 = \tau + 1$, is referred to as the *golden mean*. It was already known to the Ancients for its aesthetic properties, and used e.g. in architecture.

2.2 Quasicrystal models and superspace formalism

Quasicrystals have been a very active field of research over the past decade. The reader is invited to consult the overviews on this subject.[28,29,30,31] We first present an introduction to the geometry of quasicrystals, viewed from the standpoint of the superspace formalism, before we turn to our main topic, i.e., self-similar chains and tilings. A more extensive presentation has been given elsewhere by the author.[32]

Let us begin with a brief introduction to almost-periodicity and quasiperiodicity. Let $f(x)$ be a function of one real variable, and let

$$G(Q) = \int f(x)\exp(-iQx)\mathrm{d}x \tag{38}$$

be its Fourier transform. The function f is said to be *almost-periodic*, according to H. Bohr,[33] if its Fourier transform $G(Q)$ is discrete, namely

$$G(Q) = \sum_{q\in\mathcal{M}} C_q\delta(Q-q). \tag{39}$$

The delta functions are referred to as *Bragg peaks*, or *Bragg diffractions*. Their positions form the discrete support of the Fourier transform. This support, or rather the module over the integers spanned by it, is denoted by $\mathcal{M}$, and called the *Fourier module* of the function f. A quantitative description of diffraction spectra will be given in section 2.7.

The following particular cases of almost-periodicity are of interest:

- *Periodicity*: the Fourier module

$$\mathcal{M} = \mathcal{L} = q_0\mathbf{Z} \tag{40}$$

is a lattice. The function f is periodic, with primitive period $a = 2\pi/q_0$, i.e., $f(x) = f(x + a)$.

- *Quasiperiodicity*: the Fourier module

$$\mathcal{M} = \mathcal{L}_1 \oplus \cdots \oplus \mathcal{L}_n \tag{41}$$

is the sum of n lattices of the form $\mathcal{L}_m = q_m \mathbf{Z}$ $(m = 1, \cdots, n)$, with the generators q_m being linearly independent over the integers. The module $\mathcal{M}$ is finitely generated, so that any Bragg diffraction $Q = k_1 q_1 + \cdots + k_n q_n$ is indexed by n integers $\{k_m\}$. The function f can be viewed as a suitable incommensurate section of a function on the torus $\mathbf{T}^n$, i.e., a function of n real variables, separately periodic in each of them.

- *Limit-periodicity*: the Fourier module

$$\mathcal{M} = \bigcup_{m \geq 0} (b^{-m} \mathcal{L}) \tag{42}$$

is the union of all the scaled copies of a given lattice $\mathcal{L}$, the scaling factors b^{-m} being the inverse powers of an *integer* basis $b \geq 2$.

Quasicrystals are an extension of usual crystals, just as quasiperiodic functions are an extension of periodic ones. To be more specific, a quasicrystal in dimension d is usually modeled as a section of a periodic structure $\mathcal{S}$, living in an n-dimensional *superspace* $\mathbf{R}^n$ with $n > d$, so that

$$\mathbf{R}^n = E^{\parallel} \oplus E^{\perp}, \tag{43}$$

where $E^{\parallel}$ is the d-dimensional *physical* (or *parallel*) space, also called the *cut*, in which the quasicrystal lives, while $E^{\perp}$ is the *internal* (or *perpendicular*) space, whose dimension, $n - d$, is called the *co-dimension* of the structure. This framework has been developed to describe incommensurate modulated structures.[34] The choice of the superspace and of the cut is often guided by group-theoretical considerations.[35,36,37] In the case of icosahedral quasicrystalline phases, $\mathbf{R}^6$ shows up naturally, because the smallest crystallographic representation Γ_6 of the icosahedral group Y has dimension 6. Being reducible as a real representation, Γ_6 has two three-dimensional representation spaces, to be identified with $E^{\parallel}$ and $E^{\perp}$.

Structural models of quasicrystals are most commonly based on either of the following algorithms:

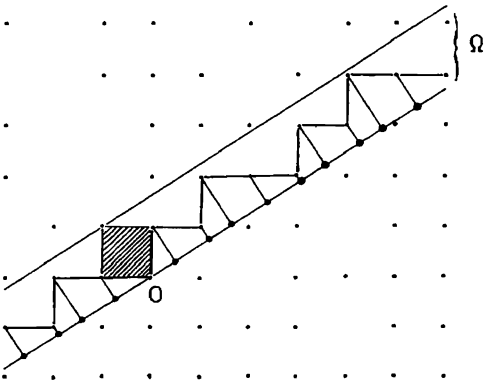


Figure 5: Cut-and-project algorithm for the canonical one-dimensional quasicrystal, built from a two-dimensional superspace.

- *Cut and projection*

  This approach, first investigated by N. de Bruijn,[38] has had a great importance in the early days of quasicrystals.[39,40,41] Figure 5 illustrates this method in the simplest case of a one-dimensional quasicrystal built from a two-dimensional superspace. The cut $E^{\parallel}$ makes an angle θ with the horizontal axis. A strip Ω is obtained by sweeping the unit square along $E^{\parallel}$. The atomic positions are the orthogonal projections onto $E^{\parallel}$ of the points of the lattice $\mathbf{Z}^2$ contained in Ω. Notice that these points can be joined by a unique infinite broken line. A binary chain covering $E^{\parallel}$ is thus obtained: the distances between neighboring atoms take either of the two values $\cos\theta$ or $\sin\theta$. A quantitative description of this canonical one-dimensional quasicrystal is given below.

- *Atomic surfaces*

  An equivalent construction of the above structure, illustrated in Figure 6, consists in decorating the vertices of the lattice $\mathbf{Z}^2$ by identical straight line segments contained in perpendicular space $E^{\perp}$, with a suitably chosen length. Nowadays most structural models for quasicrystals are built along these lines. In the simplest case a bounded, flat $(n-d)$-dimensional domain $\mathcal{A} \subset E^{\perp}$, called an *atomic surface*, or an *acceptance domain*, is attached to every vertex of an appropriate lattice $\mathcal{L} \subset \mathbf{R}^n$. The superspace structure is thus a periodic array of atomic surfaces, of the type $\mathcal{S} = \mathcal{L} \oplus \mathcal{A}$. The intersection $\mathcal{S} \cap E^{\parallel}$ consists in discrete points, giving the atomic positions in the quasicrystal.

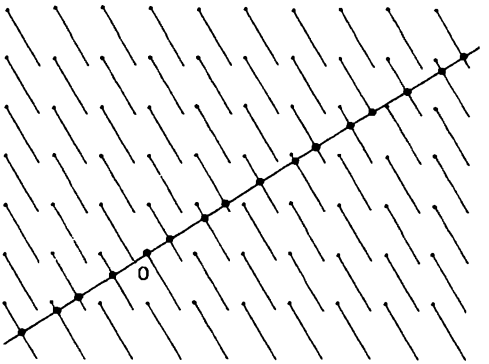


Figure 6: Atomic surfaces of the canonical one-dimensional quasicrystal, built from a two-dimensional superspace.

There are more complicated situations: the lattice $\mathcal{L}$ can be different from $\mathbf{Z}^n$; several atomic surfaces can be used in order to describe the different chemical species; yet other models involve finite decorations of the lattice $\mathcal{L}$, such as centerings.

Atomic surfaces can be viewed as closed surfaces, with the topology of a torus $\mathbf{T}^{n-d}$.[42,43] This property has the physical consequence that the atoms only hop over bounded distances if the cut $E^{\parallel}$ differs from flat space by a localized, smooth deformation. The corresponding degrees of freedom, that may become dynamical in some quasicrystals, are called *phasons*.

We close up this section by giving a quantitative description of the canonical one-dimensional quasicrystal,[44,32,45] namely the binary chain whose construction has been shown in Figures 5 and 6. If $t = \tan\theta$ denotes the slope of $E^{\parallel}$, the strip Ω is defined by the inequalities $0 < y - tx < t + 1$. As a consequence, the co-ordinates of the n-th point of $\mathbf{Z}^2$ along the broken line read

$$M_x = n - 1 - \mathrm{Int}(n\omega), \quad M_y = 1 + \mathrm{Int}(n\omega), \tag{44}$$

where $\mathrm{Int}(x)$ is the integer part of a real number x, and

$$\omega = \frac{t}{t+1} = \frac{\sin\theta}{\sin\theta + \cos\theta} \tag{45}$$

is the *incommensurability ratio*. If $t = p/q$ (irreducible fraction) is rational, then $\omega = p/(p+q)$ is also rational, and the structure is periodic, with $p + q$

atoms in a cell. To the contrary, if t, or equivalently ω, is irrational, the binary chain is a genuine one-dimensional quasicrystal.

By projecting (44) onto $E^{\parallel}$, we obtain an explicit expression for the abscissa of the n-th point of the structure, i.e.,

$$u_n = na + g_n, \tag{46}$$

where

- na is the *average lattice* of the structure, characterized by a mean interatomic distance
$$a = \frac{1}{\sin\theta + \cos\theta}. \tag{47}$$
- g_n is the *modulation*, or *fluctuation*, of the structure with respect to its average lattice. In the present case it reads
$$g_n = \gamma(n\omega), \tag{48}$$
where γ is the *modulation function*, or *hull function*, of the structure. It is a periodic function, with unit period, namely
$$\gamma(x) = (\cos\theta - \sin\theta)\big(\mathrm{Frac}(x) - 1\big), \tag{49}$$
with $\mathrm{Frac}(x) = x - \mathrm{Int}(x)$ being the fractional part of a real number x.

The Fourier transform of the structure is formally defined as

$$G(Q) = \sum_n \exp\big(-iQu_n\big). \tag{50}$$

The above results (46)-(49) on the atomic positions yield

$$G(Q) = \sum_{j,k} C_{j,k}\,\delta\left(\frac{Qa}{2\pi} - j - k\omega\right). \tag{51}$$

It is advantageous to put this result in perspective with the superspace formalism. Let $\mathbf{Q} = 2\pi(J, K)$ be a vector of the reciprocal superspace lattice $\widetilde{\mathcal{L}} = (2\pi\mathbf{Z})^2$. Its projections onto the spaces $E^{\parallel}$ and $E^{\perp}$ are

$$Q^{\parallel} = 2\pi(J\cos\theta + K\sin\theta), \quad Q^{\perp} = 2\pi(K\cos\theta - J\sin\theta). \tag{52}$$

The positions of the Bragg peaks in Eq. (51) coincide with $Q^{\parallel}$, up to the identification $j \equiv K - J$, $k \equiv J$. The corresponding amplitudes read

$$C_{j,k} = C(Q^{\perp}) = \frac{2a}{Q^{\perp}} \exp\left(\frac{iQ^{\perp}}{2a}\right) \sin\left(\frac{Q^{\perp}}{2a}\right). \tag{53}$$

The above results illustrate general features of the superspace description of quasicrystals. The Fourier module $\mathcal{M}$ is the projection of the reciprocal superspace lattice $\widetilde{\mathcal{L}}$ onto the cut $E^{\parallel}$, and the amplitudes of the Bragg diffractions are given by an amplitude function $C(\mathbf{Q}^{\perp})$, which is the Fourier transform of a single atomic surface $\mathcal{A} \subset E^{\perp}$, in suitable units.

This last property underlines a major difference between quasicrystals and other aperiodic objects, such as incommensurate modulated structures. In the case of an incommensurate structure, the hull function is usually regular (analytic), so that the amplitudes of the Bragg peaks exhibit a strong hierarchical ordering, with main diffractions and satellites. To the contrary, in the case of quasicrystals, the atomic surfaces have a sharp boundary. The hull functions, which can be used to describe quasicrystals with co-dimension one ($n = d+1$), are discontinuous. In any case the Fourier amplitudes fall off slowly, so that there is no way of distinguishing between main diffraction and satellites in quasicrystalline diffraction patterns, such as that shown in Figure 3. In the above example, the hull function $\gamma(x)$ of Eq. (49) is discontinuous, and the amplitude function $C(Q^{\perp})$ of Eq. (53) has a slow power-law decay, as $1/|Q^{\perp}|$.

2.3 The Penrose tiling

The Penrose tiling is a two-dimensional analogue of icosahedral quasicrystals. It shares their most salient features: quasiperiodicity, fivefold point symmetry, scale invariance by a factor equal to the golden mean τ. This tiling of the plane, shown in Figure 7, is made of two species of triangles, namely Robinson's P and Q, already defined in section 1.2. It can alternatively be described in terms of darts ($D = 2P$) and kites ($K = 2Q$), or in terms of the Penrose rhombs: the fat one, $L = 2P + 2Q$, with inner angles 2 and 3, and the skinny one, $S = 2P$, with inner angles 1 and 4 (in units of $\pi/5$). This tiling has had a considerable historical importance. It admits several definitions, and several construction algorithms.[46,47,48]

Here we wish to put the emphasis on the construction of the Penrose tiling by means of *inflation rules*, which ensure the *self-similarity* of the structure. Figure 8 shows how the Robinson triangles can be combined so as to form similar triangles, τ times larger. This procedure, referred to as *inflation*, can be iterated in order to generate an infinite tiling. To be complete, one has to distinguish between left and right triangles.[48,49].

If we forget for a while about geometry, and only keep track of the numbers of pieces involved, the inflation rules illustrated in Figure 8 can be written as follows:

$$\sigma_{\mathrm{P}} : \begin{vmatrix} P \to 2P + Q, \\ Q \to P + Q. \end{vmatrix} \tag{54}$$

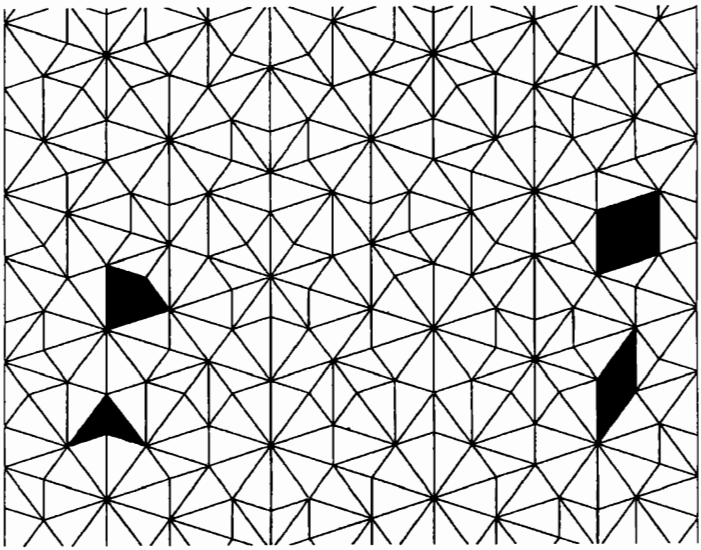


Figure 7: Penrose tiling. Open domains: Robinson's P and Q triangles. Dark areas: darts and kites (left), Penrose rhombs (right).

Such a formal transformation is referred to as a *substitution*.[50] It acts on abstract symbols called *letters*. Define the associated *counting matrix* as

$$\mathbf{M_P} = \begin{pmatrix} \#\text{ of } P\text{'s in } \sigma_{\mathbf{P}}(P) & \#\text{ of } P\text{'s in } \sigma_{\mathbf{P}}(Q) \\ \#\text{ of } Q\text{'s in } \sigma_{\mathbf{P}}(P) & \#\text{ of } Q\text{'s in } \sigma_{\mathbf{P}}(Q) \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}. \tag{55}$$

The characteristic polynomial of $\mathbf{M_P}$ is

$$\Pi(\lambda) = \det(\lambda \mathbf{1} - \mathbf{M_P}) = \lambda^2 - 3\lambda + 1 = (\lambda - \tau^2)(\lambda - \tau^{-2}). \tag{56}$$

The leading (Perron-Frobenius) eigenvalue τ^2 has a simple interpretation: τ is the linear scaling factor of the inflation rules, so that τ^2 describes the scaling factor of areas. The subleading eigenvalues of the substitution matrices associated with self-similar structures will play an important role in the following.

We end up by an observation, which seems to have been first made by L. Levitov.[51] It has been already noticed that fivefold symmetry is the first non-trivial point symmetry which is absent from Table 2, i.e., non-crystallographic. The Penrose tiling provides an example of a quasiperiodic structure with fivefold symmetry and scale invariance by a factor τ.

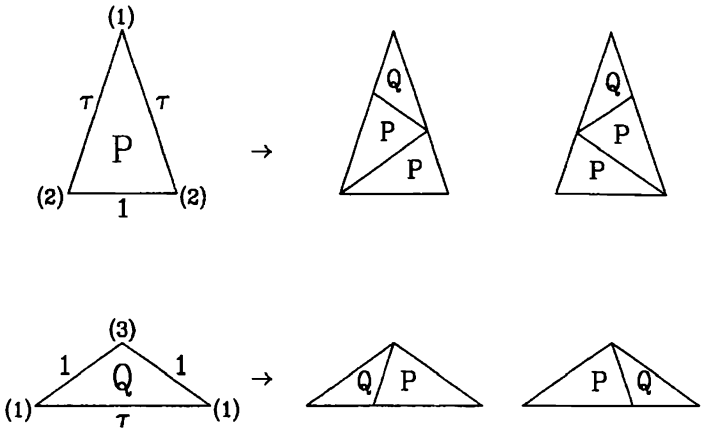


Figure 8: Inflation rules of the Penrose tiling in terms of Robinson's triangles. The numbers in parentheses give inner angles, in units of $\pi/5$.

Table 3: Classification of point symmetries of self-similar binary tilings of the plane.

N	x	structure
5	$(\sqrt{5}-1)/2 = \tau^{-1}$	Penrose
8	$\sqrt{2}$	octagonal
10	$(\sqrt{5}+1)/2 = \tau$	decagonal
12	$\sqrt{3}$	dodecagonal

More generally, which are the point symmetries of self-similar binary tilings of the plane?

This question can be answered as follows. Consider a structure with N-fold symmetry, built from a substitution acting on two letters. The associated Perron-Frobenius eigenvalue λ, a quadratic algebraic integer, reads $\lambda = \mu^2$, with μ being the linear scaling factor. Furthermore, it can be argued that μ and

$$x = 2\cos\frac{2\pi}{N} \tag{57}$$

are rationally related. Notice the similarity between Eqs. (36) and (57). But it is known from the theory of cyclotomic fields that x is an algebraic integer, with degree $\Phi(N)/2$, where $\Phi(N)$ is Euler's function, counting how many integers are smaller than N and prime to N. The answer to our question is thus

given by the condition $\Phi(N) = 4$, that easily yields the classification of Table 3. The remarkable feature of this list is that it exactly contains the point symmetries of *all the quasicrystalline phases observed so far*. In other words, quasicrystals are a realization of the next level of complexity after crystals, with quadratic algebraic integers as scaling factors, instead of natural integers. The smallest integer that is absent from both Tables 2 and 3 is $N = 7$. Tilings of the plane with sevenfold point symmetry have been investigated.[52,53] The irrational $x = 2\cos(2\pi/7)$, naturally associated with these structures, is a cubic integer (it obeys $x^3 + x^2 - 2x - 1 = 0$).

2.4 The Fibonacci chain

The Fibonacci chain is a one-dimensional analogue of the Penrose tiling and of icosahedral quasicrystals, sharing their quasiperiodicity and their self-similarity involving the golden mean τ. It is also the simplest example of a self-similar object, which we shall use for illustrating general techniques.

In the 13th century, Fibonacci introduced the following model for the growth of a population of rabbits. During a year, each adult (A) gives birth to a baby (B), whereas babies become adults. The population thus evolves according to deterministic rules, in the form of the following substitution:

$$\sigma_{\mathrm{F}} : \begin{vmatrix} A \to AB, \\ B \to A. \end{vmatrix} \tag{58}$$

The associated counting matrix

$$\mathbf{M}_{\mathrm{F}} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \tag{59}$$

is a square-root of the Penrose matrix (55). Its eigenvalues are $\lambda_1 = \tau$ and $\lambda_2 = -\tau^{-1}$.

By acting iteratively with the Fibonacci substitution σ_{F} on the initial letters A and B, we generate *words* $A_k = \sigma_{\mathrm{F}}^k(A)$, $B_k = \sigma_{\mathrm{F}}^k(B)$, which obey the concatenation relations

$$A_{k+1} = A_k B_k, \quad B_{k+1} = A_k. \tag{60}$$

In the particular case of the Fibonacci substitution, it is possible to consider the words A_k only, which obey the recursion

$$A_{k+2} = A_{k+1} A_k \quad (A_{-1} = B,\ A_0 = A). \tag{61}$$

Let ν_k^A and ν_k^B denote the numbers of letters of the words A_k and B_k. They obey the linear recursion formula

$$\begin{pmatrix} \nu_{k+1}^A \\ \nu_{k+1}^B \end{pmatrix} = \mathbf{M}_{\mathrm{F}} \begin{pmatrix} \nu_k^A \\ \nu_k^B \end{pmatrix} \quad (\nu_0^A = \nu_0^B = 1), \tag{62}$$

where $\mathbf{M}_{\mathrm{F}}$ is the Fibonacci matrix (59). It follows that the word $A_k = B_{k+1}$ consists of F_{k+2} letters, F_{k+1} of them being A's, and F_k of them being B's. In these expressions, the F_k are the Fibonacci numbers, which obey the recursion

$$F_{k+2} = F_{k+1} + F_k \quad (F_0 = 0,\ F_1 = 1), \tag{63}$$

and are given by

$$F_k = \frac{\tau^k - \left(-\tau^{-1}\right)^k}{\sqrt{5}}, \tag{64}$$

in terms of the eigenvalues of the Fibonacci matrix.

The *Fibonacci sequence* is the limit of the words A_k, i.e.,

$$A_\infty = ABAABAABABAABABAABAABABAABAABABAABAB\cdots \tag{65}$$

The *Fibonacci chain* is the binary structure obtained by "stringing beads" according to this sequence: pointlike atoms are placed at abscissas u_n, with $u_0 = 0$, so that the bond lengths assume two values, namely

$$\ell_n = u_n - u_{n-1} = \ell^A \text{ or } \ell^B, \tag{66}$$

depending on the type of the n-th letter in the sequence (65). Both the sequence and the chain are self-similar, with the irrational scaling factor τ. The quasiperiodicity of these objects will be demonstrated in a while.

To close up, we mention that the Fibonacci chain coincides with the canonical one-dimensional quasicrystal, whose construction is shown in Figures 5 and 6, if the slope of the cut $E^{\parallel}$ is $t = \tau^{-1}$, so that the incommensurability ratio is $\omega = \tau^{-2}$. More generally, for any irrational slope t, the binary structure generated by the cut-and-project algorithm can be described by concatenation rules.[44,54,55,45] The construction involves the *continued fraction expansion* of the slope t. Assuming $0 < t < 1$ for definiteness, t can be uniquely expanded as[20]

$$t = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}} = [a_1, a_2, a_3, \cdots], \tag{67}$$

where the integers $a_k \geq 1$ are called the *quotients* of the continued fraction expansion. The binary sequence encoding the canonical quasicrystal (A for a long bond of length $\ell^A = \cos\theta$, B for a short bond of length $\ell^B = \sin\theta$) is then obtained as the limit of a sequence of words $\{W_k\}$, which obey the concatenation rule

$$W_k = \begin{cases} W_{k-1}^{a_k} W_{k-2} & (k \text{ even}) \\ W_{k-2} W_{k-1}^{a_k} & (k \text{ odd}) \end{cases} \qquad (W_{-1} = B,\ W_0 = A). \tag{68}$$

The inverse golden mean τ^{-1} is the simplest irrational number: all its quotients read $a_k = 1$. As a consequence, the rules (68) amount to the Fibonacci substitution (58), up to a different choice of origin. This provides an elementary proof of the quasiperiodicity of the Fibonacci chain. More generally, we shall say that the canonical quasicrystal is self-similar if the rules (68) amount to a fixed substitution. This occurs if, and only if, the quotients are eventually periodic, i.e., $a_k = a_{k+s}$ for k large enough, with a fixed period s. This is equivalent to saying that the slope t is a quadratic algebraic number.[20] Quadratic numbers thus appear once more to play a special role in discrete geometry.

2.5 A classification of self-similar structures

We turn to the classification of self-similar structures in discrete geometry, i.e., chains and tilings generated by inflation rules, such as (54) or (58). This classification concerns geometrical characteristics of structures, as well as their diffraction spectrum. The following definition will play a central part:

A substitution has the Pisot-Vijayaraghavan *property (or* Pisot *for short) if its counting matrix* $\mathbf{M}$ *has all its eigenvalues* $\lambda_2, \cdots, \lambda_n$ *smaller than unity in modulus, except the Perron-Frobenius eigenvalue* $\lambda_1 > 1$.

In the case of an irreducible counting matrix, whose characteristic polynomial cannot be factorized over the integers, our definition coincides with the statement that the Perron-Frobenius eigenvalue λ_1 is a Pisot-Vijayaraghavan number, namely an algebraic integer, real and greater than unity, so that all its conjugates are smaller than unity in modulus.[56,57]

In order to motivate the classification given below, we first consider the modulation g_n of self-similar chains with respect to their average lattice, in the sense of Eq. (46). We restrict ourselves to binary chains, for definiteness. Let σ be a substitution acting on two letters, A and B. Consider the words A_k and B_k, defined as above. The total numbers of letters ν_k^A and ν_k^B contained in these words obey the recursion relation (62). The lengths of the finite patches of the chain associated with these words according to the rule (66), denoted by ℓ_k^A and ℓ_k^B, also obey (62), albeit with different initial values, i.e., $\ell_0^A = \ell^A$,

$\ell_0^B = \ell^B$. As a consequence, the modulations of the words A_k and B_k behave as

$$g_k^A = \ell_k^A - \nu_k^A a \sim g_k^B = \ell_k^B - \nu_k^B a \sim \lambda_2^k, \tag{69}$$

where λ_2 is the second eigenvalue of the counting matrix $\mathbf{M}$. The role of λ_2 is thus underlined.

Second, we consider the Fourier transform of self-similar structures. Taking the example of the Fibonacci chain for simplicity, we introduce the Fourier amplitudes $G_k^A(Q)$ and $G_k^B(Q)$ of the finite patches of the chain described by the words A_k and B_k, according to Eq. (80) below. The concatenation rule (60) implies the existence of recursion relations between the Fourier amplitudes, which read, in matrix form

$$\begin{pmatrix} G_{k+1}^A(Q) \\ G_{k+1}^B(Q) \end{pmatrix} = \begin{pmatrix} 1 & \exp\left(-iQ\ell_k^A\right) \\ 1 & 0 \end{pmatrix} \begin{pmatrix} G_k^A(Q) \\ G_k^B(Q) \end{pmatrix}. \tag{70}$$

The structure of the above equations is quite general: Fourier amplitudes are determined iteratively from linear recursion relations involving known phase factors.[54,55,58,59,60,45,32,61] The initial conditions encode the decoration of the structure, i.e., the details on the distribution of matter under consideration. In the present case of pointlike atoms, we have $G_0^A(Q) = \exp\left(-iQ\ell^A\right)$, $G_0^B(Q) = \exp\left(-iQ\ell^B\right)$. Now comes the key question:

Which self-similar structures in discrete geometry possess Bragg peaks in their diffraction spectrum?

E. Bombieri and J. Taylor have addressed it first, in the case of chains.[62] They noticed that linear recursions of the form (70) produce a Bragg peak under the generic circumstances that all the phase factors involved in these relations go asymptotically to unity. These conditions are, roughly speaking, of the form

$$x\lambda_1^k \to 0 \pmod 1, \tag{71}$$

where $\lambda_1 > 1$ is the Perron-Frobenius eigenvalue of the substitution, and x is proportional to the wavevector Q. The condition (71) has been studied by Pisot.[56] Its non-trivial solutions are classified as follows: λ_1 is a Pisot number, defined above, while x belongs to $\mathcal{M}(\lambda_1)$, a known module over the integers, depending only on λ_1.

The above arguments on the modulation and on the Fourier transform of self-similar chains and tilings are quite general. They have led to the following classification of self-similar structures generated by inflation rules, in terms of the eigenvalues of the counting matrix.[58,59,45,32,61] Let us emphasize that the statements which follow are generic, in the sense that they admit many particular cases, and many exceptions.

- *Pisot structures* ($|\lambda_2| < 1$)

⋆ The modulation of Pisot chains with respect to their average lattice is *bounded.*

⋆ Pisot structures generically have a purely discrete Fourier spectrum: they are *almost-periodic.* A more detailed classification is as follows:

  - $\det \mathbf{M} = \pm 1$: Such a structure is generically *quasiperiodic*, with a Fourier module spanned by n linearly independent vectors, with n being the number of types of letters (bonds, tiles) in the structure. The *atomic surfaces* of its superspace description are generically complicated objects, with fractal boundaries.[45,63] *Example*: octagonal tilings[63] [see section 2.6].
  - $\det \mathbf{M} \neq \pm 1$ and $\lambda_1 \in \mathbf{Z}$: Such a structure is *limit-periodic*, with a basis $b = \lambda_1$. *Examples*: the L-tiling and the sphinx tiling[48,64,52].
  - $\det \mathbf{M} \neq \pm 1$ and $\lambda_1 \notin \mathbf{Z}$: Such a structure is *limit-quasiperiodic*, a term due to F. Gähler[61].

- *Non-Pisot structures* ($|\lambda_2| > 1$)

⋆ The modulation of non-Pisot chains with respect to their average lattice generically exhibits a power-law divergence, of the form

$$g_n \sim n^{\beta}, \tag{72}$$

where the *wandering exponent* reads

$$\beta = \frac{\ln |\lambda_2|}{\ln \lambda_1}, \tag{73}$$

with λ_2 being the second largest eigenvalue in modulus of the counting matrix. The power law (72), which agrees with the estimate (69), is usually multiplied by an oscillatory prefactor, of the form $P(\ln n / \ln \lambda_1)$, where P is a periodic function of its argument, with unit period, that is *fractal*, i.e., continuous but nowhere differentiable.[65] In the case of tilings, there is no obvious lattice to which the structure is to be compared. Counting the vertices of a tiling contained in a domain of fixed shape and growing size can be misleading. Indeed, the seemingly simple problem of counting the points of the square lattice in a disk of radius R is essentially equivalent to the number-theoretical topic of sums of squares, described in section 1.3. In order to circumvent this difficulty, it is preferable to weigh the vertices with a smooth test function.[66]

⋆ The diffraction spectra of non-Pisot structures do not contain any (non-trivial) Bragg peak. Their Fourier intensities are generically purely singular continuous measures, with multifractal statistics.[58,59] *Example*: the fivefold binary tiling of C. Godrèche and F. Lançon[67,68,53] [see section 2.7].

• *Marginally non-Pisot structures* ($|\lambda_2| = 1$)

This marginal situation encompasses several kinds of exceptional cases. *Example*: circle-map sequences;[59,54,55] these are quasiperiodic sequences with a weakly unbounded fluctuation, while the associated chains according to the rule (66) usually have a singular continuous Fourier spectrum.

2.6 Octagonal tilings

According to the above classification, self-similar chains and tilings, generated by inflation rules with Pisot property and unit determinant, are generically quasiperiodic. The analysis of their superspace description has revealed an unexpected complexity in the morphology of their atomic surfaces, which generically possess a fractal boundary, and often infinitely many connected components.[45,63]

Self-similar chains have been investigated in a systematic way, in connection with counting systems.[45] In the case of binary chains, every atomic surface $\mathcal{A}$ generically consists of an infinity of segments of the line $E^{\perp}$, whose boundary is a self-similar, fractal Cantor set, with a non-trivial dimension d_{b}, with $0 < d_{\mathrm{b}} < 1$. As a consequence, the intensities of the Bragg diffractions fall off slower than (53), namely

$$\left|C(Q^{\perp})\right|^2 \sim \left|Q^{\perp}\right|^{-(2-d_{\mathrm{b}})}. \tag{74}$$

The only known examples of chains with a smooth atomic surface correspond to the canonical quasicrystals described in section 2.2, with a quadratic slope. In the case of chains generated by substitutions acting on $n \geq 3$ letters, atomic surfaces are $(n-1)$-dimensional objects, with fractal boundaries as a general rule, and often infinitely many connected components. No single example is known with a smooth atomic surface.

The atomic surfaces of self-similar quasiperiodic tilings of the plane have only been investigated in some examples. A characteristic case is provided by the three tilings with eightfold symmetry, shown in Figure 9.[63] These tilings are made of the same two species of tiles, the square S and the 45°-rhomb R. It is advantageous to consider the square as being made of two rectangular-isosceles triangles T. As far as counting properties are concerned, the three

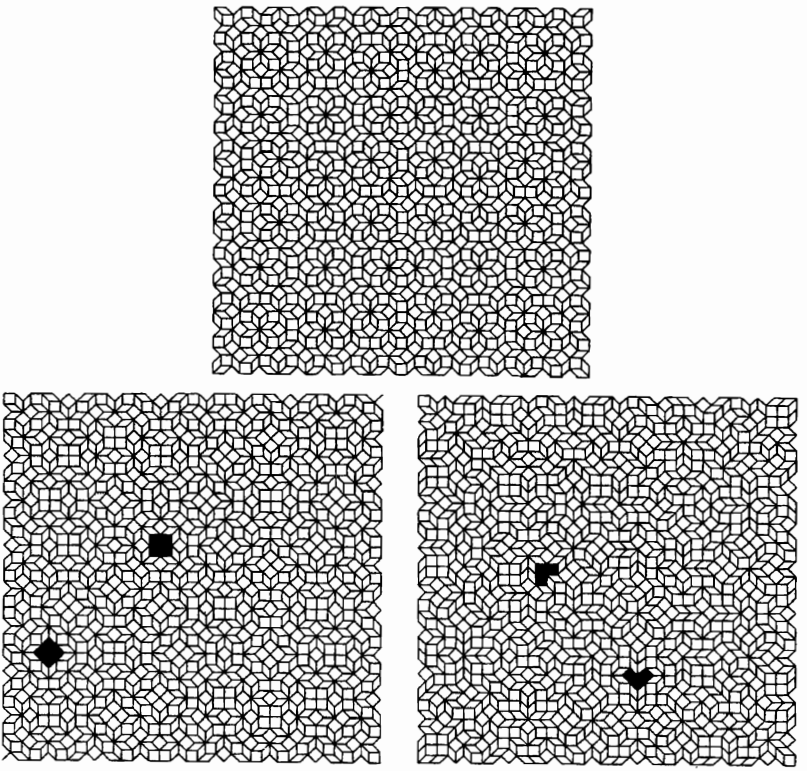


Figure 9: The three octagonal tilings investigated in Ref. 63: A (top), B (left), C (right). The dark patterns show specific local environments.

tilings are generated by the same substitution rules, i.e.,

$$\sigma : \begin{vmatrix} R \to 3R + 4T, \\ T \to 2R + 3T. \end{vmatrix} \tag{75}$$

The associated counting matrix

$$\mathbf{M} = \begin{pmatrix} 3 & 4 \\ 2 & 3 \end{pmatrix} \tag{76}$$

has eigenvalues $\lambda_1 = 3 + 2\sqrt{2} = \mu^2$ and $\lambda_2 = 3 - 2\sqrt{2} = \mu^{-2}$, with $\mu = 1 + \sqrt{2}$ being the linear scaling factor. It is therefore a Pisot matrix, with unit determinant.

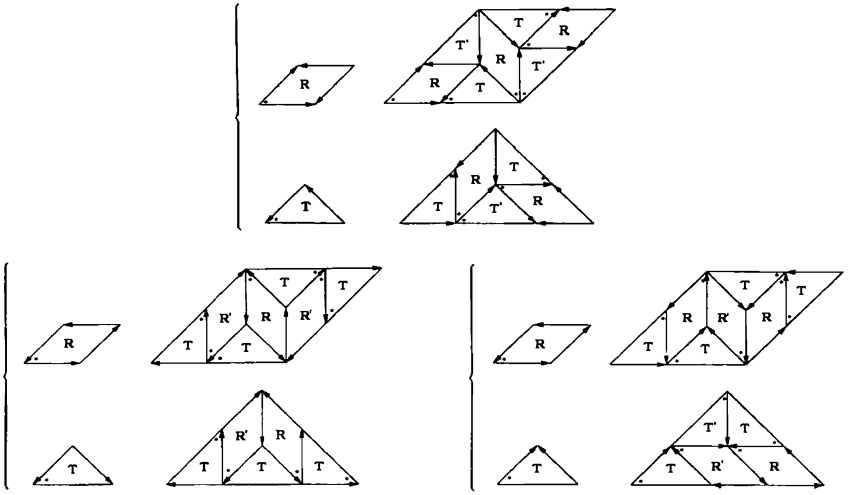


Figure 10: Inflation rules of the octagonal tilings shown in Figure 9: A (top), B (left), C (right).

Both tiles R and T occur with the same frequencies in the three tilings, $\rho^R = \sqrt{2} - 1$ and $\rho^T = 2 - \sqrt{2}$. Each tiling, however, is defined by its own arrangement of original tiles inside the inflated ones, shown in Figure 10, which leads to specific local environments.

These quasiperiodic tilings admit a four-dimensional superspace description, with a common lattice $\mathcal{L} = \mathbf{Z}^4$. Their two-dimensional atomic surfaces, shown in Figure 11, are also specific to each tiling:[53]

- *A-tiling*: this is the celebrated Ammann octagonal tiling[48]. Its atomic surface $\mathcal{A}_A$ is an octagon: it has a smooth boundary, and possesses the full eightfold symmetry of the tiling.

- *B-tiling*: its atomic surface $\mathcal{A}_B$ is a connected domain, with only fourfold symmetry. The full superspace structure is obtained by hanging a copy of the surface $\mathcal{A}_B$ at each even vertex of the lattice $\mathbf{Z}^4$, and a copy of $\mathcal{A}_B$ rotated by 45° at each odd vertex: this is a simple example of a centering, mentioned in section 2.2. The boundary of the atomic surface $\mathcal{A}_B$ is a fractal closed curve, with dimension

$$d_B = \frac{\ln 3}{\ln(1+\sqrt{2})} = 1.246477. \tag{77}$$

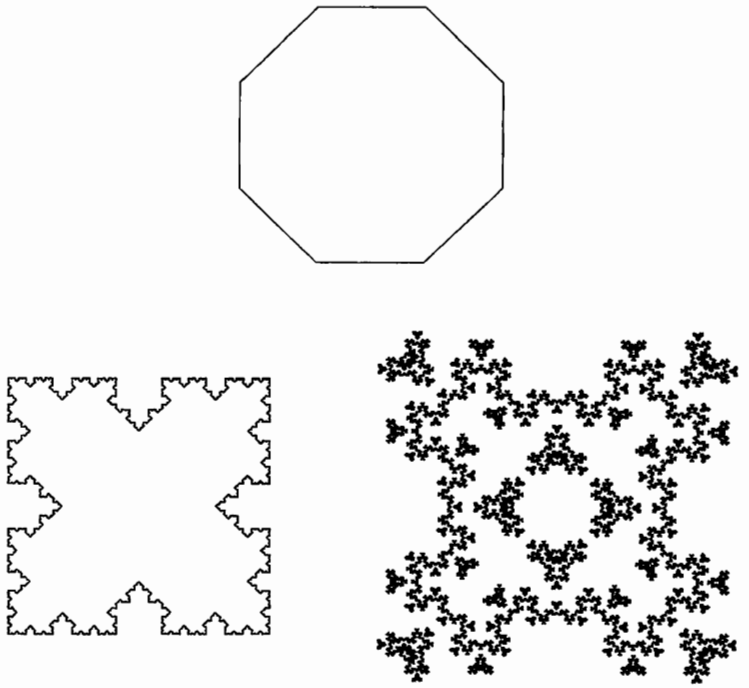


Figure 11: Boundaries of the atomic surfaces of the octagonal tilings shown in Figure 9: A (top), B (left), C (right).

- *C-tiling*: its atomic surface $\mathcal{A}_C$ consists of an infinity of connected components. It has fourfold symmetry, and the centering procedure of the B-tiling still applies. The boundary of $\mathcal{A}_C$ is a fractal set of closed curves, with altogether a dimension

$$d_C = \frac{\ln(1+\sqrt{10})}{\ln(1+\sqrt{2})} = 1.618000. \tag{78}$$

2.7 Non-Pisot structures and their diffraction spectra

According to the classification of section 2.5, self-similar chains and tilings, generated by non-Pisot substitutions, generically have an unbounded fluctua-

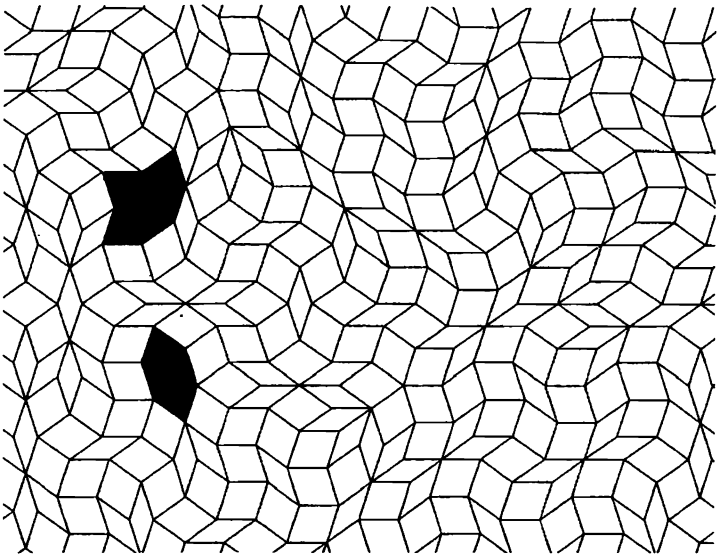


Figure 12: Fivefold binary tiling investigated in Ref. 67. Dark patterns are (up) the crown C and (down) the box B.

tion with respect to their average lattice, and a singular continuous, multifractal diffraction spectrum. In this section we shall illustrate the above general statements, and make them more precise.

We first describe a characteristic example of a non-Pisot structure, namely the binary tiling with fivefold symmetry investigated by C. Godrèche and F. Lançon.[67,68,53] This tiling, shown in Figure 12, consists of the same rhombs L and S as the Penrose tiling described in section 2.3. It is self-similar, since it can be constructed by means of inflation rules, shown in Figure 13, which transform the fat rhomb L into a crown $C = 3L + S$, and the skinny one S into a box $B = L + 2S$. The associated counting matrix

$$\mathbf{M} = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \tag{79}$$

has both its eigenvalues $\lambda_1 = 2+\tau$ and $\lambda_2 = 3-\tau$ larger than unity. This tiling is therefore a non-Pisot structure. Its wandering exponent reads $\beta = 0.251574$, according to Eq. (73).

Let us turn to a general description of the diffraction spectra of non-Pisot structures. In the following, we focus onto the one-dimensional case, again for

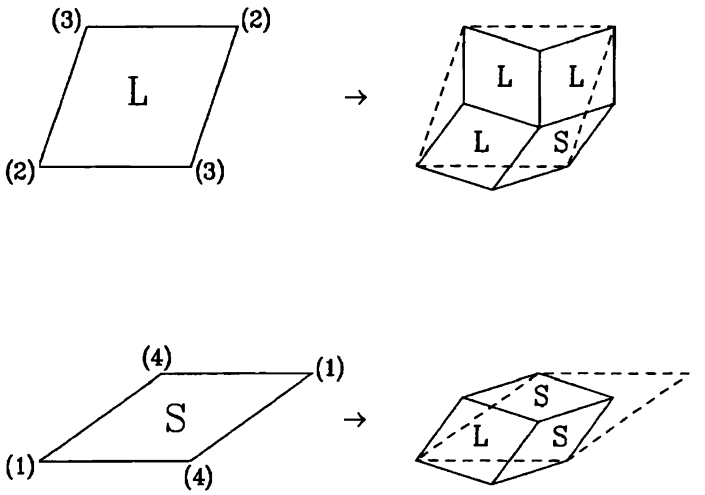


Figure 13: Inflation rules of the fivefold binary tiling shown in Figure 12. The numbers in parentheses give inner angles, in units of $\pi/5$.

the sake of simplicity. Consider an aperiodic chain, made of pointlike atoms situated at abscissas $\{u_n\}$. In order to make the definition (50) more precise, with the first N atoms we associate the *Fourier amplitude*

$$G_N(Q) = \sum_{n=1}^{N} \exp(-iQu_n), \tag{80}$$

and the *Fourier intensity*

$$S_N(Q) = \frac{1}{N}\left|G_N(Q)\right|^2 = \frac{1}{N}\sum_{m,n=1}^{N} \exp\bigl(iQ(u_m - u_n)\bigr). \tag{81}$$

The latter quantity usually has a limit for an infinite sample, referred to as the *(static) structure factor*, which reads formally

$$S(Q) = \left\langle \sum_{n} \exp\bigl(iQ(u_n - u_0)\bigr) \right\rangle, \tag{82}$$

where the brackets denote a sliding average over the origin 0. In the case of structures with configurational disorder, an equivalent procedure consists in averaging over the quenched randomness in the atomic positions.

The structure factor $S(Q)$ is observable, with some limited resolution, by the diffraction of various probes, such as electrons or X-rays. It must, however, be emphasized that $S(Q)$ is a generalized function (distribution). Indeed, in a mathematically rigorous setting, the Fourier intensity is a positive measure, referred to as the *Fourier (intensity) measure*, entirely characterized by its distribution function

$$H(Q) = \int_0^Q S(Q')\mathrm{d}Q'. \tag{83}$$

The Fourier measures of aperiodic substitutional structures usually obey scaling laws in Q-space, both locally and globally, which reflect their self-similarity in real space, albeit in a much more intricate way.

- *Local scaling*

  For some value Q_0 of the wavevector Q, the Fourier amplitude may obey a finite-size scaling law, which is most often a power law of the form

  $$G_N(Q_0) \sim N^\gamma, \tag{84}$$

  where γ is a local scaling exponent, depending on the wavevector Q_0. It can then be argued that the Fourier measure has a local singularity of the form

  $$H(Q_0 \pm \varepsilon) - H(Q_0) \sim \pm\varepsilon^\alpha, \tag{85}$$

  with

  $$\alpha = 2(1-\gamma). \tag{86}$$

  One has $\gamma \leq 1$ and $\alpha \geq 0$, by construction. The structure factor $S(Q_0)$ has a *peak*, i.e., it is divergent ($\alpha < 1$), whenever $\gamma > 1/2$.

  The scaling law (84) has an illuminating geometrical interpretation, inspired by number-theoretical investigations,[70] and referred to as the *Fresnel representation*.[55,32] It consists in plotting the successive amplitudes of Eq. (80), namely $G_0 = 0$, $G_1 = \exp(-iQu_1)$, $G_2 = \exp(-iQu_1) + \exp(-iQu_2)$, and so on, as points in the complex plane. A *Fresnel walk*, consisting of unit steps, is formed by joining these points. The power law (84) means that the dimension of this walk reads

  $$d_{\mathrm{walk}} = \frac{1}{\gamma}. \tag{87}$$

  In some examples the Fresnel walk is a nice, strictly self-similar fractal curve in the plane.[55,32]

- *Global scaling (multifractality)*

Global aspects of the scaling behavior of the Fourier intensity[58] are best described in terms of multifractal analysis.[23,24] Roughly speaking, a diffraction spectrum is multifractal if the distribution function $H(Q)$ can be attributed a singularity of the form (85) for every value of the wavevector Q, with a local exponent $\alpha(Q)$, and moreover if, for any given α in some range $[\alpha_{\min}, \alpha_{\max}]$, the set S_α of wavevectors Q so that $\alpha(Q) = \alpha$ is a fractal set, with a dimension

$$\dim S(\alpha) = f(\alpha). \tag{88}$$

The function $f(\alpha)$ is referred to as the *multifractal spectrum* of the intensity measure.

From a practical viewpoint, and especially for the purpose of a numerical evaluation, the curve $f(\alpha)$ is determined by means of the following *thermodynamical formalism*.[23,24] The measure is first regularized at the scale ε, by considering the intensities $\{w_n\}$ supported by intervals of length ε, i.e.,

$$w_n = H(n\varepsilon) - H((n-1)\varepsilon) = \int_{(n-1)\varepsilon}^{n\varepsilon} S(Q)\mathrm{d}Q. \tag{89}$$

These weights are then tested by considering the partition function

$$Z(q,\varepsilon) = \sum_{n=1}^{n_{\max}} w_n^q. \tag{90}$$

In this formula, q is a real parameter, analogous to inverse temperature in Statistical Mechanics, and the total range $Q_{\max} = n_{\max}\varepsilon$ is kept fixed. Multifractality manifests itself in the form of the scaling law

$$Z(q,\varepsilon) \sim \varepsilon^{\tau(q)} \quad (\varepsilon \to 0). \tag{91}$$

It is worth noticing that the scaling laws (34) and (91) are similar, up to the identification $k \equiv q$ and $1/\varepsilon \equiv \ln \mathcal{N}$.

The scaling exponent $\tau(q)$ yields, via the relation

$$\tau(q) = (q-1)D_q, \tag{92}$$

the spectrum of *Rényi dimensions* D_q of the intensity measure, which generalize the dimension D_0 of its topological support, the dimension of

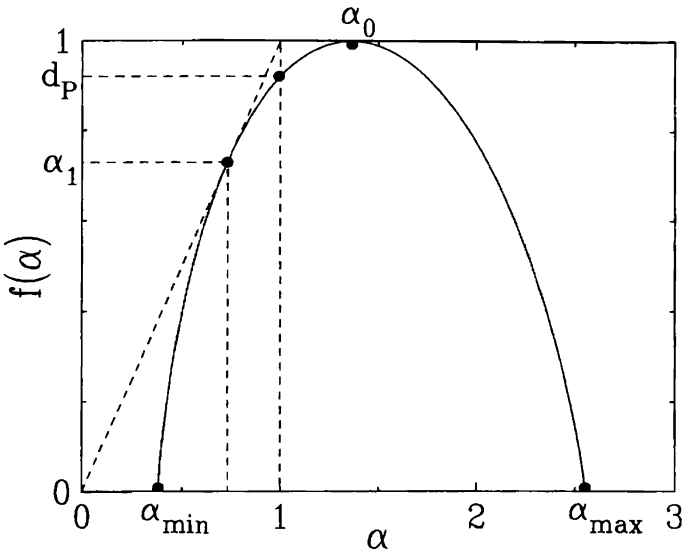


Figure 14: Qualitative sketch of the multifractal spectrum of the Fourier intensity of a typical non-Pisot chain. Particular values are described in the text.

information D_1, and so on. Furthermore, the functions $\tau(q)$ and $f(\alpha)$ are the Legendre transforms of each other, namely

$$\tau(q) + f(\alpha) = q\alpha, \quad \alpha = \frac{\mathrm{d}\tau}{\mathrm{d}q}, \quad q = \frac{\mathrm{d}f}{\mathrm{d}\alpha}. \tag{93}$$

Figure 14 shows a schematical plot of the multifractal spectrum of the Fourier intensity of a typical non-Pisot chain. The following characteristic values are of interest:[58,59,32]

- the bounds $\alpha_{\min} = D_{+\infty}$ and $\alpha_{\max} = D_{-\infty}$ determine the range of local scaling exponents α which show up with a "reasonable" probability in the diffraction spectrum;
- the abscissa α_0 of the top of the curve, corresponding to $q = 0$, represents the local exponent at a generic value of the wavevector Q. We have $f(\alpha_0) = D_0 = 1$, since the topological support of the intensity measure is the full Q-line;
- the value $d_P = f(1)$ represents the dimension of the set of peaks in the diffraction spectrum, in the general sense of singular scattering, exposed above.

Table 4: Scaling behavior of Fourier intensity: local exponents, dimension of Fresnel walk, type of intensity, measure-theoretical component.

γ	α	d_{walk}	Fresnel walk	intensity	measure
1	0	1	ballistic	Bragg	discrete
1/2	1	2	Brownian	diffuse	abs. cont.
$\in]1/2, 1[$	$\in]0, 1[$	$\in]1, 2[$	fractal	singular	sing. cont.

For a non-trivial multifractal diffraction spectrum, such as that of a generic non-Pisot self-similar chain or tiling, the particular values described above, and shown schematically in Figure 14, obey the inequalities

$$0 < \alpha_{\min} < \alpha_1 < d_P < 1 < \alpha_0 < \alpha_{\max}. \tag{94}$$

Finally, the three measure-theoretical components of the Fourier intensity, according to the Lebesgue decomposition theorem, can be heuristically described in terms of the local and global scaling properties described above (see Table 4).[58,59,68,32]

- *Bragg peaks* (discrete component): Bragg peaks are characterized by the maximal value $\gamma = 1$ of the local scaling exponent. They correspond to delta functions, i.e., a discrete component in the Fourier measure. The Fresnel walk at Bragg peaks is ballistic. According to the classification of section 2.5, the Fourier intensity of almost-periodic structures consists of Bragg peaks: it is purely discrete.

- *Diffuse scattering* (absolutely continuous component): diffuse scattering is typically observed in amorphous structures, such as glasses, which have configurational entropy, and randomness in their atomic positions. The structure factor $S(Q)$ is a smooth function, corresponding to the absolutely continuous component of the Fourier measure. For a generic wavevector Q, the Fresnel walk is a Brownian motion, whose dimension $d_{\text{walk}} = 2$ is a manifestation of the statistical law of large numbers.

- *Singular scattering* (singular continuous component): singular scattering corresponds to local scaling exponents in the range $1/2 < \gamma < 1$. It therefore appears as an intermediate kind of behavior between Bragg peaks and diffuse scattering. According to the classification of section 2.5, non-Pisot self-similar structures generically have a purely singular continuous Fourier intensity, with a non-trivial multifractal spectrum.

References

1. M.J. Giannoni, A. Voros, and J. Zinn-Justin (Editors), *Chaos and Quantum Physics, Proceedings of Les Houches Summer School, session LII* (North-Holland, Amsterdam, 1991).
2. P.J. Richens and M.V. Berry, *Physica* **2** *D*, 495 (1981).
3. V.I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, New York, 1978).
4. E. Aurell and C. Itzykson, *J. Geom. Phys.* **5**, 191 (1988).
5. C. Itzykson, *Int. J. Mod. Phys. A* **1**, 65 (1986).
6. R. Balian and C. Bloch, *Ann. Phys.* **60**, 401 (1970); **64**, 271 (1971); **69**, 76 (1972); **85**, 514 (1974).
7. A. Voros, *Comm. Math. Phys.* **110**, 439 (1987); *Advanced Studies in Pure Mathematics* **21**, 327 (1992).
8. F. Steiner, *Fortschritte der Physik* **35**, 87 (1987).
9. B. Duplantier and F. David, *J. Stat. Phys.* **51**, 327 (1988).
10. C. Itzykson and J.B. Zuber, *Nucl. Phys. B* **275**, 580 (1986).
11. C. Itzykson, P. Moussa, and J.M. Luck, *J. Phys. A* **19**, L 111 (1986).
12. A.M. Polyakov, *Phys. Lett.* **103** B, 207 (1981).
13. W.I. Weisberger, *Comm. Math. Phys.* **112**, 633 (1987).
14. E. Aurell and P. Salomonson, *Comm. Math. Phys.* **165**, 233 (1994).
15. M. Berry, *Ann. Phys.* **131**, 163 (1981).
16. C. Itzykson and J.M. Luck, *J. Phys. A* **19**, 211 (1986).
17. M.G. Lamé, *Leçons sur la théorie mathématique de l'élasticité des corps solides* (Bachelier, Paris, 1852).
18. J.W.S. (Lord) Rayleigh, *The Theory of Sound* (Dover, New York, 1945).
19. A. Weil, *Number Theory. An Approach through History, from Hammurapi to Legendre* (Birkhäuser, Boston, 1984).
20. G.H. Hardy and E.M. Wright, *An Introduction to the Theory of Numbers* (Clarendon, Oxford, 1979).
21. M. Waldschmidt, P. Moussa, J.M. Luck, and C. Itzykson (Editors), *From Number Theory to Physics* (Springer, Berlin, 1992).
22. P. Cartier, in Ref. 21.
23. G. Paladin and A. Vulpiani, *Phys. Rep.* **156**, 147 (1987).
24. J. Feder, *Fractals* (Plenum, New York, 1988).
25. S. Aubry, *J. Physique (France)* **44**, 147 (1983).
26. D. Shechtman, I. Blech, D. Gratias, and J.W. Cahn, *Phys. Rev. Lett.* **53**, 1951 (1984).
27. D. Levine and P.J. Steinhardt, *Phys. Rev. Lett.* **53**, 2477 (1984).
28. P.J. Steinhardt and S. Ostlund, *The Physics of Quasicrystals* (World

Scientific, Singapore, 1987).
29. M.V. Jaric and D. Gratias (Editors), *Aperiodicity and Order* (Academic Press, New York, three volumes published since 1988).
30. D.P. DiVincenzo and P.J. Steinhardt, *Quasicrystals: the State of the Art, Directions in Condensed Matter Physics, vol.* **11** (World Scientific, Singapore, 1991).
31. C. Janot, *Quasicrystals: a Primer* (Clarendon, Oxford, 1992).
32. J.M. Luck, in *Fundamental Problems in Statistical Mechanics VIII*, edited by H. van Beijeren and M.H. Ernst (Elsevier, Amsterdam, 1994).
33. H. Bohr, *Almost Periodic Functions* (Chelsea, New York, 1947).
34. A. Janner and T. Janssen, *Phys. Rev. B* **15**, 643 (1977); T. Janssen and A. Janner, *Adv. Phys.* **36**, 519 (1987).
35. T. Janssen, *Acta Cryst. A* **42**, 261 (1986); *Phys. Rep.* **168**, 55 (1988).
36. P. Kramer and R.W. Haase, in Ref. 29.
37. N.D. Mermin, in Ref. 30.
38. N.G. de Bruijn, *Nederl. Akad. Wetensch. Proc. A* **43**, 39 (1981).
39. M. Duneau and A. Katz, *Phys. Rev. Lett.* **54**, 2688 (1985); *J. Physique (France)* **47**, 181 (1986).
40. V. Elser, *Phys. Rev. B* **32**, 4892 (1985).
41. P.A. Kalugin, A.Yu. Kitayev, and L.S. Levitov, *J. Physique (France) Lett.* **46**, L 601 (1985); *J.E.T.P. Lett.* **41**, 145 (1985).
42. A. Katz, in Ref. 21.
43. L.S. Levitov, in Ref. 30.
44. J.M. Luck and D. Petritis, *J. Stat. Phys.* **42**, 289 (1986).
45. J.M. Luck, C. Godrèche, A. Janner, and T. Janssen, *J. Phys. A* **26**, 1951 (1993).
46. R. Penrose, *Math. Intell.* **2**, 32 (1979).
47. M. Gardner, *Scientific American* **236**, 110 (1977).
48. B. Grünbaum and G.C. Shephard, *Tilings and Patterns* (Freeman, New York, 1987).
49. C. Godrèche and J.M. Luck, *J. Stat. Phys.* **55**, 1 (1989).
50. M. Queffélec, *Substitution Dynamical Systems. Spectral Analysis* (Springer, Berlin, 1987).
51. L.S. Levitov, *Europhys. Lett.* **6**, 517 (1988).
52. C. Godrèche and J.M. Luck, in *Proceedings of the Anniversary Adriatico Research Conference on Quasicrystals*, edited by M.V. Jaric and S. Lundqvist (World Scientific, Singapore, 1990).
53. F. Lançon and L. Billard, *Phase Transitions* **44**, 37 (1993).
54. S. Aubry, C. Godrèche, and J.M. Luck, *Europhys. Lett.* **4**, 639 (1988).
55. S. Aubry, C. Godrèche, and J.M. Luck, *J. Stat. Phys.* **51**, 1033 (1988).

56. C. Pisot, *Ann. Scuola Norm. Sup. Pisa (Italy)* **7**, 205 (1938).
57. J.W.S. Cassels, *An Introduction to Diophantine Approximation* (Cambridge University Press, 1957).
58. C. Godrèche and J.M. Luck, *J. Phys. A* **23**, 3769 (1990).
59. C. Godrèche and J.M. Luck, *Phys. Rev. B* **45**, 176 (1992).
60. M. Kolar, B. Iochum, and L. Raymond, *J. Phys. A* **26**, 7343 (1993).
61. F. Gähler and R. Klitzing, in *The Mathematics of Aperiodic Order*, edited by R.V. Moody and J. Patera (Kluwer, Dordrecht, 1996).
62. E. Bombieri and J.E. Taylor, *J. Physique (France) Coll.* **C3**, 19 (1987); *Contemp. Math.* **64**, 241 (1987).
63. C. Godrèche, J.M. Luck, A. Janner, and T. Janssen, *J. Physique (France) I* **3**, 1921 (1993).
64. C. Godrèche, *J. Phys. A* **22**, L 1163 (1989).
65. J.M. Dumont, in *Number Theory and Physics, Springer Proceedings in Physics* **47**, edited by J.M. Luck, P. Moussa, and M. Waldschmidt (Springer, Berlin, 1990).
66. J.M. Luck, *Europhys. Lett.* **24**, 359 (1993).
67. C. Godrèche and F. Lançon, *J. Physique (France) I* **2**, 207 (1992).
68. C. Godrèche, *Phase Transitions* **32**, 45 (1991).
69. C. Godrèche, J.M. Luck, and F. Vallet, *J. Phys. A* **20**, 4483 (1987).
70. M. Dekking, M. Mendès France, and A. van der Poorten, *Math. Intell.* **4**, 130; 173; 190 (1983).

PHYSICS AND ARITHMETIC CHAOS IN THE FOURIER TRANSFORM

M.C. GUTZWILLER
IBM, T.J. Watson Research Center
Yorktown Heights, NY 10598

The Sinai billiard is the archetype of a classically chaotic system. It can be greatly generalized in two ways: i) replace the original square boundary with an arbitrary parallelogram and periodic boundary conditions, ii) replace the circular hard wall of radius R at the center by any circularly symmetric system, compact or non-compact. At a fixed energy E, both the classical trajectory and a propagating wave alternate between two regimes: i) constant direction of motion θ, ii) constant angular momentum L, which are complementary in quantum mechanics. The physics in either region is integrable and to a large extent arbitrary. Any chaos in the combined system is due to the transition from one to the other, through a Fourier transform F in quantum mechanics, which is discrete and finite of dimension $\cong 2\sqrt{2m\bar{E}}R/\hbar$. The spectrum of F is highly degenerate, with eigenvalues ± 1 in the real form (cosine or sine transform) occurring in almost the same multiplicity (with differences of 1 according as odd dimension). The matrix Fcan be simplified to a row of 2 by 2 blocks along the diagonal with the help of elementary number theory. But the directions of the eigenvectors in each block are shown to be distributed on the unit circle in an apparently uncorrelated manner. Are the two eigenspaces randomly oriented independently of the way they were calculated?

1 Introduction

Quantum mechanics is full of surprises that may provoke different responses in different people. As soon as a new problem is approached, and several investigators get busy trying to find a solution, they will concentrate on different aspects as the crucial ingredient for their method. Although the endresult is bound to be the same, the various explanations might sound quite different. This diversity may look strange to an outsider, but it is quite natural particularly in quantum mechanics where it appears built into the system from the very beginning and at a very deep level.

The purpose of this article is to point out yet another feature that enters into some typical problems and can be blamed for the some of our difficulties in finding a good base for understanding the solution. We like to invoke the idea of chaos every time we find ourselves unable not so much to obtain some numerical answer as to see through the numbers and comprehend where they come from. We call nature itself chaotic, although very often it is just one special step in our theory that seems to contradict our expectations of orderly

and rational behavior.

The special step to be called chaotic in this paper is the finite discrete Fourier transform. It causes, in my individual interpretation, all the trouble that makes a whole class of simple-looking problems hard to solve. This class starts from a model in classical mechanics that was first discussed by Sinai in 1968;[10] he succeeded in calculating the Kolmogoroff entropy which turned out to be positive as expected from more intuitive arguments. This entropy is defined to agree with Botzmann's original expression for the thermodynamic entropy as a measure for the (negative) logarithm of the probability for a particular arrangment of molecules. Evidently, the larger the entropy the more disorderly the arrangement, or, in our more recent manner of speaking, the more chaotic the classical motion.

The quantum-mechanical Sinai-billiard was first treated in a 1981 paper by Berry[2] that proposes many clever and useful ideas, and offers some numerical results about the energy levels. This special problem, a particle moving freely inside a unit-square with a circular obstacle of radius R at the center, can be generalized in many ways. Both, Sinai's proof of a positive entropy and Berry's method of calculating the spectrum, can be carried over without any change. The chaos seems to arise from the transition between the two complementary symmetries, conserved angular momentum in the neighborhood of the circular obstacle and conserved direction of motion inside the square.

In its quantum-mechanical version, this transition requires a finite discrete Fourier transform, i.e., the linear unitary transformation F where a complex vector-space of $2L+1$ dimensions is tranformed by the matrix,

$$F_{\ell m} = \frac{1}{\sqrt{2L+1}} \exp(2w\pi i \frac{\ell m}{2L+1}), \tag{1}$$

where w and L are positive integers. The value of L depends only on the mass m_0 and the energy E of the particle as well as the radius R of the circular obstacle independently of its precise nature. If we define a wave number $k = \sqrt{2m_0E}/\hbar$, then we can choose $L \leq kR$ provided we put $w = 2$ in (1). The indices ℓ and m designate angular momenta in multiples of $\hbar$, and the upper limit comes from the fact that a classical particle with angular momentum larger than kR cannot penetrate into the circular obstacle.

1.1 Outline of this article

This paper consists of a somewhat sketchy first part where the physical arguments for the presence of chaos are presented. Its main purpose is to show that even if the shape of the square is changed appropriately and the circular

obstacle at the center is modified, neither Sinai's result nor Berry's calculations will undergo any significant change. The Fourier transform will remain the cause of our difficulties.

The second part is a more careful, but elementary discussion of the Fourier transform, quite independently of its physical cricumstances. The main problem is purely mathematical: Can we diagonalize the matrix (1), not only by finding its spectrum which is fairly easy, but also by determining its eigenvectors? Remarkably, I have found almost no literature concerning the last point, except for some learned hints from a colleague at the IBM TJ Watson Research Center, Ephraim Feig, who has been very helpful and encouraging.

The matrix (1) will first be reorganized by renumbering its rows and columns with the help of the "discrete logarithm modulo $2L + 1$". Then F can be broken up into two by two blocks along the diagonal using a finite discrete Fourier transform in a space of L dimensions. Each block contains two eigenvectors with different eigenvalues, either ± 1 or $\pm i$, and it is trivial to calculate their orientation. But the corresponding angles follow no recognizable pattern, and seem distributed on the unit-circle without any apparent correlations.

Several conclusions will be drawn from this excercise in number theory related to the Fourier transform, all of them in the form of general questions or even challenges to the theoreticians. Most obviously, can the statistics of energy levels, in the Sinai billiard and its generalizations, be understood in terms of the apparent random behavior in the spectrum of the Fourier transform? More generally in quantum mechanics, since the Fourier transform describes the transition between non-commuting observables such as the kinetic and the potential energy in Schroedinger's equation, can this analysis be of any help?

1.2 Finite Quantum-Mechanics

A short note was published in the Comptes Rendus of the French Academy of Sciences of 1986 by Balian and Itzykson[1] on a system that leads almost exactly to the same mathematical problem as our view of the Sinai billiard. M.L. Mehta[8] then published a paper where he acknowledges Itzykson as suggesting the problem to find the eigenvalues and the eigenvectors of the finite Fourier transform. Since this meeting is dedicated to the memory of Claude Itzykson it gives me great satisfaction to rephrase his work in the present context although I discovered this connection only recently.

Let the wave function in the neighborhood of the circular obstacle be

written in the form,

$$f(\phi) = \frac{1}{\sqrt{2\pi}} \sum_{-L}^{L} a_\ell e^{2i\ell\phi}, \tag{2}$$

where we can think of $\ell\hbar$ as the angular momentum of the outgoing wave, and ϕ as the angular coordinate. The factor 2 in the exponential was inserted to take into account the center of symmetry, which remains even when the square boundary is replaced by a parallelogram.

The most highly localized wave-function of this kind, the equivalent of Dirac's δ-function, would be $\sin(2L+1)\phi/\sqrt{2\pi(2L+1)}\sin\phi$. If we test our wave function $f(\phi)$ with $2L+1$ such δ-functions that are evenly spaced around the circle, at $2\ell\pi/(2L+1)$, the result would be a vector $(f_{-L}, \ldots, f_0, \ldots, f_L)$. This vector can be directly expressed in terms of the amplitudes a_ℓ by transforming the vector $(a_{-L}, \ldots, a_0, \ldots, a_L)$ by the above Fourier transform with w=2.

In the nomenclature of Balian and Itzykson, each basis vector in the vector-space of the f's is an eigenvector of the operator Q, with the eigenvalue ω^ℓ where $\omega = \exp(2i\pi/(2L+1))$. Indeed, these are the locations of the test-points that were chosen on the rim of circular obstacle. The (angular) momentum operator P shifts one basis vector to the next higher one in a circular manner; it has the same eigenvalues, ω^λ, and its eigenvectors in the f-space are given by $f_\ell = \omega^{-\ell\lambda}$.

Balian and Itzykson investigate the Heisenberg group which is obtained by the multiplication of the operators P, Q, and ωI. They construct a unitary representation of this group as well as for the canonical group which is related to its automorphisms. Although I have not been able to understand the details, the formulas contain all the number-theoretical tools that I will use in the second part of this paper. All this leads to the diagonalization of the Fourier matrix (1) which is also my goal, and I can only assume that my results are the same although my calculations are quite elementary. The present article seems to go further by numerically evaluating the relevant formulas, and pointing out their apparent chaos. Mehta's work on the hand gives a very different construction; but he does not obtain a set of orthonormal eigenvectors.

2 The Many Versions of the Sinai Billiard

2.1 Different Shapes and Phase factors

The original billiard of Sinai was designed to imitate, in the most simple-minded manner, a gas of hard spherical balls which bounce around inside

a finite enclosure. The formidable technical difficulties of this fundamental problem were boiled down to the shape of a square for the enclosure, and the collisions between the balls were reduced to a single point particle hitting a circular hard wall at the center of the enclosure.

The particle was allowed to take advantage of periodic boundary conditions in the square. The formulas simplify thereby because the particle maintains its direction of motion as long as it stays away from the circular obstacle. Nothing changes in this respect if the square is replaced by any parallelogram with the basic vectors $\vec{\sigma}$ and $\vec{\tau}$, as long as the circular obstacle fits inside without touching the boundaries.

The quantum-mechanical analog allows for one more generalization which has some importance in calculating the band-structure in crystals, a connection that was already pointed out in Berry's paper. The wave function of the particle in the parallelogram may correspond to an overall crystal-momentum $\hbar\vec{K}$. When the particle leaves the parallelogram through the side $\vec{\tau}$ in the direction of $\vec{\sigma}$, then its wave function acquires a factor $\exp i(\vec{K},\vec{\sigma})$, and similarly with the sides $\vec{\sigma}$ and $\vec{\tau}$ interchanged.

The crystal-momentum can vary in the whole two-dimensional plane, but its variation can be limited to one Brillouin zone, i.e., without loss of generality one can impose the limits $0 \leq (\vec{K},\vec{\sigma}) < 2\pi$ and $0 \leq (\vec{K},\vec{\tau}) < 2\pi$. The crystal momentum $\hbar\vec{K}$ is simply one more conceivable parameter in the generalized Sinai billiard in quantum mechanics. Its presence in the problem, however, will break the reflection symmetry at the origin.

2.2 An Assortement of Circular Obstacles

The hard circular wall of radius R has the effect of spreading the motion of the particle of mass m_0 into different directions, called defocusing for simplicity's sake. Consider two trajectories that hit the critical circle in the same place, but at angles differing by $\delta\theta$. On the average, the particle will travel a distance $D = (Area\ of\ parallelogram)/2R$ before hitting the circle again. By this time the two trajcetories will be a distance $d \cong D\delta\theta$ away from each other. Therefore, the difference in their direction of motion after the second collision will increase to $\delta\Theta \cong d/R \cong (Area/R^2)\delta\theta$.

This argument seems somewhat paradoxical because it suggests that the smaller the circular obstacle the more effective it is in defocusing the classical trajectories. If we estimate the increasing divergence of the trajectories per unit length of trajectory, and divide $\delta\Theta$ by the average travel-distance D between collisions, the divergence $\delta\theta$ is found to increase at a rate inversely proportional to R, i.e., the smaller circle the bigger the chaos! Of course, quantum mechanics

does not allow this conclusion because the obstacle does not scatter any longer when its radius becomes smaller than the de Broglie wave-length $2\pi/k$ of the particle.

This superficial explanation for the chaos in the Sinai billiard remains valid if the obstacle is replaced by all kinds of different contraptions as long as the circular symmetry is preserved. Instead of a hard wall, one can think of any potential that depends only on the distance from the center of the circle, e.g., an attractive Coulomb field. The particle will penetrate into the inside of the critical circle where it follows along the appropriate ellipse, parabola, or hyperbola according as its energy E. Eventually it will emerge again at a different place, but with the same angular momentum M. There might actually be bound states that are modified by the presence of the parallelogram.

The main difference with the Sinai billiard is the shift $\omega(E, M)$ in the angle from the place where the particle hits the circle to the place where it energes back into the parallelogram. The angular momentum M is the same before and after the interlude inside the circular obstacle. Therefore, the change in the direction of motion is again symmetric with respect to the local radius, exactly as in the hard collision of the Sinai billiard, but with the additional complication of the shift $\omega(E, M)$.

This picture can be further modified in the following manner which has been discussed at some length by the author.[5] The inside of the critical circle can be replaced by an exponential horn or a pseudosphere so that the particle continues its motion on such a two-dimensional surface of circular symmetry. It is convenient to choose a surface of constant, negative Gaussian curvature $-1/R^2$ because the classical trajectories can be obtained by elementary constructions, and the solutions of Schroedinger's equation are no worse than Legendre functions for which many convenient expansions are known.

Since the inside of the critical circle is no longer compact, it may happen that the particle does not return to the inside of the parallelogram at all. It may escape from the parallelogram provided its angular momentum M allows it to do so. For instance in the case of the exponential horn $M = 0$, i.e., only the particle hitting exactly along the local radius is able to escape through the horn.

A pseudosphere of area $2\pi R^2$ can be joined to a cylindrical pipe of radius smaller than R, whereas a pseudosphere of the full area $4\pi R^2$ can be joined to an open Euclidean plane. The idea of the Sinai billiard can then be inverted: the particle approaches an obstacle in one or two dimensions, and gets caught inside a parallelogram; it will emerge again in quantum mechanics with a well-defined phase-shift $\eta(E, M)$ depending on energy and angular mementum, although in classical mechanics the phenomenon is more complicated.

This last version of the Sinai billiard is expected to yield a phase-shift that is completely different from the results of the usual potential scattering in quantum mechanics. It should resemble quite closely the author's calculations [4] of scattering on the punctured torus (leaky torus), i.e., a torus of constant negative curvature with an exceptional point. In the simplest case, the scattering phase-shift $\eta(E,0)$, since only particles with $M = 0$ escape, is given by the logarithm of the Riemann zeta-function on the line in the complex plane parallel to the critical line, but with the real part equal to 1. The amazing properties of this almost-periodic function were discussed again by the author.[6]

2.3 Quantum Mechanics Near the Circular Obstacle

The purpose of this section is to remind the reader of the method for calculating the spectrum in problems of this kind. Many of the details can be found in the papers by Berry [2] and the author.[5]

The easy part is the solution of the stationary Schroedinger equation just outside of the circular obstacle in polar coordinates (r,θ). For each angular momentum $M = \ell\hbar$, one has an outgoing and an incoming wave which are Hankel functions of the first and the second kind of index ℓ and argument kr, where the wave number is always $k = \sqrt{2m_0E}/\hbar$. Both are multiplied by $\exp(i\ell\theta)$ for their angular dependence. The complete wave-function is expanded in the form,

$$\phi = \sum_{\ell=-\infty}^{+\infty} [a_\ell H_\ell^{(1)}(kr) + b_\ell H_\ell^{(2)}(kr)]e^{i\ell\theta}, \tag{3}$$

where the coefficients a_ℓ and b_ℓ are complex numbers.

The exact nature of the circular obstacle determines the ratio a_ℓ/b_ℓ for each ℓ. In the usual Sinai billiard the wave function ϕ has to vanish for $r = R$, so that $a_\ell/b_\ell = -H_\ell^{(2)}(kR)/H_\ell^{(1)}(kR)$. This ratio has the absolute value 1; when kR becomes larger than the absolute value of ℓ, this ratio becomes equal to 1.

More generally, we can say that the vectors $a' = (\ldots, a_{-1}, a_0, a_{+1}, \ldots)$ and $b' = (\ldots, b_{-1}, b_0, b_{+1}, \ldots)$ are tied to each other through the unitary matrix D,

$$a = -Db,\ D_{\ell\ell} = \exp(-2i\delta_\ell(E)),\ D_{\ell m} = 0, \tag{4}$$

for $\ell \neq m$. The phase angle $\delta_\ell(E)$ in the diagonal elements is usually called the phase-loss; it tells us all we have to know about the wave's penetration into the circular obstacle.

If this obstacle is non-compact, as in the case of the exponential horn or the pseudosphere, some of these diagonal elements may contain other parameters while their value is still on the unit circle. As shown in author's work,[5] the element D_{00} depends on the ratio of the amplitudes α and β for the incoming and the outgoing waves of angular momentum 0 very far out on the exponential horn. In the nomenclature at the end of the preceding section, $\alpha/\beta = \exp(i\eta(E, 0)$.

2.4 The Waves in the Parallelogram

The discussion of the waves leaving the circular obstacle, and eventually returning to it after being scattered by the boundaries of the parallelogram, is more involved. Only the essential points will be mentioned, and the reader is referred again to the work of Berry[2] and the author (1993 and in preparation). Our language will sound as if the wave propagates in time: "the outgoing wave leaves the obstacle, gets reflected off the boundaries of the parallelogram, and returns as an incoming wave to the obstacle." Of course, the interpretation of these words is purely symbolic and formal because everything is assumed to happen at a constant energy E.

The boundary conditions are enforced by following construction. The same outgoing wave, $\sum_\ell c_\ell H_\ell^{(1)}(kr)\exp(i\ell\theta)$, is produced simultaneously by all the equivalent obstacles in the complete lattice for the parallelogram, i.e., from all the points with coordinates $\vec{\gamma} = \mu\vec{\sigma} + \nu\vec{\tau}$ where μ and ν run through the integers from $-\infty$ to $+\infty$. These waves come from the lattice points outside the origin. They have to be expanded in the neighborhood of the origin with the help of the addition theorem for Bessel functions. Finally, they have to be disentangled into incoming and outgoing waves as shown in (3).

Both the vector a and the vector b in (3) can be expressed in terms of the assumed vector c quite formally as

$$2a^{prime} = c'(I + iY),\ 2b' = c'(I - iY), \tag{5}$$

where I is the identity, or unit matrix. The matrix Y has the elements

$$Y_{\ell m} = \sum_{\gamma \neq 0} Y_{\ell-m}(k\gamma)e^{i(\ell-m)\theta}, \tag{6}$$

where the Bessel function $Y_{\ell-m}(k\gamma)$ is real for real values of its argument. The reflexion symmetry of the lattice with respect to the origin shows that $Y_{\ell-m} = 0$ for odd $\ell - m$.

These relations are purely formal because the matrix-elements $Y_{\ell m}$ are given by an expansion over the lattice sites $\vec{\gamma}$ that converges conditionally.

The above formulas and their consequences can be made rigorous, as will be shown in a forthcoming paper by the author. It will also be proved that all the matrices can be truncated, as mentioned earlier, for the absolute values of ℓ and m exceeding some bound like $2kR$. After eliminating the vector c from (5),

$$b = -Ta, T = \frac{I - iY}{I + iY}, \tag{7}$$

where the matrix T is again unitary since Y is basically real and symmetric. Notice that T transforms even (odd) numbered components of a into even (odd) numbered components of b.

Formula (6) shows explicitly that the matrix T is a Toeplitz matrix, i.e., it has the same value for all elements along a parallel to the diagonal, where $\ell - m$ is constant. Therefore, T can be diagonalized by a Fourier transform F. Since the even and odd numbered components are completely independent, let us only look at the even ones. The Fourier transform of a vector $g = (\ldots, g_{-2}, g_0, g_{+2}, \ldots)$ is a function $f(\chi)$ of period π of an angle χ. The matrix T can, therefore, be written in the form,

$$FTF^{-1} = \delta(\chi - \zeta)\Theta(\chi, E), \tag{8}$$

where we have used the Dirac δ-function. The function $\Theta(\chi, E)$ depends on the angle χ and on the energy E in a rather complicated way which reflects the resonances of the parallelogram.

2.5 Calculation of the spectrum

The spectrum of the generalized Sinai-billiard results from combining the conditions (4) and (7) into the secular equation,

$$\det(I - U) = 0,\ U = TD = F^{-1}\Theta FD, \tag{9}$$

where only the diagonal matrices Θ and D depend on E. Since there are no further parameters if the obstacle is compact, and the determinant is presumably a smooth function of E, this equation yields a discrete set of eigenvalues E_j. If the obstacle is not compact, external parameters, such as the ratio α/β appear in U besides E. Equation (8) then gives this ratio, or equivalently the phase shift $\eta(E, 0)$, as a function of the energy.

The Fourier transformation F can be truncated for practical purposes in the manner indicated in (1). The value $w = 2$ in (1) comes directly from the matrix Y in (6) which does not mix even and odd angular momenta. Although the truncation depends on E if we set $L \cong kR = \sqrt{2m_0 E}R/\hbar$, it really arises

automatically from the diagonal matrices D and Θ. Also the truncation of F to an odd number of components is natural because there is a natural pairing of positive and negative angular momenta.

It is now evident that all the physics is hidden in the diagonal unitary matrices D and Θ. At the same time, we can claim with some justification that the chaos in the generalized Sinai-billiard does not come from either of these matrices (or their classical equivalents). Each matrix describes a different and unrelated, perfectly integrable problem, i.e., on one hand, the waves inside a circularly symmetric, compact or non-compact enclosure, and on the other hand, the waves inside an arbitrary parallelogram with periodic boundary conditions (including some phase conditions as in a crystal).

The chaotic features in the solutions of the Sinai billiard arise from combining these two integrable problems with complementary symmetries, conserved angular momentum and conserved direction of motion. The Fourier transform F is the "glue" that holds these two discrepant ingredients together in one physical problem. If F could be decomposed into its eigenspaces in a simple and intuitively appealing procedure, then we would not have to invoke chaos for understanding the spectrum of the Sinai billiard in any of its many realizations.

The second part of this paper will try to show that the Fourier transform as given by the matrix F cannot be thrown into a form where its effect in the Sinai billiard is easily understood. Sofar, the tricks of mathematicians have not provided the physicists with any help toward a better idea than staring at the results of numerical computations. Statistics seems to be our only available future!

3 The Diagonalization of the Discrete Finite Fourier Transform

3.1 The Real and Imaginary Part of the Complex DFFT

The matrix F in (1) is unitary, i.e., if multiplied by its Hermitian conjugate it yields the identity. Therefore, its eigenvalues lie on the unit-circle. Moreover, it can be checked that its fourth power is also the identity, so that its eigenvalues can only be ± 1 or $\pm i$. The first question is concerned with the multipicities of these eigenvalues.

The $(2L+1)$-dimensional complex vector-space with a typical vector $a' = (a_{-L}, \ldots, a_0, \ldots, a_{+L})$ gets transformed into the vector $f' = (f_{-L}, \ldots, f_0, \ldots, f_{+L})$ in the usual manner $f = Fa$.

Let us define the $(L+1)$-dimensional vector-space of the even components $b' = (b_0, \ldots, b_L)$ where $b_0 = a_0$ and $b_m = (a_{-m} + a_{+m})/\sqrt{2}$, and the L-

dimensional vector-space of the odd components $c' = (c_1, \ldots, c_L)$ where $c_m = (a_{+m} - a_{-m})/\sqrt{2}$. And let us represent in exactly the same manner the vector f in terms of its even components g and its odd components h.

Then we have for the even components $g = Cb$ in terms of the matrix,

$$C_{\ell m} = \frac{2}{\sqrt{2L+1}} \cos(2w\pi \frac{\ell m}{2L+1}), \tag{10}$$

for $\ell > 0$ and $m > 0$, along with $C_{00} = 1/\sqrt{2L+1}$ and $C_{\ell 0} = C_{0m} = \sqrt{2}/\sqrt{2L+1}$. Similarly, for the odd components $h = iSc$ in terms of the matrix,

$$S_{\ell m} = \frac{2}{\sqrt{2L+1}} \sin(2w\pi \frac{\ell m}{2L+1}). \tag{11}$$

Both of the matrices are real and orthogonal so that their eigenvalues are again on the unit circle. But moreover, they are symmetric so that their eigenvalues are real. Therefore, their eigenvalues are ± 1. There are $L+1$ eigenvalues ± 1 corresponding to the even components and L eigenvalues $\pm i$ in F corresponding to the odd components. The next question is concerned with the degeneracies in C and in S.

3.2 Decomposition of the 2L+1 into Prime Factors

Suppose that we can decompose $2L+1 = (2L_1+1)(2L_2+1)$ where the two factors are prime to each other. The integer ℓ running from $-L$ to $+L$ can then be represented by a couple (ℓ_1, ℓ_2) where $\ell_1 = \ell$ modulo $2L_1+1$ and $\ell_2 = \ell$ modulo $2L_2+1$; we can again choose $-L_1 \leq \ell_1 \leq +L_1$ and $-L_2 \leq \ell_2 \leq +L_2$.

The solution of these two Diophantine equations is guaranteed by the Chinese remainder-theorem. First, the two equations $u_1 = 1$ modulo $2L_1+1$, and $u_1 = 0$ modulo $2L_2+1$ are solved for u_1 where clearly, $u_1 = v_1(2L_2+1)$. Second, the two equations $u_2 = 1$ modulo $2L_2+1$, and $u_2 = 0$ modulo $2L_1+1$ are solved for u_2 where $u_2 = v_2(2L_1+1)$. Then we can put $\ell = \ell_1 u_1 + \ell_2 u_2 = \ell_1 v_1(2L_2+1) + \ell_2 v_2(2L_1+1)$ modulo $2L+1$, and check our conditions in the preceding paragraph.

This expression along with the similar one, $m = m_1 v_1(2L_2+1) + m_2 v_2(2L_1+1)$ modulo $2L+1$, can be inserted into matrix elements of the Fourier matrix (1). In this manner we find that,

$$\exp(\frac{2\pi i w \ell m}{2L+1}) = \exp(\frac{2\pi i w \ell_1 m_1}{2L_1+1} v_1^2(2L_2+1)) \exp(\frac{2\pi i w \ell_2 m_2}{2L_2+1} v_2^2(2L_1+1)), \tag{12}$$

so that the matrix decays into a Cartesian product of two similar matrices, $2L_1+1$ by $2L_1+1$ and $2L_2+1$ by $2L_2+1$. The important feature to observe,

however, is the new factor $v_1^2(2L_2+1)$ in the exponent of the first matrix, and the factor $v_2^2(2L_1+1)$ in the second. Factors of this kind cannot be removed.

The decomposition of $2L+1$ into prime factors simplifies the diagonalization of the Fourier matrix (1), but one cannot avoid the appearance of factors like w in (1), even if $w=1$ for the Fourier matrix before the decomposition of $2L+1$ into prime factors. They are the left-overs from the decomposition, and can have any arbitrary value depending on the circumstances, as long as w is prime to $2L+1$.

3.3 The Multiplicities of the Eigenvalues in the DFFT

After the decomposition into even and odd components, all we have to do now is to compute the trace of the matrix F. Therefore, we write

$$Trace(F) = \frac{1}{\sqrt{2L+1}} \sum_{\ell=-L}^{+L} \exp(w\frac{2\pi i\ell^2}{2L+1}) = \frac{1}{\sqrt{2L+1}} G(w, 2L+1), \quad (13)$$

where the second expression follows immediately because $(-\ell)^2 = \ell^2 = (2L+1-\ell)^2$ modulo $2L+1$, and the second line defines the Gauss sum.

The explicit evaluation of $G(w, 2L+1)$ is one of the marvels of number theory, and can be found in many advanced textbooks. It involves the theory of quadratic residues, i.e., the problem of solving the Diophantine equation $x^2 = w$ modulo $2L+1$. Its solution requires the computation of the Legendre symbol (w/p) for any prime that divides $2L+1$. The Legendre symbol equals 1 if the equation has a solution, 0 if p divides w, and -1 if the equation has no solution. (w/p) can be computed with relative ease by using the quadratic-recipocity law of Euler-Legendre-Gauss-Jacobi.

Evidently, this route is not so simple to travel, and it would be far too long for any detailed discussion. Therefore, we quote only the results of direct interest in the present context. If $L = 4\kappa + \lambda$, we get the simple formulas,

$$G(1, 2L+1) = i^{L^2}\sqrt{2L+1},\ G(2, 2L+1) = i^{-\lambda}\sqrt{2L+1}. \quad (14)$$

According to these formulas for the Gauss sums, the unit-vectors in (13) add up as if they were oriented at random, since the absolute value of their sum is equal to the sqare root of their number. Nevertheless, they have a well defined phase which gave even Gauss a lot of trouble.

Concerning the spectrum of the Fourier matrix (1), we conclude that its four eigenvalues, ± 1 and $\pm i$, occur essentially with the same multiplicities. More precisely, the multiplicities of three of them are equal, and the multiplicity of the fourth differs by ± 1 depending on the value of L modulo 4. The formula

(14) can also be applied to find the multiplicities of the eigenvalues for the matrix (10) of the cosine transform and (11) of the sine transform.

It should mentioned that the same statements about multiplicities were derived without the help of number theory by McClellan and Parks [7] in the IEEE Transactions on Audio and Electroacoustics as well as by Dickinson and Steiglitz [3] in the IEEE Transactions on Acoustics, Speech, and Signal Processing. Although their arguments are closer to everybody's mathematical training, I don't find them any simpler, and to some extent more arbitrary. Also, they do not adress the problem of finding the eigenspaces which is our main interest.

3.4 Renumbering the Rows and Columns in the Fourier matrix F

All the further discusssion will be based on the two following assumptions: i) the number of dimensions $2L+1$ cannnot be decomposed into prime factors, i.e., there is a prime number p such that $2L+1 = p^{\kappa}$ with κ a positive integer; ii) in all the explicit calculations we will set $\kappa = 1$. The first assumption seems umavoidable, and not very drastic because the decomposition of the number of dimensions into prime factors is straightforward. The second assumption simplifies the notations and computations, and the generalization to $\kappa > 1$ follows the same method as $\kappa = 1$.

As far the applications to the Sinai billiard is concerned, the number of dimensions $2L+1$ increases with the energy E. Our assumptions restrict, therefore, this number to increase somewhat haphazardly by picking only the prime numbers p. While this feature is not quite satisfactory, it constitutes a vast improvement over the practice of using "Fast Fourier Transform" algorithms whose dimensions have to be an integer power of 2. In other words, there are many more prime numbers than powers of 2; it would be impossible to understand any of the underlying physics if we had jump from one power of 2 to the next.

The center column in F as well as its center row, both numbered 0 in our nomenclature, have all their matrix elements equal to 1. The four L by L matrices in the corners have a total of $4L^2 = (p-1)^2$ matrix elements whose values depend on the product ℓm modulo p where both ℓ and m differ from 0. According to a basic theorem in number theory, which probably goes back to Fermat and was first proved by Euler, there are exactly $p-1$ combinations (ℓ, m) whose product modulo p is a fixed integer in the interval from 1 to $p-1$.

The basic trick in diagonalizing F consists in renumbering the rows and columns so that the matrix elements are arranged according as the sums of their indices rather than according as their products. This conversion can be

achieved by introducing the idea of a generator g in the multiplicative group of integers $(1, 2, \ldots, p-1)$ modulo p. Again, Fermat and Euler have provided all the required details so that we can represent any integer of this group in the form,

$$\ell = g^{\lambda} \text{ modulo } p, \tag{15}$$

where λ is ordinarily made to vary in the additive group of integers $(1, 2, \ldots, p-1)$ modulo $p-1$, where $p-1$ assumes the role of 0 according to Fermat's "little theorem".

Formula (15) provides a one-to-one map from the integers $1 \leq \ell \leq p-1$ into themselves. Jacobi published the first comprehensive table of this map for the primes up to 1000, but nowadays such tables are easily generated even with the simplest program such as BASIC on DOS. Going from λ to ℓ represents a kind of discrete exponential function, whereas going from ℓ to λ is called the discrete logarithm modulo p. The traditional notation is $\lambda = I(\ell)$, the index function modulo p for a fixed generator g.

This one-to-one map is profoundly chaotic to the point where it is the most common tool for the generation of random numbers as well as the key in many cryptographic enterprises. Moreover, there seem to exist unsolved mysteries in the nature of the smallest generator g for a given prime, although their number is easy to figure out. Given any one of them, the others are obtained with the help of (15) where λ runs through all the integers that are prime relative to $p-1$.

The required renumbering of the rows and columnns in F is accomplished by a permutation matrix. It leaves the center row and the center column untouched so that one can ignore them for the time being. The new $p-1$ by $p-1$ matrix has the row indices λ and the column indices μ. Its matrix elements depend only on the sum $\lambda + \mu$ modulo $p-1$, i.e., most emphatically, the same matrix elements are not arranged along the lines parallel to the main diagonal, but to the cross diagonal. This is not a Toeplitz matrix.

Nevertheless, an ordinary Fourier transform of this $p-1$ by $p-1$ matrix almost succeeds in the diagonalization. As will be shown in the next section, the full p by p matrix is reduced thereby to a set of 2 by 2 matrices along the main diagonal. Each "box" of this kind contains a pair of eigenvalues, either ± 1 or $\pm i$, and the corresponding eigenvectors are trivial to compute in terms of a rotation angle. The chaos resides in the distribution of these angles, and it is reasonable to suspect that this chaos is a direct consequence of the renumbering. The trouble remains serious, however, as long as no other scheme for the diagonalization of the Fourier transform is available.

3.5 Explicit Construction of the Eigenvectors

The most straightforward method to find the eigenspaces of F is to define the following orthonormal base: a vector ψ_0 with the components $(\psi_0)_\ell = \delta_{\ell 0}$, and the remaining $p-1$ vectors given by the formula,

$$(\psi_\rho)_0 = 0,\ (\psi_\rho)_\ell = \frac{1}{\sqrt{p-1}} \exp(\frac{2\pi i \rho I(\ell)}{p-1}) \text{ for } \ell \neq 0, \tag{16}$$

where $1 \leq \rho \leq p-1 = 2L$. For the time being $-L \leq \ell \leq +L$, whereas $0 \leq \rho \leq 2L$. Notice that the vectors ψ_ρ and $\psi_{2L-\rho}$ are complex conjugate to each other.

Since $g^L = -1$ modulo p (a necessary condition for g to be a generator!), it follows that $g^{L+\lambda} = -g^\lambda = -\ell$. Therefore, $I(-\ell) = L + I(\ell)$ always with $L = (p-1)/2$, so that $(\psi_\rho)_{-\ell}$ is $(-1)^\rho$ times the complex conjugate of $(\psi_\rho)_\ell$. The vectors ψ_ρ for even ρ are associated with eigenvectors of the cosine transform, whereas for odd ρ they are associated with the eigenvectors of the sine transform.

The following computations are elementary,

$$F\psi_0 = \frac{1}{\sqrt{p}}\psi_0 + \frac{\sqrt{p-1}}{\sqrt{p}}\psi_{p-1},\ F\psi_{p-1} = \frac{\sqrt{p-1}}{\sqrt{p}}\psi_0 - \frac{1}{\sqrt{p}}\psi_{p-1}, \tag{17}$$

$$F\psi_\rho = \triangle_\rho \psi_{2L-\rho} \text{ for} 1 \leq \rho \leq 2L-1 = p-2, \tag{18}$$

with the definition,

$$\triangle_\rho = \frac{1}{\sqrt{p}} \sum_{n=-L}^{+L}{}' \exp(2w\pi i \frac{n}{p}) \exp(2\pi i \frac{\rho I(n)}{p-1}), \tag{19}$$

where the prime on the sum symbol implies that $n = 0$ is left out.

Therefore, the vectors ψ_ρ and $\psi_{2L-\rho}$ are coupled by the Fourier transform quite generally for $0 \leq \rho \leq 2L = p-1$. But there are two special cases: i) $\rho = 0$ with $\rho = 2L$ where (17) gives an explicit 2 by 2 matrix with real entries and the determinant -1, so that the eigenvalues are ± 1 and we can immediately find the two corresponding eigenvectors. ii) $\rho = L = (p-1)/2$ which is coupled to itself, i.e., ψ_L is already an eigenvector of F with the eigenvalue $\triangle_L$ given by (19).

For the remaining values of ρ, again an elementary computation shows that

$$\triangle_\rho^* = (-1)^\rho \triangle_{2L-\rho}. \tag{20}$$

Therefore, the determinant of the unitary 2 by 2 matrix that couples ψ_ρ with $\psi_{2L-\rho}$ equals -1 for even ρ and yields a pair of eigenvectors, one each for ± 1. For for odd ρ, however, the determinant is $+1$ and yields a pair of eigenvectors, one for each $\pm i$. These conclusions rest on the fact that the absolute value of $\triangle_\rho$ equals 1, whose direct proof requires one more elementary but rather tricky computation.

The eigenvectors from ψ_0 and ψ_{p-1} are

$$\pm\sqrt{p-1}\psi_0 + (\sqrt{p} \pm 1)\psi_{p-1} \text{ for } \pm 1. \tag{21}$$

The eigenvectors of F and their eigenvalues for even ρ are,

$$\triangle_\rho^* \psi_\rho \pm \psi_{2L-\rho} \text{ for } \pm 1, \tag{22}$$

while the eigenvectors for odd ρ and their eigenvalues are,

$$\triangle_\rho^* \psi_\rho \mp i\psi_{2L-\rho} \text{ for } \pm i. \tag{23}$$

That leaves the eigenvector ψ_L whose eigenvalue is $\triangle_L$, whose sign contains the whole mystery of the phase in the Gauss sum.

3.6 Changing the Generator g modulo p

Consider two generators g_1 and g_2 modulo p with the corresponding index functions $I_1(n)$ and $I_2(n)$, and assume that the second generator can be represented in terms of the first through,

$$g_2 = {g_1}^h \text{ modulo } p, \tag{24}$$

where h is any integer that is prime to $p-1$. It follows that

$$I_1(n) = h\, I_2(n) \text{ modulo } p-1. \tag{25}$$

Let the vectors χ_σ be generated by the same formula (16) as the vectors ψ_ρ, with $g_2, I_2(\ell), \sigma$ replacing $g = g_1, I_1(\ell) = I(\ell), \rho$. The main question is: how are the vectors χ_σ and ψ_ρ related to each other? Again, an elementary calculation shows that

$$\chi_\sigma = \psi_\rho iff \sigma = h\rho \text{ modulo } p-1. \tag{26}$$

Similarly, the two $\triangle$-functions, $\triangle_\sigma^{(2)}$ with respect to the generator g_2 and $\triangle_\rho^{(1)}$ with respect to the generator g_1, are related through,

$$\triangle_\sigma^{(2)} = \triangle_\rho^{(1)} \text{ where } \sigma = h\rho \text{ modulo } p-1. \tag{27}$$

The eigenvectors of F remain in the same pairing if the generator is changed from g_1 to g_2, and the formulas (22) and (23) are invariant under such a change. It can happen for a fixed change of generators h, however, that $\sigma = 2L - \rho$ (and $2L - \sigma = \rho$) for some particular values of ρ, but not for all of them. The formulas (22) and (23) remain the same, of course, but we find that,

$$\triangle^{\star}_{2L-\rho}\psi_{2L-\rho} \pm \psi_{\rho} = \pm(\triangle^{\star}_{\rho}\psi_{\rho} \pm \psi_{2L-\rho}), \tag{28}$$

for even ρ, while for odd ρ,

$$\triangle^{\star}_{2L-\rho}\psi_{2L-\rho} \mp i\psi_{\rho} = \pm i(\triangle^{\star}_{\rho}\psi_{\rho} \mp i\psi_{2L-\rho}). \tag{29}$$

There is no invariant way to distinguish ψ_{ρ} and $\psi_{2L-\rho}$. The vectors ψ_{ρ} don't change at all when the generator is changed according to (26), nor do the $\triangle$ functions according to (27). Since the the absolute value of the $\triangle_{\rho}$ is 1, only its phase angle is required for determining the eigenvectors. Because of the indeterminacy with respect to ρ and $2L - \rho$, however, that phase angle can either refer to $\triangle_{\rho}$ or to $\triangle_{2L-\rho} = \triangle^{\star}_{\rho}$. Therefore, we will always take this angle to be in the upper half of the unit circle.

A short remark concerns the dependence of the eigenvectors on the special factor w in the Fourier transform F in (1). The eigenvectors are given by (16) independently of w, and their transformation by F leads to (18) with (19). The factor w in the exponent of the first factor in (19), however, can be shifted to the second factor, and appears eventually as factor $\exp(-2\pi i\rho I(w)/(p-1))$ in front of $\triangle_{\rho}$ as computed with $w = 1$. In other words, the phase angle of $\triangle_{\rho}$ gets decreased by $2\pi\rho I(w)/(p-1)$. Of course, such a shift will not affect the more complicated correlations between the points $\triangle_{\rho}$ on the upper unit circle.

3.7 Computing the Eigenvectors of the DFFT

Formulas (22) and (23) as well as the discussion in the preceding section show that the eigenvectors of the DFFT require only the knowledge of the phase α_{ρ} in $\triangle_{\rho} = \exp(\pm i\alpha_{\rho})$ where $0 \leq \alpha_{\rho} \leq \pi$. Formula (19) can be simplified for the explicit calculation at the price of having slightly different expressions for even and odd values of ρ.

With the help of $I(-n) = L + I(n)$ with $L = (p-1)/2$ (as pointed out at the beginning of 3.5) we find that for even ρ,

$$\triangle_{\rho} = \frac{2}{\sqrt{p}}\sum_{n=1}^{L}\cos(2w\pi i\frac{n}{p})\ \exp(2\pi i\frac{\rho I(n)}{p-1}) = \exp(\pm i\alpha_{\rho}), \tag{30}$$

whereas for odd ρ,

$$\triangle_\rho = \frac{2i}{\sqrt{p}} \sum_{n=1}^{L} \sin(2w\pi i \frac{n}{p}) \exp(2\pi i \frac{\rho I(n)}{p-1}) = i \exp(\pm i\beta_\rho) = \tilde{\triangle}_\rho, \qquad (31)$$

where the expression defines β_ρ again in the range $0 \leq \beta_\rho \leq \pi$.

The real and imaginary parts of $\triangle_\rho$ are now obtained by decomposing $\exp(2\pi i \frac{\rho I(n)}{p-1})$ into cosine and sine. The whole calculation becomes elementary provided a simple routine for computing the index function $I(n)$ is available; but that requires no more than the "mod-function". A valuable check for large primes p comes from the absolute of value of $\triangle_\rho$ equal to 1. Even the generator g can be obtained by trial and error in a simple routine that checks whether $g^L = -1$ modulo p. The BASIC interpreter which is part of every DOS operating system does all these things, and has been used for all the data to be discussed.

The most convenient choice of a generator g for numerical calculations is the smallest, usually a single-digit integer, although some strange things can happen. Of the 167 prime numbers below 1000, only 13 require a double-digit smallest generator; the first of these is 191 with 19 as the smallest generator; the next four primes are 193, 197, 199 with the smallest generators 5, 2, 3. In enumerating the pairs of eigenvectors we will use the index ρ belonging to the smallest generator.

The $L+1$ eigenvalues ± 1 contain one trivial pair which follows from (21) for $\rho = 0$. It is not very interesting and will not be part of the data because its direction in the (ψ_0, ψ_{p-1}) space converges to $\pm\pi/4$ for large primes p. As a general rule, there is also an isolated eigenvalue which is ± 1 for even L, and $\pm i$ for odd L. That leaves $[(L-1)/2]$ non-trivial pairs ± 1 and $[L/2]$ pairs of $\pm i$, where the symbol $[X]$ indicates the largest integer that does not exceed X.

If we set $w = 1$ for simplicity's sake, according to (14) for even L, i.e., $p = 1$ modulo 4, there is a single eigenvalue $+1$ plus $L/2$ pairs of each, ± 1 and $\pm i$ eigenvalues. For odd L, however, i.e., $p = 3$ modulo 4, there is a single eigenvalue $+i$ plus $(L+1)/2$ pairs of ± 1 pairs and $(L-1)/2$ pairs of $\pm i$ eigenvalues. The data will list the angles α_ρ or β_ρ for each set of pairs separately.

3.8 The distribution of α's and β's

The points on the unit circle representing $\triangle_\rho$ for even ρ = 2, 4,..., 2L-2, and $\tilde{\triangle}_\rho$ for odd ρ = 1, 3,..., 2L-1 may not be randomly distributed. Indications of

some residual order can be seen when the following sums are computed,

$$\sum_{even\rho} \Delta_\rho = \frac{1}{\sqrt{p}}(1 + (p-1)\cos(2\pi w/p)), \tag{32}$$

$$\sum_{odd\rho} \tilde{\Delta}_\rho = \frac{p-1}{\sqrt{p}}\sin(2\pi w/p)). \tag{33}$$

Obviously, for large values of p the sum over the points with even index grows as $\sqrt{p}$, while sum over the points with odd index goes to 0 as $2\pi w/\sqrt{p}$.

Nevertheless, both of these behaviors are consistent with a random distribution on the unit circle, because the real value for the sum of these unit vectors is dictated by the symmetry of the distribution with respect of the real axis. The fixed value of the sum as function of p, however, indicates a certain degree of correlation. But it seems hard to discover any subtle order simply by looking at the numerical data, e.g., in the two tables that are given at the end of this paper.

Both tables were calculated for the discrete finite Fourier transform (DFFT) of the prime dimension $p = 199$, and with $w = 1$ in (1). This particular choice was made because p is large enough to justify an effort toward some statistical analysis, and yet small enough to present the results in detail. With $L = 99 = 3$ modulo 4, the formula (14) for the Gauss sum indicates 50 pairs of eigenvalues ± 1 and 49 pairs of eigenvalues $\pm i$ with an isolated eigenvalue $+i$. Table 1 gives the angles α_ρ from (30) for the 49 non-trivial pairs of the cosine transform, while Table 2 gives the angles β_ρ from (31) for the 49 pairs plus the isolated eigenvalue of the sine transform.

The upper third of each table lists the increasing values of the α's and β's in ordinary degrees from 0^0 to 180^0. The middle third of each table lists the differences between the consecutive angles as they appear in the upper third, again in ordinary degrees; the first angle in the upper table is again found at the head of the middle table. The last value in the middle table gives the angle that is needed to return from the last value of the upper table to the first one. Finally the lower table list the differences in the middle table in increasing order. Table 2 lists the isolated eigenvalue $+1$ for the sine transform with the angle $.003^0$, which is the result of the calculation and gives an idea of the numerical precision in general.

At the end of the line between the upper and the middle table, the location for the sum of all 98 unit vectors Δ_ρ for even ρ is given, resp. for the 99 unit vectors $\tilde{\Delta}_r ho$ for odd ρ. The line between the middle and the lower table gives some additional information: first as a check, the sum of all the differences in the middle and lower table is seen to be 180; then the number of entries in each

table is divided into 180 to give the average distance between the points on the unit circle. Finally, a measure for the mean-square deviation is offered: (sum of angular differences squared / sum of angular differences) - average angular difference.

A simple model for the statistics is to assume that the points on the unit circle are located as if thrown there independently by a pure random-process. Then the distribution of the distances between nearest neighbors is a Poisson distribution. The meaurements of length in the interval from 0 to 180 are normalized to the average distance, which we assume to be $180/50 = 3.6$ for simplicity's sake. The probability for finding the nearest neighbor on the way up a distance x away in an interval of length dx is given by $\exp(-x)dx$.

For a simple test we divide the distances into bins of a length equal to half the average distance, i.e., 1.8 starting at 0. The most probable numbers in the first 10 bins are then given by the series 19.67, 11.93, 7.24, 4.39, 2.66, 1.61, .98, .59, .36, .22. The corresponding numbers from the Table 1 for the even eigenvalue-pairs are 18, 14, 4, 5, 4, 2, 1, 1, 0, 0, whereas for the odd eigenvalue-pairs one finds 19, 11, 8, 5, 4, 1, 0, 2, 0. At this primitive level of statistical analysis one can hardly avoid the conclusion that the angles are independently distributed.

3.9 Some Conclusions

The main claim of this article is a statement concerning the spectrum of the discrete finite Fourier transform (DFFT): It is well known that not only the multiplicities of the eigenvalues, but also the eigenspaces can be constructed explicitly with the help of relatively simple results of number theory. This construction eventually boils down to diagonalizing many unitary 2 by 2 matrices, each of which is characterized by an angle. The striking result for these angles is that they are apparently distributed like so many independent points on the unit circle.

As long as there does not exist any other construction for the spectrum and the eigenspaces of the DFFT, one can say with some justification that the Fourier transform is responsible for the chaos in problems like the generalized Sinai billiard. One can further venture the claim that this kind of chaos is unavoidable in quantum mechanics where the interplay between complementary quantities, such as the direction of motion and the angular momentum in the Sinai billiard, is essential. It would be interesting to investigate to what extent the mild symptoms of chaos in quantum mechanics (in contrast tó the more virulent symptoms in classical mechanics) can be reduced to the trouble in the DFFT.

A suggestion for a large field of future work is the generalization of the Sinai billiard to three dimensions where the same dichotomy is prevalent: conserved direction of motion for one part of the motion, and conserved angular momentum for the other. In either part we have a large choice of possibilities, all of them integrable and represented by diagonal unitary matrices in quantum mechanics. Again they are tied together by a Fourier transform, which occurs on a sphere this time, however; but the questions concerning its spectrum may be even more difficult because we are now dealing with a non-commutative group of transformations, whereas the group of rotations of a circle is Abelian.

Finally, I am not sure to what extent the effects of this apparent chaos in the DFFT can be tracked down in number theory. Most of the arithmetic in the second part of this paper can be found in any textbook of algebra that treats the study of the cyclotomic fields (cf. Rademacher[9]). Actually, the idea of these fields and most of the above arithmetic goes back two centuries, to Lagrange and Gauss who was thereby led to his famous construction of the hepta-dekagon. Nevertheless, are there in number theory the kind of statistical features that seem rather obvious in the DFFT ?

References

1. R. Balian and C. Itzykson, *C.R.Acad.Sc.Paris* Serie I **303**, 773 (1986).
2. M.V. Berry, *Ann. Phys.* (New York) **131**, 163r (1981).
3. B.W. Dickinson and K. Steiglitz, *IEEE Trans* ASSP - **30**, 25 (1982).
4. M.C. Gutzwiller, *Physica* D **7**, 341 (1983).
5. M.C. Gutzwiller, *Chaos* **3**, 591 (1993).
6. M.C. Gutzwiller, *Prog. Theor. Phys.* Suppl. **No 116**, 1 (1994).
7. J.H. McClellan and T.W. Parks, *IEEE Trans* AU - **20**, 66 (1972).
8. M.L. Mehta, *Journal of Mathematical Physics* **28**, 781 (1987).
9. H. Rademacher in *Lectures on Elementary Number Theory* (Publishing Co., New York, 1964).
10. Y.G. Sinai, *Funct. Anal. Appl.* **2**, 61, 245 (1968).

Table 1: The Angles α of the Cosine Transform

6.406	25.660	60.518	96.950	142.941
6.654	30.112	63.663	100.440	150.741
9.986	31.996	67.520	109.745	152.620
10.616	32.680	67.991	118.110	157.276
11.557	32.755	69.135	126.018	157.649
12.544	39.360	70.162	133.112	163.124
14.701	40.023	72.854	134.344	163.315
15.922	40.948	76.241	137.750	176.340
17.665	46.996	79.333	140.253	178.067
20.029	56.989	84.441	142.618	0
6.406				14.09985
.248	4.452	3.145	3.489	7.799
3.332	1.884	3.857	9.306	1.879
.630	.684	.471	8.365	4.656
.940	.048	1.144	7.908	.373
.987	6.605	1.028	7.093	5.475
2.157	.663	2.692	1.233	.191
1.221	.925	3.387	3.405	13.082
1.743	6.048	3.092	2.504	1.670
2.364	9.994	5.108	2.364	8.339
5.631	3.529	12.509	.324	0
180	49	3.673	2.894	
.075	.940	2.157	3.490	7.093
.191	.987	2.364	3.529	7.799
.248	1.028	2.364	3.857	7.908
.324	1.144	2.504	4.452	8.340
.373	1.221	2.692	4.656	8.365
.471	1.233	3.092	5.108	9.306
.630	1.670	3.145	5.475	9.993
.663	1.743	3.332	5.631	12.509
.684	1.879	3.387	6.048	13.082
.925	1.884	3.405	6.605	0

Table 2: The Angles β of the Sine Transform

.003	28.959	74.332	123.946	146.731
2.291	28.987	78.142	125.632	146.808
5.342	32.501	79.039	127.279	148.498
12.182	36.579	79.998	128.228	150.060
13.571	44.118	82.289	129.341	155.906
14.247	44.431	86.200	129.348	157.726
19.960	49.499	94.160	132.148	161.645
24.132	52.481	104.536	132.605	164.702
26.048	65.996	106.341	140.558	172.240
27.826	71.727	110.737	145.806	178.003
.003				1.44332
2.288	.028	3.810	1.686	.077
3.050	3.514	.897	1.647	1.690
6.840	4.077	.959	.949	1.561
1.390	7.540	2.290	1.112	5.846
.676	.313	3.912	.007	1.820
5.713	5.068	7.959	2.799	3.919
4.172	2.982	10.377	.458	3.057
1.915	13.515	1.805	7.953	7.538
1.778	5.731	4.395	5.248	5.763
1.133	2.605	13.209	.925	2.000
180	50	3.6	2.72466	
.007	1.112	1.915	3.810	5.763
.028	1.133	2.000	3.912	5.846
.077	1.390	2.288	3.919	6.840
.313	1.561	2.290	4.077	7.538
.458	1.647	2.605	4.172	7.540
.676	1.686	2.799	4.395	7.953
.897	1.690	2.982	5.068	7.959
.925	1.778	3.050	5.248	10.376
.949	1.805	3.057	5.713	13.209
.959	1.820	3.514	5.731	13.515

QUANTUM AND OPTICAL ARITHMETIC AND FRACTALS

M.V. BERRY
H. H. Wills Physics Laboratory,
Tyndall Avenue, Bristol BS8 1TL, UK

Three waves depend on a common mathematical structure. The waves are light beyond a diffraction grating with sharp-edged slits, initially transversely uniform microwaves propagating along a strip guide, and the evolving probability amplitude for a quantum particle in a one-dimensional box where the initial state is a constant. The mathematical structure is the indefinite integral over ξ of

$$K(\xi,\tau)=\sum_{n=-\infty}^{\infty}\exp\left\{\mathrm{i}\pi\left[2\xi\left(n+\frac{1}{2}\right)-\tau\left(n+\frac{1}{2}\right)^2\right]\right\}$$

which is a theta function (Gauss sum) on its natural boundary. For rational τ, the integral is piecewise constant and describes fractional quantum revivals and the fractional Talbot effect. For irrational τ, the graph of the wave intensity as a function of ξ is a fractal with dimension 3/2. As a function of τ, the graph has dimension 7/4. On some diagonal lines (space-time sections in quantum mechanics), e.g. $\xi=(1-\tau)/2$, the dimension is 5/4.

1 Introduction

Claude Itzykson excelled in finding physical applications for sophisticated concepts and techniques from number theory. I think he would have enjoyed the physics described here, in which arithmetic and fractal geometry appear in a way that was surprising at first and then gave new insights into the simplest quantum time-dependence and the distribution of light in a waveguide and beyond a diffraction grating. I will give only a brief description; details and generalizations can be found in two recent papers[1,2].

My aim is to describe geometrical structures in the sum

$$\Psi(\xi,\tau)=\frac{2}{\pi}\sum_{n=0}^{\infty}\frac{(-1)^n}{(n+\frac{1}{2})}\cos\left\{2\pi\xi\left(n+\frac{1}{2}\right)\right\}\exp\left\{-\mathrm{i}\pi\tau\left(n+\frac{1}{2}\right)^2\right\} \quad (1)$$

In a first interpretation, this satisfies the time-dependent Schrödinger equation

$$\mathrm{i}\partial_\tau\Psi(\xi,\tau)=-\frac{1}{4\pi}\partial_\xi^2\Psi(\xi,\tau) \quad (2)$$

with the boundary condition

QUANTUM AND OPTICAL ARITHMETIC AND FRACTALS

M.V. BERRY
H. H. Wills Physics Laboratory,
Tyndall Avenue, Bristol BS8 1TL, UK

Three waves depend on a common mathematical structure. The waves are light beyond a diffraction grating with sharp-edged slits, initially transversely uniform microwaves propagating along a strip guide, and the evolving probability amplitude for a quantum particle in a one-dimensional box where the initial state is a constant. The mathematical structure is the indefinite integral over ξ of

$$K(\xi,\tau)=\sum_{n=-\infty}^{\infty}\exp\left\{\mathrm{i}\pi\left[2\xi\left(n+\frac{1}{2}\right)-\tau\left(n+\frac{1}{2}\right)^2\right]\right\}$$

which is a theta function (Gauss sum) on its natural boundary. For rational τ, the integral is piecewise constant and describes fractional quantum revivals and the fractional Talbot effect. For irrational τ, the graph of the wave intensity as a function of ξ is a fractal with dimension 3/2. As a function of τ, the graph has dimension 7/4. On some diagonal lines (space-time sections in quantum mechanics), e.g. $\xi=(1-\tau)/2$, the dimension is 5/4.

1 Introduction

Claude Itzykson excelled in finding physical applications for sophisticated concepts and techniques from number theory. I think he would have enjoyed the physics described here, in which arithmetic and fractal geometry appear in a way that was surprising at first and then gave new insights into the simplest quantum time-dependence and the distribution of light in a waveguide and beyond a diffraction grating. I will give only a brief description; details and generalizations can be found in two recent papers[1,2].

My aim is to describe geometrical structures in the sum

$$\Psi(\xi,\tau)=\frac{2}{\pi}\sum_{n=0}^{\infty}\frac{(-1)^n}{\left(n+\frac{1}{2}\right)}\cos\left\{2\pi\xi\left(n+\frac{1}{2}\right)\right\}\exp\left\{-\mathrm{i}\pi\tau\left(n+\frac{1}{2}\right)^2\right\} \quad (1)$$

In a first interpretation, this satisfies the time-dependent Schrödinger equation

$$\mathrm{i}\partial_\tau\Psi(\xi,\tau)=-\frac{1}{4\pi}\partial_\xi^2\Psi(\xi,\tau) \quad (2)$$

with the boundary condition

$$\Psi\left(\pm\frac{1}{2},\tau\right)=0 \tag{3}$$

and the initial condition

$$\Psi(\xi,0)=\frac{2}{\pi}\sum_{n=0}^{\infty}\frac{(-1)^n}{\left(n+\frac{1}{2}\right)}\cos\left\{2\pi\xi\left(n+\frac{1}{2}\right)\right\}=1\quad\left(|\xi|<\frac{1}{2}\right) \tag{4}$$

It follows that Ψ describes what is perhaps the simplest nonstationary quantum bound state, where an initially spatially constant wavefunction evolves inside a unit box (infinite potential well). Taken together, (3) and (4) imply that the initial state is discontinuous at the walls $\xi=\pm1/2$, a fact that will have interesting consequences.

A different interpretation of (2) is as the paraxial wave equation in two-dimensions, with τ as a longitudinal spatial variable and ξ the transverse variable. This gives solutions to the Helmholtz equation with wavenumber k and wavelength $\lambda=2\pi/k$, namely

$$\left(\partial_x^2+\partial_z^2+k^2\right)\phi(x,z)=0 \tag{5}$$

with the variables related by

$$\begin{gathered}x=a\xi,\quad z=z_{\mathrm{T}}\tau,\quad \phi(x,z)\approx\exp(\mathrm{i}kz)\,\Psi(\xi,\tau)\\ \text{where } z_{\mathrm{T}}\equiv\tfrac{a^2}{\lambda}\text{ and } a\text{ is a length}\end{gathered} \tag{6}$$

Here, the approximation is paraxial, that is it describes waves travelling close to the z axis, in a sense that will be made more precise later. With this reinterpretation, (1) describes the wave propagating inside a waveguide in the form of a unit strip with Dirichlet boundary conditions at the edges $x=\pm a/2$, where the wave is spatially uniform as it enters the guide.

A further interpretation is provided by the wave

$$\Psi_1(\xi,\tau)\equiv\frac{1}{2}\left[1+\Psi(2\xi,4\tau)\right] \tag{7}$$

which also satisfies (2) and with (5) and (6) can be interpreted as a paraxial wave in the half-plane $(0\le x\le\infty)$, $(-\infty<z<\infty)$. This is periodic in ξ with period 1, and satisfies the condition

$$\Psi_1(\xi,0) = 1 \left(\text{if } |\xi \bmod 1| < \frac{1}{4}\right) \qquad (8)$$
$$= 0 \left(\text{if } |\xi \bmod 1| > \frac{1}{4}\right)$$

It follows that Ψ_1 describes the paraxial propagation of a plane wave of light, initially travelling in the ζ direction, after passage at $\zeta = 0$ through a diffraction grating with period $\Delta\xi = 1$, that is $\Delta x = a$, consisting of opaque and transparent strips with equal widths; this is called a Ronchi grating.

In what follows, repeated use will be made of the following obvious symmetries of the wave (1):

$$\Psi(\xi,-\tau) = \Psi^*(\xi,\tau);\ \Psi(\xi,\tau+1) = \exp\{-i\pi/4\}\,\Psi(\xi,\tau);$$
$$\text{i.e. } \Psi(\xi,1-\tau) = \exp\{-i\pi/4\}\,\Psi^*(\xi,\tau);$$
$$\Psi(-\xi,\tau) = \Psi(\xi,\tau);\ \Psi(\xi+1,\tau) = -\Psi(\xi,\tau); \qquad (9)$$
$$\text{i.e. } \Psi(1-\xi,\tau) = -\Psi(\xi,\tau)$$

2 Revivals and Talbot images

It follows from (9) that at integer times τ the probability density

$$P(\xi,\tau) \equiv |\Psi(\xi,\tau)|^2 \qquad (10)$$

repeatedly reconstructs its initially constant form. This is the simplest example[3] of the much more general phenomenon of *quantum revivals*[4,5], in which a wide class of initial quantum states (for example, representing electrons in atoms) get approximately reconstructed at integer multiples of a time that depends on the spectrum and the form of the initial state.

For the diffraction grating wave (7), the analogous statement is

$$\Psi_1(\xi,\tau+1) = \frac{1}{2}\left[1 - \Psi(2\xi,4\tau)\right] = \Psi_1\left(\xi+\frac{1}{2},\tau\right) \qquad (11)$$

Therefore at integer multiples p of the distance z_T in (6), the grating profile (8) is reconstructed by the diffracted light, with a half-period shift if p is odd. This repeated self-imaging is the *Talbot effect*[6-8], and z_T is the Talbot distance. The Talbot effect is also a more general phenomenon that is embodied in (7),

because (paraxially) perfect imaging occurs not only for the Ronchi grating (8) but for any profile.

Now consider the quantum wave Ψ at rational times $\tau = p/q$, where p and q are mutually prime integers. It is helpful to write this as the integral over the propagator K for the particle in a box. Thus

$$\begin{aligned} \Psi(\xi,\tau) &= \int_{-1/2}^{1/2} d\xi_0 K(\xi-\xi_0,\tau), \\ \text{where } K(\xi,\tau) &= \sum_{n=-\infty}^{\infty} \exp\left\{ i\pi \left[2\xi\left(n+\frac{1}{2}\right) - \tau\left(n+\frac{1}{2}\right)^2 \right] \right\} \end{aligned} \quad (12)$$

Now set $\tau = p/q$, and split the sum into groups of q terms by defining[9]

$$n = lq + s \quad (-\infty < l < \infty, 1 \le s \le q) \quad (13)$$

The crucial observation is that the exponential involving l^2 can be simplified, because

$$\exp\left\{-i\pi pql^2\right\} = \exp\left\{-i\pi q e_p l\right\}$$

$$\text{where } e_p = 1\,(p \text{ even}) \text{ or } 0\,(p \text{ odd}) \quad (14)$$

This enables the sum over l to be evaluated as a series of δ functions, to give, after some reduction,

$$K\left(\xi,\frac{p}{q}\right) = \frac{1}{\sqrt{q}} \sum_{n=-\infty}^{\infty} A_n(q,p)\,\delta\left(\xi - \frac{1}{2}\left(e_p + \frac{p}{q}\right) - \frac{n}{q}\right) \text{ where} \quad (15)$$

$$A_n(q,p) = \frac{1}{\sqrt{q}} \exp\left\{ i\pi\left(\frac{p}{4q} + \frac{e_p}{2} + \frac{n}{q}\right)\right\} \sum_{s=1}^{q} \exp\left\{ i\frac{\pi}{q}\left((2n+qe_p)\,s - ps^2\right)\right\}$$

When combined with the integral (12), this result shows that the wave for these rational times is the (piecewise constant) superposition of q shifted overlapping copies (labelled by n) of the initial wave in the box (set equal to zero outside). These are *fractional quantum revivals*[3,4]. The nature of the superposition is determined by the A_n, which are easily shown to be pure phase factors, that is

$$A_n\,(q,p) = \exp\{\mathrm{i}\Phi_n\,(q,p)\} \tag{16}$$

The phases Φ_n can be evaluated by recognizing the formula for A_n in (15) as a variant of the Gauss sum of number theory; explicit formulae can be found elsewhere[1,2,9].

Figure 1 shows fractional revivals of the probability density for increasing q, as $\tau = p/q$ takes values given by successive approximations to the golden mean. These were computed as the sum of q shifted copies with the phases Φ_n. Confirmation of the correctness of the analysis was obtained by computing Ψ for the same τ with the eigenfunction series (1), which gave exactly the same pictures (apart from Gibbs oscillations smoothing the discontinuities, caused by truncating the series).

The optical analogues of the fractional revivals are the *fractional Talbot images*[10, 11] at distances $z = z_{\mathrm{T}}\ p/q$ from the grating. In each unit cell (e.g. $0 \leq x \leq a$), these images are superpositions, with phases Φ_n, of q copies of the grating, with amplitudes reduced by $1/\sqrt{q}$. The above-mentioned explicit formulae for the phases considerably facilitate the calculation of these images.

3 Fractal waves

Rational values of τ are special. For typical (i.e. irrational) τ, and, more generally, as a function of the variables ξ and τ, the Schrödinger wave Ψ defined by (1) - and also the diffraction grating wave Ψ_1 defined by (7) - possesses rich fractal properties. These are conveniently described by the fractal dimensions of the graphs of the probability density (10) along lines in spacetime, that is, in the ξ, τ plane.

In calculating the various fractal dimensions[12], I will apply to $\Psi(\xi,\tau)$ a result for Fourier series

$$f\,(u) = \sum_m a_m \exp\{\mathrm{i}mu\} \tag{17}$$

where the a_m have random or pseudorandom phases. If the power spectrum $|a_m|^2$ has the asymptotic form

$$|a_m|^2 \ \sim |m|^{-\beta}\ as\ |m| \to \infty, \text{where } 1 < \beta \leq 3, \tag{18}$$

then the graphs of $\mathrm{Re}f$ and $\mathrm{Im}f$ are continuous but nondifferentiable, with fractal dimension

$$D_{\mathrm{f}} = \frac{1}{2}\,(5-\beta) \tag{19}$$

Thus $\beta = 3$ corresponds to a (just) differentiable curve with $D_f = 1$, and $\beta = 1$ would correspond to an area-filling curve with $D_f = 2$. Equation (17) can be obtained (by simple dimensional analysis) from the result[13] that if f has dimension D_f the mean square increment of $f(u)$ over a infinitesimal distance Δu is proportional to $(\Delta u)^{4-2D_f}$; for a straightforward derivation of this result, see[14]. This applies when D_f represents the capacity dimension, but in the present context I expect it to hold for the Hausdorff and other fractal dimensions as well[15]. Elsewhere[2] I have argued that the fractal dimension of the graph of $|f(u)|^2$ is, almost always, also D_f. Therefore the fractal properties of $\mathrm{Re}\,\Psi$ and $\mathrm{Im}\,\Psi$ are inherited by the probability density $P(\xi, \tau)$.

At irrational times τ, when quantum revivals do not occur, the quantum wave Ψ, regarded as a function of position ξ in the box, has the form (17), with Fourier components $m = (n+1/2)$ and pseudorandom phases $\pi\tau m^2$. The power spectrum is, from (1), proportional to m^{-2}, so that $\beta = 2$ in (18) and, from (19), the fractal dimension of the probability density is $D_\xi = 3/2$. The same argument gives the dimension of the graph of wave intensity across the strip waveguide, and of the light intensity beyond a Ronchi grating in almost all planes, where z/z_T is irrational and there are no fractional Talbot images.

Figure 1 shows how the spatial quantum fractal for $\tau = \tau_G = (3 - \sqrt{5})/2$ emerges as the limit of sequences of fractional revivals corresponding to the continued-fraction (Fibonacci) approximants to τ_G. Figure 2 shows the spatial fractal for $\tau = 1/2\pi$ in greater detail, and a magnification illustrating the self-similarity. Pictures for other irrational τ, and for the Talbot image intensity $|\Psi_1|^2$, are similar.

Now consider Ψ as a function of time τ, at fixed position ξ. As has been explained, the probability density derived from the wave (1) is periodic in τ. Its Fourier series contains longitudinal frequencies restricted to the values $m = (n+1/2)^2$. For such a lacunary series, the power spectrum is $|a_n|^2 \mathrm{d}n/\mathrm{d}m$, which is proportional to $m^{-3/2}$. In the argument following (17) we now have $\beta = 3/2$, giving the unexpected fractal dimension $D_\tau = (5 - \beta)/2 = 7/4$. This is not restricted to rational ξ, and indeed we think the probability density is a fractal function of time everywhere, except at the walls $\xi = \pm 1/2$. Figure 3 shows one of these time fractals, and a magnification illustrating its self-similarity. The greater fractal dimension, reflected in the greater irregularity of these curves in comparison with those in figure 2, is clear.

The time and space fractals can be seen together in figure 3, which is a landscape plot of $P(\xi, \tau)$. Evidently this is not an amorphous fractal, but contains much additional structure. In particular, unanticipated diagonal lines can be discerned, corresponding to particular spacetime slices through the P landscape (one is visible issuing from the rear right of the picture; with

other views of the landscape[2], more are visible). As I have described in detail elsewhere[2], these form part of an infinite set of lines, namely

$$\xi = m\tau + n + \frac{1}{2}\ (m \neq 0)\,,\quad \xi = \left(m + \frac{1}{2}\right)\tau + \frac{1}{2}n \quad (m, n \text{ integer}) \qquad (20)$$

on which partial destructive interference between the terms in (1) reduces the fractal dimension to $D_{\text{diag}} = 5/4$. One of these spacetime fractals is shown in figure 5; it is noticeably less irregular than either of the previous fractals, reflecting its smaller dimension. The lines (20) also appear as 'canals' in the analogous computations for Gaussian wavepackets[3].

Further analysis of the sum (1) may reveal more fractal treasures, for example spacetime lines (not necessarily straight) on which more complicated interference between groups of terms leads to fractal probability densities with dimensions different from D_τ, D_ξ, or D_{diag}.

4 Discussion.

In the wave behind a diffraction grating, several of the phenomena so far discussed, namely the Talbot images for rational z/z_T, the transverse fractals with dimension 3/2, for irrational z/z_T, and the 'longitudinal' fractals with dimension 7/4, as z varies for fixed x, have been observed in a recent experiment[1]. At first it seems surprising that it is possible to see the transverse fractal dimension 3/2, because any misorientation of the plane of observation, such as must surely occur in reality, ought to give a light intensity pattern whose fractal dimension is that of a generic section, namely the larger value 7/4. That the transverse fractal can in fact be seen is the result of a curious combination of two circumstances, and could hold a more general message for the interpretation of experiments involving fractals.

The first is that the Talbot fractals possess infinitely fine detail only in the limit of perfect paraxiality, unlike the quantum fractals that are exact solutions of Schrödinger's equation. A detailed analysis[1] shows that this is the limit $\lambda/a \to 0$. Finite values of λ/a give rise to postparaxial blurring of the transverse and longitudinal detail in the fractals (and the edges of the rational Talbot images) according to the same function that gives diffraction smoothing of the cusp caustic of geometrical optics. (This echo of geometrical asymptotics is curious - albeit mathematically unsurprising - in view of the fact that the sharp detail that is being postparaxially blurred is itself the result of (paraxial) wave interference).

The second circumstance is that the natural 'unit cell' of the Talbot effect, namely $\Delta_x = a$, $\Delta z = z_T$, is enormously elongated relative to the dimensionless unit cell $\Delta\xi = 1$, $\Delta\tau = 1$ (by a factor a/λ, which in the experiment reported in [1] was 803). This effect causes any misorientation of the observation plane in x, z space to be greatly reduced in ξ, τ space, often to such a degree that it falls within the postparaxial τ blurring. So, paradoxically, the transverse fractal is rendered observable because fine detail in the longitudinal fractal is obscured. The same argument suggests -again paradoxically- that the elongation of the natural cell should make it more difficult to see the dimensionally dominant longitudinal fractal; nevertheless, observation was possible, by careful alignment exploiting the symmetry of the Talbot images about $\xi = 0$.

The Talbot fractals are not restricted to Ronchi gratings, but will occur, with the same dimensions $D_\xi = 3/2$ and $D\tau = 7/4$, whenever the transmission function has discontinuities in its amplitude and phase.

For the quantum fractals, the situation described here can be generalized enormously[2], to quantum waves evolving in arbitrary D-dimensional enclosures with $D-1$ dimensional boundaries, from arbitrary initial states with discontinuities. It would seem that the chaology of geodesics (trivially integrable in a one-dimensional box) might dominate the fractal geometry of the evolution, since fine detail depends on high-lying eigenstates, that is on semiclassical asymptotics. But in fact the results have a wider universality: whatever the chaology, the graph of the probability density as a function of time at a fixed position is a fractal curve with dimension 7/4, and the graph as a function of position for fixed time is a $D + 1/2$ dimensional fractal hypersurface. These results do not apply if the boundary of the enclosure is itself a fractal, with dimension $D - 1 + \gamma$ (where $0 < \gamma < 1$); then, a physical argument suggests that the time fractal dimension is $(7 + \gamma)/4$, and the space fractal dimension is $D + (1 + \gamma)/2$. Further generalizations can easily be envisaged.

The simple wave $\Psi(\xi, \tau)$ (equation 1) is the indefinite integral (cf. 12) of a Jacobi theta function (infinite Gauss sum)[16], on the natural boundary of its domain of convergence: if τ had a small negative imaginary part, the sum in (12) would converge. Although the theta function does not converge, the singular behaviour on its natural boundary is reflected in the rich structure of fractals and quantum revivals in its integral Ψ, which does converge. The theta functions themselves would arise in the evolution of quantum waves from a δ-function initial condition, or from a grating of infinitely narrow slits. In the optical case there are two principal effects that would make the analogue of the sum converge. First, there is non-paraxiality, mentioned already. Second, there is the effect of the finite number N of slits in any real grating, giving rise to a wave described by a finite Gauss sum. As $N \to \infty$, the wave in irrational

planes z/z_T gets infinitely complicated, in a way that can be fully described by a chaotic renormalization transformation[17].

Finally, it is worth remarking how different are the superficially analogous problems involving the heat equation, obtained by analytic continuation of τ to the negative imaginary axis. Then (1) describes the evolution of temperature in a bar where the initial temperature is constant and the ends are maintained at a different constant temperature. In this situation the sum (1), and the associated theta function, converge exponentially, and there are no revivals and no fractals.

References

1. Berry, M.V. *J. Modern Optics* in press, (1996).
2. Berry, M.V. *J. Phys. A* submitted, (1996).
3. Stifter, P., Leichtle, C., Lamb, W.E. & Schleich, W.P. *in preparation* (1996).
4. Averbukh, I.S. & Perelman, N.F. *Physics Letters A* **139**, 449-453 (1989).
5. Mallalieu, M. & Stroud, C.R.J. *Phys. Rev. A.* **51**, 1827-1835 (1995).
6. Talbot, H., F. *Phil. Mag.* **9**, 401-407 (1836).
7. Rayleigh, L. *Phil. Mag.* **11**, 196-205 (1881).
8. Patorski, K. *Progress in Optics* **27**, 1-108 (1989).
9. Hannay, J.H. & Berry, M.V. *Physica* **1D**, 267-290 (1980).
10. Hiedemann, E.A. & Breazale, M.A. *J. Opt. Soc. Amer.* **49**, 372-375 (1959).
11. Winthrop, J.T. & Worthington, C.R. *J. Opt. Soc. Amer.* **55**, 373-381 (1965).
12. Mandelbrot, B.B. *The Fractal Geometry of Nature* (Freeman, San Francisco, 1982).
13. Orey, S. Z. *Warsch. verw. Geb.* **15**, 249-256 (1970).
14. Berry, M.V. & Lewis, Z.V. *Proc.Roy.Soc.* **A370**, 459-484 (1980).
15. Falconer, K. *The geometry of fractal sets: mathematical foundations and applications* (Wiley, New York, 1990).
16. Gradshteyn, I.S. & Ryzhik, I.M. *Table of Integrals, Series and Products* (Academic Press, New York and London, 1980).
17. Berry, M.V. & Goldberg, J. *Nonlinearity* **1**, 1-26 (1988).

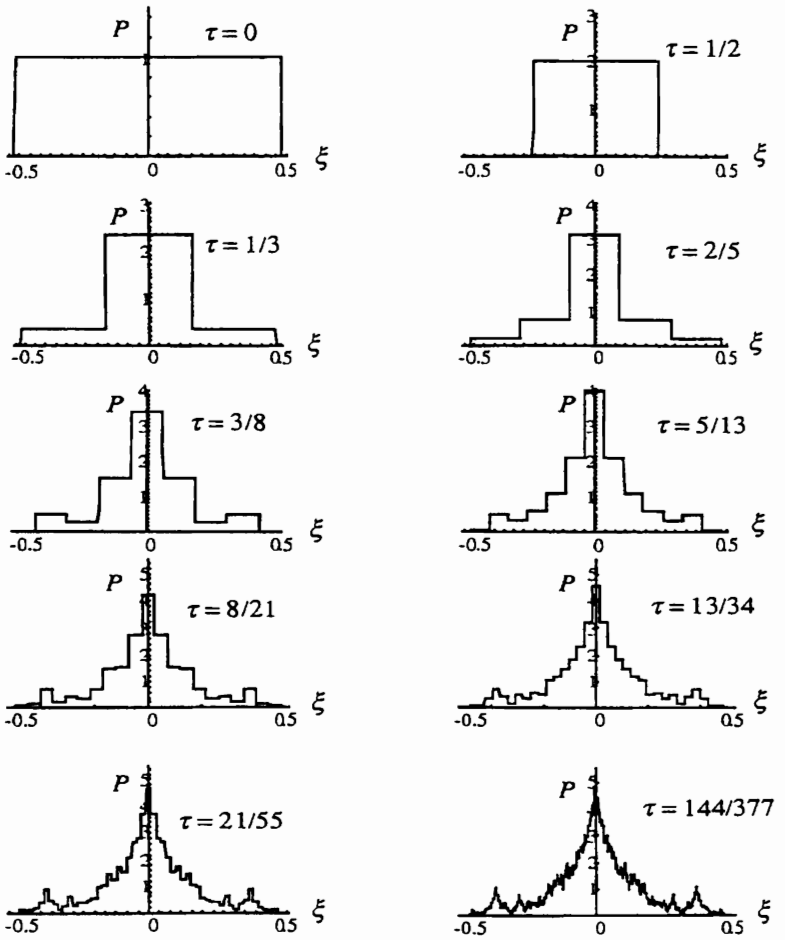


Figure 1. Probability density $P(\xi,\tau)=|\Psi(\xi,\tau)|^2$ for a particle in a box, at the indicated times τ, approximating the golden mean $\tau_G=(3-\sqrt{5})/2 = 0.381966$ ($\approx 144/377+3.15\times10^{-6}$), showing fractional quantum revivals accumulating to give the spatial fractal, with dimension $D_\xi=3/2$.

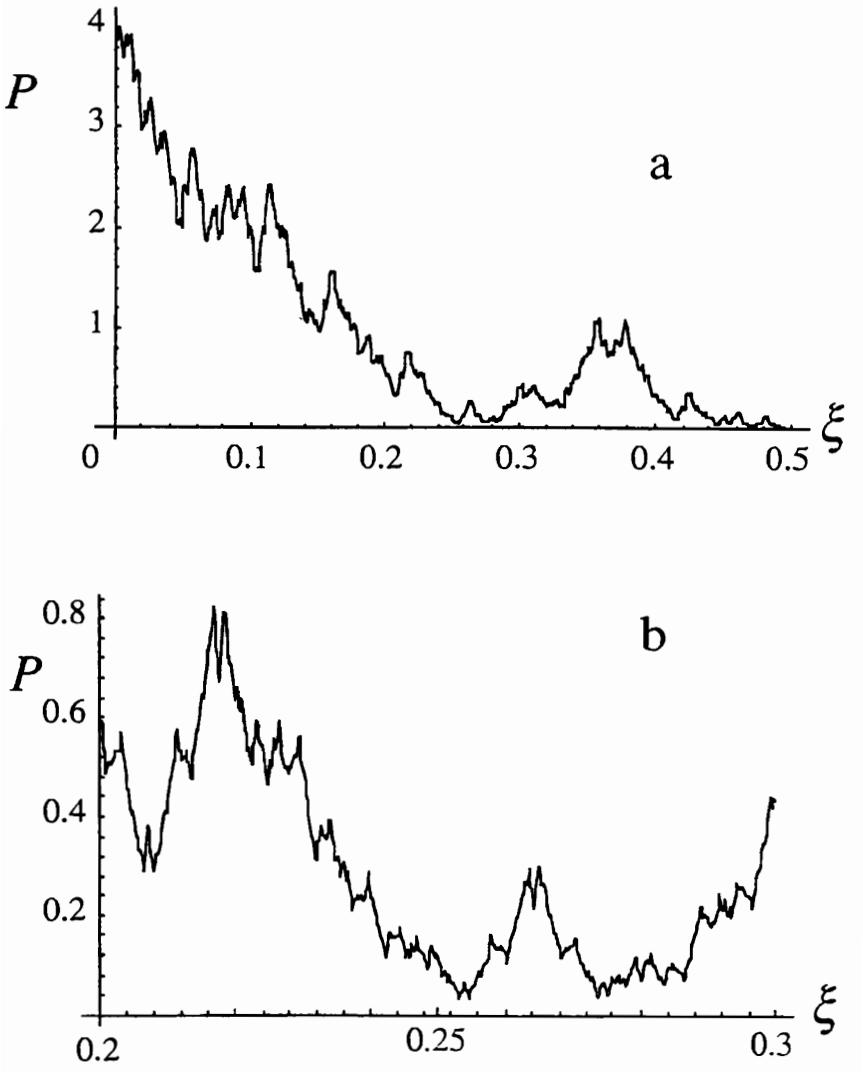


Figure 2. Quantum spatial fractal for $\tau=1/e$. (a): over the irreducible range $0\leq\xi\leq 1/2$, and (b): magnified to show the self-similarity.

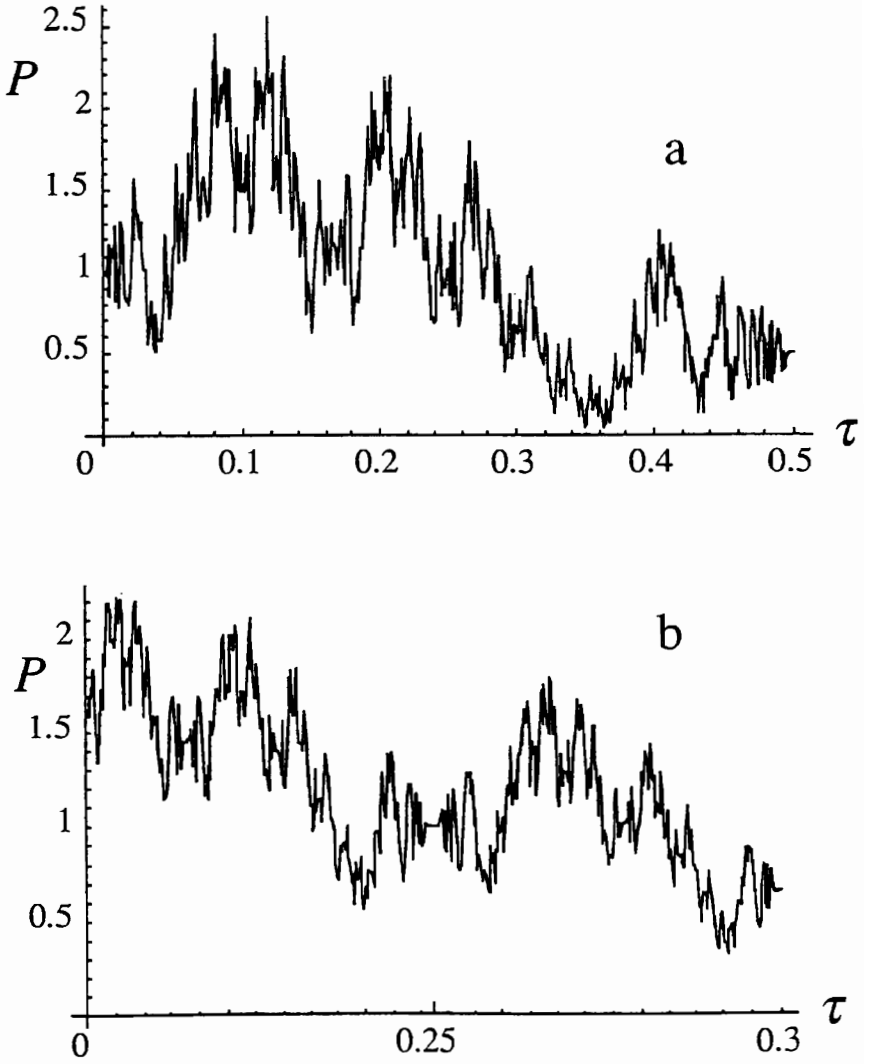


Figure 3. Quantum time fractal, with dimension $D_\tau = 7/4$, for $\xi = 0.25$; (a): over the irreducible range $0 \leq \tau \leq 1/2$, and (b): magnified to show the self-similarity.

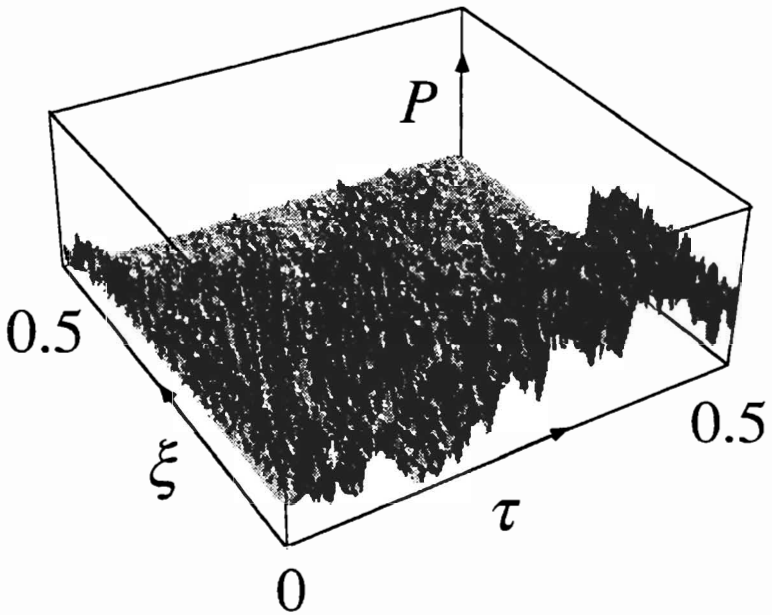


Figure 4. Illuminated landscape plot of the fractal probability density $P(\xi,\tau)$ in time and space for a particle in a box.

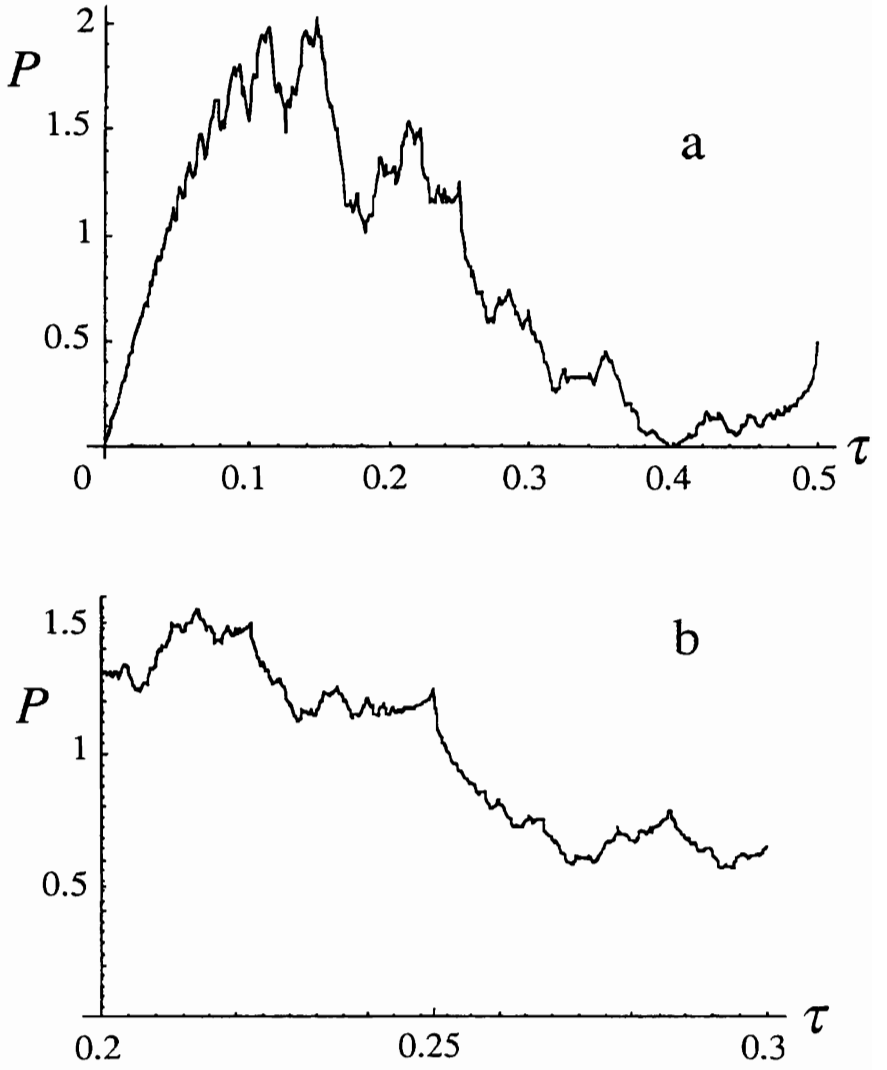


Figure 5. Quantum diagonal fractal, with dimension $D_{\text{diag}}=5/4$, for the spacetime slice $\xi=(5\tau+1)/2$; (a): over the range $0\leq\tau\leq 1/2$, and (b): magnified to show the self-similarity.

CORRELATIONS AND TRANSPORT IN ONE DIMENSIONAL QUANTUM IMPURITY PROBLEMS

F. LESAGE, H. SALEUR†
Department of Physics, University of Southern California
Los Angeles, CA 90089-0484

We review a set of exact results about correlations and transport in one dimensional quantum impurity problems that we have obtained in the last three years . These include the spin two-point function in the double well problem of dissipative quantum mechanics (or equivalently the anisotropic Kondo problem), the DC conductance and noise and the AC conductance for the tunneling between edges in the fractional quantum Hall effect. Few technical details are given; rather, we try to outline the principles of the methods, the nature of the results, and to explain what makes some questions more difficult than others.

1. Introduction.

One dimensional quantum impurity problems have been for many years a prominent subject in theoretical and mathematical physics. The reason is that these problems exhibit very non trivial physical features, have experimental applications, and present big technical challenges, whose solution however appears possible. Important theoretical steps in the study of these problems include the renormalization group analysis [1], [2], the Bethe ansatz solution [3], and, more recently, the application of conformal field theory [4].

In this paper, we review the results we have obtained on these problems in the last three years, in collaborations including P. Fendley, A. Ludwig, S. Skorik and N. Warner. Our main focus has been on the exact computation of correlation functions and related dynamical properties, all the way from small to large distances. The Bethe ansatz had led so far only to the computation of thermodynamical quantities [3]. The use of conformal field theory had led to interesting results but only in the vicinity of the small and large distance conformal invariant fixed points [4]. Even numerical studies had, until very recently, not been completely satisfactory (see however [5], [6], [7] for recent progress). On the other hand, correlation functions are the most interesting physical quantities: for instance, in the Kondo problem, the question of the screening cloud [8] is directly related with the spin one point function, in the double well problem of dissipative quantum mechanics [9] the transition from coherent to incoherent regime is signalled by the behaviour of the spin two

† Packard Fellow

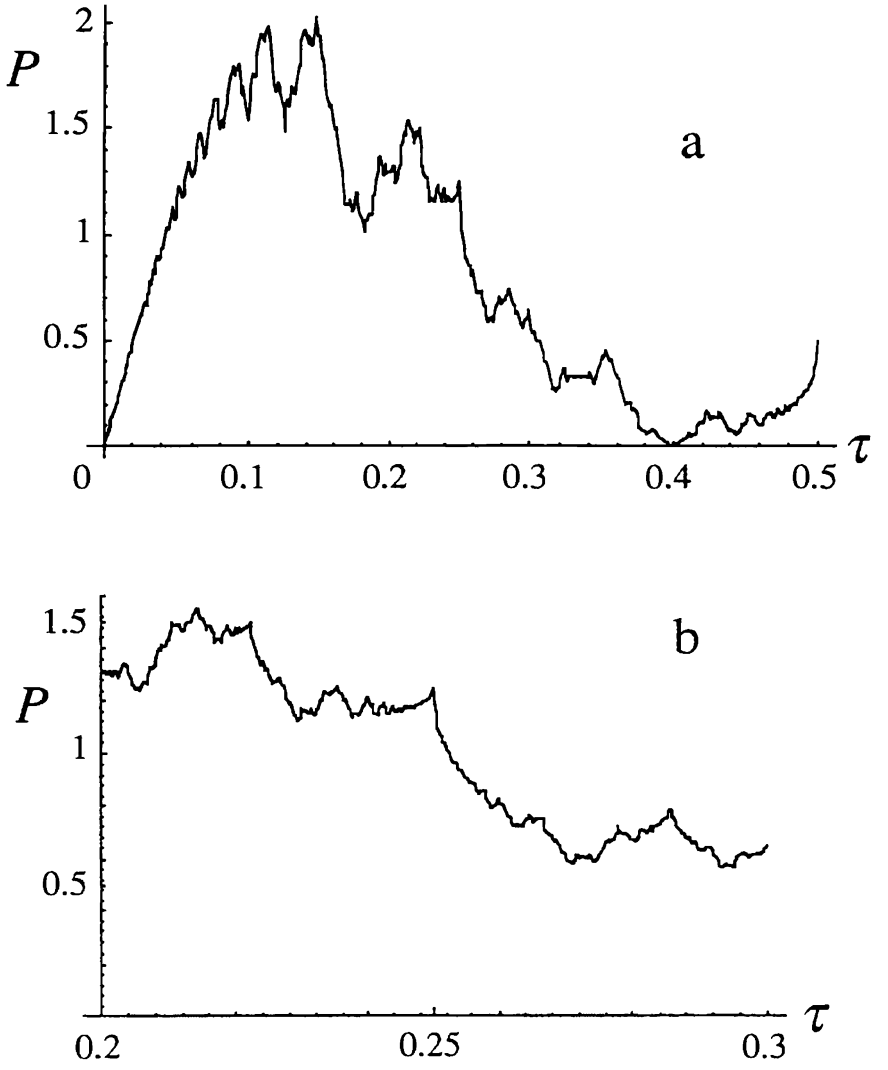


Figure 5. Quantum diagonal fractal, with dimension $D_{\text{diag}}=5/4$, for the spacetime slice $\xi=(5\tau+1)/2$; (a): over the range $0\leq\tau\leq 1/2$, and (b): magnified to show the self-similarity.

CORRELATIONS AND TRANSPORT IN ONE DIMENSIONAL QUANTUM IMPURITY PROBLEMS

F. LESAGE, H. SALEUR†
Department of Physics, University of Southern California
Los Angeles, CA 90089-0484

We review a set of exact results about correlations and transport in one dimensional quantum impurity problems that we have obtained in the last three years . These include the spin two-point function in the double well problem of dissipative quantum mechanics (or equivalently the anisotropic Kondo problem), the DC conductance and noise and the AC conductance for the tunneling between edges in the fractional quantum Hall effect. Few technical details are given; rather, we try to outline the principles of the methods, the nature of the results, and to explain what makes some questions more difficult than others.

1. Introduction.

One dimensional quantum impurity problems have been for many years a prominent subject in theoretical and mathematical physics. The reason is that these problems exhibit very non trivial physical features, have experimental applications, and present big technical challenges, whose solution however appears possible. Important theoretical steps in the study of these problems include the renormalization group analysis [1], [2], the Bethe ansatz solution [3], and, more recently, the application of conformal field theory [4].

In this paper, we review the results we have obtained on these problems in the last three years, in collaborations including P. Fendley, A. Ludwig, S. Skorik and N. Warner. Our main focus has been on the exact computation of correlation functions and related dynamical properties, all the way from small to large distances. The Bethe ansatz had led so far only to the computation of thermodynamical quantities [3]. The use of conformal field theory had led to interesting results but only in the vicinity of the small and large distance conformal invariant fixed points [4]. Even numerical studies had, until very recently, not been completely satisfactory (see however [5], [6], [7] for recent progress). On the other hand, correlation functions are the most interesting physical quantities: for instance, in the Kondo problem, the question of the screening cloud [8] is directly related with the spin one point function, in the double well problem of dissipative quantum mechanics [9] the transition from coherent to incoherent regime is signalled by the behaviour of the spin two

† Packard Fellow

point function, and, in the most recent quantum impurity problem - the edge tunneling in the fractional quantum Hall effect - the only quantities of interest are transport properties [10].

We have been concerned with three physical problems: the (anisotropic) Kondo problem, dissipative quantum mechanics, and edge tunneling in the fractional quantum Hall effect. These problems are all based on a hamiltonian of the form :

$$H = \frac{1}{2}\int_{-\infty}^{0} dx\, [8\pi g \Pi^2 + \frac{1}{8\pi g}(\partial_x \phi)^2] + \mathcal{B}. \tag{1.1}$$

The model is defined on the negative axis and has interaction at $x = 0$. We will use the euclidian coordinates $z = x + iy$ and $\bar{z} = x - iy$ with $y = -it$ (Fig. 1).

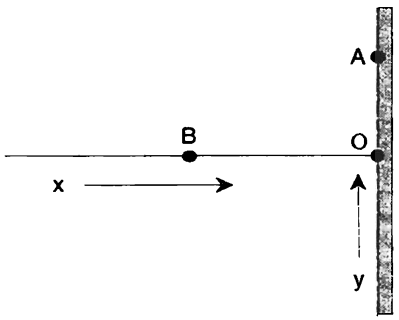


Fig. 1: Geometry of the problem.

Here the boundary term $\mathcal{B}$ describes the interaction of the fields with a boundary degree of freedom, which we chose to be a spin :

$$\mathcal{B} = \lambda \left[\sigma_+ e^{i\phi(0)/2} + \sigma_- e^{-i\phi(0)/2}\right], \tag{1.2}$$

where $\sigma_\pm$ are taken in a representation of $su(2)_q$ with $q = e^{i\pi g}$:

$$[\sigma_z, \sigma_\pm] = \pm 2\sigma_\pm, \quad [\sigma_+, \sigma_-] = \frac{q^{\sigma_z} - q^{-\sigma_z}}{q - q^{-1}}. \tag{1.3}$$

At first sight, this choice of boundary term might seem unnatural, but it reduces to well known cases with different choices of representations. For example, the spin 1/2 representation of $su(2)_q$ is isomorphic to that of $su(2)$ and the previous hamiltonian therefore describes the usual anisotropic Kondo model. Higher spin have little physical interest, but

their formal consideration can be quite useful. When g is rational, and a cyclic representation is chosen, it is possible to map the system on the massless boundary sine-Gordon model [11] with :

$$\mathcal{B} = 2\lambda \cos \phi(0)/2. \tag{1.4}$$

Instead of a free boson, one could consider a more complicated theory in the bulk. Such problems can also be of physical interest - for instance a theory with central charge $c = \frac{3}{2}$ would correspond to the two channel Kondo problem - but we will not discuss this here.

Under a renormalisation group transformation, the coupling constant flows according to :

$$\frac{d\lambda(l)}{dl} = (1-g)\lambda(l), \tag{1.5}$$

which shows that for $g < 1$ the perturbation is relevant, while it is irrelevant for $g > 1$. In the following we restrict ourselves to $0 < g < 1$. In the simplest case of a spin $1/2$, the system flows from a free spin in the UV to a screened spin in the IR. For the boundary sine-Gordon model, the flow is from Neumann boundary condition on the field ϕ in the UV to Dirichlet boundary conditions in the IR. The problem is to compute the properties, including the correlation functions, for all couplings λ.

Before launching into technical details, we would like to stress that the hamiltonian (1.1) has also a fundamental interest [12]. It leads to a quantum field theoretic version of the "quantum monodromy operator", well known in the theory of the Yang Baxter equation, and is therefore deeply related to the integrable structure of conformal field theories. To understand that, we first observe that in the massless case, one can easily "unfold" the system. Instead of having a boundary where right movers are transformed into left movers, one can think of the system as having no boundary, but instead an impurity at $x = 0$, through which say left movers scatter. Then, consider the partition function Z_j of this problem on a cylinder of circumference $\beta = 1/T$. The hamiltonian (1.1) (or its unfolded version) corresponds to a propagation in the y direction. We can alternatively compute the same partition function within a modular transformed point of view, by considering propagation in the x direction instead. Then, the hamiltonian is simply the free boson one defined on a circle, and the term $\mathcal{B}$ in (1.1) becomes a simple impurity at $x = 0$. In other words, the partition function can be written

$$Z_j = < 0|\mathrm{tr}\ e^{i\pi P_L \sigma_z} L_j(\lambda)|0 >, \tag{1.6}$$

where

$$L_j(\lambda) = \Pi_j \left\{ e^{i\pi P_L \sigma_z} \mathcal{P} \exp \left[q^{-1/2} \lambda \int_0^{1/T} dy \left(e^{-2i\phi_L(y)} q^{\sigma_z/2} \sigma_+ + e^{2i\phi_L(y)} q^{-\sigma_z/2} \sigma_- \right) \right] \right\}, \tag{1.7}$$

where ϕ_L is the left component of the field ϕ, P is the momentum operator in the canonical decomposition of ϕ. Π_j indicates that the spin operators are taken in the spin-$j/2$ representation, $\mathcal{P}$ indicates path ordering, and the exponentials are normal-ordered.

The operator $L_j(\lambda)$ acts both on the degrees of freedom of the external spin and on the degrees of freedom of the free boson. If one thinks of the former as "horizontal" and the latter as "vertical", it is clear that L_j is a continuum limit of the usual monodromy operator in the XXZ chain [13]. By working out this analogy a little more closely [12], [14], one shows in particular that it solves the Yang Baxter equation

$$R^{j,j'}(\lambda,\mu) L_j(\lambda) L_{j'}(\mu) = L_{j'}(\mu) L_j(\lambda) R^{j,j'}(\lambda,\mu), \tag{1.8}$$

where $R^{jj'}$ is the usual R-matrix solution of the Yang Baxter equation, acting in the tensor product of spin j and spin j' representations [13]. In particular, we see that the coupling λ is now interpreted as a spectral parameter. It is well known how the monodromy operator in the XXZ case is related with the conserved quantities and the whole integrable structure. Similarly, L_j here is deeply related with the integrable structure of the conformal field theory, so quantum impurity problems provide a "probe" of this structure.

The determination of correlation functions in these problems had been outstanding for many years. Our (partial) success comes from the combination of several techniques and ideas that were evolved only recently, within the spectacular developments in the field of integrable systems (for progress on correlation functions in other problems, and using different ideas, see [15]).

A standard approach to the problem is perturbation theory [2]. Within an imaginary time formalism, it is easily checked that the partition function is the same as that of a Coulomb gas with positive and negative charges interacting on a circle. The charges can take any relative positions for the boundary sine- Gordon model, but they have to be alternating for the spin 1/2 Kondo problem. While this reformulation was well known, it involves very impressive Coulomb integrals that were usually considered as untractable. Thanks to recent developments in the theory of symmetric polynomials [16], we have actually been able [17], [11] to obtain these integrals in "closed form" - more precisely in

the form of infinite series of rational functions of gamma functions. This allows extremely quick numerical evaluations of the thermodynamical quantities for any g and λ. More interestingly, this allows a control on the analytical behaviour of these quantities as a function of g - even a formal continuation of quantitites beyond $g = 1$, which is of key importance in the study of "duality". Finally, although we do not completely understand why, these series have natural generalizations that readily provide some of the dynamical properties, for instance the voltage dependent conductance in the tunneling problem.

The other approach is based on integrability. That the hamiltonian (1.2) is integrable has of course been known for a long time ((1.4) became of interest more recently [18]). So far however, the integrability had been used mostly to compute thermodynamic properties. To address correlation functions and transport properties , an easy but important step is to realize that integrability allows to describe the system in a new basis (plane waves behave in a very complicated way at the boundary). This basis is made of quasi particle states, the quasi particles being massless solitons/anti-solitons [1] and breathers, that scatter non trivially [19]. This scattering is conveniently encoded in the Fadeev-Zamolodchikov relations [20] :

$$\begin{aligned} Z^{\epsilon_1}(\theta_1)Z^{\epsilon_2}(\theta_2) &= S^{\epsilon_1\epsilon_2}_{\epsilon'_1\epsilon'_2}(\theta_1-\theta_2)Z^{\epsilon'_2}(\theta_2)Z^{\epsilon'_1}(\theta_1) \\ Z^*_{\epsilon_1}(\theta_1)Z^*_{\epsilon_2}(\theta_2) &= S^{\epsilon'_1\epsilon'_2}_{\epsilon_1\epsilon_2}(\theta_1-\theta_2)Z^*_{\epsilon'_2}(\theta_2)Z^*_{\epsilon'_1}(\theta_1) \\ Z^{\epsilon_1}(\theta_1)Z^*_{\epsilon_2}(\theta_2) &= S^{\epsilon'_2\epsilon_1}_{\epsilon_2\epsilon'_1}(\theta_1-\theta_2)Z^*_{\epsilon'_2}Z^{\epsilon'_1}(\theta_1) + 2\pi\delta^{\epsilon_1}_{\epsilon_2}\delta(\theta_1-\theta_2), \end{aligned} \tag{1.9}$$

where S is a solution of the Yang-Baxter equation, θ is the rapidity parametrizing energy and momentum (see below) and Z^ϵ (resp. Z^*_ϵ) the annihilation (resp. creation) operator of a particle of type ϵ . In this basis the boundary interaction is then simply given by a reflection matrix $R^{\epsilon'}_{\epsilon}$. This R matrix has in turn to solve the boundary Yang-Baxter equation.

This description could have been used many years ago [19]. One has however to overcome a slight conceptual difficulty: while L and R moving massless particles will scatter trivially because the theory is massless , the L particles (and similarly the R) will have a non trivial scattering. Since two such particles are both moving in the same direction at the speed of light, their "scattering" is difficult to imagine physically. In fact, the S matrix has to be interpreted as giving the monodromy properties of the wave function, and an

[1] We sometimes designate these by kink and antikink.

approach based on massless particles would not make much sense if the theory was not integrable.

In a first series of papers, we have generalized the Landauer-Büttiker transport theory [21] (that deals usually with free massless electrons) to massless particles with factorized scattering This allowed us to determine exactly DC transport properties in the tunneling problem: the DC conductance [22], [23], the DC noise [24], [25], both out of equilibrium (with a voltage).

The massless basis can also be used to compute time and space dependent correlators (this was actually done first, in the slightly different context of flows between bulk theories, in [26]). Indeed, the matrix elements of physical operators in this basis can also be determined using the tools of integrable systems: in fact they follow easily from computations done previously in the massive case [27]. Because the theory is truly interacting, there are infinitely many such matrix elements to take into account: for instance, the current acting on the vacuum can create any neutral state, with an arbitrary number of massless particles. Fortunately, few of these matrix elements are enough to obtain the correlators with arbitrary accuracy all the way from small to large distances. In a second series of papers [28],[29], we have used this method to determine, for instance, the $T = 0$ spin correlator in the dissipative quantum mechanics problem, the $T = 0$ screening cloud in the anisotropic Kondo problem, the $T = 0$ AC conductance in the tunneling problem. All these quantities were determined without magnetic field (voltage). The consideration of the correlators at $T \neq 0$ and (or) with a magnetic field (a voltage) is more difficult. We have recently obtained results [30] for the AC noise at $T = 0$ out of equilibrium in the tunneling problem.

Before proceeding, we would like to comment on our use of the word "exact". We usually call exact result the expression of a physical quantity either in terms of a small number of known functions, or in terms of the solutions of a small number of integral or differential equations. In that respect, all the DC results obtained by the combination of TBA and the Landauer Büttiker approach are exact. We have also used the vocable exact for the AC properties. This is stretching its meaning a bit, since an exact expression for the correlators would involve an infinite number of terms. More precisely, our result is that we have an expression involving a small number of integrals, which provides arbitrary accuracy all the way from small to large distances. This is to be contrasted with perturbation theory where more and more terms are necessary to get accurate results at large coupling. For

this reason, the perturbative approach is not exact. However, we call exact the expressions for the Coulomb gas integrals in terms of series of gamma functions. In our opinion, they deserve this label because the sums can be very quickly evaluated to an arbitrary accuracy, and because these sums give access to non trivial information - eg the continuation beyong $g = 1/2$ or $g = 1$.

In the following, we set $e = \hbar = 1$.

2. The Physical problems.

2.1. $\nu = 1/3$ Hall effect.

Edge excitations are one of the most exciting aspects of the fractional quantum Hall effect. For the sake of the discussion, let us limit ourselves to the "simple" filling fractions $\nu = 1/t$ with t an odd integer. When we place ourselves *on* a plateau, the Hall law and current conservation hold classically. These classical equations can be derived from an action principle, thus leading to an effective action for the electromagnetic potential A :

$$S(A) = \frac{\nu}{2} \int_{\mathcal{J}} A \wedge dA, \tag{2.1}$$

with $\mathcal{J}$ the 2+1 space bounded by the sample geometry. The presence of finite geometry generates an anomaly under gauge transformations and one is forced to add a boundary term to cancel this anomaly[31], [32]. This is where the boundary currents come into the picture. They are described by a relativistic action, which in a bosonised form looks like :

$$S(\phi) = \frac{1}{8\pi\nu} \int_{-\infty}^{\infty} dx \, [(\partial_x \phi)^2 + (\partial_t \phi)^2]. \tag{2.2}$$

If we choose a geometry with currents emanating from reservoirs at the top and bottom of the sample (see Fig. 2) we have two such actions describing the right and left moving currents at the bottom and top respectively. In that picture the creation and destruction of Hall quasiparticles are described by operators of the form :

$$\mathcal{O}_m(x) = e^{im\phi(x)}, \tag{2.3}$$

for charges $Q_m = \nu m$, thus $m = 1$ correspond to the Laughlin quasi-particle. Now we can define the tunneling problem we wish to address [33]: Imagine a Hall system to which we

add an obstruction at $x = 0$. Then there is a possibility for transfer of charges by the quasiparticles at that point (Fig. 2), which is modelled by a term of the form :

$$\mathcal{B} = \sum_{m=1}^{\infty} \lambda_m \, e^{im[\phi_L(0) - \phi_R(0)]} + \text{c.c.} \tag{2.4}$$

here each term describes the transfer of a quasiparticle of charge νm from one edge to the other. It turns out that for the specific choice, $\nu = 1/3$, only the transfer of Laughlin quasi-particles of charge $Q = 1/3$ is a relevant perturbation [10].

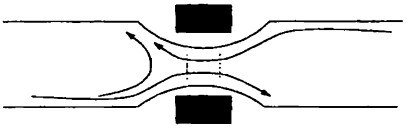


Fig. 2: Sketch of the tunneling experiment.

So the system we are trying to describe has the following hamiltonian :

$$H = \frac{1}{2} \int_{-\infty}^{\infty} dx \, [4\pi\nu\Pi^2 + \frac{1}{4\pi\nu}(\partial_x\varphi)^2] + \lambda\delta(x)\cos(\varphi_L - \varphi_R), \tag{2.5}$$

where the L and R components depend on x, t as $\varphi_L(x+t), \varphi_R(x-t)$. It was shown in [22] how by using even and odd field basis, this problem can be mapped on a problem on a half line with only the odd field interacting with the impurity. The odd hamiltonian then reads :

$$H = \frac{1}{2} \int_{-\infty}^{0} [8\pi g(\Pi^o)^2 + \frac{1}{8\pi g}(\partial_x\phi^o)^2] + \lambda\delta(x)\cos\frac{1}{2}\phi^o, \tag{2.6}$$

and in the following we will write $\phi \equiv \phi^o$ and g instead of ν. Thus, for this problem, $\mathcal{B} = \lambda\cos\frac{1}{2}\phi(x=0,t)$. The problem now is of the form (1.1). Without the impurity, the charges Q_L and Q_R were conserved individually on each edge. Now, transfer of charge is possible and $Q_R + Q_L$ is conserved but not $\Delta Q = Q_L - Q_R$. But these expressions are just the charges for the even and odd bosons used in the transformation (up to a factor of $\sqrt{2}$). Thus the even field, which has no interaction will not play any role in the following whereas the odd field contains all the information about the backscattering current created by the impurity: I_B.

Experimentally there will be resonances controlled by the gate voltage V_g at $x = 0$, where the conductance will be $G = 1/3$[2]. Then changing the gate voltage takes the

[2] We have put an additional factor 2π in our definition of the conductance so that it takes this simple form, even with the convention $\hbar = 1$.

system off resonance by an amount $V_g - V_g^* \propto \lambda$: this thus provides an experimental setting for (1.4). Of course, quantum field theory gives control of the universal results only. This means that one has to look for experimental values of the physical temperature and $V_g - V_g^*$ as small as possible, and to build universal curves as a function of scaled variables, for instance $(V_g - V_g^*)^{3/2}/T$ (for $g = 1/3$). Quantities of physical interest include the DC and AC conductance and noise.

2.2. The Kondo problem, and the double well problem.

We will assume that the reader is well aware of the physics of the (anisotropic) Kondo problem itself. We just recall that it is originally a three dimensional problem; however only s waves couple to the impurity, and this allows reduction to one dimension, the radial coordinate.

We will mostly discuss the dissipative quantum mechanics version of this problem: the double well system with Ohmic dissipation [9] [3]. This dissipation is created by the environment, and one is interested in dynamical properties of the double well. An archetype hamiltonian to describe this system is :

$$H = -\frac{\Delta}{2}\sigma_x + \frac{\epsilon}{2}\sigma_z + \sum_\alpha \left[\frac{p_\alpha^2}{2m_\alpha} + \frac{1}{2}m_\alpha\omega_\alpha^2 \left(x_\alpha - \frac{C_\alpha}{m_\alpha\omega_\alpha^2}\sigma_z \right)^2 \right]. \tag{2.7}$$

In this expression, the Pauli matrices σ_x, σ_z act on the two dimensional space of states. Δ is a tunnelling matrix element and ϵ denotes a bias between the two states. Dissipation comes from the coupling of this two states system to an environment of oscillators described by the last part of the hamiltonian. Each of these individual oscillators has a different mass, frequency and coupling to the two states. It was shown [9] that the system can be mapped, in the so called ohmic dissipation case, to an anisotropic Kondo model. In that case the oscillators are chosen so that :

$$J(\omega) = \frac{\pi}{2}\sum_\alpha \frac{C_\alpha^2}{m_\alpha\omega_\alpha}\delta(\omega - \omega_\alpha) = \eta\omega e^{-\omega/\omega_c}, \tag{2.8}$$

with ω_c a cut-off. The conduction electrons in the latter play the role of dissipation and the z value of the spin is associated with the two states. The bias, ϵ would correspond to a magnetic field in the Kondo problem. The parameter in the Kondo model is proportional

[3] The boundary sine-Gordon problem also describes a dissipative quantum mechanics problem, this time of a particle moving in a "washboard" potential [34].

to the strength of the dissipation $g \propto \eta$. The coupling to the impurity, is related to the tunelling term in (2.7), the precise relations given in [9].

The question of interest is the coherence between the two states as a function of the strenght of the dissipation, or as a function of g in (1.1). When $g > 1$ the system is localised in one of the wells. When $g < 1$, then if the state is initially in the "up" position, it will relax to an equal probability to be in either states as a function of time. There are two possibilities: it can either relax monotically, or decay in an oscillatory fashion. A mean to study the relaxation is to look at the following correlator :

$$C(t) = \frac{1}{2} < [S_z(t), S_z(0)] > . \tag{2.9}$$

It describes the probability to be in a state $S_z(t)$ given that the system was in state $S_z(0)$ at $t = 0$. By means of a simple transformation, this correlator can be related to a current correlation.

3. A perturbative approach.

All orders in perturbation theory can actually be computed in these problems, for several quantities, thanks to the theory of symmetric polynomials.

The partition function at finite temperature can be deduced from the form of the two point function :

$$\langle e^{i\phi(y)/2} e^{-i\phi(y')/2} \rangle = \left| \frac{\kappa}{\pi T} \sin[\pi T(y - y')] \right|^{-2g} , \tag{3.1}$$

where κ is a cut-off. Then at each order in perturbation theory, there is a contribution I_{2n} which is given by the Boltzmann weight of a 2D Coulomb gas on a circle with an equal number of positive and negative charges [17] :

$$I_{2n}(\{u_i\}, \{u_i'\}) = \int_0^{2\pi} \prod_i \frac{du_i}{2\pi} \frac{du_i'}{2\pi} \left| \frac{\prod_{i<j} 4 \sin \frac{u_{ij}}{2} \sin \frac{u'_{ij}}{2}}{\prod_{i,j} 2 \sin \frac{(u_i - u_j')}{2}} \right|^{2g} \tag{3.2}$$

In these notations the partition function is just given by :

$$\mathcal{Z}_{BSG} = 1 + \sum_{n=1}^{\infty} x^{2n} I_{2n}, \tag{3.3}$$

with $x = \lambda / T(2\pi T/\kappa)^g$ an effective coupling. Remarkably, while the integrals (3.2) had been outstanding for many years, they can in fact be computed after some few manipulations using Jack polynomials. One ends up with the result :

$$I_{2n} = \Gamma(g)^{-2n} \sum_{m} \prod_{i=1}^{n} \left[\frac{\Gamma(m_i + g(n - i + 1))}{\Gamma(m_i + g(n - i) + 1)} \right]^2 . \tag{3.4}$$

Here $m = (m_1, m_2, ..., m_n)$ is a partition with at most n rows.

A similar perturbative approach is also well known for the anisotropic Kondo model. The main difference is the presence of the $\sigma_\pm$ in the boundary term, which constrains the charges to be alternating on the circle [2]. The resulting integrals are then ordered and the method based on Jack polynomials fails. However, one can use the integrability of the problem by means of fusion relations. Using that the tensor product of a cyclic representation with a spin 1/2 is a sum of two cyclic representations, one finds the following relation between the partition functions of the Kondo and boundary sine-Gordon model [11]:

$$\mathcal{Z}_1[(q-q^{-1})x] = \frac{\mathcal{Z}_{BSG}(qx) + \mathcal{Z}_{BSG}(q^{-1}x)}{\mathcal{Z}_{BSG}(x)}, \tag{3.5}$$

with x the effective coupling constant. Thus at each order, the coefficient of the Kondo problem are determined by those of the boundary sine-Gordon model. Equivalently, ordered integrals can be expressed in terms of unordered ones, a result which is easy to understand at lower orders.

A pessimistic reader might object that (3.4) is not so exciting since after all what we have been doing is simply replace a multidimensional integral by a multidimensional sum. In fact, although the coefficients (3.4) have a rather bulky expression, they lead to very quick numerical evaluation (infinitely quicker than Monte Carlo computation of the integrals). Moreover, while the integrals are well defined only for $g < 1/2$, the expression (3.4) can be continued beyond $g = 1/2$, leading to a dimensionally regularized version of the theory, with no UV cut-off. In fact, the expressions (3.4) can even be continued beyond $g = 1$, providing a dimensionally regularized version of irrelevant perturbation theory. This is very useful to address the intriguing issue of duality.

Let us illustrate this in the simplest case [35]. Take the second cumulant of $\mathcal{F} = -T \log \mathcal{Z}_{BSG}$, then the sum can be done explicitely and we find :

$$f_2 = -\frac{\Gamma(1-2g)}{[\Gamma(1-g)]^2}. \tag{3.6}$$

This term is finite all the way to $g = 1$ apart from a pole at $g = 1/2$ and there is no branch point anywhere. Thus the analytic continuation is perfectly well defined. This generalises to the higher terms, one can check that in the term $x^{2n} f_{2n}$ there is a pole at $g = 1 - 1/(2n)$ with known residue. These poles result in logarithmic terms of the form $(-1)^n T_B/(2\pi n T) \log T_B/T$ in the free energy, where T_B is a boundary energy scale (see

next section). (3.6) vanishes at $g = 1$, and is well defined for $g > 1$ by analytic continuation of the Gamma function.

As we observed earlier, thermodynamic properties are not the most interesting ones. For the boundary sine-Gordon case, the most important quantity would be the non equilibrium DC conductance. To evalate it, a possible approach would be to handle the model out of equilibrium using the Keldysh formalism. We have not done so, but we have found a remarkable formula for this conductance (whose validity we checked against other approaches) that indicates deep simplifications. Introduce the quantity $Z_{BSG}(V)$ defined as in (3.3) but with V dependent integrals

$$I_{2n}(V) = \Gamma(g)^{-2n} \sum_m \prod_{i=1}^{n} \frac{\Gamma[m_i + g(n-i+1)]\Gamma[p + m_i + g(n-i+1)]}{\Gamma[m_i + 1 + g(n-i)]\Gamma[p + m_i + 1 + g(n-i)]}, \tag{3.7}$$

where $i\frac{Vg}{T} = 2\pi p$. Then we have found [17],[35] the following expression for the conductance

$$G(x, V) = g - \frac{ig\pi x}{2} \frac{\partial}{\partial(V/2T)} \frac{\partial}{\partial x} \ln\left(\frac{Z_{BSG}(V)}{Z_{BSG}(-V)}\right). \tag{3.8}$$

Hence, a bit mysteriously, equilibrium quantities lead very simply to non equilibrium ones. Since for the moment we do not completely understand why (3.8) holds, we need a more reliable method to evaluate this conductance as well as other quantities.

4. A new basis.

The hamiltonian (1.1) describes free bosons interacting with an impurity. While plane waves solve the bulk problem readily, their interaction with the impurity is very complicated. Instead, we will think of the free boson as a limit of the sine-Gordon theory :

$$H = \frac{1}{2} \int_{-\infty}^{0} dx \, [8\pi g \Pi^2 + \frac{1}{8\pi g} (\partial_x \phi)^2] + \Lambda \int_{-\infty}^{0} \cos\phi(x), \tag{4.1}$$

when the bulk coupling Λ (hence also the mass) goes to zero. The spectrum of the sine-Gordon model, which is a quantum integrable theory, is well known [20]. In the limit $\Lambda \to 0$, one simply obtains right and left moving massless solitons/anti-solitons and breathers. The energy and momentum of solitons for example, is described in terms of a rapidity by :

$$e = \pm p = \mu e^{\theta} \tag{4.2}$$

with μ an arbitrary energy scale. The scattering between left or right movers is the same as the massive scattering, ie it is given by Zamolodchikov's S-matrix for the sine-Gordon model. The scattering between left and right movers is a constant phase taken to be one in the following.

This somewhat artificial method to describe a free theory is very convenient in the presence of the boundary interaction. The latter translates simply into a reflection matrix which is a solution of the massless boundary Yang-Baxter equation, $R_c^{c'}(p/T_B)$. This matrix includes a scale, $T_B = \mu e^{\theta_B}$ describing the impurity strength $T_B \propto \lambda^{1/(1-g)}$. In that picture, $T_B = 0$ corresponds to the UV fixed point, and $T_B \to \infty$ to the IR fixed point. In this massless case, one can as well "unfold" the system, and instead of boundary scattering, deal with impurity scattering [36], [37]. In fact, the reflection matrix is closely related with higher spin solutions of the Yang Baxter equation of the type $S^{1/2,j}$, j a higher $su(2)_q$ spin.

5. DC transport properties

5.1. Linear conductance

Let us consider tunneling in the quantum Hall effect. Of physical interest is the linear conductance in the presence of the point contact. From a field theory point of view, one usually needs to compute the two point function of the currents in order to find the conductance. A standard way of representing it at zero temperature is through the Kubo formula :

$$G(\omega_M) = -\frac{1}{8\pi\omega_M L^2}\int_{-L}^{L} dx\, dx' \int_{-\infty}^{\infty} dy\; e^{i\omega_M y} < j(x,y)j(x',0) >, \tag{5.1}$$

where ω_M is a Matsubara frequency, y is imaginary time, $y = it$. One gets back to real physical frequencies by letting $\omega_M = -i\omega$. In (5.1), j is the physical current in the unfolded system, $j = \partial_t(\varphi_L - \varphi_R)$. Without impurity, the AC conductance of the Luttinger liquid is frequency independent, $G = g$. When adding the impurity, it becomes $G = \frac{g}{2} + \Delta G$. After some simple manipulations using the folding, one finds :

$$\begin{aligned}\Delta G(\omega_M) = \frac{1}{8\pi\omega_M L^2}\int_{-L}^{0} dx dx' \int_{-\infty}^{\infty} dy e^{i\omega_M y} \\ [< \partial_z\phi(x,y)\partial_{\bar{z}'}\phi(x',0) > + < \partial_{\bar{z}}\phi(x,y)\partial_{z'}\phi(x',0) >],\end{aligned} \tag{5.2}$$

where $z = x + iy$. Then the DC conductance is found by taking the limit $\omega \to 0$. A finite temperature computation for the DC conductance can be done along similar lines by working on a cylinder.

This previous formulation using correlation functions will be useful in the next subsection when computing finite frequency properties, but there is a much simpler method to compute the DC conductance in the finite temperature case: it will be computed from a rate, or Boltzmann, equation.

The starting point is the reflection matrix for the solitons and anti-solitons. Scattering of a single kink by the point contact is described by a one-particle R matrix with elements $R_+^+(p/T_B) = R_-^-(p/T_B)$ for kink $\to$ kink, and antikink $\to$ antikink, as well as $R_+^-(p/T_B) = R_-^+(p/T_B)$ for kink $\to$ antikink, and vice versa. These were derived exactly in [18] :

$$\begin{aligned} R_+^+(p/T_B) &= \frac{(p/T_B)^{(1/g)-1}}{1+i(p/T_B)^{(1/g)-1}} \exp[i\alpha_g(p/T_B)] \\ R_+^-(p/T_B) &= \frac{1}{1+i(p/T_B)^{(1/g)-1}} \exp[i\alpha_g(p/T_B)], \end{aligned} \tag{5.3}$$

where α_g is a phase. If we go back to the original theory, there are charges on each edge with a possibility of transfer at $x = 0$. The description used here is in term of a "folded" theory. The total charge of the system, $Q_L + Q_R$ is conserved but difference $\Delta Q = Q_R - Q_L$ is not conserved because of the boundary. Every time a soliton is reflected into an anti-soliton, the corresponding physical process is the transfer of charge at the point contact, or tunneling. Without tunneling, the (dimensionless) conductivity is $G = 1/3$, but in the presence of charge transfer, there is going to be a tunneling current described by :

$$I_B = \partial_t(\frac{1}{2}\Delta Q) = \partial_t(\frac{1}{\sqrt{2}}Q^o), \tag{5.4}$$

where Q^o denotes the odd field charge. The presence of a voltage provides a chemical potential difference for solitons and anti-solitons. A positive voltage implies that there will be more solitons, and when scattering on the impurity, there will be more solitons turning into anti-solitons than the inverse process. Since the scattering is elastic, it is possible to describe the charge transport at the impurity in terms of the probabilities of finding solitons and anti-solitons at momentum p at the point contact and the transition probability $|R_-^+|^2$. Let us now define by $n_V(p)$ the allowed orbitals at momentum p or the density of states and by $f_\pm(p)$ the occupation number for the solitons and anti-solitons at that same momentum. Then since the solitons have charge +1 and the anti-solitons -1 in

our normalisations, we have that :

$$\frac{\langle \Delta Q \rangle_V}{2L} = \int_0^\infty dp\ n_V(p)\ [f_+(p,V) - f_-(p,V)], \tag{5.5}$$

where L is the length of the system. The backscattering current then follows from a rate (Boltzmann) equation. The number of kinks of momentum p which scatter into antikinks per unit time is given by $|R_+^-|^2 n_V f_+[1-f_-]$; the factor $[1-f_-]f_+$ accounts for the probabilities of the initial state being filled and the final state being open. The rate at which antikinks scatter to kinks is likewise proportional to $[1-f_+]f_-$, so the charge changes at a rate proportional to $[1-f_-]f_+ - [1-f_+]f_- = f_+ - f_-$. Using (5.4) and (5.5) we have :

$$I_B(V) = -\int_0^\infty dp n_V(p)\ v_F |R_+^-(p/T_B)|^2 [f_+(p,V) - f_-(p,V)]. \tag{5.6}$$

we obtain the desired backscattering contribution to the linear-response conductance:

$$G_B = \lim_{V\to 0} \frac{1}{V} I_B(V) = -2\int_0^\infty dp\ n_0(p) |S_{+-}(p/T_B)|^2 \partial_V f_+(p,V)\big|_{V=0}. \tag{5.7}$$

The density of states $n_V(p)$ has been evaluated at $V=0$ because $f_+(p,V) - f_-(p,V)$ is already proportional to V. The total conductance is thus $G = \nu + G_B$.

The only thing left is the computation of the state densities and occupation numbers. This is the standard thermodynamic Bethe ansatz (TBA) [38]: in this picture, away from the impurity we have a "gas" of massless solitons/anti-solitons and breathers, with Yang-Baxter scattering, and the requirement of a periodic boundary condition results in an equation which relates the densities n_j and f_j of all these quasiparticles :

$$n_j(p) = 1 + \frac{1}{2\pi p} \sum_k \int_0^\infty dp' \Phi_{jk}(p/p') n_k(p') f_k(p'), \tag{5.8}$$

where $\Phi_{jk}(p) = -i(d/d\ln p) \ln S_{jk}^{bulk}(p)$ (the impurity would change these equations by terms of order $1/L$ which are negligible for the conductance). For $\nu = 1/3$ there is only one breather (b) and we have :

$$\Phi_{bb}(x) = 2\Phi_{++}(x) = 2\Phi_{+-}(x) = -\frac{4x}{x^2+1}$$

$$\Phi_{b+}(x) = \Phi_{+b}(1/x) = -\frac{4x^3+8x}{x^4+4},$$

where the others follow from the symmetry $+ \leftrightarrow -$.

One defines an auxiliary pseudoenergy variable ϵ_j to parametrize f_j via $f_j \equiv 1/(1 + \exp(\epsilon_j - \mu_j/T))$, where the μ_j are the chemical potentials: $\mu_+ = -\mu_- = V/2$; $\mu_b = 0$. By demanding that the free energy at temperature T (expressible in terms of f_j and n_j) be minimized, we find an equation for ϵ_j in terms of the (known) bulk S matrix elements:

$$\epsilon_j(p/T, V/T) = \frac{p}{T} - \sum_k \int_0^\infty \frac{dp'}{2\pi p'} \Phi_{jk}(p/p') \ln[1 + e^{\mu_k/T} e^{-\epsilon_k(p'/T, V/T)}]. \tag{5.9}$$

Solving this equation for ϵ_j gives the functions f_j. Even though the breather does not appear in (5.7), it interacts with the kink and antikink and affects the calculation of $f_\pm$. We can now evaluate the conductance explicitly, After a few manipulations one finds :

$$\begin{aligned} G &= -\int_0^\infty dp \frac{p^4}{p^4 + T_B^4} \frac{\partial \epsilon_+(p/T, 0)}{\partial p} \frac{\partial f(p)}{\partial \epsilon_+} \\ &= \int_0^\infty dx\, x^3 \frac{4(T_B/T)^4}{(x^4 + (T_B/T)^4)^2} \frac{1}{1 + e^{\epsilon_+(x,0)}}, \end{aligned} \tag{5.10}$$

with the variable $x = p/T$. The resulting conductance is compared on Fig. 3 to the Monte Carlo computations of [10] and to the experiment where this measurement was done [39]. One can see that the curve agrees reasonably well with the experiment and very well with the numerical results.

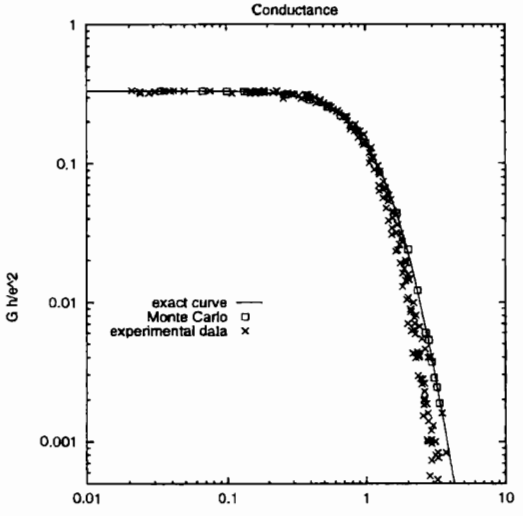


Fig. 3: Conductance for $g = 1/3$.

In the limit $T_B/T \to \infty$, $G \propto (T/T_B)^4$ as expected from [33]. It should be stressed that the universal form of the curve is heavily dependent on the Luttinger liquid behaviour of the system, and therefore, it is a signature of the Laughlin quasi-particle transfer between the edges.

Observe that to obtain G we need to solve for the ϵ_j's. While this is rather easy for $g = 1/t$, the TBA system becomes considerably involved for g a rational number, and depends on the decomposition of g in continued fractions. The perturbative formula (3.8) is easier to use in that case. Even if the problem is of little interest in the context of the quantum Hall effect at g generic, (5.7) also describes the mobility for a particle in the washboard potential, where g can be arbitrary [34].

It is instructive to compare our result with the one for free fermions. Formula (5.7) is actually well known and derives from the Landauer Büttiker approach to transport. The main difference with free fermions is in the filling fractions: the pseudo-energy ϵ is not the bare energy, but a complicated function, solution of integral equations.

The previous computation can easily be extended out of equilibrium, ie in the presence of a finite Hall voltage [23].

5.2. DC noise

This is not to say that the difference with free fermions is only in the filling fractions. More differences appear when we look at more complicated quantities. For instance one can consider the DC noise. It will involve not only the shot noise at the impurity, but the populations fluctuations in the bulk. These can be obtained by looking around the TBA saddle point, and take a rather complicated form. As an example, at zero temperature, the following relation holds [24]:

$$\langle I^2 \rangle = \frac{g}{2(1-g)} V^2 \partial_V \frac{I}{V} = \frac{g}{2(1-g)} (VG - I). \tag{5.11}$$

More complicated formula are available for $T \neq 0$ [25].

Contrary to the current, the noise provides a measure of the charge of the carriers. One checks from (5.11) that it is indeed Laughlin quasi-particles which tunnel at weak backscattering, and electrons at strong backscattering.

5.3. Duality?

The zero temperature limit becomes much simpler because the integral equations are linear. This lead to an explicit solution for the conductivity in that limit [23]. Using this

solution one can prove an exact duality between the IR and the UV limit, if we express the conductivity as :

$$G = g + \sum_{n=1}^{\infty} G_{2n}(g) \left(\frac{T_B}{T}\right)^{2n(1-g)} \tag{5.12}$$

in the UV and as :

$$G = \sum_{n=1}^{\infty} K_{2n}(g) \left(\frac{T}{T'_B}\right)^{2n(1/g-1)}, \tag{5.13}$$

in the IR (with T'_B proportional to T_B), it is possible to show that $K_{2n}(g) = G_{2n}(1/g)$. Physically, this is a duality between Laughlin quasi-particles UV and electrons in the IR.

There are strong indications that such a duality extends to finite temperature (for instance, by using the perturbative expressions of section 3 continued beyond $g > 1$, one can check it at the first few orders), but no proof yet. The standard arguments [40] rely on an instanton expansion, where fluctuations are neglected. The existence of the duality means that this instanton expansion has to be in fact exact, for some yet unknown reason.

6. Dynamical Properties.

The difference with free fermions becomes even more dramatic when one considers time dependent properties. For instance, as already mentioned in the introduction, the current operator acting on the ground state can, for generic g, create any state that is neutral, ie made of an arbitrary number of breathers, and soliton/antisoliton pairs. Clearly, the resulting expressions will bear very little resemblance with the ones of the Landauer Büttiker transport theory (the DC case was more favorable because, for massless relativistic particles, DC properties are the same as global, space integrated properties, and the x-integral of the current is the charge, which behaves in much the same way for free fermions and for our solitons/antisolitons). There is no choice then but to evaluate the correlators in (5.2) by: (i) inserting a complete set of states,

$$1 = \sum_{n=0}^{\infty} \sum_{\epsilon_i} \int \frac{d\beta_1 ... d\beta_n}{(2\pi)^n n!} |\beta_1, ..., \beta_n >_{\epsilon_1,...,\epsilon_n}{}^{\epsilon_n,...,\epsilon_1} < \beta_n, ..., \beta_1|, \tag{6.1}$$

(ii) computing each matrix element - the so called form-factors, (iii) performing the corresponding infinite sum [41],[42].

The form-factors can easily be derived from previous work on the massive sine-Gordon model [27] (we recall they follow from the solution of a set of axioms dependent on the

S-matrix. See also [41]). The infinite sum would be a priori a terrible obstacle. However, for g not too close to 1, it turns out that this sum converges extremely fast, and only a few form-factors (ie a few intermediate states in (6.1)) are necessary to obtain the current-correlator with arbitrary precision, all the way from the UV to the IR. We note that a similar observation was made for massive theories in [43], and for flows between bulk theories in [26]. In the present case, one can quantify this convergence rather easily [29]: unfortunately, one finds that more and more terms will be needed as $g \to 1$.

We now illustrate the method in the case of the double well problem of dissipative quantum mechanics. The bulk theory is the same as before and the main difference with the tunneling problem described previously lies in the reflection matrices. In this case the boundary preserve charge and the reflection matrices are given by :

$$R_{\pm}^{\pm}(\theta) = \tanh(\frac{\beta}{2} - i\frac{\pi}{4}). \tag{6.2}$$

Another difference is that we are interested in the behaviour of the spin at the boundary as a function of time. This is a priori puzzling, because the massless description is based on the IR fixed point where the spin is screened. So it does not appear at all in the integrable description. But when looking at the perturbative expansion in imaginary time for the spin-spin correlation one can find the following relation :

$$\langle \sigma^z(y)\sigma^z(0)\rangle - 1 = \langle I(y)I(0)\rangle_{T_B} - \langle I(y)I(0)\rangle_0, \tag{6.3}$$

with the subscript in the correlators denoting the boundary interaction in which they are taken and I given by :

$$I(y) = \int_{-\infty}^{0} dx\ \partial_x \phi(x, y). \tag{6.4}$$

Thus, the problem of computing spin correlations is reduced to a problem of current correlators where we can use the exact same technique as we did in the previous section (apart from the new reflection matrices). As we have mentionned in the introduction, we are interested in the coherence of the dissipative system which is probed by the spin-spin function or its Fourier transform :

$$\chi''(\omega) = \frac{1}{4\pi} \int dt\ e^{i\omega t} [\sigma^z(t), \sigma^z(0)]. \tag{6.5}$$

Let us first give the result for $g = 1/2$ which is straightforward since only the two particle

form factor is non zero. One finds :

$$\chi''(\omega) = \frac{4}{\pi^2}\frac{T_B}{\omega}\mathrm{Im}\frac{1}{\omega + 2iT_B}\ln\left(\frac{\omega + iT_B}{iT_B}\right)$$
$$= \frac{1}{\pi^2}\frac{4T_B^2}{\omega^2 + 4T_B^2}\left[\frac{1}{\omega}\ln\left(\frac{T_B^2 + \omega^2}{T_B^2}\right) + \frac{1}{T_B}\tan^{-1}\frac{\omega}{T_B}\right]. \tag{6.6}$$

We observe that this is strictly decreasing in ω and therefore has no maximum away from $\omega = 0$.

For other values of g, all expressions are rather complicated and can be found in [29]. Here, we just give some curves for the standard quantity $S(\omega) = \chi''(\omega)/\omega$ at different values of g. Physically [9] at $g = 1$ one expects a peak centered at $\omega = 0$ describing a localized system. At the opposite point, $g = 0$, the so called classical limit, there are two peaks away from zero describing the oscillations. We found that the peaks $g = 1/3$. In Fig. 4 we show the spectral function $S(\omega)$ for different values of g as computed from form factors. These curves have been confirmed numerically in [7].

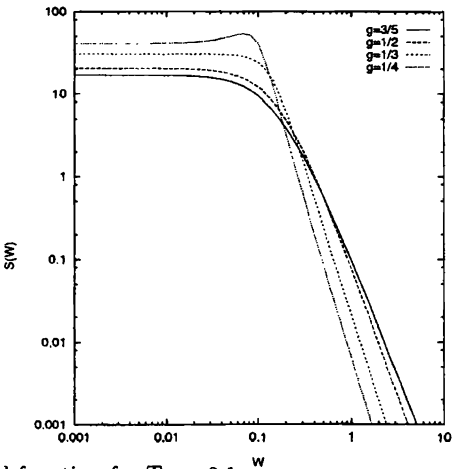


Fig. 4: Spectral function for $T_B = 0.1$.

Similar results are available for the conductance $G(\omega)$ at $T = 0$ and $V = 0$.

7. Conclusions.

We have used the form-factors technique for two other problems of interest: the equivalent of the Friedel oscillations caused by an impurity in a Luttinger liquid [44] at

$T = 0$ (or, in close relation, the screening cloud problem in the anisotropic Kondo problem [8]) [45], and the AC noise at $T = 0$ in the presence of a voltage [46],[30].

One might have a slight discomfort in using a Landauer Büttiker type of approach with integrable quasi particles. Ideally, we would like to recover the DC results only via solid quantum field theory methods and the Keldysh formalism. We hope this is possible, but we have not yet done so. As a step in that direction, we would like to mention that the formula for the linear conductance obtained in section 5 through the Landauer Büttiker approach can actually be recovered using the Kubo formula and form-factors.

At the present time, what is lacking to have a complete set of results is a way to compute dynamical properties at $T \neq 0$. To solve this problem, one has presumably to understand more deeply integrable quantum field theories at finite temperature. Some preliminary results in this direction have been obtained in [47].

Acknowledgements.

We thank P. Fendley, A. Ludwig, S. Skorik and N. P. Warner for continuous collaborations on various aspects of this work. We benefitted from discussions with many people, including I. Affleck, C. de C. Chamon, R. Egger, M. P. Fisher, D. Freed, A. Leclair, C. Mak, G. Mussardo, D. Serban, S. Zamolodchikov. This work was supported by the Packard Foundation, the National Young Investigator Program and the DOE. F. Lesage was also partly supported by a Canadian NSERC postdoctoral Fellowship.

Nous voudrions dédier les travaux résumés ici à la mémoire de Claude Itzykson. Nous remercions le CEA, le SPhT, et tout particulièrement Jean Michel Drouffe et Jean Bernard Zuber, d'avoir organisé cette conférence.

References

[1] J. Kondo, Prog. Th. Phys. 32 (1964) 37.

[2] P.W. Anderson, G. Yuval and D.R. Hamman, Phys. Rev. B1 (1970) 4464.

[3] N. Andrei, K. Furuya and J. Lowenstein, Rev. Mod. Phys. 55 (1983) 331; A.M. Tsvelick, P.B. Wiegmann, Adv. Phys. 32 (1983) 453.

[4] I. Affleck and A.W.W. Ludwig, Nucl. Phys. B360 (1991) 641.

[5] K. Leung, R. Egger and C. Mak, Phys. Rev. Lett. 75 (1995) 3344, cond-mat/9509078.

[6] S. Chakravarty and J. Rudnick, Phys. Rev. Lett. 75 (1995) 501 .

[7] T. A. Costi, C. Kieffer, Phys. Rev. Lett. 76 (1996) 1683.

[8] I. Affleck, E. Sorensen, cond-mat/9508030 and cond-mat/9511031 to appear in Phys. Rev. B.

[9] A.J. Leggett, S. Chakravarty, A.T. Dorsey, M.P.A. Fisher, A. Garg and W. Zwerger, Rev. Mod. Phys. 59 (1987) 1.

[10] K. Moon, H. Yi, C.L. Kane, S.M. Girvin and M.P.A. Fisher, Phys. Rev. Lett. 71 (1993) 4391, cond-mat/9408068.

[11] P. Fendley and H. Saleur, Phys. Rev. Lett. 75 (1995) 4492, cond-mat/9506104.

[12] V. Bazhanov, S. Lukyanov and A.B. Zamolodchikov, Comm. Math. Phys. 11 (1996) 381, hep-th/9412229; V. Bazhanov, S. Lukyanov and A.B. Zamolodchikov, "Integrable Structure of conformal Field Theory II", hep-th/9604044.

[13] L. Faddeev, Sov. Sci. Rev. Math. C1, 107 (1980), and references therein

[14] H. Saleur, unpublished.

[15] V. E. Korepin, N.M Bogoliubov, A.G. Izergin, "Quantum Inverse Scattering Method and Correlation Functions", Cambridge University Press (1993) and references therein.

[16] I.G. Macdonald, Seminaire Lotharingien, Publ. I.R.M.A., Strasbourg 1988; R.P. Stanley, Adv. in Math. 77 (1989) 76; K.W.J. Kadell, Compos. Math. 87 (1993) 5.

[17] P. Fendley, F. Lesage and H. Saleur, J. Stat. Phys. 79 (1995) 799, hep-th/9409176.

[18] S. Ghoshal, A. Zamolodchikov, Int. J. Phys. A9, (1994) 3841.

[19] L.D. Faddeev and L.A. Takhtajan, Phys. Lett. A85 (1981) 375; A.B. Zamolodchikov and Al.B. Zamolodchikov, Nucl. Phys. B379 (1992) 602; P. Fendley, H. Saleur and Al.B. Zamolodchikov, Int. J. Mod. Phys. A8 (1993) 5751; N.Yu. Reshetikhin and H. Saleur, Nucl. Phys. B419 (1994) 507.

[20] A.B. Zamolodchikov and Al.B. Zamolodchikov, Ann. Phys. 120 (1979) 253.

[21] R. Landauer, Physica D38 (1989) 594; G.B. Lesovik, JETP Lett. 49 (1989) 594; M. Büttiker, Phys. Rev. Lett. 65 (1990) 2901.

[22] P. Fendley, A.W.W. Ludwig and H. Saleur, Phys. Rev. Lett. 74 (1995) 3005, cond-mat/9408068

[23] P. Fendley, A.W.W. Ludwig and H. Saleur, Phys. Rev. B52 (1995) 8934, cond-mat/9503172.

[24] P. Fendley, A. Ludwig and H. Saleur, Phys. Rev. Lett. 75 (1995) 2196, cond-mat/9505031.

[25] P. Fendley, H. Saleur, to appear in Phys. Rev. B., cond-mat/9601117

[26] G. Delfino, G. Mussardo, P. Simonetti, Phys. Rev. D51 (1995) 6620.

[27] F. Smirnov, "Form factors in completely integrable models of quantum field theory", World Scientific and references therein.

[28] F. Lesage, H. Saleur, S. Skorik, Phys. Rev. Lett. 76, (1996) 3388, cond-mat/9512087

[29] F. Lesage, H. Saleur, S. Skorik, to appear in Nucl. Phys. B, cond-mat/9603043.

[30] F. Lesage, H. Saleur, to appear

[31] J. Fröhlich, T. Kerler, Nucl. Phys. B42 (1990) 8133.

[32] X.G. Wen, Int. J. Mod. Phys. B6 (1992) 1711.

[33] C.L. Kane and M.P.A. Fisher, Phys. Rev. B46 (1992) 15233.

[34] A. Schmid, Phys. Rev. Lett. 51 (1983) 1506; M.P.A. Fisher and W. Zwerger, Phys. Rev. B32 (1985) 6190; F. Guinea, V. Hakim and A. Muramatsu, Phys. Rev. Lett. 54 (1985) 263.

[35] P. Fendley, F. Lesage, H. Saleur, to appear in J. Stat. Phys. , cond-mat/9510055.

[36] P. Fendley, Phys. Rev. Lett. 71 (1993) 2485, cond-mat/9304031.

[37] P. Fendley, H. Saleur, N. Warner, Nucl. Phys. B 430 (1994) 577, hep-th/9406104.

[38] C.N. Yang and C.P. Yang, J. Math. Phys. 10 (1969) 1115.

[39] F.P. Milliken, C.P. Umbach and R.A. Webb, "Indications of a Luttinger Liquid in the Fractional Quantum Hall Regime", to appear in Solid State Communications.

[40] A. Schmid, Phys. Rev. Lett. 51 (1983) 1506.

[41] G. Mussardo, "Spectral representation of correlation functions in two-dimensional quantum field theories", Talk Int. Coll. QFT II, Tata insitute.

[42] A.Fring, G. Mussardo, P. Simonetti, Nucl. Phys. **B**393 (1993) 413; G. Mussardo, A. Koubek, Phys. Lett. B.311 (1993), 193; G. Mussardo, A. Koubek, P. Simonetti, Int. J. Mod. Phys. A 9 (1994), 3307.

[43] J. Cardy, G. Mussardo, Nucl. Phys. B 410 (1993), 451.

[44] R. Egger, H. Grabert, Phys. Rev. Lett. 75 (1995) 3505 .

[45] F. Lesage, H. Saleur, to appear

[46] C. de C. Chamon, D.E. Freed and X.G. Wen, Phys. Rev. B51 (1995) 2363, cond-mat/9408064.

[47] A. Leclair, F. Lesage, S. Sachdev, H. Saleur, to appear in Nucl. Phy. B., cond-mat/9606104.

LYAPUNOV EXPONENTS AND HODGE THEORY

M. KONTSEVICH
I.H.E.S., Bures-sur-Yvette, France

Claude Itzykson was fascinated (among other things) by the mathematics of *integrable* billiards (see [AI]). My talk is devoted to new results about the *chaotic* regime. These results were obtained in collaboration with Anton Zorich.

We started from computer experiments with simple one-dimensional ergodic dynamical systems, and quite unexpectedly ended with topological string theory. The result is a formula connecting fractal dimensions in one dimensional "conformal field theory" and explicit integrals over certain moduli spaces. Also a new analogy arose between ergodic theory and complex algebraic geometry.

We will finish the preface with a brief summary of what is left behind the scene. Our moduli spaces are close relatives of those arising in Seiberg-Witten approach to the supersymmetric Yang-Mills theory. The integrals in the main formula can also be considered as correlators in a topological string theory with $c = 1$. Probably, there is way to calculate them in terms of a matrix model and an integrable hierarchy. In the derivation we use some identity in Kähler geometry which looks like a use of $N = 2$ supersymmetry.

1. Interval exchange transformations

Let us consider a classical mechanical system with the action of a quasi-periodic external force. Mathematically such a system can be described as a symplectic manifold (X, ω) and a closed 1-form α on it. The Hamiltonian of the system is a multivalued function H such that $dH = \alpha$. Branches of H differ from each other by additive constants. One can write the equations of motions $dF(x(t))/dt = \{F, H\}(x(t))$, $F \in C^\infty(X)$, as usual. In contrast with the case of globally defined Hamiltonians, the system does not have first integrals in general. More precisely, one still can make a reduction to codimension 1 near local minima or maxima of H. Nevertheless, the dynamics on an open part of X is expected to be ergodic.

Many physical systems produce after averaging multivalued Hamiltonians. Examples include celestial mechanics, magnetic surfaces, motions of charged particles on Fermi surfaces in crystals etc. (see the survey by S. P. Novikov [Nov]).

We consider the simplest case of 2-dimensional phase space. Thus we have a closed oriented surface Σ with an area element $\omega \in \Omega^2(\Sigma)$ and an area-

preserving vector field ξ:

$$i_\xi \omega = \alpha \in \Omega^1(\Sigma), \quad d\alpha = 0, \quad Lie_\xi(\omega) = 0$$

The main feature of the 2-dimensional case is that the system depends essentially on a finite number of parameters. Generically, the surface splits into a finite number of components filled with periodic trajectories and a finite number of minimal components, where every trajectory is dense. We can associate with every minimal component a so-called *interval exchage transformation* T (see [CFS]).

First of all, we choose an interval I on Σ transversal to the vector field ξ. The transformation T is defined as the first return map (the Poincaré map) from I to itself. The form α defines a measure dx and an orientation on I. The map T preserves both dx and the orientation. Also, it is easy see that generically T has a finite number of discontinuity points $a_1, \dots a_{k-1}$ where k is the number of intervals of continuity of T. Thus we can identify I with an interval in $\mathbf{R}$ and write T as follows:

$$I = [0, a] \subset \mathbf{R}, \quad 0 = a_0 < a_1 < a_2 \dots < a_{k-1} < a = a_k$$

$$T(x) = x + b_i \quad for \quad a_i < x < a_{i+1}$$

where $b_i \in \mathbf{R}$, $i = 0, \dots, k-1$ are some constants. Moreover, intervals $(a_0, a_1), (a_1, a_2), \dots, (a_{k-1}, a_k)$ after the application of this map will be situated on I in an order described by a permutation $\sigma \in S_k$ and without overlapping. Thus numbers b_i can be reconstructed uniquely from the numbers a_i and the permutation σ.

We did not use the area element on the surface Σ in this construction. Everything is defined in terms of a closed 1-form α and an orientation on Σ. Thus it is enough to have an oriented foliation with a transversal measure (and with finitely many singularities) on an oriented surface. It is easy to go back from interval exchange transformations to oriented surfaces with measured foliations. The systems which we will get by the inverse construction correspond to multivalued Hamiltonians H without local minima and maxima.

Also we can consider possibly non-orientable foliations with transversal measures on possibly non-orientable surfaces. This leads to the consideration of mechanical systems with various additional symmetries. The first return map is defined on an interval in an appropriate ramified double covering of the surface.

The permutation σ is called *irreducible* if, for any j, $1 \leq j \leq k-1$, one has

$$\sigma(\{1, 2, \dots, j\}) \neq \{1, 2, \dots, j\} \quad and \quad \sigma(j+1) \neq \sigma(j) + 1 \quad .$$

First return maps for ergodic flows give irreducible permutations.

Theorem (H. Masur [M], W. Veech [V1], 1982). *Let us consider the interval exchange map T for an irreducible permutation σ and generic values of continuous parameters a_i (generic with respect to the Lebesgue measure on the parameter space $\mathbf{R}_+^k = \{(a_1, \ldots, a_k)\}$). Then the map T is ergodic with respect to the Lebesgue measure dx. The entropy of the map T is* 0.

An analogous result is true for non-orientable measured foliations on *orientable* surfaces. The case of non-orientable surfaces is always degenerate. In such case foliations almost always have non-trivial families of closed leaves (see [N]). Presumably, the interesting (ergodic) part is always reduced to measured foliations on orientable surfaces. In order to simplify the exposition we will mainly consider here the case when both the surface and the foliation are orientable.

2. Error terms: first results and computer experiments

Several years ago A. Zorich began the study of the error term in the ergodic theorem for the map T. Let x be a generic point on I and let (y_1, y_2) be a generic subinterval of I. Since the map T is ergodic (for generic values of lengths of subintervals) we have the following equality

$$\#\{i : 1 \leq i \leq N,\ T^i(x) \in (y_1, y_2)\} = (y_2 - y_1)N + o(N)$$

as $N \longrightarrow +\infty$. It was first observed in computer experiments (see [Z1]) that this error term (denoted above by $o(N)$) typically has the growth of a power of N,

$$error\ \ term = O(N^\lambda) \ .$$

Here $\lambda < 1$ is an universal exponent depending only on the permutation σ (see [Z2] for the proof of the related statement).

In the case of 2 or 3 intervals, one has $\lambda = 0$. In these cases the genus of the surface is 1 and the transformation itself is equivalent to the generic irrational rotation of a circle.

In the case of 4 intervals for all irreducible permutations one has

$$\lambda = 0.33333+ \sim (10^{-6})$$

In the case of 5 intervals for all irreducible permutations one has

$$\lambda = 0.50000+ \sim (10^{-6})$$

These two cases correspond to surfaces of genus 2.

If we have 6 intervals (surfaces of genus 3), then the number λ depends on the permutation:

$$\lambda = 0,6156\ldots \text{ or } 0.7173\ldots \ .$$

These two numbers are probably irrational.

Also computer experiments show (see [Z1]) that a generic closed 1-form on a surface Σ defines a filtration on $H_1(\Sigma, \mathbf{R})$ ("fractal Hodge structure") by subspaces

$$H_1(\Sigma, \mathbf{R}) \supset F^{\lambda_g} \supset \ldots \supset F^{\lambda_2} \supset F^{\lambda_1} \supset 0, \quad g = \textit{genus of } \Sigma, \quad \dim(F^{\lambda_j}) = j$$

where $1 = \lambda_1 > \lambda_2 > \ldots \lambda_g > 0$ are some universal constants depending only on the permutation. The number λ which gives the error term in the ergodic theorem is the second exponent λ_2. The highest term of the filtration F^{λ_g} is a Lagrangian subspace of $H_1(\Sigma, \mathbf{R})$.

One can see numbers λ_i geometrically. Let us consider a generic trajectory of the area-preserving vector field ξ on Σ. We consider a sequence of pieces of this trajectory $x(t)$ of lengths $l_j \longrightarrow +\infty$, $j = 1, 2, \ldots$ such that $x(l_j)$ is close to the starting point $x(0)$. We connect two ends of these pieces by short intervals and get a sequence of closed oriented curves C_j on Σ. Homology classes of curves C_j are elements $\mathbf{v}_j = [C_j]$ in the group $H_1(\Sigma, \mathbf{Z})$.

Vectors $\mathbf{v}_j$ at the first approximation are close to a one-dimensional space,

$$\mathbf{v}_j = \mathbf{u}\, l_j + o(l_j)$$

where $\mathbf{u}$ is a non-zero element of $H_1(\Sigma, \mathbf{R})$. Homology class $\mathbf{u}$ is Poincaré dual to the cohomology class $[\alpha]$ of the 1-form α. The lowest non-trivial term in Zorich's filtration is

$$F^1 = F^{\lambda_1} = \mathbf{R} \cdot \mathbf{u} \ .$$

After the projection to the quotient space $H_1(\Sigma, \mathbf{R}) \longrightarrow H_1(\Sigma, \mathbf{R})/\mathbf{R} \cdot \mathbf{u}$ we get again a sequence of vectors. It turns out that for large j these vectors are again close to a 1-dimensional subspace $\mathcal{L}$. Also these vectors mostly will have size $(l_j)^{\lambda_2 + o(1)}$. We define 2-dimensional space $F^{\lambda_2} \subset H_1(\Sigma, \mathbf{R})$ as the inverse image of the 1-dimensional space $\mathcal{L}$. We can repeat the procedure g times. On the last step we get a chaotic sequence of vectors of bounded length in the g-dimensional quotient space $H_1(\Sigma, \mathbf{R})/F^{\lambda_g}$ (see also [Z3]).

There is, presumably, an equivalent way to describe numbers λ_i. Namely, let ϕ be a smooth function on Σ. Assume for simplicity that the multi-valued Hamiltonian has only non-degenerate (Morse) singularities. Then, for generic trajectory $x(t)$, we expect that the number $\int_0^T \phi(x(t))dt$ for large T with high probability has size $T^{\lambda_i + o(T)}$ for some $i \in \{1, \ldots, g\}$. Exponent $\lambda_1 = 1$ appears for all functions with non-zero average value,

$$\int_\Sigma \phi\, \omega \neq 0 \ .$$

The next exponent, λ_2, should work for functions in a codimension 1 subspace in $C^\infty(\Sigma)$ etc.

We discovered in computer experiments (more than 100 cases) that the sum of numbers λ_j is rational,

$$\lambda_1 + \ldots + \lambda_g \in \mathbf{Q} \ .$$

For example, if genus of Σ is 3 and we have 4 simple saddle points for the foliation, then

$$\lambda_1 + \lambda_2 + \lambda_3 = 1. + .5517\ldots + .3411\ldots = 53/28 \ .$$

Also, our observation explains why the case of genus 2 is exceptional. If we have two numbers first of which is equal to 1 and the sum is rational, then the second number is rational too.

3. Moduli spaces

We want to study the renormalization procedure for interval exchange maps. For example, we can define a map from the space of parameters

$$\{(a_1, \ldots, a_k; \sigma)\} = \mathbf{R}_+^k \times S_k$$

to itself considering the first return map of the half $[0, a/2]$ of the original interval $I = [0, a]$. There are also other ways, but the most elegant is the one which we describe at the end of this section. In order to do it we introduce, following W. Veech certain moduli space.

The space $\Omega^1_{closed}(\Sigma)/Diff(\Sigma)$ of equivalence classes of closed 1-forms on a surface Σ is non-Hausdorf. In order to cure it we consider a "doubling" of this space consisting of the space of closed complex-valued 1-forms $\alpha_{\mathbf{C}}$ satisfying the condition

$$Re\alpha \wedge Im\alpha_{|_x} > 0$$

for almost all points $x \in \Sigma$. The notion of positivity here is well-defined because the surface Σ is oriented.

The real-valued 1-form α whose leaves we considered before is the real part, $Re\alpha_{\mathbf{C}}$, of the complex-valued form $\alpha_{\mathbf{C}}$. First of all, we should be sure that we didn't restrict ourselves to some special class of closed 1-forms. It follows from results E. Calabi (see [C]), or from resluts of A. Katok (see [K]) that, for any closed real 1-form α giving an ergodic foliation, there exists at least one closed 1-form $\alpha' \neq 0$ such that $\alpha \wedge \alpha' \geq 0$ everywhere except points where α vanishes. Thus we have a complex valued closed 1-form $\alpha_{\mathbf{C}} = \alpha + i\alpha'$.

Any such complex-valued 1-form defines a complex structure on Σ. Locally outside of zeroes of $\alpha_{\mathbf{C}}$ there is a complex-valued coordinate $z : \Sigma \longrightarrow \mathbf{C}$ such that $dz = \alpha_{\mathbf{C}}$. Holomorphic functions are defined as continuous functions holomorphic in coordinate z. Also, there is a canonical flat metric $(Re\alpha_{\mathbf{C}})^2 + (Im\alpha_{\mathbf{C}})^2$ on Σ with singularities at zeroes of $\alpha_{\mathbf{C}}$.

Let us fix a sequence of non-negative integers $\mathbf{d} = (d_1, \ldots, d_n)$ such that $\sum_i d_i = 2g - 2$ where $g \geq 2$ is the genus of the surface. We denote by $\mathcal{M}_{\mathbf{d}}$ the moduli space of triples $(C; p_1, \ldots, p_n; \alpha_{\mathbf{C}})$ where C is a smooth complex curve of genus g, p_i are pairwise distinct points of C, and $\alpha_{\mathbf{C}}$ is a holomorphic 1-form on C which vanishes up to order d_i at p_i and is non-zero at all other points of C. From this definition it is clear that $\mathcal{M}_{\mathbf{d}}$ is a Hausdorf complex analytic (and algebraic) space (see [V3]).

First of all, $\mathcal{M}_{\mathbf{d}}$ is a complex orbifold of dimension $2g - 1 + k$. Let us consider the period map from a neighborhood of a point $(C, \alpha_{\mathbf{C}})$ of $\mathcal{M}_{\mathbf{d}}$ into the cohomology group $H^1(C, \{p_1, \ldots, p_n\}; \mathbf{C})$. Closed form $\alpha_{\mathbf{C}}$ defines an element into the relative cohomology group $H^1(C, \{p_1, \ldots, p_n\}; \mathbf{C})$ by integration along paths connecting points p_i. In a neighborhood of any point $C, (p_i), \alpha_{\mathbf{C}})$ of $\mathcal{M}_{\mathbf{d}}$, we can identify cohomology groups $H^1(C', \{p'_1, \ldots, p'_n\}; \mathbf{C})$ with $H^1(C, \{p_1, \ldots, p_n\}; \mathbf{C})$ using the Gauss-Manin connection.

Thus we get a map (the period map) from this neighborhood into a vector space. An easy calculation shows that the deformation theory is not obstructed and we get locally a one-to-one correpondence between $\mathcal{M}_{\mathbf{d}}$ and an open domain in the vector space $H^1(C, \{p_1, \ldots, p_n\}; \mathbf{C})$.

We claim that $\mathcal{M}_{\mathbf{d}}$ has structures 1), 2), 3), 4) listed below.

1) *a holomorphic affine structure on $\mathcal{M}_{\mathbf{d}}$ modelled on the vector space* $H^1(C, \{p_1, \ldots, p_n\}; \mathbf{C})$,

2) *a smooth measure μ on $\mathcal{M}_{\mathbf{d}}$,*

3) *a locally quadratic non-holomorphic function $A : \mathcal{M}_{\mathbf{d}} \longrightarrow \mathbf{R}_+$,*

4) *a non-holomorphic action of the group $GL_+(2, \mathbf{R})$ on $\mathcal{M}_{\mathbf{d}}$.*

The first structure we already defined using the period map.

The tangent space to $\mathcal{M}_{\mathbf{d}}$ at each point contains a lattice,

$$H^1(C, \{p_1, \ldots, p_k\}; \mathbf{C}) = H^1(C, \{p_1, \ldots, p_k\}; \mathbf{R} \oplus i\mathbf{R}) \supset$$

$$\supset H^1(C, \{p_1, \ldots, p_k\}; \mathbf{Z}) \oplus i \cdot H^1(C, \{p_1, \ldots, p_k\}; \mathbf{Z}) .$$

The Lebesgue measure (= the Haar measure) on the tangent space to $\mathcal{M}_{\mathbf{d}}$ can be uniquely normalized by the condition that the volume of the quotient torus is equal to 1. Thus we defined the density of a measure μ at each point of $\mathcal{M}_{\mathbf{d}}$.

We define the function $A : \mathcal{M}_{\mathbf{d}} \longrightarrow \mathbf{R}_+$ by the formula $A(C, \alpha_{\mathbf{C}}) = \frac{i}{2} \int_C \alpha_{\mathbf{C}} \wedge \overline{\alpha_{\mathbf{C}}}$. In other terms, it is the area of C for the flat metric associated with $\alpha_{\mathbf{C}}$.

The group $GL_+(2,\mathbf{R})$ of 2×2-matrices with positive determinant acts by linear transformations with constant coefficients on the pair of real-valued 1-forms $(Re(\alpha_{\mathbf{C}}), Im(\alpha_{\mathbf{C}}))$. In the local affine coordinates, this action is the action of $GL_+(2,\mathbf{R})$ on the vector space

$$H^1(C, \{p_1, \ldots, p_n\}; \mathbf{C}) \simeq \mathbf{C} \otimes H^1(\ldots; \mathbf{R}) \simeq \mathbf{R}^2 \otimes H^1(\ldots; \mathbf{R})$$

through the first factor in the tensor product. From this description it is clear that the subgroup $SL(2,\mathbf{R})$ preserves the measure μ and the function A.

On the hypersurface $\mathcal{M}_{\mathbf{d}}{}^{(1)} = A^{-1}(1)$ (the level set of the function A) we define the induced measure by the formula

$$\mu^{(1)} = \frac{\mu}{dA}$$

The group $SL(2,\mathbf{R})$ acts on $\mathcal{M}_{\mathbf{d}}{}^{(1)}$ preserving $\mu^{(1)}$.

Theorem (H. Masur, W. Veech). *The total volume of* $\mathcal{M}_{\mathbf{d}}{}^{(1)}$ *with respect to the measure* $\mu^{(1)}$ *is finite.*

Let us denote by $\mathcal{M}$ any connected component of $\mathcal{M}_{\mathbf{d}}$ and by $\mathcal{M}^{(1)}$ its intersection with $\mathcal{M}_{\mathbf{d}}{}^{(1)}$.

Theorem (H. Masur, W. Veech). *The action of the* 1*-parameter group* $\{diag(e^t, e^{-t})\} \subset SL(2,\mathbf{R})$ *on* $(\mathcal{M}^{(1)}, \mu^{(1)})$ *is ergodic.*

The action in this theorem is in fact the renormalization group flow for interval exchange maps. Another name for this flow is the "Teichm'ueller geodesic flow" because it gives the Euler-Lagrange equations for geodesics for the Teichmüller metric on the moduli space of complex curves. Notice that this metric is not a Riemannian metric, but only a Finsler metric.

The intuitive explanation of the ergodicity is that the group $\{diag(e^t, e^{-t})\}$ expands leaves of the foliation by affine subspaces parallel to $H^1(\ldots, \mathbf{R})$ and contracts leaves of the foliation by subspaces parallel to $H^1(\ldots, i\mathbf{R})$.

4. Topology of the moduli space

In the last theorem from the pervious section we consider *connected components* of the moduli space $\mathcal{M}_{\mathbf{d}}$. From the first glance it seems to be not necessary because normally moduli spaces are connected. It is not true in our case. W. Veech and P. Arnoux discovered by direct calculations in terms of permutations that there are several connected components. The set of irreducible permutations is decomposed into certain equivalence classes called Rauzy classes. These classes correspond to connected components of spaces $\mathcal{M}_{\mathbf{d}}$. For a long time the geometric origin of non-connectedness was not clear.

Recently we have obtained the complete classification of connected components. First of all, there are two series of connected components of $\mathcal{M}_{\mathbf{d}}$ consisting of hyperelliptic curves such that the set of singular points is invariant under the hyperelliptic involution. The first series corresponds to curves with one singular point, $\mathbf{d} = 2g - 2$ for $g \geq 2$. The second series correponds to curves of genus $g \geq 2$ with two singular points, $\mathbf{d} = (g-1, g-1)$.

If all orders of zeros are even numbers, we have a spin structure on C given by a half of the canonical divisor

$$S = \sum_i \frac{d_i}{2}[p_i] \in Pic(C) \ .$$

It is well-known that spin structures have a topological characteristic (parity) which doesn't change under continuous deformations (see [A]). The parity of a spin-structure is the parity of the dimension of the space of global sections of the corresponding holomorphic line bundle.

Classification theorem. *There are hyperelliptic and non-hyperelliptic connected components of the moduli space of holomorphic 1-forms. For non-hyperelliptic components there are two cases: the vector* $\mathbf{d}$ *is divisible by* 2, *or not. If* $\mathbf{d}$ *is divisible by* 2 *then there are two components corresponding to even and odd spin structures. There are exceptional cases when we get an empty set: 1) for* $g = 2$: *all non-hyperelliptic strata; 2) for* $g = 3$: *non-hyperelliptic strata with* $\mathbf{d}$ *divisible by* 2 *and even spin structure.*

We have analogous results for the moduli space of quadratic differentials. At the moment we do not know anything about the topology of connected components except for the hyperelliptic locus.

Conjecture. *Each connected component* $\mathcal{M}$ *of* $\mathcal{M}_{\mathbf{d}}$ *has homotopy type* $K(\pi, 1)$, *where* π *is a group commensurable with some mapping class group.*

5. Lyapunov exponents

We recall here the famous multiplicative ergodic theorem.

Theorem (V. Oseledets [O]). *Let* $T_t : (X, \mu) \longrightarrow (X, \mu)$, $t \in \mathbf{R}_+$, *be an ergodic flow on a space* X *with finite measure* μ; *let* V *be an* $\mathbf{R}_+$*-equivariant measurable finite-dimensional real vector bundle. We also assume that a (non-equivariant) norm* $|\ |$ *on* V *is chosen such that, for all* $t \in \mathbf{R}_+$,

$$\int_X log(1 + |T_t : V_x \longrightarrow V_{T_t(x)}|)\mu < +\infty \ .$$

Then there are real constants $\lambda_1 > \lambda_2 > \ldots > \lambda_k$ *and an equivarinat filtration of the vector bundle* V

$$V = V^{\lambda_1} \supset \ldots \supset V^{\lambda_k} \supset 0$$

such that, for almost all $x \in X$ *and all* $v \in V_x \setminus \{0\}$, *one has*

$$|T_t(v)| = e^{\lambda_j t + o(t)}, \quad t \longrightarrow +\infty$$

where j *is the maximal value for which* $v \in (V^{\lambda_j})_x$. *The filtration* V^{λ_j} *and numbers* λ_j *do not change if we replace norm* $|\ |$ *by another norm* $|\ |'$ *such that*

$$\int_X log\left(max_{v \in V_x \setminus \{0\}}\left(max(|v|/|v|', |v|'/|v|)\right)\right) \mu < +\infty$$

Analogous statement is true for systems in discrete time $\mathbf{Z}_+$.

Numbers λ_j are called Lyapunov exponents of the equivariant vector bundle V. Usually people formulate this theorem using language of matrix-valued 1-cocycles instead of equivariant vector bundles. This is equivalent to the formulation above because any vector bundle on a measurable space can be trivialized on the complement to a subset of measure zero.

If our system is reversible, we can change the positive direction of the time. Lyapunov exponents will be replaced by negative Lyapunov exponents. A new filtration will appear. This new filtration is opposite to the previous one, and they together define an equivariant splitting of V into the direct sum of subbundles.

Lyapunov exponents are, in general, very hard to evaluate other than numerically. We are aware only about two examples of explicit formulas. One example is the geodesic flow on a locally symmetric domain and V being a homogeneous vector bundle. In this case one can explicitly construct the splitting of V. The second example is the multiplication of random independent matrices whose entries are independent equally distributed gaussian random variables. In this case one can calculate Lyapunov exponents using rotational invariance and the Markov property.

Our calculation seems to be the first calculation of Lyapunov exponents in a non-homogeneous situation. As the reader will see later, our proof uses a replacement of a deterministic system by a Markov process.

Let us define a vector bundle H^1 over $\mathcal{M}$ by saying that its fiber at point $(C, \alpha_{\mathbf{C}})$ is the cohomology group $H^1(C, \mathbf{R})$. We apply the multiplicative ergodic theorem to the action of $\{diag(e^t, e^{-t})\}$ on $\mathcal{M}^{(1)}$ and to the bundle H^1. The action of the group on this bundle is defined by the lift using the natural flat connection (Gauss-Manin connection). We will not specify for a moment the norm on H^1 because all natural choices are equivalent in the sense specified in our formulation of the multiplicative ergodic theorem.

The structure group of the bundle H^1 is reduced to $Sp(2g, \mathbf{R}) \subset GL(2g, \mathbf{R})$. One can see easily that in this case Lyapunov exponents form a symmetric subset of $\mathbf{R}$. Also, in all experiments we do not have degenerate Lyapunov exponents, i.e. the picture is

$$\lambda_1 \geq \lambda_2 \geq \ldots \lambda_g >\geq 0 \geq \lambda_{g+1} = -\lambda_g >\geq \ldots \geq \lambda_{2g} = -\lambda_1$$

Theorem (A. Zorich). *1) The highest Lyapunov exponent λ_1 is equal to 1 and has multiplicity one. The corresponding 1-dimensional subbundle is $Re(\alpha)\mathbf{R} \subset H^1$. 2) The second Lyapunov exponent λ_2 governs the error term in the ergodic theorem for interval exchange maps. 3) The filtration on H^1 related with the positive time dynamics depends locally only on the cohomology class $[Re\alpha] \in H^1(\Sigma, \{p_1, \ldots, p_n\}; \mathbf{R})$.*

The first part is quite easy. At least, the growth of the norm for the 1-dimensional bundle $Re(\alpha)\mathbf{R} \subset H^1$ is exponential with the rate 1.

The second part looks more mysterious. We compare two different dynamical systems, the original flow on the surface and the renormalization group flow on the moduli space. The time in one system is morally an exponent of the time in another system. The technical tool here is a mixture of the ordinary (additive) and the multiplicative ergodic theorem for an action of the group $Aff(\mathbf{R}^1)$ of affine transformation of line. We are planning to write in a future a detailed proof. A rather technical proof of related statement can be found in [Z2].

The third part is not hard, but surprising. In fact, the positive-time filtration on H^1 coincides with the filtration for *real-valued* closed 1-forms described in section 2. Thus it is independent on the choice of the imaginary part.

In computer experiments we observed that the spectrum of Lyapunov exponents is simple. In the rest of the paper wee will assume for simplicity that the non-degeneracy holds always. The general reason to beleive in it is that there is no additional symmetry in the system which can force the Lyapunov spectrum to be degenerate.

6. Analogy with the Hodge theory

We see that our moduli space locally is decomposed into the product of two manifolds

$$H^1(\ldots; \mathbf{R}) \times H^1(\ldots; i\mathbf{R}) \ .$$

More precisely, we have two complementary subbundles in the tangent bundle satisfying the Frobenius integrability condition. This is quite analogous the geometry of a complex manifold. If N is an almost complex manifold, then we have two complementary subbundles $T^{1,0}$ and $T^{0,1}$ in the *complexified* tangent

bundle $T_N \otimes \mathbf{C}$. The integrability condition of the almost-complex structure is equivalent to the formal integrability of distributions $T^{1,0}$ and $T^{0,1}$.

Also, if we have a family of complex manifolds X_b, $b \in B$, paramertized holomorphically by a complex manifold B, then for every integer k we have a holomorphic vector bundle over B with the fiber $H^k(X_b; \mathbf{C})$. This bundle carries a natural flat connection, and a holomorphic filtration by subbundles coming from the standard spectral sequence.

This picture (variations of Hodge structures, see [G]) is parallel to the situation in the multiplicative ergodic theorem applied to a smooth dynamical system. Let M denote the underlying manifold of the system. The tangent bundle T_M is an equivariant bundle. Thus, in the case of ergodicity and convergence of certain integrals we get a canonical measurable splitting of T_M into the direct sum of subbundles indexed by Lyapunov exponents. It is well known in many cases (and is expected in general) that these subbundles, and also all terms of both filtrations are integrable, i.e. they are tangent to leaves of (non-smooth) foliations on M. Two most important foliations (expanding and contracting foliations) correspond to terms of filtrations associated with all positive or all negative exponents. It is known that if the invariant measure is smooth then the sum of positive exponents is equal to the entropy of the system (Pesin formula).

7. Formula for the sum of exponents

The main result of our work is an explicit formula for the sum of positive Lyapunov exponents $\lambda_1 + \ldots + \lambda_g$ for the equivarinat bundle H^1 over the connected component $\mathcal{M}$ of moduli spaces of curves with holomorphic 1-forms.

We want to warn the reader that this equivariant bundle is not the whole tangent bundle $T_{\mathcal{M}}$. Lyapunov exponents for $T_{\mathcal{M}}$ can be calculated easily through numbers λ_j. The entropy of the Teichmüller geodesic flow is equal by the Pesin formula to the complex dimension of $\mathcal{M}$. In short, what we are computing there is more delicate information than the entropy of the system.

Hypersurface $\mathcal{M}^{(1)}$ is isomorphic to the quotient space $\mathcal{M}/\mathbf{R}_+^*$, where $\mathbf{R}_+^*$ is identified with subgroup $\{diag(e^t, e^t)\}$ of $GL_+(2, \mathbf{R})$. We denote by $\mathcal{M}^{(2)}$ the quotient space

$$\mathcal{M}^{(1)}/SO(2, \mathbf{R}) \simeq \mathcal{M}/\mathbf{C}^* \ .$$

This space is a complex algebraic orbifold.

Orbits of the group $GL_+(2, \mathbf{R})$ define a 4-dimensional foliation on $\mathcal{M}$. It induces a 3-dimensional foliation on $\mathcal{M}^{(1)}$ by orbits of $SL(2, \mathbf{R})$, and a 2-dimensional foliation $\mathcal{F}$ on $\mathcal{M}^{(2)}$. Leaves of $\mathcal{F}$ are complex curves in $\mathcal{M}^{(2)}$, but the foliation itself is not holomorphic.

3-dimensional foliation on $\mathcal{M}^{(1)}$ carries a natural transversal measure. This measure is the quotient of $\mu^{(1)}$ by the Haar measure on $SL(2,\mathbf{R})$. The transversal measure on $\mathcal{M}^{(1)}$ induces a transversal measure on $\mathcal{M}^{(2)}$. We have natural orientations on $\mathcal{M}^{(2)}$ and on leaves of $\mathcal{F}$ arising from complex structures. Thus we can construct differential form β such that

$$\beta \in \Omega^{dim_{\mathbf{R}}\mathcal{M}^{(2)}-2}(\mathcal{M}^{(2)}),\ d\beta = 0,\quad Ker\,\beta = \mathcal{F}\ .$$

The natural projection $\mathcal{M} \longrightarrow \mathcal{M}^{(2)}$ is a holomorphic $\mathbf{C}^*$-bundle with a Hermitean metric given by the function A. Thus we have a natural curvature form $\gamma_1 \in \Omega^2(\mathcal{M}^{(2)})$, $d\gamma_1 = 0$ representing the first Chern class $c_1(\mathcal{M} \longrightarrow \mathcal{M}^{(2)})$. This form is given locally by the formula

$$\gamma_1 = \frac{1}{2\pi i}\partial\overline{\partial}\, log\,(A(s))$$

where s is a non-zero holomorphic section of the line bundle $\mathcal{M} \longrightarrow \mathcal{M}^{(2)}$.

We also have another holomorphic vector bundle on $\mathcal{M}^{(2)}$. The fiber of this bundle (denoted by $H^{(1,0)}$) is equal to $H^0(C,\Omega^1_C)$, the term of the Hodge filtration in $H^1 \otimes \mathbf{C}$. This holomorphic bundle carries a natural hermitean metric coming from the polarization in Hodge theory. The formula for the metric is

$$|\omega|^2 := \frac{1}{2\pi i}\int_C \omega \wedge \overline{\omega},\quad \omega \in \Gamma(C,\Omega^1_{hol})$$

This metric defines again a canonical closed 2-form γ_2 representing the characteristic class $c_1(H^{(1,0)})$.

Main Theorem.

$$\lambda_1 + \ldots + \lambda_g = \frac{\int_{\mathcal{M}^{(2)}} \beta \wedge \gamma_2}{\int_{\mathcal{M}^{(2)}} \beta \wedge \gamma_1}$$

In this formula, we can not go directly to the cohomology, because the orbifold $\mathcal{M}^{(2)}$ is not compact. In order to overcome this difficulty, we constructed a compactification $\overline{\mathcal{M}}^{(2)}$ of $\mathcal{M}^{(2)}$ with toroidal singularities. All three differential forms $\beta, \gamma_1, \gamma_2$ in the formula seem to be smooth on $\overline{\mathcal{M}}^{(2)}$. Both γ_1 and γ_2 represent classes in $H^2(\overline{\mathcal{M}}^{(2)}, \mathbf{Q})$. It seems that β also represents a rational cohomology class although we don't have a proof yet. In the case of *one* critical point of 1-form α it is true because, by invariance reasons, the form β is proportional to a power of γ_1. Another possible explanation of rationality is that $[\beta]$ is *proportional* to a rational class because the part of $H^{dim-2}(\overline{\mathcal{M}}^{(2)}, \mathbf{R})$ consisting of classes vanishing on boundary divisors, can be one-dimensional. In any case, we almost explained the rationality of $\sum_{j=1}^{g} \lambda_j$ observed in experiments.

8. Proof of the formula

Any leaf of the foliation $\mathcal{F}$ carries a natural hyperbolic metric. The generic leave is a copy of the upper half-plane $SL(2,\mathbf{R})/SO(2,\mathbf{R})$. We are studying the behavior of the monodromy of the Gauss-Manin connection in H^1 along *a long geodesic going in a random direction on a generic leave of* $\mathcal{F}$. It was an old idea of Dennis Sullivan to replace the walk along random geodesic by a random walk on the hyperbolic plane (the Brownian motion). The trajectory of the random walk goes to infinity in a random direction with approximately constant speed.

The meaning of the sum $\lambda_1+\ldots\lambda_g$ is the following. We move using the parallel transport a generic Lagrangian subspace L in the fiber of H^1 and calculate the average growth of the volume element L associated with the Riemannian metric on L induced from the natural metric (polarization) on H^1.

As we discuss above, we can replace the geodesic flow by the Brownian motion. We will approximate the random walk by a sequence of small jumps of a fixed length in random uniformly distributed directions on the hyperbolic plane.

Identity. *Fix $x \in \mathcal{M}^{(2)}$ and identify the leaf $\mathcal{F}_x$ of $\mathcal{F}$ passing through x with the model of the Lobachevsky plane in unit disc $\{z \in \mathbf{C} |\ |z|<1\}$ in such a way that $x \mapsto z = 0$. Also, we trivialize the vector bundle H^1 over $\mathcal{F}_x$ using the Gauss-Manin connection. Then, for any Lagrangian subspace $L \subset H^1_0$, and for any ϵ, $0<\epsilon<1$ the following identity holds:*

$$\frac{1}{2\pi}\int_0^{2\pi} d\theta\ log\left(\frac{volume\ on\ L,\ for\ metric\ in\ H^1_{\epsilon e^{i\theta}}}{volume\ on\ L,\ for\ metric\ in\ H^1_0}\right) =$$

$$\int\limits_{disc\ |z|\le\epsilon} log\left(\frac{\epsilon^2}{|z|^2}\right)\gamma_2\ .$$

The proof of this identity follows. Let us choose a locally constant basis $l_1,\ldots,l_g$ of L and a basis $v_1(z),\ldots,v_g(z)$ in $H^{1,0}_z$ depending holomorphically on $z \in \mathcal{F}_x$. Then we have

$$|l_1\wedge\ldots\wedge l_g|^2_z = \frac{(l_1\wedge\ldots\wedge l_g\wedge v_1\wedge\ldots\wedge v_g)\otimes(l_1\wedge\ldots\wedge l_g\wedge\overline{v}_1\wedge\ldots\wedge\overline{v}_g)}{(v_1\wedge\ldots\wedge v_g\wedge\overline{v}_1\wedge\ldots\wedge\overline{v}_g)\otimes(v_1\wedge\ldots\wedge v_g\wedge\overline{v}_1\wedge\ldots\wedge\overline{v}_g)}$$

where the numerator and the denominator are considered as elements of the one dimensional complex vector space $(\wedge^{2g}(H^1_0\otimes\mathbf{C}))^{\otimes 2}$.

If we apply the Laplace-Beltrami operator $\Delta = (1/2\pi i) \times \partial_z \overline{\partial_z}$ to the logarithms of both sides of the formula from above, we get that

$$\Delta(log(|l_1 \wedge \ldots \wedge l_g|_z^2)) =$$

$$\Delta(holomorphic\ \ function) + \Delta(antiholomorphic\ \ function) + \gamma_{2|\mathcal{F}_x}$$

Application of the simple formula

$$\frac{1}{2\pi}\int_0^{2\pi} f(\epsilon e^{i\theta})\, d\theta \ - \ f(0) = \frac{1}{2\pi i}\int_{|z|\leq\epsilon} log\left(\frac{\epsilon^2}{|z|^2}\right)\partial_z\overline{\partial_z}(f)dz \wedge d\overline{z}, \ \ \forall f \in C^\infty$$

gives the main identity.

The mean value on the disc $\{z|\,|z| < \epsilon\}$ of the function $z \mapsto log\,(\epsilon^2/|z|^2)$ is equal to 1. The main identity implies that the average growth of the volume on L depends not on L but only on the position of the point $x \in \mathcal{M}^{(2)}$. Because of ergodicity we can average over the invariant probability measure $Z^{-1} \times \mu^{(2)}$, where $Z = \int_{\mathcal{M}^{(2)}} \mu^{(2)}$ is the total volume of $\mathcal{M}^{(2)}$. The invariant measure $\mu^{(2)}$ is proportional to $\beta \wedge \gamma_1$. This explains the denominator in the formula for $\lambda_1 + \ldots + \lambda_g$.

9. Generalizations

In our proof we treat the higher-dimensional moduli space $\mathcal{M}^{(2)}$ as a "curve with hyperbolic metric". In general, in many situations ergodic foliations with transversal measures and certain differential-geometric structures along leaves can be considered as virtual manifolds with the same type of geometric structure. Also, ergodic actions of groups can be considered as virtual discrete subgroups (Mackey's philosophy).

Our proof works literally in a different situation. Let C be a complex curve of genus $g > 0$ parametrizing polarized Abelian varieties A_x, $x \in C$ of complex dimension G.

We endow C with the canonical hyperbolic metric and consider the geodesic flow on it. It gives us an ergodic dynamical system. For the equivariant bundle we will take the symplectic local system H^1 over C with fibers $H^1(A_x, \mathbf{R})$. Again, the sum of positive Lyapunov exponents is rational:

$$\lambda_1 + \ldots + \lambda_G = \frac{deg(H^{1,0})}{2g-2}$$

References

[A] M. Atiyah, "Riemann surfaces and spin structures", Ann. scient. Éc. Norm. Sup., 4^e série 4(1971), 47–62.

[AI] E. Aurell, C. Itzykson, "Rational billiards and algebraic curves", Journal of Geometry and Physics, vol. 5, no. 2, (1988).

[C] E. Calabi, "An intrinsic characterization of harmonic 1-forms", Global Analysis, Papers in Honor of K.Kodaira, (D.C.Spencer and S.Iyanaga, ed.), 1969, pp. 101–107.

[CFS] I. P. Cornfeld, S. V. Fomin, Ya. G. Sinai, "Ergodic Theory", Springer-Verlag 1982.

[G] P. Griffiths, "Variation of Hodge structure", in "Topics in transcendental algebraic geometry", Ed. by P. Griffiths, Princeton University Press 1984, 3-29.

[K] A. B. Katok, "Invariant measures of flows on oriented surfaces", Soviet Math. Dokl., 14 (1973), 1104-1108.

[M] H. Masur, "Interval exchange transformations and measured foliations", Ann. Math., 115-1(1982), 169-200.

[N] A. Nogueira, "Almost all interval exchange transformations with flips are nonergodic", Ergodic Theory and Dynamical Systems, 9 (1989), no. 3, 515-525.

[Nov] S. Novikov, "Hamiltonian formalism and a multi-valued analogue of Morse theory", Russian Math. Surveys, **37:5** (1982), 1-56.

[O] V. I. Oseledetc, "A Multiplicative Ergodic Theorem. Ljapunov characteristic numbers for dynamical systems", Trans. Moscow Math. Soc. **19**, (1968), 197–231.

[V1] W. Veech, "Gauss measures for transformations on the space of interval exchange maps", Ann, Math. 115 (1982), 201-242.

[V2] W. Veech, "The Teichmüller geodesic flow", Ann. Math. 124 (1986), 441-530.

[V3] W. Veech, "Moduli spaces of quadratic differentials", Journal d'Analyse Math., 55 (1990), 117-171.

[Z1] A. Zorich, "Asymptotic flag of an orientable measured foliation on a surface", in "Geometric Study of Foliations", World Sci., 1994, 479-498.

[Z2] A. Zorich, "Deviation for interval exchange transformations", Ergodic Theory adn Dynamical Systems, to appear.

[Z3] A. Zorich, "On hyperplane sections of periodic surfaces", in "Solitons, Geometry, and Topology: on the Crossroad", V. M. Bukhshtaber and S. P. Novikov (eds.), Transactions of the AMS, ser. 2, **179**, AMS, Providence,RI, 1997, pp. 173-189.

GAUGE DYNAMICS AND COMPACTIFICATION TO THREE DIMENSIONS

N. SEIBERG[1]
Department of Physics and Astronomy
Rutgers University
Piscataway, NJ 08855-0849
E. WITTEN[2]
School of Natural Sciences, Institute for Advanced Study
Olden Lane, Princeton NJ 08540

We study the compactification and dimensional reduction of four-dimensional $N = 2$ supersymmetric gauge theories to three dimensions. The vacuum structure can be determined quite precisely. Compactification on a circle of radius R gives a theory that interpolates in an interesting way between the four-dimensional result for $R \to \infty$ and the three-dimensional result for $R \to 0$.

1. Introduction

In [1,2], the dynamics of the Coulomb branch of $N = 2$ super Yang-Mills theory was analyzed using general constraints of supersymmetry and low energy effective field theory – extended, crucially, by allowing for the possibility of duality transformations. The purpose of the present paper is to study the same theory compactified or reduced to three dimensions.

Compactification to three dimensions means that one formulates the quantum theory on $\mathbf{R}^3 \times \mathbf{S}^1_R$, where $\mathbf{S}^1_R$ is a circle of circumference $2\pi R$. For $R \to \infty$ one should recover the four-dimensional solution of [1,2].

Dimensional reduction means instead that at the classical level, one takes the fields to be independent of the fourth dimension, and then one quantizes the resulting three-dimensional theory. Intuitively, one would expect that this three-dimensional theory should be equivalent to the small R limit of compactification. After all, the energetic cost of excitations that carry non-zero momentum along $\mathbf{S}^1_R$ diverges as $R \to 0$.

In section two of this paper, the Coulomb branch of the three-dimensional theory will be analyzed, for gauge groups $SU(2)$ and $U(1)$. In fact, drawing upon ideas of [3,4], results on this subject have been inferred recently from string theory [5]. Here we will show what can be learned about the problem using some simple arguments of field theory, and in particular we recover many of the results of [5]. In section three, we analyze the

[1] Supported in part by DOE DE-FG02-96ER40559

[2] Supported in part by NSF PHY95-13835

References

[A] M. Atiyah, "Riemann surfaces and spin structures", Ann. scient. Éc. Norm. Sup., 4^e série 4(1971), 47–62.

[AI] E. Aurell, C. Itzykson, "Rational billiards and algebraic curves", Journal of Geometry and Physics, vol. 5, no. 2, (1988).

[C] E. Calabi, "An intrinsic characterization of harmonic 1-forms", Global Analysis, Papers in Honor of K.Kodaira, (D.C.Spencer and S.Iyanaga, ed.), 1969, pp. 101–107.

[CFS] I. P. Cornfeld, S. V. Fomin, Ya. G. Sinai, "Ergodic Theory", Springer-Verlag 1982.

[G] P. Griffiths, "Variation of Hodge structure", in "Topics in transcendental algebraic geometry", Ed. by P. Griffiths, Princeton University Press 1984, 3-29.

[K] A. B. Katok, "Invariant measures of flows on oriented surfaces", Soviet Math. Dokl., 14 (1973), 1104-1108.

[M] H. Masur, "Interval exchange transformations and measured foliations", Ann. Math., 115-1(1982), 169-200.

[N] A. Nogueira, "Almost all interval exchange transformations with flips are nonergodic", Ergodic Theory and Dynamical Systems, 9 (1989), no. 3, 515-525.

[Nov] S. Novikov, "Hamiltonian formalism and a multi-valued analogue of Morse theory", Russian Math. Surveys, **37:5** (1982), 1-56.

[O] V. I. Oseledetc, "A Multiplicative Ergodic Theorem. Ljapunov characteristic numbers for dynamical systems", Trans. Moscow Math. Soc. **19**, (1968), 197–231.

[V1] W. Veech, "Gauss measures for transformations on the space of interval exchange maps", Ann, Math. 115 (1982), 201-242.

[V2] W. Veech, "The Teichmüller geodesic flow", Ann. Math. 124 (1986), 441-530.

[V3] W. Veech, "Moduli spaces of quadratic differentials", Journal d'Analyse Math., 55 (1990), 117-171.

[Z1] A. Zorich, "Asymptotic flag of an orientable measured foliation on a surface", in "Geometric Study of Foliations", World Sci., 1994, 479-498.

[Z2] A. Zorich, "Deviation for interval exchange transformations", Ergodic Theory adn Dynamical Systems, to appear.

[Z3] A. Zorich, "On hyperplane sections of periodic surfaces", in "Solitons, Geometry, and Topology: on the Crossroad", V. M. Bukhshtaber and S. P. Novikov (eds.), Transactions of the AMS, ser. 2, **179**, AMS, Providence,RI, 1997, pp. 173-189.

GAUGE DYNAMICS AND COMPACTIFICATION TO THREE DIMENSIONS

N. SEIBERG[1]
Department of Physics and Astronomy
Rutgers University
Piscataway, NJ 08855-0849
E. WITTEN[2]
School of Natural Sciences, Institute for Advanced Study
Olden Lane, Princeton NJ 08540

We study the compactification and dimensional reduction of four-dimensional $N = 2$ supersymmetric gauge theories to three dimensions. The vacuum structure can be determined quite precisely. Compactification on a circle of radius R gives a theory that interpolates in an interesting way between the four-dimensional result for $R \to \infty$ and the three-dimensional result for $R \to 0$.

1. Introduction

In [1,2], the dynamics of the Coulomb branch of $N = 2$ super Yang-Mills theory was analyzed using general constraints of supersymmetry and low energy effective field theory – extended, crucially, by allowing for the possibility of duality transformations. The purpose of the present paper is to study the same theory compactified or reduced to three dimensions.

Compactification to three dimensions means that one formulates the quantum theory on $\mathbf{R}^3 \times \mathbf{S}^1_R$, where $\mathbf{S}^1_R$ is a circle of circumference $2\pi R$. For $R \to \infty$ one should recover the four-dimensional solution of [1,2].

Dimensional reduction means instead that at the classical level, one takes the fields to be independent of the fourth dimension, and then one quantizes the resulting three-dimensional theory. Intuitively, one would expect that this three-dimensional theory should be equivalent to the small R limit of compactification. After all, the energetic cost of excitations that carry non-zero momentum along $\mathbf{S}^1_R$ diverges as $R \to 0$.

In section two of this paper, the Coulomb branch of the three-dimensional theory will be analyzed, for gauge groups $SU(2)$ and $U(1)$. In fact, drawing upon ideas of [3,4], results on this subject have been inferred recently from string theory [5]. Here we will show what can be learned about the problem using some simple arguments of field theory, and in particular we recover many of the results of [5]. In section three, we analyze the

[1] Supported in part by DOE DE-FG02-96ER40559

[2] Supported in part by NSF PHY95-13835

four-dimensional quantum theory on $\mathbf{R}^3 \times \mathbf{S}^1_R$ using some simple field theory arguments, among other things verifying that the large R limit gives back the four-dimensional theory while the small R limit gives the three-dimensional theory. In section four we recover and explain results of section three from the standpoint of string theory.

2. The Three-Dimensional Theory

2.1. The Problem

We will here be discussing three-dimensional supersymmetric gauge theories which have $N = 4$ supersymmetry in the three-dimensional sense (corresponding to $N = 2$ in four dimensions). They can be constructed by dimensional reduction of six-dimensional $N = 1$ super Yang-Mills theory to three dimensions. This is a convenient starting point in understanding the field content and symmetries of the models. First we consider the pure gauge theories, without matter hypermultiplets.

In six dimensions, the fields are the gauge field A and Weyl fermions ψ in the adjoint representation of the gauge group G. There is an $SU(2)_R$ symmetry that acts only on the fermions; the fermions and supercharges transform as doublets of $SU(2)_R$.

Upon dimensional reduction to three dimensions – that is, taking the fields to be independent of three coordinates $x^{4,5,6}$ – one obtains a theory with the following additional structures. The last three components of A become in three dimensions scalar fields ϕ_i, $i = 1, 2, 3$, in the adjoint representation. These scalars transform in the vector representation under the group of rotations of the $x^{4,5,6}$; we will call the double cover of this group $SU(2)_N$. Note that in reduction to four dimensions, only two such scalars appear, and instead of $SU(2)_N$, one gets only a $U(1)$ symmetry of rotations of the $x^{5,6}$ plane. This symmetry is often called $U(1)_R$, and has an anomaly involving four-dimensional instantons. In three dimensions, because the group $SU(2)_N$ is simple, there is no possibility of such an anomaly. Finally, three dimensional Euclidean space $\mathbf{R}^3$ has a group of rotations whose double cover we will call $SU(2)_E$.

Under $SU(2)_R \times SU(2)_N \times SU(2)_E$, the fermions transform as $(\mathbf{2}, \mathbf{2}, \mathbf{2})$, as do the supercharges (so that $SU(2)_N$ is a group of R symmetries just like $SU(2)_R$), while the scalars transform as $(\mathbf{1}, \mathbf{3}, \mathbf{1})$.

Now to formulate the problem of the Coulomb branch, the starting point is the potential energy for the scalars. This arises by dimensional reduction from the F^2 kinetic

energy of gauge fields in six dimensions, and is

$$V = \frac{1}{4e^2} \sum_{i<j} \mathrm{Tr}[\phi_i, \phi_j]^2 \tag{2.1}$$

where e is the gauge coupling. For the classical energy to vanish, it is necessary and sufficient that the ϕ_i should commute. One can consequently take them to lie in a maximal commuting subalgebra of the Lie algebra of G. If G has rank r, the space of zeroes of V, up to gauge transformation, has real dimension $3r$. A generic set of commuting ϕ_i breaks G to an Abelian subgroup $U(1)^r$. In addition to the ϕ_i, there are then r massless photons. Since a photon is dual to a scalar in three space-time dimensions, there are in all $4r$ massless scalars – $3r$ components of ϕ_i and r duals of the photons.

Are these $4r$ scalars really massless in the quantum theory? The $N = 4$ supersymmetry makes it impossible to generate a superpotential, so there are only two rather special ways to have masses. One possibility is to include a three-dimensional Chern-Simons interaction, with a quantized integer-valued coupling k. For non-zero k, the modes described above do indeed get masses, and the problem we will pose in this paper of studying the Coulomb branch does not arise. (There is an interesting question of whether the theory with $k \neq 0$ has a supersymmetric vacuum; at least for large k, the answer can be seen to be "yes" by using perturbation theory in $1/k$.) If the gauge group G has $U(1)$ factors, it is possible to include Fayet-Iliopoulos D-terms (transforming as $(\mathbf{3},\mathbf{1},\mathbf{1})$ under $SU(2)_R \times SU(2)_N \times SU(2)_E$), again giving mass to some modes. In this paper, we will mainly consider the case that G is semi-simple, so that D-terms are impossible; but even when we consider $G = U(1)$, we will focus on the case that the D-terms are absent.

With these restrictions, then, the $4r$ scalars are really massless and parametrize a family of vacuum states. (This is also true later when we include hypermultiplets.) Moreover, by considering the region of large ϕ_i, we know that for a generic vacuum in this family, the physics is free in the infrared and can be described by a conventional low energy effective field theory. The most general low energy effective action for $4r$ massless scalars in three dimensional $N = 4$ supersymmetry is a sigma model with a target space that is a hyper-Kähler manifold of quaternionic dimension r. Thus, the moduli space $\mathcal{M}$ of vacua is to be understood as such a hyper-Kähler manifold.

In this paper, we will only consider in detail the cases $G = SU(2)$ and $G = U(1)$, for which $r = 1$, and $\mathcal{M}$ is simply a hyper-Kähler manifold of real dimension four. Moreover, this manifold has a non-trivial action of $SU(2)_N$, which highly constrains the problem; the

hyper-Kähler manifolds we need are (with one easy exception, the reason for which will emerge) to be found in the classification in [6] of certain four-dimensional hyper-Kähler manifolds with $SO(3)$ symmetry.

So far we have discussed the pure gauge theories. It is also possible to include matter hypermultiplets. For $G = SU(2)$, we will consider in some detail the case of matter hypermultiplets in the doublet or two-dimensional representation of G. The basic such object is a multiplet that contains four real scalars that transform as $(\mathbf{2},\mathbf{1},\mathbf{1},\mathbf{2})$ under $SU(2)_R \times SU(2)_N \times SU(2)_E \times G$, along with fermions transforming as $(\mathbf{1},\mathbf{2},\mathbf{2},\mathbf{2})$. For somewhat quirky reasons, such a multiplet is sometimes called a half-hypermultiplet. In [1], the $G = SU(2)$ theory was studied (in four dimensions) with any number N_f of doublet hypermultiplets, or in other words $2N_f$ half-hypermultiplets. With this notation, it appears that we should allow for the case in which N_f is a half-integer rather than an integer, but at this point some subtleties involving global anomalies intervene. In four dimensions, given the fermion content of the half-hypermultiplet, the theories with half-integral N_f are simply inconsistent because of a $\mathbf{Z}_2$ global anomaly [7]. In three dimensions, the situation is somewhat different. The theories with half-integral N_f exist, but for those theories the Chern-Simons coupling k cannot vanish, and the Coulomb branch that we will be studying in this paper does not exist. In fact, because of a global anomaly (see p. 309 of [8]), k is congruent to N_f modulo $\mathbf{Z}$, and can vanish only if N_f is integral.[3] So we will only consider integer N_f in this paper.

For the other case $G = U(1)$, we will consider the behavior with an arbitrary number M of hypermultiplets of charge one.

Until further notice, all of our hypermultiplets will have zero bare mass. After understanding the case of zero bare mass, we will make brief remarks on the role of the bare masses.

2.2. Behavior At Infinity

The starting point of the analysis is to understand what happens in the semi-classical

[3] In terms immediately relevant to this paper, the global anomaly pointed out in [8] would show up as follows. If N_f is half-integral, then the number of fermion zero modes in a monopole field would be odd. This appears to lead to a contradiction as amplitudes in a monopole field would change sign under a 2π rotation. The resolution of the paradox is not that the theory does not exist, but that when N_f is odd, k is half-integral and in particular non-zero; as non-zero k gives the photon a mass, finite action monopoles do not exist.

region of large $|\phi|$.

For the potential energy V to vanish means that the ϕ_i commute and so can be simultaneously diagonalized by a gauge transformation. This means for $SU(2)$ that one can take

$$\phi_i = \begin{pmatrix} a_i & 0 \\ 0 & -a_i \end{pmatrix} \tag{2.2}$$

for some a_i. The a_i are defined up to a Weyl transformation, which exchanges the two eigenvalues of the ϕ_i, and so acts as $a_i \to -a_i$. The space of zeroes of V is thus a copy of $\mathbf{R}^3/\mathbf{Z}_2$. For a complete description of the moduli space of vacua, one must also include an extra circle, parametrizing a fourth scalar σ which is dual to the photon. The Weyl group (which acts by charge conjugation) multiplies also the fourth scalar by -1. So the space of vacua at the classical level is $(\mathbf{R}^3 \times \mathbf{S}^1)/\mathbf{Z}_2$, where the $\mathbf{Z}_2$ multiplies all four coordinates by -1. The classical metric on the moduli space is a flat metric

$$ds^2 = \frac{1}{e^2}\sum_i d\phi_i^2 + e^2 d\sigma^2. \tag{2.3}$$

The factor of $1/e^2$ for the ϕ_i reflects the fact that (like the whole classical Lagrangian) the ϕ kinetic energy is of order $1/e^2$. The photon kinetic energy is likewise of order $1/e^2$, but after duality this turns into e^{+2} for σ. (Some constants in (2.3), omitted in this section for simplicity, are worked out in detail in section three.)

For $G = U(1)$, there is no Weyl group and the classical moduli space is simply $\mathbf{R}^3 \times U(1)$. For simplicity and to treat the two cases in parallel, we will postpone dividing by the Weyl group until the end of the discussion, and formulate the following as if classically one is on $\mathbf{R}^3 \times \mathbf{S}^1$. The region at infinity in $\mathbf{R}^3$ is homotopic to a two-sphere. Thus, topologically we have at infinity a product $\mathbf{S}^2 \times \mathbf{S}^1$ at the classical level. As one goes to infinity, the $\mathbf{S}^2$ grows (radius proportional to $|\phi|$) but the $\mathbf{S}^1$ has a fixed circumference of order e. The $\mathbf{S}^2$ is visible classically, but the $\mathbf{S}^1$, which appears via duality, is a more subtle part of the quantum story. The possibility exists that in the quantum theory, instead of a product $\mathbf{S}^2 \times \mathbf{S}^1$ at infinity, one has an $\mathbf{S}^1$ fiber bundle over $\mathbf{S}^2$. In fact, to describe such a fiber bundle, as noted in [5], the classical metric should be changed to something like

$$ds_Q^2 = \frac{1}{e^2}\sum_i d\phi_i^2 + e^2(d\sigma - sB_i(\phi)d\phi^i)^2, \tag{2.4}$$

where here B is the Dirac monopole $U(1)$ gauge field over $\mathbf{S}^2$, and *a priori* s is any integer. Because (2.4) differs from the classical metric only in terms of order e^2, quantum loop

corrections can be responsible for changing (2.3) to (2.4) and so for generating $s \neq 0$. In fact, if A is the undualized $U(1)$ gauge field, then the integer s would show up prior to duality in an interaction $s\epsilon^{\lambda\mu\nu}A_\lambda\epsilon_{ijk}\hat{\phi}^i\partial_\mu\hat{\phi}^j\partial_\nu\hat{\phi}^k$, where $\hat{\phi}^i = \phi^i/(\phi\cdot\phi)^{\frac{1}{2}}$; because it multiplies no power of e, this interaction could arise as a one-loop effect.

The integer s could thus, as was proposed in [5], be computed from a one-loop diagram. We will instead compute it mainly by counting fermion zero modes in a monopole field.

Non-Trivial $\mathbf{S}^1$ *Bundles Over* $\mathbf{S}^2$

As background, and to help in interpreting the results, let us recall the detailed description of non-trivial $\mathbf{S}^1$ bundles over $\mathbf{S}^2$. An $\mathbf{S}^1$ bundle over any base B (with oriented fibers) is classified topologically by the Euler class of the bundle, which takes values in $H^2(B,\mathbf{Z})$; as $H^2(\mathbf{S}^2,\mathbf{Z}) \cong \mathbf{Z}$, the possible bundles over $\mathbf{S}^2$ are labeled by an integer s, which was introduced in (2.4). For $B = \mathbf{S}^2$, the possible non-trivial bundles may be described in the following standard fashion.

The basic example is simply the three-sphere, regarded as a fiber bundle over $\mathbf{S}^2$. Let u_α, $\alpha = 1,2$ be two complex numbers with

$$|u_1|^2 + |u_2|^2 = 1. \tag{2.5}$$

The possible u_α parametrize a copy of $\mathbf{S}^3$. If we set

$$\vec{n} = \overline{u}\vec{\sigma}u, \tag{2.6}$$

with $\vec{\sigma}$ the usual Pauli σ matrices, then in a standard fashion one can show by consequence of (2.5) that $\vec{n}^2 = 1$. Thus the map from u to $\vec{n}$ is a map from $\mathbf{S}^3$ to $\mathbf{S}^2$. All $\vec{n}$'s arise, and for given $\vec{n}$, u is unique up to a $U(1)$ transformation

$$u_\alpha \to e^{i\theta}u_\alpha, \quad 0 \leq \theta \leq 2\pi. \tag{2.7}$$

Thus the space of u's for given $\vec{n}$ is a copy of $U(1) = \mathbf{S}^1$; the map from $\mathbf{S}^3$ to $\mathbf{S}^2$ exhibits $\mathbf{S}^3$ as a fiber bundle over $\mathbf{S}^2$ with fiber $\mathbf{S}^1$.

To introduce an arbitrary integer s, we begin now with $\mathbf{S}^3 \times \mathbf{S}^1$, labeling the $\mathbf{S}^1$ by an angle ψ ($0 \leq \psi \leq 2\pi$), and divide by a $U(1)$ group that acts by

$$u_\alpha \to e^{i\theta}u_\alpha, \quad \psi \to \psi + s\theta. \tag{2.8}$$

Let L_s be the quotient $(S^3 \times S^1)/U(1)$ with the given $U(1)$ action. Then L_s maps to S^2 by forgetting ψ; as we have noted above, the quotient of u-space by $u \to e^{i\theta}u$ is S^2. The fiber of the map to S^2 is a circle, so L_s is a circle bundle over S^2, for any s.

Let us next work out the topology of L_s. We note that L_0 is the trivial bundle $S^2 \times S^1$; in this case, the $U(1)$ in (2.8) does not act on the second factor in $S^3 \times S^1$, and dividing by it projects the first factor to S^2. In general, L_{-s} is mapped to L_s by $\psi \to -\psi$, so they have the same topology. Finally, for any $s > 0$, L_s is isomorphic to the "lens space" S^3/Z_s obtained by dividing S^3 by $u_\alpha \to e^{2\pi ik/s}u_\alpha$, $k = 0, 1, \ldots, s-1$. One sees this by using the θ in (2.8) to "gauge away" ψ, leaving a residual Z_s gauge symmetry that acts on u.

The lens space L_s has a manifest $SU(2) \times U(1)$ symmetry, where the $SU(2)$ acts in the standard fashion on the u_α and the $U(1)$ acts by $\psi \to \psi +$ constant. Any circle bundle over S^2 with $SU(2) \times U(1)$ symmetry will be equivalent to L_s with some value of s; we want a practical way to determine s. Suppose one is sitting at some point on S^2, say $\vec{n} = (0, 0, 1)$. In a standard basis of the Pauli matrices, this corresponds to $u_\alpha = (1, 0)$. The point $\vec{n} = (0, 0, 1)$ is invariant under a $U(1)$ subgroup of $SU(2)$, consisting of rotations about the third axis; on the u_α this acts by

$$J = \frac{i}{2}\left(u_1\frac{\partial}{\partial u_1} - u_2\frac{\partial}{\partial u_2}\right). \tag{2.9}$$

The $1/2$ is present because the u_α are in the spin one-half representation of $SU(2)$, and is consistent with the fact that $e^{2\pi J} = 1$ in acting on $\vec{n}$. Sitting at the point $u = (1, 0)$, that transformation is equivalent (modulo a "gauge transformation" (2.8)) to that generated by

$$\tilde{J} = -\frac{s}{2}\frac{\partial}{\partial\psi}. \tag{2.10}$$

So we get our criterion for determining the value of s: *a rotation around a given point $P \in S^2$ acts with charge $-s/2$ on the S^1 fiber over P.* In particular, such a rotation shifts ψ by πs, so that $SU(2)$ acts faithfully on L_s if s is odd, but $SU(2)/Z_2 = SO(3)$ acts if s is even.

Since, in the case of gauge group $G = SU(2)$, we are interested in dividing by the Weyl group, we should also discuss S^1 bundles over $\mathbf{RP}^2 = S^2/Z_2$. The transformation $\vec{n} \to -\vec{n}$ corresponds in terms of u_α to

$$\alpha : (u_1, u_2) \to (\overline{u}_2, -\overline{u}_1). \tag{2.11}$$

In the quantum field theories we want to study, the Weyl group also acts on ψ (the dual of the photon) by $\alpha(\psi) = -\psi$ (and this is in any case needed for consistency with the "gauge

invariance" (2.8)), so the circle bundles M_s over $\mathbf{RP}^2$ that we want are obtained simply by dividing L_s by a $\mathbf{Z}_2$ that acts as (2.11) on u and multiplies ψ by -1. We recall that in turn $L_s = \mathbf{S}^3/\mathbf{Z}_s$, where $\mathbf{Z}_s$ is generated by $\beta : u_\alpha \to e^{2\pi i/s} u_\alpha$. So M_s is the quotient of $\mathbf{S}^3$ by the group generated by α and β. There is no loss of generality in assuming that s is even, say $s = 2k$, since if s is odd, by replacing the group generators α and β by α and $\alpha\beta$, one can reduce to the even s case (the point being that if β is of odd order, then $\alpha\beta$ is of even order). The group generated by α and β is then a dihedral group Γ_k characterized by the relations

$$\begin{aligned} \alpha^2 = \beta^k &= -1 \\ \alpha\beta &= \beta^{-1}\alpha, \end{aligned} \tag{2.12}$$

where in the first relation -1 (which in our realization of the group acts by $u_\alpha \to -u_\alpha$) is understood as a central element of Γ_k. In the correspondence between finite subgroups of $SU(2)$ and the $A - D - E$ series of Lie groups, the group Γ_k corresponds to D_{k+2}, that is, to $SO(2k+4)$.

2.3. Behavior In A Monopole Field

One of the key aspects of $2+1$ dimensional gauge theories is that, as first explained by Polyakov twenty years ago [9], magnetic monopoles in unbroken $U(1)$ subgroups of the gauge group can appear as instantons.

The contribution of such an instanton is obviously proportional to e^{-I}, where I is the action of the instanton. A more subtle fact is that [9] if σ is the scalar dual to the $U(1)$ gauge field, then the instanton contribution also has a factor of $e^{-i\sigma}$, incorporating in the dual description the long range fields of the instanton. Beyond these general factors of $e^{-(I+i\sigma)}$, there may be additional factors coming, for instance, from fermion zero modes. For example [10], in $N = 2$ super Yang-Mills theory, with the instanton being a solution of the Bogomol'nyi-Prasad-Sommerfeld (BPS) monopole equation, the instanton is invariant under half of the four supercharges; the others generate two fermion zero modes. The field $I + i\sigma$ is the bosonic part of a chiral superfield. The effect of the fermion zero modes is that the function $e^{-(I+i\sigma)}$ must be integrated over chiral superspace, and is a superpotential rather than an ordinary potential.

In the present context of $N = 4$ super Yang-Mills theory, there are eight supercharges, of which half annihilate a supersymmetric instanton. As in [11,12], a supersymmetric solution in such a context will (if additional fermion zero modes are absent or can be absorbed)

generate a correction to the metric on moduli space, rather than a superpotential. We first consider the minimal $N = 4$ theory, without hypermultiplets, in which the fermion zero modes are generated entirely by the unbroken supersymmetries.

As usual in instanton physics, it is essential to analyze the symmetries of the instanton amplitude. We recall that the $N = 4$ gauge theory in three dimensions has a symmetry group $SU(2)_R \times SU(2)_N \times SU(2)_E$, with the supercharges transforming as $(\mathbf{2},\mathbf{2},\mathbf{2})$. The BPS monopole is invariant under the rotation group $SU(2)_E$ (mixed with a gauge transformation) and under $SU(2)_R$ (which only acts on fermions). However, the choice of a vacuum expectation value of the ϕ_i breaks $SU(2)_N$ to a subgroup $U(1)_N$ even before one considers monopoles; the BPS monopole is constructed using only a single real scalar in the adjoint, which can be chosen to be the field with an expectation value at infinity, and so is invariant under $U(1)_N$.

Under the unbroken group $SU(2)_R \times SU(2)_E \times U(1)_N$, the supercharges transform as $(\mathbf{2},\mathbf{2})^{1/2} \oplus (\mathbf{2},\mathbf{2})^{-1/2}$, where the superscript is the $U(1)_N$ charge, which takes half integral values on the supercharges because they transform as spin one-half under $SU(2)_N$. The BPS monopole is invariant under half of the supercharges in an $SU(2)_R \times SU(2)_E \times U(1)_N$-invariant fashion, so the unbroken supersymmetries must be, if we pick the sign of the $U(1)_N$ generator appropriately, the piece transforming as $(\mathbf{2},\mathbf{2})^{-1/2}$. The fermion zero modes therefore have the quantum numbers $(\mathbf{2},\mathbf{2})^{1/2}$. The instanton amplitude is schematically

$$\psi\psi\psi\psi\, e^{-(I+i\sigma)}, \tag{2.13}$$

where the ψ's are fermions of $U(1)_N = 1/2$. Note that if we consider antimonopoles instead of monopoles, the zero modes transform as $(\mathbf{2},\mathbf{2})^{-1/2}$, and (2.13) is replaced by

$$\widetilde{\psi}\widetilde{\psi}\widetilde{\psi}\widetilde{\psi}\, e^{-(I-i\sigma)}, \tag{2.14}$$

with $\widetilde{\psi}$ being fermions of $U(1)_N = -1/2$.

The $\psi\psi\psi\psi$ vertex carries $U(1)_N$ charge $4 \cdot (1/2) = 2$. One might be tempted to conclude that there is an anomaly in $U(1)_N$ conservation in a monopole field, but this is impossible as $U(1)_N$ is a subgroup of the simple group $SU(2)_N$. Rather, we must assign a transformation law to σ so that the instanton amplitude is invariant. Clearly, this means that the $U(1)_N$ generator must act on σ as $+2\partial/\partial\sigma$, meaning that in the notation (2.10) (including the factor of $1/2$ present there), $s = -4$ for the pure $N = 4$ gauge theory. The moduli space of the pure $N = 4$ theory therefore does not look at infinity like $\mathbf{S}^2 \times \mathbf{S}^1$ but like the lens space L_{-4} described in the last subsection.

Now, let us determine the value of s if one includes hypermultiplets in the two-dimensional representation of $SU(2)$. A doublet half-hypermultiplet in a monopole field has a single fermion zero mode (for the relevant index theorem see [13]), with the opposite sign of $U(1)_N$ from that of the vector multiplet zero modes. So with N_f hypermultiplets ($2N_f$ half-hypermultiplets), there are $2N_f$ zero modes, giving[4]

$$s = -4 + 2N_f. \tag{2.15}$$

For future use, we can also now work out the value of s for a $U(1)$ theory with hypermultiplets. There are no monopoles in the pure $U(1)$ gauge theory, but by thinking of s as the coefficient of a one-loop amplitude, and the fields of the $U(1)$ theory as a subset of the fields of an $SU(2)$ theory, one can infer the result for $U(1)$ from that for $SU(2)$. The $U(1)$ theory without hypermultiplets is free, so the vector multiplet contributes nothing. The hypermultiplet contribution in the $SU(2)$ theory with doublet hypermultiplets can be inferred from a one-loop diagram with the hypermultiplet running around the loop and external fields being vector multiplets. If we simply restrict the external fields to be in a $U(1)$ subalgebra, then the $SU(2)$ diagram with the internal fields being a doublet half-hypermultiplet turns into the $U(1)$ diagram with the internal fields being a hypermultiplet of charge one. (In particular, if we embed $U(1)$ in $SU(2)$ so that the doublet of $SU(2)$ has $U(1)$ charges ± 1, then a half-hypermultiplet of $SU(2)$ reduces to an ordinary charge one hypermultiplet of $U(1)$.) The value of s for a $U(1)$ theory with M hypermultiplets of charge 1 is thus obtained by replacing 4 by 0 and $2N_f$ by M in (2.15):

$$s = M. \tag{2.16}$$

Going back to the $SU(2)$ theory, we see from (2.15) that s is always even. This means (as noted following (2.10)) that it is not $SU(2)_N$ but $SU(2)_N/\mathbf{Z}_2$, which we will call $SO(3)_N$, that acts faithfully on the moduli space $\mathcal{M}$ of vacua. Furthermore, $s \neq 0$ except

[4] It is curious that in four-dimensional $N = 2$ super Yang-Mills theory, the analogous counting of zero modes in an instanton field gives a factor of $-8+2N_f$, instead of $-4+2N_f$. The difference arises because the half-hypermultiplet has the same number of fermion zero modes in a three-dimensional monopole or four-dimensional instanton, but the vector multiplet has twice as many zero modes in the four-dimensional case – four generated by ordinary supersymmetries that have an analog in the three-dimensional problem, and four more by superconformal symmetries that do not.

for $N_f = 2$. When $s \neq 0$, $SO(3)_N$ acts non-trivially on the scalar σ that is dual to the photon. This means that the generic $SO(3)_N$ orbit is three-dimensional. Also, because $SU(2)_N$ is a group of R symmetries, the three complex structures of the hyper-Kähler manifold $\mathcal{M}$ are rotated by the $SO(3)_N$ action. In [6], four-dimensional hyper-Kähler manifolds with an $SO(3)$ action that rotates the complex structures and has generic three-dimensional orbits were classified. From what has just been said, all of our metrics will appear on their list except for $N_f = 2$.

2.4. The Metric On Moduli Space

Before comparing to results of [6], and to expectations from string theory, let us ask what sort of metrics we expect on the moduli space $\mathcal{M}$, for various N_f. First we consider the case of gauge group $SU(2)$. The starting point is the classical answer, the flat metric on $(\mathbf{R}^3 \times \mathbf{S}^1)/\mathbf{Z}_2$. There is then a one loop correction to the structure at infinity, for $N_f \neq 2$. The effect of this correction is that "infinity" for $N_f \neq 2$ looks not like $(\mathbf{S}^2 \times \mathbf{S}^1)/\mathbf{Z}_2$ but like $L_s/\mathbf{Z}_2$, with $s = 2N_f - 4$.

Perturbation corrections to the metric on $\mathcal{M}$ are entirely determined by the one-loop correction plus the non-linear terms in the Einstein equations. (This is analogous to the fact that in four dimensions, perturbative corrections beyond one loop are forbidden by holomorphy.) This may be proved as follows. A "new" k-loop correction to the metric would be a self-dual solution of the linearized Einstein equations on $\mathcal{M}$ (since hyper-Kähler metrics automatically obey the Einstein equations and are self-dual) and would be $SU(2)_N \times U(1)$ invariant (since perturbation theory has this symmetry). Imposing the $U(1)$ (which acts by translation of σ, the dual of the photon) gives a dimensional reduction of the Einstein equations to three-dimensional scalar-Maxwell equations on $\mathbf{R}^3$, with $SU(2)_N$ acting by rotations. The only rotationally-invariant mode of the Maxwell field in three dimensions is the "magnetic charge," the integer s that we already encountered at one loop. The s-wave mode of the scalar is related by self-duality of the metric to the "magnetic charge" so is likewise determined at one loop. Thus, the whole perturbation series is determined by the one-loop term plus the equations of hyper-Kähler geometry.

As in four dimensions, however, there can be instanton corrections to the metric, the relevant instantons here being BPS monopoles. For $N_f = 0$, it is clear that instantons contribute to the metric. In fact, the non-derivative $\psi\psi\psi\psi e^{-(I+i\sigma)}$ vertex described above is part of the supersymmetric completion of a correction to the metric. So there is a one-instanton contribution to the metric for $N_f = 0$. What happens for $N_f > 0$? There will

be hypermultiplet zero modes in a monopole field, so that the one-instanton field gives a vertex $\psi^4\chi^{2N_f}e^{-(I+i\sigma)}$ (χ being fermion components of the hypermultiplet, of opposite $U(1)_N$ charge from ψ), which has too many fermions to be related by supersymmetry to the metric on $\mathcal{M}$. A correction to the metric still might arise from an r-instanton contribution with $r > 1$. Since the $U(1)_N$ charge carried by vector or hypermultiplet zero modes could be determined from an index theorem and is proportional to r, an r-instanton contribution will give in the first instance a vertex $\psi^{4r}\chi^{2rN_f}e^{-r(I+i\sigma)}$. However, in integrating over bosonic collective coordinates and computing various quantum corrections, ψ and χ zero modes of opposite charge might pair up and be lifted. This process might generate a vertex $\psi^4 e^{-r(I+i\sigma)}$ – which would be related by supersymmetry to a correction to the metric – if $2rN_f = 4r - 4$ or in other words

$$r = \frac{1}{1 - N_f/2}. \tag{2.17}$$

But we also need r to be a positive integer, since BPS monopoles only exist for such values of r. (Considering anti-monopoles instead of monopoles reverses all quantum numbers and leads to the same restriction on r; in fact, since the metric is real, there is an anti-monopole contribution if and only if there is a monopole contribution.) So the only cases are $N_f = 0$ and $r = 1$, or $N_f = 1$ and $r = 2$.[5] (The fact that only one value of r appears we take to mean that the exact metric is determined by this one contribution together with the non-linear Einstein equations.)

In sum, then, for $N_f = 0$ we expect a metric with a perturbative contribution that gives $s = -4$, plus monopole corrections, and for $N_f = 1$ we expect a metric with a perturbative correction that gives $s = -2$, plus monopole corrections. For $N_f = 2$, the perturbative and monopole corrections both vanish, and the quantum metric should very plausibly coincide with the classical metric, that is, the flat metric on $(\mathbf{R}^3 \times \mathbf{S}^1)/\mathbf{Z}_2$. For $N_f > 2$, there is a perturbative correction at infinity, with $s = 2N_f - 4$, and the monopole corrections vanish.

String Theory And Field Theory

[5] For $N_f > 0$, there is a symmetry reason that only even r can contribute to the metric. The relevant symmetry is the one that changes the sign of just one of the half-hypermultiplets (and so extends $SO(2N_f)$ to $O(2N_f)$). Since in a one-monopole field the fermion zero mode measure is odd under this symmetry, the symmetry must be defined to shift σ by π. The χ zero modes are odd under this $\mathbf{Z}_2$ for odd r and $N_f > 0$, implying that they cannot be lifted.

Let us now recall the expectations from string theory [5]:

(1) For $N_f = 0, 1$, the metric on moduli space is expected to be complete and smooth.

(2) For $N_f \geq 2$, one expects the metric to have a D_{N_f} singularity.

To clarify the meaning of the second statement, recall that for $N_f > 2$, the D_{N_f} singularity is the singularity obtained by dividing $\mathbf{C}^2$ by the dihedral group Γ_{N_f-2}. This group was introduced earlier and is generated by elements α, β with $\alpha^2 = \beta^{N_f-2} = -1$ (the symbol -1 simply denotes a central element of the group), and $\alpha\beta = \beta^{-1}\alpha$. For $N_f = 2$, something special happens: D_2 is the same as $A_1 \times A_1$, or $SU(2) \times SU(2)$, so a D_2 singularity should be simply a pair of A_1 singularities, that is, $\mathbf{Z}_2$ orbifold singularities.

Let us now make a preliminary comparison of the string theory statements with what we have learned from field theory. For $N_f > 2$ we have found that *topologically* the moduli space $\mathcal{M}$ looks near infinity like $\mathbf{C}^2/\Gamma_{N_f-2}$. (The *metric* near infinity on $\mathcal{M}$ does not look like the obvious flat metric on $\mathbf{C}^2/\Gamma_{N_f-2}$.) We actually want to express the singularity near the origin rather than the behavior at infinity in terms of Γ_{N_f-2}; we will do this momentarily. Likewise, for $N_f = 2$, the moduli space that we claim, namely $(\mathbf{R}^3 \times \mathbf{S}^1)/\mathbf{Z}_2$, indeed has a pair of $\mathbf{Z}_2$ orbifold singularities (from the two $\mathbf{Z}_2$ fixed points on $\mathbf{R}^3 \times \mathbf{S}^1$) as expected.

For $N_f = 2$, a more precise comparison of the string theory and field theory results is possible. In fact, from string theory one can see why the moduli space should be $(\mathbf{R}^3 \times \mathbf{S}^1)/\mathbf{Z}_2$ with the flat metric, just as we have found from field theory. There are many possible approaches to this result, but a quick way is to compactify M-theory on $\mathbf{R}^7 \times$ K3 and consider a two-brane whose world-volume fills out $\mathbf{R}^3 \times \{p\}$, where $\mathbf{R}^3$ is a linear subspace of $\mathbf{R}^7$ and p is a point in K3. Consider the quantum field theory on the world-volume of this two-brane. The moduli space of vacua of this theory is the K3 manifold itself, which parametrizes the choice of p. By arguments as in [5], in various limits in which heavier modes decouple, this theory will reduce at low energy to the three-dimensional $N = 4$ super Yang-Mills theory with gauge group $SU(2)$. In particular, in K3 moduli space, there is a locus in which the K3 looks like $(\mathbf{T}^3 \times \mathbf{S}^1)/\mathbf{Z}_2$ with the flat metric. Taking the $\mathbf{T}^3$ to be large and restricting to a neighborhood of a $\mathbf{Z}_2$ fixed point in $\mathbf{T}^3$, one gets a piece of the K3 that looks like a flat $(\mathbf{R}^3 \times \mathbf{S}^1)/\mathbf{Z}_2$. In this piece of the K3, there are two A_1 singularities, giving on $\mathbf{R}^7$ a gauge symmetry $SU(2) \times SU(2) = SO(4)$, which will be observed as a global symmetry along the two-brane world-volume. The global symmetry means that the world-volume theory is the $N_f = 2$ theory, and by construction

its moduli space is $(\mathbf{R}^3 \times \mathbf{S}^1)/\mathbf{Z}_2$ with flat metric, as was claimed above.

Comparison To Exact Metrics

To learn more, we compare now to what is known [6] about four-dimensional hyper-Kähler manifolds with an $SO(3)$ symmetry of the appropriate kind. Assuming that one wants a metric with at most isolated singularities, the possibilities are extremely limited. For a smooth manifold with these properties, there are only two possibilities. One (sometimes called the Atiyah-Hitchin manifold; it was studied in [6] because of its interpretation as the two-monopole moduli space) is a complete hyper-Kähler manifold $\mathcal{N}$, with fundamental group $\mathbf{Z}_2$. Topologically, $\mathcal{N}$ looks like a two-plane bundle over $\mathbf{RP}^2$. The structure at infinity looks like $L_{-4}/\mathbf{Z}_2$, corresponding to a one-loop correction with $s = -4$. The other possibility, which we will call $\overline{\mathcal{N}}$, is the simply-connected double cover of $\mathcal{N}$; it is topologically a two-plane bundle over $\mathbf{S}^2$, and the structure at infinity looks like $L_{-2}/\mathbf{Z}_2$, corresponding to a one-loop correction with $s = -2$. Since we found $s = -4$ and $s = -2$ for the two cases – $N_f = 0, 1$ – for which we expect a smooth metric, we propose that the $N_f = 0$ theory has moduli space $\mathcal{N}$, and the $N_f = 1$ theory has moduli space $\overline{\mathcal{N}}$.[6] We will discuss in more detail the fundamental group and its physical interpretation later.

Now let us discuss the possible singular metrics. According to [6], a hyper-Kähler metric with the requisite sort of symmetry and only isolated singularities is severely constrained. Such a manifold is topologically $\mathbf{C}^2/\Gamma$, where Γ is a cyclic or dihedral subgroup of $SU(2)$ (or if the metric is flat, Γ may be any finite subgroup). As for the metric on $\mathbf{C}^2/\Gamma$, it may be flat, but there is a more general possibility. As the space at infinity looks like $\mathbf{S}^3/\Gamma$, which is an $\mathbf{S}^1$ bundle over $\mathbf{S}^2$ or $\mathbf{RP}^2$, one can have a metric – a variant of the Taub-NUT metric – in which the $\mathbf{S}^1$ approaches at infinity an arbitrary radius R.[7] R can be varied simply by multiplying the metric by a constant; the flat metric on $\mathbf{C}^2/\Gamma$ is

[6] The extra $\mathbf{Z}_2$ symmetry of $\overline{\mathcal{N}}$ which we mod out by to get $\mathcal{N}$ is the global symmetry of the microscopic $N_f = 1$ theory, mentioned earlier, that prevents a one-monopole correction to the metric for $N_f = 1$. That this symmetry acts freely on the moduli space – even in the strong coupling region – is related to the discussion of confinement that we give later.

[7] There are a few subtleties here relative to assertions in [6] that come reflect the fact that the authors of [6] wanted smooth metrics with an $SO(3)_N$ action, rather than $SU(2)_N$. They therefore construct the Taub-NUT metric with a $\mathbf{Z}_2$ orbifold singularity, and do not make explicit that it has a smooth double cover (acted on by $SU(2)_N$ instead of $SO(3)_N$) that can be divided by any cyclic or dihedral group Γ (the quotient is acted on by $SO(3)_N$ except in the case that Γ is cyclic of odd order). We here need these slight generalizations.

obtained in the $R \to \infty$ limit. In the present problem, we want R of order e, since that is the circumference of the circle obtained by dualizing the photon.

Given that the $SU(2)$ gauge theory with $N_f > 2$ hypermultiplets has moduli space $\mathbf{C}^2/\Gamma$ for some Γ, all that really remains is to identify Γ. But we have determined that at infinity the structure looks like $\mathbf{S}^3/\Gamma_{N_f-2}$, so $\Gamma = \Gamma_{N_f-2}$. Hence the moduli space has a Γ_{N_f-2} orbifold singularity at the origin. Since, in the association of subgroups of $SU(2)$ with $A - D - E$ groups, $SO(2N_f) = D_f$ corresponds to Γ_{N_f-2}, we have confirmed from field theory the string theory claim [5] that the theory with N_f hypermultiplets has a D_{N_f} singularity.

It is easy to consider $U(1)$ gauge theories in a similar way. We saw that the $U(1)$ gauge theory with M charge one hypermultiplets has a one-loop correction with $s = M$, and that the moduli space at infinity looks like $\mathbf{C}^2/\mathbf{Z}_M$. Hence in this case, $\Gamma = Z_M$, and there is a $\mathbf{Z}_M$ orbifold singularity at the origin. This confirms the claim [5] that the $U(1)$ theory with M charge one hypermultiplets has a $\mathbf{Z}_M$ (or A_{M-1}) singularity in the strong coupling region. For $M = 1$, this means that the moduli space is completely smooth. The metric for $M = 1$ is uniquely determined by the symmetries, smoothness, and asymptotic behavior to be the smooth Taub-NUT metric.

The Taub-NUT-like metrics on $\mathbf{C}^2/\Gamma$ have a very simple structure. They are given by an elementary closed formula ([6], p. 76). In fact, in addition to the $SO(3)$ symmetry, the Taub-NUT metrics have an extra $U(1)$ symmetry that acts by translation of the scalar σ which is dual to the photon; this is a precise statement of the absence of monopole corrections. On the other hand, the metric on $\mathcal{N}$ or its double cover, while exponentially close to a Taub-NUT type metric at infinity, has ([6], p. 77) exponentially small corrections which violate the extra $U(1)$ and which we interpret as monopole corrections.

It may seem somewhat odd that the metric for $N_f > 1$ is so different from what it is for $N_f \leq 1$. It is perhaps comforting, therefore, that ([6], p. 56) in a sense, the manifold $\overline{\mathcal{N}}$ is a kind of analytic continuation of the D_{N_f} space to $N_f = 1$. In fact, as a complex manifold, the Taub-NUT space for D_{N_f} is described by the equation

$$y^2 = x^2 v - v^{N_f - 1}. \tag{2.18}$$

This has a D_{N_f} singularity at $y = x = v = 0$, for $N_f \geq 2$, and two A_1 singularities (at $y = v = 0$, $x = \pm 1$) for $N_f = 2$. If one simply sets $N_f = 1$, the same formula does give the complex structure of $\overline{\mathcal{N}}$ – though there is no longer a singularity. We will return to this formula for the complex structure in section three.

2.5. Some Physical Properties

We will use these results to discuss some physical properties of these models.

First we consider symmetry breaking. For any $N_f \neq 2$, on the generic orbit $SO(3)_N$ is broken to a finite subgroup. (For $N_f = 2$, the generic unbroken group is $O(2)$.) What happens in the strong coupling region? For $N_f \geq 2$, the $SO(3)_N$ is completely restored at the strong coupling orbifold points. For $N_f = 0, 1$, this is not so. The most degenerate $SO(3)_N$ orbit in $\mathcal{N}$ is a copy of $\mathbf{RP}^2$; in $\overline{\mathcal{N}}$ the most degenerate orbit is a copy of $\mathbf{S}^2$. So the maximal unbroken subgroup of $SO(3)_N$ is $O(2)$ or $SO(2)$ for $N_f = 0$ and $N_f = 1$.

We now turn to consider the significance of the fundamental group of $\mathcal{N}$ and $\overline{\mathcal{N}}$.

The $N_f = 0$ theory has no fields with half-integral gauge quantum numbers, so it can be meaningfully probed with external charges in such a representation. Let us consider the fields that would be produced by such a charge. In terms of the photon, an external charge produces in $2 + 1$ dimensions an electric field varying as $1/r$; to be more precise, in Cartesian coordinates x_a, $a = 1, 2$ with $r = \sqrt{x_1^2 + x_2^2}$, the electric field is $E_a \sim x_a/r^2$. After performing a duality transformation, the external charge becomes a vortex for the dual scalar σ; that is, σ jumps by 2π in circumnavigating the external charge. The energy of such a vortex has a potential logarithmic infinity both at short distances and at large distances. The behavior at short distances should be cut off for our present purposes, but the behavior at long distances is physically significant; it reflects logarithmic confinement of electric charge in weakly coupled $2 + 1$-dimensional QED.

To describe this situation in a more general language, we can say that along a circle that runs around the external charge, the fields make a loop in the moduli space $\mathcal{M}$ of vacua. If this loop is trivial in $\pi_1(\mathcal{M})$, then even in the low energy theory one can see that the "vorticity" produced by the external charge is not really conserved, and that the external charge can be screened. If the loop is non-trivial in $\pi_1(\mathcal{M})$, then the external charge cannot be screened in the low energy theory, though it is still conceivable that it can be screened by massive modes that have been integrated out in deriving the low energy theory.

For $N_f = 1$, the loop produced by an external charge is automatically trivial in $\pi_1(\mathcal{M})$ since in fact $\mathcal{M} = \overline{\mathcal{N}}$ is simply connected. This is in accord with the fact that the $N_f = 1$ theory has isospin one-half fields, so that external charges can be screened. For $N_f > 1$, in order to make this argument, one has to decide how a low energy physicist would understand the singularities. However, at least for $N_f > 2$ where the moduli space

(being a cone $\mathbf{C}^2/\Gamma$) is contractible, it is plausible that a physicist knowing only the low energy structure would determine that the external charges can be screened.

For $N_f = 0$, however, the answer is quite different. The loop C produced by an external charge is the generator of $\pi_1(\mathcal{N})$, as we will see momentarily, so the external charge cannot be screened either in the low energy theory or microscopically. In showing that C is the generator of $\pi_1(\mathcal{N})$, the point is that in the analysis in chapter nine of [6], the fundamental group at infinity in the moduli space is generated by two circles, defined respectively by one-forms that were called σ_1 and σ_2. (Loops wrapping once around these circles give our standard generators α and β of the fundamental group at infinity, which for $\mathcal{N}$ is what we called Γ_2.) Moreover, the metric was described in terms of functions a, b, and c. Since at infinity in the moduli space, b approaches a limit and and a and c diverge, it is the circle defined by σ_2 that corresponds in the semi-classical region to the photon and so to the loop C. On the other hand, on the exceptional $\mathbf{RP}^2$ orbit, $a = 0$ and $b \neq 0$. Hence the σ_1 circle can be contracted in the interior of $\mathcal{N}$, and the σ_2 circle – that is the loop produced by the external charge – survives as the generator of $\pi_1(\mathcal{N})$, as we wanted to show.[8]

One might ask, for $N_f = 0$, what sort of confinement is observed in this theory. As long as the vacua parametrized by $\mathcal{N}$ are precisely degenerate, the energy of a pair of external charges separated a distance ρ will grow only as $\log \rho$, since the energy of a vortex configuration of massless fields has only a logarithmic divergence in the infrared; such a vortex configuration will form between the two external charges. However, suppose that one makes a generic small perturbation of the $N_f = 0$ theory that lifts enough of the vacuum degeneracy so that a loop that generates $\pi_1(\mathcal{N})$ cannot be deformed into the space of exact minima of the energy. (It does not matter whether the perturbation preserves some supersymmetry.) Then the fields on a contour that encloses one external charge but not the other cannot be everywhere at values that exactly minimize the energy. In such a situation, a sort of string will form between the external charges (one might think of it as a domain wall ending on them), and the energy will grow linearly in ρ. Thus, like the four-dimensional $N = 2$ theory [1] with $N_f = 0$, the three-dimensional $N = 4$ theory with $N_f = 0$ does not have linear confinement but gives linear confinement after a generic small

[8] This also means that the $\mathbf{Z}_2$ symmetry of $\overline{\mathcal{N}}$, by which one would divide to get $\mathcal{N}$, is a π shift of the scalar σ dual to the photon; as explained in connection with (2.17), this symmetry must be accompanied by a sign change of one half-hypermultiplet.

perturbation.

Finally, note that the association here of confinement with $\pi_1(\mathcal{N})$ is somewhat analogous to the association in some four-dimensional $SO(N)$ gauge theories of confinement with π_2 of a moduli space [14].

Another issue of physical interest stems from π_2 of the moduli spaces. Since these groups are non-trivial, the low energy theory on the moduli spaces can have solitons. There is no reason to expect these solitons to be BPS-saturated at the generic vacuum on the moduli space. Furthermore, their detailed properties can depend on higher dimension operators which are not considered in this paper. Nevertheless, the topology of the moduli spaces supports solitons which are localized excitations in the three-dimensional theory. Their interest is related to the fact that most of the global symmetry of the theory does not act on the Coulomb branch of the moduli space. For example, all the light fields are invariant under the global $SU(N_f)$ symmetry of the $U(1)$ gauge theories or the $SO(2N_f)$ of the $SU(2)$ gauge theories. We claim that these solitons are in the adjoint representations of these groups. This is easiest to establish using the string theory viewpoint [3-5]. The M-theory two-brane can wrap non-trivial two cycles to yield zero-branes which are $SU(N_f)$ or $SO(2N_f)$ gauge bosons. Our solitons can be interpreted as bound states of such a gauge boson with a two brane at a generic point in its moduli space. From a three-dimensional viewpoint, these solitons are bound states of the elementary hypermultiplets. They are bound by the logarithmic Coulomb forces to neutral composites. This situation is similar to current algebra in four dimensions. There, the non-trivial π_3 of the moduli space leads to solitons. Their topological charge is identified with the global $U(1)$ baryon number [15,14], which exists in the microscopic theory. In both situations the global symmetry of the microscopic theory manifests itself through the topology of the moduli space.

2.6. Incorporation Of Bare Masses

We will now try to discuss the incorporation of bare masses for the hypermultiplets.

In four dimensions, the bare mass of a hypermultiplet is a complex parameter, with two real components, while in three dimensions a third parameter appears. This arises as follows. In four dimensions, the group that we have called $SO(3)_N$ is reduced to an $SO(2)$ group, usually called $U(1)_R$. A complex hypermultiplet mass parameter carries $U(1)_R$ charge, or equivalently, its real and imaginary parts transform as a vector of $SO(2)$. In three dimensions, as the $SO(2)$ is extended to $SO(3)_N$, the mass vector gets a third component to fill out the vector representation of $SO(3)_N$. It is easy to reach the same conclusion

by viewing the masses as expectation values of background fields in vector multiplets that gauge some of the flavor symmetries [16]. Since all the bosons in the vector multiplets originate from gauge fields in six dimensions and since the masses are scalars in three dimensions, they must be in a vector representation of $SO(3)_N$. This interpretation also makes it obvious that they are in the adjoint representation of the flavor group $SO(2N_f)$ ($SU(N_f)$ in the $U(1)$ gauge theory). Requiring that the background fields should preserve supersymmetry means that they can all be gauged to a common maximal torus of the flavor group, and this is why there are precisely N_f triplets of mass parameters.

In general, in four-dimensional $N = 2$ theories, the moduli space of the Coulomb branch of vacua parametrizes [1] a family of complex tori. The total space of the family is a complex manifold $\mathcal{M}'$ with a holomorphic two-form ω, and, according to section 17 of [2], the dependence of $\mathcal{M}'$ on the masses is determined by the requirement that the periods of ω vary linearly in the masses.

The moduli space $\mathcal{M}$ of vacua in three dimensions is a hyper-Kähler manifold which in fact is the analog of $\mathcal{M}'$; this relation will be elucidated in the next section. The analogs of ω are the three covariantly constant two-forms ω_a, $a = 1, 2, 3$ of the hyper-Kähler manifold $\mathcal{M}$ (two of which correspond to the real and imaginary parts of ω). These transform in the vector representation of $SO(3)_N$. We normalize them in the semi-classical region of large ϕ to be independent of the hypermultiplet bare masses.

The natural three-dimensional analog of the four-dimensional statement that the periods of ω vary linearly in the masses is then a three-dimensional statement that the periods of the ω_a should vary linearly with the masses. Notice that such an assertion is compatible with $SO(3)_N$, as both the mass parameters and the two-forms transform as $SO(3)_N$ vectors. A direct field theory justification of this principle in three-dimensional $N = 4$ models is not clear at the moment. [9] We will here simply accept this principle and discuss its implementation for the $SU(2)$ theory with N_f doublets.

First we consider the case $N_f = 0$. The moduli space $\mathcal{M}$ that we proposed is homotopic to the two-manifold $\mathbf{RP}^2$. As this is unorientable, the two-dimensional homology of this manifold has rank zero, and a closed two-form has no periods. Thus, there is no

[9] But note that for those three-dimensional $N = 4$ models that have been related to string theory [5], which include those studied in detail in this paper, the fact that the periods of ω_a vary linearly in the masses follows from the fact that the periods of the ω_a are the natural coordinates parametrizing hyper-Kähler metrics on K3.

way to perturb this model to include mass parameters. That is just as well, since no hypermultiplets are present in the model.

Now consider $N_f = 1$. The moduli space $\overline{\mathcal{N}}$ is homotopic to $\mathbf{S}^2$; a closed two-form on this manifold has a single period, the integral over $\mathbf{S}^2$. Thus, a single "mass vector" can be introduced, compatible with the fact that the model has $N_f = 1$. In fact, the hyper-Kähler metric that is the appropriate deformation of $\overline{\mathcal{N}}$ to include masses has been described explicitly by Dancer [17]. Dancer constructs a deformation $\overline{\mathcal{N}}_{\vec{\lambda}}$ of the hyper-Kähler manifold $\mathcal{N}$ depending on an $SO(3)_N$ vector $\vec{\lambda}$. That the periods of $\vec{\omega}$ vary linearly with $\vec{\lambda}$ is a consequence of Dancer's construction of $\overline{\mathcal{N}}_{\vec{\lambda}}$ as a $U(1)$ hyper-Kähler quotient (of a hyper-Kähler eight-manifold) with $\vec{\lambda}$ as the constant term in the moment map. We will return to Dancer's manifold in section three.

For $N_f > 1$, the real homology of the resolution of the D_{N_f} singularity is known to have two-dimensional homology of rank N_f, so that N_f mass vectors can be introduced.

It is now by the way clear, even without solving for hyper-Kähler metrics as in [6], that for $N_f > 2$ the metric on the moduli space of vacua must be singular. An $SO(3)$ action with three-dimensional orbits on a four-manifold constrains the topology so much that there could not be $N_f > 2$ independent two-cycles, unless some or all are collapsed at a singularity.

Even though we have not determined the metric, it is easy to see how the masses affect the singularity of the moduli space. First, physically, we expect that if only $k < N_f$ masses vanish the singularity should be D_k. Furthermore, if n masses are equal and non-zero we expect an A_{n-1} singularity (classically, upon adjusting the Higgs field to cancel the bare mass of some of the fields, we get a $U(1)$ gauge theory with n massless hypermultiplets, which gives an A_{n-1} singularity, from which a Higgs branch emanates). This is exactly the behavior after the D_{N_f} singularity is blown up. The N_f mass parameters are the parameters labeling the blow-up of the singularity.

3. Field Theory On $\mathbf{R}^3 \times \mathbf{S}^1_R$

In the remainder of this paper, we will mainly be studying four-dimensional $N = 2$ super Yang-Mills theory formulated on a space-time $\mathbf{R}^3 \times \mathbf{S}^1_R$, where $\mathbf{S}^1_R$ is a circle of circumference $2\pi R$. We focus on the case of gauge group $G = SU(2)$, with $N_f \leq 4$ matter hypermultiplets in the two-dimensional representation. (The upper bound on N_f, which has no analog in three dimensions, ensures a non-positive beta function in the four-

dimensional theory.) We recall that the bosonic fields of the theory are the $SU(2)$ gauge field and a complex scalar ϕ in the adjoint representation.

To begin with, we consider what happens for R much greater than the natural length scale of the four-dimensional theory (which is set by an appropriate bare mass, order parameter, or by the scale parameter Λ introduced in quantizing the theory). In this regime, one can borrow four-dimensional results. The moduli space of vacua in four dimensions is [1,2] the complex u plane, where $u = \mathrm{Tr}\,\phi^2$ is the natural order parameter. The massless bosons are u and an Abelian photon, which we will call A. The effective action for A, in four dimensions, looks like

$$L = \int d^4x \left(\frac{1}{4e^2} F_{\mu\nu} F^{\mu\nu} + \frac{i\theta}{32\pi^2} F_{\mu\nu} \widetilde{F}^{\mu\nu} \right). \tag{3.1}$$

Here $\mu, \nu = 1 \ldots 4$ are space-time indices, $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, and $\widetilde{F}_{\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\alpha\beta} F^{\alpha\beta}$. e and θ are functions of u and were determined in [1,2]. A key point in computing them was to interpret e and θ as determining the complex structure of an elliptic curve E. The most natural convention in defining E, in the case $N_f \neq 0$, was explained on pp. 487-8 of [2]. E is the complex torus with τ parameter

$$\tau = \frac{\theta}{\pi} + \frac{8\pi i}{e^2}. \tag{3.2}$$

E is isomorphic in other words to $\mathbf{C}/\Gamma$, where Γ is the lattice in the complex plane generated by the complex numbers 1 and τ. For $N_f = 0$, one can also conveniently use, as in [1], an isogenous elliptic curve with τ replaced by $\tau/2$, but this is awkward if one wishes to let N_f vary.

Once we work on $\mathbf{R}^3 \times \mathbf{S}^1_R$, there is a small subtlety about defining the theory in the $N_f = 0$ case. In quantizing a gauge theory, one must divide by the group of gauge transformations. But precisely what gauge transformations do we want to divide by? Do we want to consider gauge transformations which, in going around the $\mathbf{S}^1_R$, are single-valued in $SU(2)$, or gauge transformations that would be single-valued only in $SO(3)$? For $N_f \neq 0$, since there are fields that are not invariant under the center of $SU(2)$, the gauge transformations that are single-valued only in $SO(3)$ are not symmetries, so one is forced to divide only by the smaller group. For $N_f = 0$, one is free to divide by either the larger or the smaller group; the two choices give slightly different (but obviously closely related) quantum theories, the moduli space of vacua of one being a double cover of the moduli space of vacua of the other. To obtain results that vary smoothly with N_f, in

quantizing the $N_f = 0$ theory, we will divide only by the "small" gauge group, the gauge transformations that are single-valued in $SU(2)$. It will be seen, as one might expect, that this choice will agree with (3.2), while the other choice has the effect of replacing τ by $\tau/2$.

To determine what happens in compactification on $\mathbf{R}^3 \times \mathbf{S}^1_R$ for very large R, we simply expand (3.1) in terms of fields that are massless in the three-dimensional sense. These are the fourth component A_4 of the gauge field and also a three-dimensional photon A_i, $i = 1, \dots, 3$ which is dual to another scalar σ. First of all, the gauge field A in (3.1) is normalized (see the discussion of eqn. (3.12) in [1]) so that fields in the two-dimensional representation of $SU(2)$ have half-integral charges, and the magnetic flux of a magnetic monopole is 4π. Because of the first assertion, and the fact that we are only dividing by the gauge transformations that are single-valued in $SU(2)$, $\int_{\mathbf{S}^1_R} A$ is gauge-invariant modulo 4π. We therefore write the massless scalar coming from A_4 as

$$A_4 = \frac{b}{\pi R}, \tag{3.3}$$

where b is an angular variable, $0 \leq b \leq 2\pi$.

The effective action becomes in terms of b and the three-dimensional photon

$$L = \int d^3x \left(\frac{1}{\pi R e^2} |db|^2 + \frac{\pi R}{2e^2} F_{ij}^2 + \frac{i\theta}{8\pi^2} \epsilon^{ijk} F_{ij} \partial_k b \right). \tag{3.4}$$

The next issue is to dualize the three-dimensional photon. To do so, introduce a two-form B_{ij} with (in addition to standard gauge invariance $A_i \to A_i + \partial_i w$) an extended gauge-invariance

$$A_i \to A_i + C_i, \quad B_{ij} \to B_{ij} + \partial_i C_j - \partial_j C_i \tag{3.5}$$

where C is an arbitrary connection on a line bundle, and introduce also a scalar field σ with $0 \leq \sigma \leq 2\pi$. Replace the F-dependent part of (3.4) by

$$\int d^3x \left(\frac{\pi R}{2e^2} (F_{ij} - B_{ij})^2 + \frac{i\theta}{8\pi^2} \epsilon^{ijk} (F_{ij} - B_{ij}) \partial_k b + \frac{i}{8\pi} \epsilon^{ijk} B_{ij} \partial_k \sigma \right). \tag{3.6}$$

The point of this is that if one first integrates over σ, then σ serves as a Lagrange multiplier, enabling one to set $B = 0$ modulo an extended gauge transformation (3.5); in this way one reduces (3.6) to the relevant part of (3.4). On the other hand, one can use the extended gauge invariance (3.5) to set $F = 0$, whereupon after integrating over B one gets a dual description with a massless scalar σ. The dual formula for the low energy action is in fact

$$\widetilde{L} = \int d^3x \left(\frac{1}{\pi R e^2} |db|^2 + \frac{e^2}{\pi R (8\pi)^2} \left| d\sigma - \frac{\theta}{\pi} db \right|^2 \right). \tag{3.7}$$

This is a sigma model in which the target space is a two-torus E with the τ parameter given in (3.2). (Had we chosen to divide by the "big" group of gauge transformations, b would have been replaced by $b/2$, and τ by $\tau/2$, giving formulas related to the other description of the $N_f = 0$ theory.)

Moreover, the area V_E of E is

$$V_E(R) = \frac{1}{16\pi R}. \tag{3.8}$$

An overall multiplicative constant in (3.8) depends on exactly how one writes the effective action of a sigma model in terms of the metric of the target space, but the R dependence of V_E is significant. We see immediately that near four dimensions, that is for $R \to \infty$, the torus E is small.

One can likewise work out other terms in the effective action of the theory on $\mathbf{R}^3 \times \mathbf{S}^1_R$. For instance, on $\mathbf{R}^4$, the effective action for u is given by an expression

$$\int d^4x \; g_{u\overline{u}} du \, d\overline{u}, \tag{3.9}$$

where $g_{u\overline{u}}$ is a metric on the u plane computed in [1]. After compactification on $\mathbf{S}^1_R$, one gets the three-dimensional effective action, in the large R approximation, simply by integrating over $\mathbf{S}^1_R$, giving

$$\int d^3x \, 2\pi R \, g_{u\overline{u}} du \, d\overline{u}. \tag{3.10}$$

3.1. First Look At The Moduli Space

Now we can describe the moduli space $\mathcal{M}$ of vacua of the $N = 2$ theory compactified on $\mathbf{R}^3 \times \mathbf{S}^1_R$, at least for large R. The vacua are labeled by the order parameter u, together with, for every u, an additional complex torus E_u. From what we have just seen, the relevant family of tori is the *same* family of tori that controls the u dependence of the gauge couplings in four dimensions. So we can immediately borrow results from [2]. With τ normalized as in (3.2), the appropriate family of tori is described by the algebraic equation

$$y^2 = x^3 - x^2 u + x. \tag{3.11}$$

Therefore, the moduli space of the three-dimensional theory, for large R, is given by (3.11).

Actually, there are a few imprecisions here. A minor one is that the equation (3.11), for given u, does not describe a compact torus; one point on the torus is at $x = y = \infty$. This

was not very important in the four-dimensional story, where only the complex structure of E_u, which can still be detected even if a point is projected to infinity, was of interest. But after compactification to three dimensions, every point on E_u, including the point $x = y = \infty$, is an observable vacuum state of the theory. So if we want to be more precise, we should extend x and y to a set of homogeneous coordinates x, y, and z, and write the equation for E_u in its homogeneous form:

$$zy^2 = x^3 - zx^2u + z^2x. \tag{3.12}$$

We will omit this except when it is essential.

A more far-reaching point is that while in four-dimensions it suffices to describe $x-y-u$ space as a complex manifold (since the complex structure of E_u is all that one really needs), once one is in three dimensions, the moduli space $\mathcal{M}$ has a *hyper-Kähler metric*, and merely describing it as a complex manifold, as in (3.12), does not suffice. We must complete the description by finding the metric. We know the large R limit of the hyper-Kähler metric, from (3.7) and (3.10). Let us examine some aspects of that result with the aim of giving a formulation that makes sense for arbitrary R.

Note that, as the R dependence of (3.10) is inverse to that of (3.7), *the volume form on the moduli space of three-dimensional vacua is independent of* R, at least in the approximation of dimensional reduction from four dimensions. That volume form is in fact a constant multiple of $db \wedge d\sigma \wedge g_{u\overline{u}} du \wedge d\overline{u}$. This can be put in a more convenient form as follows. The differential form dx/y is invariant under translations on E, so it is a linear combination of db and $d\sigma$, with u-dependent coefficients. Hence $|dx/y|^2 = db \wedge d\sigma \cdot f(u, \overline{u})$, for some function of u. But in fact $f(u, \overline{u}) = g_{u\overline{u}}$. For this, recall from [1] that

$$g_{u\overline{u}} = 2\mathrm{Im}\left(\frac{da}{du}\frac{d\overline{a}_D}{d\overline{u}}\right) \tag{3.13}$$

where da/du and da_D/du are the periods of dx/y. On the other hand, from the Riemann relations

$$\int_E |dx/y|^2 = 2\mathrm{Im}\left(\frac{da}{du}\frac{d\overline{a}_D}{d\overline{u}}\right). \tag{3.14}$$

The conclusion, then, is that in terms of the holomorphic two-form

$$\omega = \frac{dx \wedge du}{y} \tag{3.15}$$

on $\mathcal{M}$, the volume form, at least for large R, is just

$$\Theta = \omega \wedge \overline{\omega}. \tag{3.16}$$

3.2. R Dependence Of The Metric

Let us now go back to four dimensions as a starting point, and ask, from that point of view, what happens to the dynamics of the $N = 2$ theory when one compactifies from $\mathbf{R}^4$ to $\mathbf{R}^3 \times \mathbf{S}^1_R$? One still has ordinary, localized four-dimensional instantons. The main novelty is that one has in addition a new kind of instanton, namely a magnetic monopole (or a dyon) that wraps around $\mathbf{S}^1_R$. The action of such an instanton, for large R, is $I = 2\pi RM$, where M is the mass of the monopole in the four-dimensional sense.

The moduli space $\mathcal{M}$ of vacua is a hyper-Kähler manifold. In one of its complex structures, the one exhibited in (3.12), $\mathcal{M}$ is elliptically fibered over the complex u plane. Let us call this the distinguished complex structure.

In the distinguished complex structure, M is not a holomorphic function (rather, it is the absolute value of the holomorphic function $a_D + na$ where n is the dyon charge). Therefore, it is impossible for monopoles to correct the distinguished complex structure of the moduli space. However, monopoles do contribute to the metric on $\mathcal{M}$. In fact, for $R = 0$ these contributions were discussed in the last section, and the case $R \neq 0$ can be treated similarly.[10] Changing the metric on $\mathcal{M}$ without changing the distinguished complex structure means that the other complex structures on $\mathcal{M}$ will change.

So far, we have just given a heuristic reason in terms of monopoles that the distinguished complex structure of $\mathcal{M}$ is independent of R. Two more fundamental reasons for this can be given. (1) Picking the distinguished complex structure selects an $N = 1$ subalgebra of $N = 2$ supersymmetry. This $N = 1$ algebra relates R to a three-dimensional vector that comes from the components g_{i4}, $i = 1, \ldots, 3$ of the space-time metric tensor g; that vector is dual to a scalar η. $N = 1$ supersymmetry would require the complex structure of $\mathcal{M}$ to depend on η if it depends on R, but the zero mode of η decouples in flat space quantum field theory. (It might not decouple in the field of a gravitational instanton!) (2) The string theory approach [4,5], as we will explain in section four, makes it clear that there is R dependence in the Kähler metric of $\mathcal{M}$ but not in the distinguished complex structure.

There is actually a natural rationale for a change in the metric of $\mathcal{M}$ due to monopoles. The complex manifold $\mathcal{M}$ is smooth for $N_f = 0$ as one can verify from (3.12). But, as was discussed in [12] in a related context, the metric obtained by dimensional reduction as

[10] In section two, we found in three dimensions that there were no monopole contributions for $N_f > 1$, but this depended on a symmetry that is absent at $R > 0$.

in (3.7) and (3.10) is not smooth; there are singularities at points where the fiber E_u has a singularity. Those are points at which the monopole mass goes to zero and monopole corrections cannot be ignored; it was proposed in [12] that the effect of the monopole corrections would be to eliminate the singularities and produce a smooth hyper-Kähler metric. For $N_f \geq 2$, $\mathcal{M}$ has orbifold singularities in its complex structure, as we will review below; in that case, one would propose that with monopole corrections included, the hyper-Kähler metric is smooth except for the orbifold singularities present in the complex structure.

So at this point, we know one complex structure on $\mathcal{M}$, and we need a recipe to determine the smooth hyper-Kähler metric (or hyper-Kähler metric with orbifold singularities) for given R. Yau's theorem on existence of Ricci-flat Kähler metrics has analogs in the non-compact case [18]. The basic idea is that to determine a hyper-Kähler metric, given a complex structure, one needs (i) the non-degenerate holomorphic two-form ω, (ii) a two-dimensional class that should be the Kähler class of the metric, (iii) a specification of the desired behavior at infinity.

In the present case, we propose that these data should be as follows. (i) We take ω to be $\omega = dx \wedge du/y$, as introduced above. We ask that the hyper-Kähler metric should have $\omega \wedge \overline{\omega}$ as its volume form. (ii) We specify the Kähler class of the metric by stating that the area of E_u is (as in (3.8)) $V_E(R) = 1/16\pi R$ and that other periods of the Kähler form, if any, are independent of R. (iii) Infinity in $\mathcal{M}$ is the region of large u; we specify the metric in this region by asking that it should reduce to what was obtained in (3.7) and (3.10).

We will assume that with an appropriate non-compact version of Yau's theorem, (i), (ii), and (iii) suffice to determine a unique smooth hyper-Kähler metric on $\mathcal{M}$ (or a hyper-Kähler metric with only orbifold singularities forced by the complex structure). The most delicate question for physics is whether (i) and (ii), which we found in the large R limit, are actually exact statements about the quantum field theory. In the next section, we will use string theory to argue that this is so, but for now we take it as a plausible assumption. In particular, we assume, according to (ii), that the area of E_u diverges for $R \to 0$; we will now see that this has interesting and verifiable consequences.

3.3. Comparison To Three Dimensions

In the last subsection, a proposal was made for the description of the hyper-Kähler moduli space $\mathcal{M}$ that arises in compactification of the $N = 2$ theory on $\mathbf{S}^1_R$, for any positive

R. Formally speaking, as $R \to 0$, this should go over to the purely three-dimensional $N = 4$ theory, analyzed in section two. Our next goal is to make this connection.

Since we claim that the area of E_u is $1/16\pi R$, something must diverge in the limit $R \to 0$; the E_u cannot remain compact. We earlier exhibited the compactness of E_u's for $N_f = 0$ by writing the equation in the homogeneous form

$$zy^2 = x^3 - zx^2u + z^2x. \tag{3.17}$$

This compactness will have to disappear for $R \to 0$, if our formula for the area is correct.

Here is another reason that the compactness must be lost. At $R = 0$, the moduli space has an $SO(3)_N$ symmetry which was extensively discussed in section two. Since $SO(3)_N$ rotates the complex structures, the full $SO(3)_N$ will not be manifest once one picks a distinguished complex structure. However, a $U(1)$ subgroup, which preserves the distinguished complex structure, should be visible. In fact, one should see a $\mathbf{C}^*$ that preserves the complex structure, of which the $U(1)$ subgroup preserves the metric. But the complex surface (3.17) does not have a non-trivial $\mathbf{C}^*$ action; such a group would have to map each E_u to another $E_{u'}$ (because the holomorphic function u would have to be constant on the image of E_u) and hence to itself (since the different E_u have different j-invariants), but a torus E_u does not have a non-trivial $\mathbf{C}^*$ action. So something must be deleted in order to find the $\mathbf{C}^*$ action.

Suppose that we throw away the points with $z = 0$. After that we can scale z to 1 and reduce to affine coordinates x, y. This gives back the original description in which the points at $x = y = \infty$ are omitted:

$$y^2 = x^3 - x^2u + x. \tag{3.18}$$

Let $v = x - u$, giving

$$y^2 = x^2v + x. \tag{3.19}$$

Suddenly a $\mathbf{C}^*$ action, with weights $1, 2, -2$ for y, x, v, is apparent. Moreover,[11] (3.19) gives the complex structure of the Atiyah-Hitchin manifold $\mathcal{N}$, which we have proposed as the moduli space of the $N_f = 0$ theory in three dimensions!

[11] According to p. 20 of [6], $\overline{\mathcal{N}}$ is the complex surface $Y^2 = X^2V + 1$, and $\mathcal{N}$ is the quotient by the freely acting $\mathbf{Z}_2$ symmetry $X \to -X, Y \to -Y, V \to V$. To take the quotient, we introduce the $\mathbf{Z}_2$-invariant independent variables $x = X^2$, $y = XY$ (we need not introduce Y^2 since it equals $X^2V + 1 = xV + 1$). In terms of x, y, and $v = V$, the equation $Y^2 = X^2V + 1$ implies $y^2 = x^2v + x$, which then describes $\mathcal{N}$.

Thus, we propose that what must be deleted when $R \to 0$ and the area of $E_u \to \infty$ are simply the points $x = y = \infty$.[12] We will now give many checks showing how a similar story works for $N_f > 0$. We first consider the case of zero hypermultiplet bare mass, and then incorporate the bare mass for $N_f = 1$.

For $N_f = 1$, in affine coordinates, the result obtained in [2] was

$$y^2 = x^3 - x^2 u + 1. \tag{3.20}$$

After substituting $v = x - u$, we get

$$y^2 = x^2 v + 1, \tag{3.21}$$

which has the expected $\mathbf{C}^*$ action with weights $0, 1, -2$ for y, x, v. Moreover ([6], p. 20), (3.21) does give the complex structure of $\overline{N}$, the double cover of the Atiyah-Hitchin manifold which was proposed in section two as the moduli space of the $N_f = 1$ theory in three dimensions.

For $N_f = 2$, the result obtained in [2] was

$$y^2 = (x^2 - 1)(x - u). \tag{3.22}$$

After the substitution $v = x - u$, we get

$$y^2 = (x^2 - 1)v, \tag{3.23}$$

with the expected $\mathbf{C}^*$ action (weights $1, 0, 2$ for y, x, v) and the two A_1 singularities (at $y = v = 0$, $x = \pm 1$) expected for the three-dimensional $N_f = 2$ theory.

For $N_f = 3$, the result of [2] was

$$y^2 = x^2(x - u) + (x - u)^2. \tag{3.24}$$

The substitution $v = x - u$ gives

$$y^2 = x^2 v + v^2, \tag{3.25}$$

which is a standard form of the A_3 or equivalently D_3 singularity, as expected.

[12] Those points must be deleted before one can make the change of variables from x and u to x and v. In fact, in homogeneous coordinates a similar substitution $v = x - uz$ fails to be an invertible change of coordinates at $z = 0$, where x and v fail to be independent.

Finally, for the $N_f = 4$ theory with zero bare mass, one has

$$y^2 = (x - e_1 u)(x - e_2 u)(x - e_3 u). \tag{3.26}$$

After linear transformations of x and u (that is replacing x and u by certain linear combinations that will be called x and v), this can be put in the form

$$y^2 = x^2 v + v^3, \tag{3.27}$$

which is a standard form of the D_4 singularity. (This D_4 singularity – and the associated configuration of two-spheres after deformation of the singularity – is actually closely related to the way D_4 triality was exhibited in section 17 of [2].)

Note that all of these results depend on changes of variables – mixing x and u – that would be unnatural in four dimensions (where u is a physical field and x is a somewhat mysterious mathematical abstraction) but are natural in three dimensions where x and u are on the same footing.

Finally, let us consider the $N_f = 1$ theory with a bare mass m. According to [1,2], the appropriate object in four dimensions is described by the equation

$$y^2 = x^3 - x^2 u + 2mx + 1. \tag{3.28}$$

We proposed at the end of section two that the three-dimensional $N_f = 1$ theory should be described by Dancer's manifold, whose complex structure (see the second paper cited in [17]) is

$$y^2 = x^2 v + i\lambda x + 1. \tag{3.29}$$

These agree after the usual change of variables $v = x - u$ and an obvious identification of λ and m.

3.4. Soft Breaking To $N = 1$

One of the main tools in [1] was to consider what happens what one adds to the theory a superpotential $\Delta W = \epsilon u$, softly breaking the $N = 2$ supersymmetry to $N = 1$. The result was to produce two vacua with monopole condensation, a mass gap, and confinement.

We now want to ask what happens if one makes the same perturbation after compactification to three dimensions on $\mathbf{S}^1_R$. *A priori*, because of the mass gap in four dimensions, one should find the same two vacua after compactification on $\mathbf{S}^1_R$, at least if R is big enough.

To investigate this, we look for critical points of the superpotential

$$W = \lambda\left(y^2 - x^3 + x^2 u - x\right) + \epsilon u, \tag{3.30}$$

where the chiral superfield λ is introduced as a sort of Lagrange multiplier to enforce the constraint $F = 0$, where $F = y^2 - x^3 + x^2 u - x$ is the quantity whose vanishing is the defining condition of E_u. The equations for a critical point of W are

$$F = \frac{\partial F}{\partial y} = \frac{\partial F}{\partial x} = \frac{\partial F}{\partial u} = 0. \tag{3.31}$$

The equations (3.31) are conditions for a singularity of the fiber E_u. They are precisely the conditions found in [1] for a vacuum state in the presence of the ϵu perturbation. They have two solutions, at $y = 0$, $u = \pm 2$, $x = u/2$. So we see that the two vacua found in [1] indeed persist after compactification on $\mathbf{S}^1_R$.

In the limit, though, of $R \to 0$, a puzzle presents itself. In three dimensions, the $\Delta W = \epsilon u$ perturbation breaks $N = 4$ supersymmetry to $N = 2$, giving bare masses to fields that are not in the $N = 2$ vector multiplet. But the minimal $N = 2$ theory generates a superpotential [10]. It is uniquely determined by the symmetries of the theory to be

$$W = e^{-\Phi} \tag{3.32}$$

where Φ is an $N = 2$ chiral superfield which originates by duality from the massless vector multiplet. The superpotential (3.32) does not have a stationary point and therefore the theory does not have a vacuum – it runs off to infinity. How is this fact consistent with the above construction?

To resolve this point, we should be more precise about some of the above formulas, restoring the dependence on the four-dimensional gauge-coupling $g_4(\mu)$ and the renormalization point μ (the scale parameter Λ is determined by $\Lambda^4 = \mu^4 \exp(-8\pi^2/g_4(\mu)^2)$). In the equation $y^2 - x^3 + x^2 u - x = 0$, the term linear in x is an instanton effect. To restore the dependence on g_4, we should write

$$y^2 = x^3 - x^2 u + x\mu^4 \exp(-8\pi^2/g_4(\mu)^2). \tag{3.33}$$

Now we introduce the three-dimensional gauge coupling, defined classically by $1/g_3^2 = R/g_4^2$. (Corrections to that formula hopefully do not matter for the qualitative remarks that we are about to make.) In terms of g_3, (3.33) becomes

$$y^2 = x^3 - x^2 u + \eta x, \tag{3.34}$$

where $\eta = \mu^4 \exp(-8\pi^2/Rg_3^2)$. If we are going to compare to [10], we should keep g_3 fixed as $R \to 0$; this means taking $\eta \to 0$. That is clear intuitively; the three-dimensional theory does not have four-dimensional instantons, so in some sense the instanton factor η should be dropped as $R \to 0$. On the other hand, we do not want to simply discard the linear term in (3.34), as this would not give the Atiyah-Hitchin manifold. Instead we make a change of variables $x - u = v$, $x = \eta\widetilde{x}$, $y = \eta\widetilde{y}$, and get the Atiyah-Hitchin manifold

$$\widetilde{y}^2 = \widetilde{x}^2 v + \widetilde{x}. \tag{3.35}$$

Now the superpotential $\Delta W = \epsilon u$ is in the new variables $\Delta W = \epsilon(\eta\widetilde{x} - v)$. So, as in (3.30), we study

$$W = \eta^2 \lambda \left(\widetilde{y}^2 - \widetilde{x}^2 v - \widetilde{x}\right) + \epsilon(\eta\widetilde{x} - v). \tag{3.36}$$

Solving $\partial W/\partial\lambda = \partial W/\partial\widetilde{y} = \partial W/\partial v = 0$ for λ, $\widetilde{y}$ and v we find an effective superpotential for $\widetilde{x}$

$$W_{eff} = \epsilon(\eta\widetilde{x} + \frac{1}{\widetilde{x}}). \tag{3.37}$$

The critical points are at $\widetilde{x} = \pm\eta^{-1/2}$. So for every non-zero η there are two vacua, but as $\eta \to 0$, the vacua run away to infinity. In fact, our analysis leads to a new derivation of (3.32) for $\eta = 0$, if we identify $e^{-\Phi} = \epsilon/\widetilde{x}$.

4. String Theory Viewpoint

In this concluding section, we will use the string theory viewpoint [3-5] to explain some crucial points that entered in sections two and three:

(1) If one compactifies from four to three dimensions on $\mathbf{S}^1_R$, then varying R does not change the distinguished complex structure of $\mathcal{M}$, which is the one in which $\mathcal{M}$ is elliptically fibered over the complex u plane. On the other hand, varying R does change the Kähler metric of $\mathcal{M}$, in such a way that the area of the fibers is a multiple of $1/R$.

(2) In three dimensions, the hypermultiplet bare masses correspond to periods of the covariantly constant two-forms ω_a on the moduli space.

The starting point is to consider M-theory compactification on $\mathbf{R}^7 \times$ K3. Then one considers a two-brane whose world-volume is $\mathbf{R}^3 \times \{p\}$, where $\mathbf{R}^3$ is a signature $-++$ flat subspace of $\mathbf{R}^7$, and p is a point in K3. The quantum field theory on the two-brane world-volume is a $2+1$-dimensional theory. The moduli space of vacua of this theory is a copy of K3, since p could be any point in K3.

On the other hand, this theory is dual to the Type I or heterotic string compactified on $\mathbf{R}^7 \times \mathbf{T}^3$. Under the duality, the M-theory two-brane corresponds to a Type I five-brane wrapped over the $\mathbf{T}^3$ (to give a two-brane in $\mathbf{R}^7$). On the five-brane world-volume there is an $SU(2)$ gauge symmetry. Therefore, suitable limits of this theory can look like $SU(2)$ or (in the event of some high energy symmetry breaking) $U(1)$ gauge theories in $2+1$ dimensions.

The moduli space of M-theory on K3 is a product of two factors. One, a copy of $\mathbf{R}^+$, parametrizes the K3 volume and corresponds to the heterotic or Type I coupling constant. The other factor is as follows. The two-dimensional integral cohomology of K3 is an even self-dual lattice of signature $(19,3)$; we denote it as $\Gamma^{19,3}$. The remaining factor in the M-theory moduli space is the choice of a three-dimensional positive definite real subspace V^+ of $\mathbf{R}^{19,3} = \Gamma^{19,3} \otimes_{\mathbf{Z}} \mathbf{R}$. The choice of V^+ is equivalent to a choice of the periods of the covariantly constant two-forms ω^a in the hyper-Kähler metric on K3. By the time one gets to a limit in which one sees $2+1$-dimensional $SU(2)$ gauge theory, a piece of the K3 is interpreted as the moduli space $\mathcal{M}$ of vacua [5], and the mass parameters correspond to periods that can be measured in $\mathcal{M}$; that is the basic reason for (2) above.

As for the heterotic string on $\mathbf{T}^3$, it has a Narain lattice $\Gamma^{19,3}$, and the moduli space is the space of three-dimensional positive definite subspaces V^+ of $\mathbf{R}^{19,3}$, interpreted as the space of right-moving momenta.

If we want to see *four*-dimensional quantum field theory on $\mathbf{R}^{2,1} \times \mathbf{S}^1$, we should split the $\mathbf{T}^3$ as $\mathbf{S}^1 \times \mathbf{T}^2$, in such a way that the Wilson lines and B-field all live only on the $\mathbf{T}^2$ factor. Then we will see a five-brane compactified on $\mathbf{T}^2$ to $\mathbf{R}^3 \times \mathbf{S}^1$; by tuning the moduli of the $\mathbf{T}^2$ appropriately, we can get four-dimensional $SU(2)$ gauge theory on $\mathbf{R}^3 \times \mathbf{S}^1$, with various numbers of hypermultiplets. Splitting the $\mathbf{T}^3$ in the indicated fashion means splitting the Narain lattice as $\Gamma^{19,3} = \Gamma^{1,1} \oplus \Gamma^{18,2}$, in a way compatible with V^+; that is V^+ is the direct sum of a one-dimensional subspace of $\mathbf{R}^{1,1}$ and a two-dimensional subspace of $\mathbf{R}^{18,2}$.

In terms of M-theory on K3, this splitting can be accomplished by specializing to K3's that are elliptically fibered (over $\mathbf{P}^1$) with a section. For such a K3, the fiber F and section S obey $F \cdot F = 0$, $F \cdot S = 1$, $S \cdot S = -2$, and generate a $\Gamma^{1,1}$ subspace of the cohomology. On such a K3, there is a distinguished complex structure, the one in which the K3 is elliptically fibered. In any limit in which a piece of the K3 turns into the moduli space $\mathcal{M}$ of a field theory, $\mathcal{M}$ will inherit a distinguished complex structure in which it is

elliptically fibered, explaining part of point (1) above.

In terms of K3, the compatibility of V^+ with the splitting $\Gamma^{19,3} = \Gamma^{1,1} \oplus \Gamma^{18,2}$ means that the Kähler form is an element of $\mathbf{R}^{1,1}$ (while the real and imaginary parts of the holomorphic two-form ω lie in $\Gamma^{18,2}$). The Kähler form is therefore dual to a linear combination of F and S, leaving two parameters of which one can be regarded as the overall volume of K3, while the second is the area of the fiber F. In the constructions of [3-5], the volume of the K3 (or heterotic string coupling constant) does not correspond to an interesting modulus of the $2+1$-dimensional or $3+1$-dimensional field theories, so we just fix it. The remaining moduli are then the area of F (which is varied while keeping fixed the volume) and the choice of the complex structure of the elliptic fibration, which is equivalent to the choice of the linear subspace generated by $\omega \in \Gamma^{18,2} \otimes_{\mathbf{Z}} \mathbf{C}$.

In the duality between M-theory on K3 and the heterotic string on $\mathbf{S}^1 \times \mathbf{T}^2$, if we want the $\mathbf{S}^1$ radius to go to infinity, we must take the area of F to zero. The remaining moduli are then only the choice of ω. That is why, once one gets to four-dimensional quantum field theory, with $\mathcal{M}$ being a piece of K3, one sees precisely a complex structure on $\mathcal{M}$ in which $\mathcal{M}$ is elliptically fibered and no other data.

If, however, one want to get quantum field theory on $\mathbf{R}^3 \times \mathbf{S}^1$, with a finite radius of $\mathbf{S}^1$, one is free to vary the area of F, while keeping fixed the volume form and complex structure. So, as stated in (1) above, the extra modulus one gets upon compactification on $\mathbf{S}^1_R$ is the ability to vary the area of the elliptic fiber in the hyper-Kähler metric, while keeping fixed the volume form and distinguished complex structure on the moduli space.

The relation between the radius of $\mathbf{S}^1_R$ and the area of the fiber F can be worked out as follows. In the duality between M-theory on K3 and the heterotic string on $\mathbf{S}^1 \times \mathbf{T}^2$, the wrapping number of two-branes on F is dual to the momentum along the $\mathbf{S}^1$. A two-brane wrapped on F has an energy which is a multiple of the area of F, while a massless particle with minimum non-zero momentum along $\mathbf{S}^1$ has energy $1/R$. So under the duality, the area of F is mapped to a constant times $1/R$, explaining the last assertion in (1) above.

References

[1] N. Seiberg and E. Witten, "Electric-Magnetic Duality, Monopole Condensation, And Confinement In $N = 2$ Supersymmetric Yang-Mills Theory," Nucl. Phys. **B426** (1994) 19.

[2] N. Seiberg and E. Witten, "Monopoles, Duality, And Chiral Symmetry Breaking In $N = 2$ Supersymmetric QCD," Nucl. Phys. **B431** (1994) 484.

[3] A. Sen, "F-Theory And Orientifolds," hep-th/9605150.

[4] T. Banks, M. Douglas, and N. Seiberg, "Probing F-Theory With Branes," hep-th/9605199.

[5] N. Seiberg, "IR Dynamics On Branes And Space-Time Geometry," hep-th/9606017.

[6] M. F. Atiyah and N. Hitchin, *The Geometry And Dynamics Of Magnetic Monopoles* (Princeton University Press, 1988).

[7] E. Witten, "An $SU(2)$ Anomaly," Phys. Lett. **B117** (1982) 324.

[8] L. Alvarez-Gaumé and E. Witten, "Gravitational Anomalies," Nucl. Phys. **B234** (1983) 269.

[9] A. M. Polyakov, "Quark Confinement And The Topology Of Gauge Groups," Nucl. Phys. **B120** (1977) 429.

[10] I. Affleck, J. A. Harvey, and E. Witten, "Instantons And (Super)Symmetry Breaking In $2 + 1$ Dimensions," Nucl. Phys. **B206** (1982) 413.

[11] N. Seiberg, "Supersymmetry And Non-Perturbative Beta Functions," Phys. Lett. **206B** (1988) 75.

[12] K. Becker, M. Becker, and A. Strominger, "Fivebranes, Membranes, And Nonperturbative String Theory," Nucl.Phys. **B456** (1995) 130.

[13] C. Callias, "Index Theorems On Open Spaces," Commun. Math. Phys. **62** (1978) 213.

[14] E. Witten, "Global Aspects Of Current Algebra," Nucl. Phys. **B223** (1983) 422.

[15] T.H.R. Skyrme, Proc.Roy.Soc. **A260** (1961) 127; E. Witten, Nucl. Phys. **223** (1983) 433.

[16] P.C. Argyres, M.R. Plesser and N. Seiberg, "The Moduli Space of Vacua of $N = 2$ SUSY QCD and Duality in $N = 1$ SUSY QCD," hep-th/9603042.

[17] A. S. Dancer, "Nahm's Equations And Hyperkähler Geometry," Commun. Math. Phys. **158** (1993) 545, "A Family Of Hyperkähler Manifolds," preprint.

[18] G. Tian and S.-T. Yau, "Complete Kähler Manifolds With Zero Ricci Curvature, I, II" (preprints).