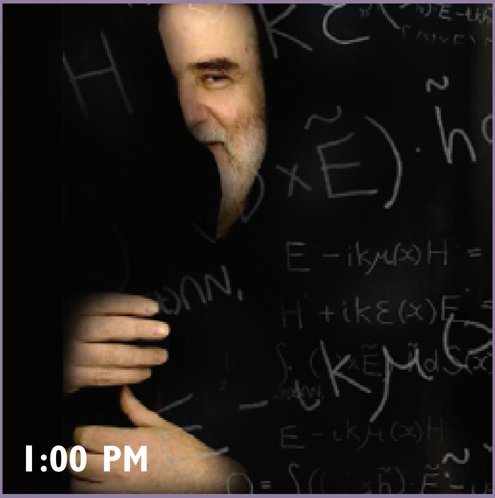
AMERICAN MATHEMATICAL SOCIETY

CURRENT EVENTS BULLETIN

Tuesday, January 8, 2008, 1:00 PM to 5:00 PM
Joint Mathematics Meetings, San Diego

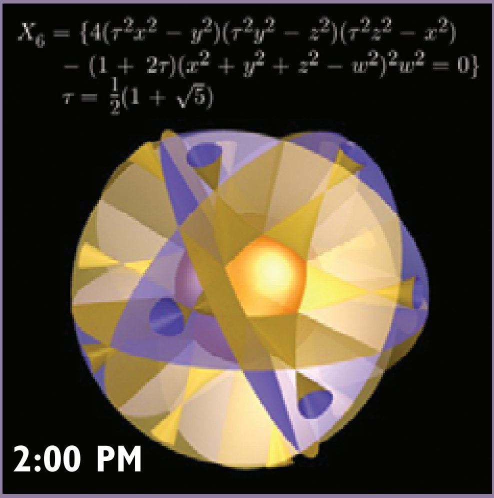
Organized by David Eisenbud, University of California, Berkeley



1:00 PM

Günther Uhlmann

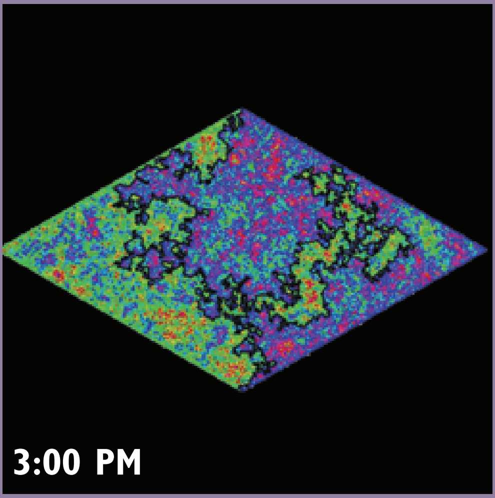
Invisibility


$X_6 = \{4(\tau^2x^2 - y^2)(\tau^2y^2 - z^2)(\tau^2z^2 - x^2)$
$- (1 + 2\tau)(x^2 + y^2 + z^2 - w^2)^2w^2 = 0\}$
$\tau = \frac{1}{2}(1 + \sqrt{5})$

2:00 PM

Antonella Grassi

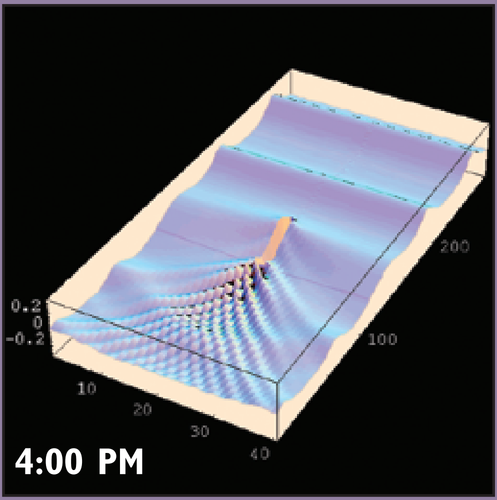
Birational Geometry:
Old and New



3:00 PM

Gregory F. Lawler

Conformal Invariance and
2-d Statistical Physics



4:00 PM

Terence C. Tao

Why are Solitons Stable?

Introduction to the Current Events Bulletin

Will the Riemann Hypothesis be proved this week? What is the Geometric Langlands Conjecture about? How could you best exploit a stream of data flowing by too fast to capture? I love the idea of having an expert explain such things to me in a brief, accessible way. I think we mathematicians are provoked to ask such questions by our sense that underneath the vastness of mathematics is a fundamental unity allowing us to look into many different corners -- though we couldn't possibly work in all of them. And I, like most of us, love common-room gossip.

The Current Events Bulletin Session at the Joint Mathematics Meetings, begun in 2003, is an event where the speakers do not report on their own work, but survey some of the most interesting current developments in mathematics, pure and applied. The wonderful tradition of the Bourbaki Seminar is an inspiration, but we aim for more accessible treatments and a wider range of subjects. I've been the organizer of these sessions since they started, but a broadly constituted advisory committee helps select the topics and speakers. Excellence in exposition is a prime consideration.

A written exposition greatly increases the number of people who can enjoy the product of the sessions, so speakers are asked to do the hard work of producing such articles. These are made into a booklet distributed at the meeting. Speakers are then invited to submit papers based on them to the *Bulletin of the AMS*, and this has led to many fine publications.

I hope you'll enjoy the papers produced from these sessions, but there's nothing like being at the talks -- don't miss them!

David Eisenbud, Organizer
University of California, Berkeley
de@msri.org

For PDF files of talks given in prior years, see
http://www.ams.org/ams/current-events-bulletin.html.
The list of speakers/titles from prior years may be found at the end of this booklet.

Invisibility

Allan Greenleaf[*]
Yaroslav Kurylev[†]
Matti Lassas,[‡]
Günther Uhlmann[§]

Abstract

We will describe recent theoretical and experimental progress on making objects invisibile to electromagnetic waves. Maxwell's equations have transformation laws that allow for design of electromagnetic parameters that would steer light around a hidden region, returning it to its original path on the far side. Not only would observers be unaware of the contents of the hidden region, they would not even be aware that something was hidden. The object would have no shadow. New advances in metamaterials have given some experimental evidence that this indeed can be made possible at certain frequencies.

1 Introduction

There have recently been many studies [AE, GKLU1, MN, Le, PSS1, MBW, W] on the possibility, both theoretical and practical, of a region or object

[*] Department of Mathematics, University of Rochester, Rochester, NY 14627, USA. Partially supported by NSF grant DMS-0551894.

[†] Department of Mathematics, University College London, Gower Street, London, WC1E 5BT, UK

[‡] Helsinki University of Technology, Institute of Mathematics, P.O.Box 1100, FIN-02015, Finland. Partially supported by Academy of Finland CoE Project 213476.

[§] Department of Mathematics, University of Washington, Seattle, WA 98195, USA. Partially supported by the NSF and a Walker Family Endowed Professorship.

being shielded, or cloaked from detection via electromagnetic waves. The interest in cloaking was raised in particular in 2006 when it was realized that practical cloaking constructions are possible using so-called metamaterials which allow fairly arbitrary specification of electromagnetic material parameters. At the present moment such materials have been implemented at microwave frequencies [Sc]. On the practical limitations of cloaking, we note that, with current technology, above microwave frequencies the required metamaterials are difficult to fabricate and assemble, although research is presently progressing on metamaterial engineering at optical frequencies [Sh]. Furthermore, metamaterials are inherently prone to dispersion, so that realistic cloaking must currently be considered as occurring at a single wavelength, or very narrow range of wavelengths.

The theoretical considerations related to cloaking were introduced already in 2003, before the appearance of practical possibilities for cloaking. Indeed, the cloaking constructions in the zero frequency case, i.e., for electrostatics, were introduced as counterexamples in the study of inverse problems. In [GLU2, GLU3] it was shown that passive objects can be coated with a layer of material with a degenerate conductivity which makes the object undetectable by electrical impedance tomography (EIT), that is, in the electrostatic measurements. This gave counterexamples for uniqueness in the Calderón inverse problem for the conductivity equation. The counterexamples were motivated by consideration of certain degenerating families of Riemannian metrics, which in the limit correspond to singular conductivities, i.e., that are not bounded below or above. A related example of a complete but noncompact two-dimensional Riemannian manifold with boundary having the same Dirichlet-to-Neumann map as a compact one was given in [LTU].

Before discussing the recent results on cloaking and counterexamples in inverse problems, let us briefly discuss the positive results for inverse problems. The paradigm problem is Calderón's inverse problem, which is the question of whether an unknown conductivity distribution inside a domain in $\mathbb{R}^n$, modelling for example the human thorax, can be determined from voltage and current measurements made on the boundary. For isotropic conductivities this problem can be mathematically formulated as follows: Let Ω be the measurement domain, and denote by σ a bounded and strictly positive function describing the conductivity in Ω. In Ω the voltage potential u satisfies the equation

$$\nabla \cdot \sigma \nabla u = 0. \tag{1}$$

To uniquely fix the solution u it is enough to give its value, f, on the boundary. In the idealized case, one measures for all voltage distributions $u|_{\partial\Omega} = f$ on the boundary the corresponding current flux, $\nu\cdot \sigma\nabla u$, through the boundary, where ν is the exterior unit normal to $\partial\Omega$. Mathematically this amounts to the knowledge of the Dirichlet–Neumann map Λ corresponding to σ, i.e., the map taking the Dirichlet boundary values of the solution to (1) to the corresponding Neumann boundary values,

$$\Lambda: \quad u|_{\partial\Omega} \mapsto \nu\cdot \sigma\nabla u|_{\partial\Omega}.$$

Calderón's inverse problem is then to reconstruct σ from Λ. The problem was originally proposed by Calderón [C] in 1980. Sylvester and Uhlmann [SyU] proved unique identifiability of the conductivity in dimensions three and higher for isotropic conductivities which are C^∞–smooth, and Nachman gave a reconstruction method [N]. In three dimensions or higher unique identifiability of the conductivity is known for conductivities with 3/2 derivatives [BT], [PPU]. In two dimensions the first global result for C^2 conductivities is due to Nachman [N1]. This was improved in [BU] to Lipschitz conductivities. Astala and Päivärinta showed in [AP] that uniqueness holds also for general isotropic conductivities merely in L^∞.

The Calderón problem with an anisotropic, i.e., matrix-valued, conductivity that is uniformly bounded from above and below has been studied in two dimensions [S, N1, SuU, ALP] and in three dimensions or higher [LaU, LeU, LTU]. For example, for the anisotropic inverse conductivity problem in the two dimensional case, it is known that Cauchy data determines the conductivity tensor up to a diffeomorphism $F : \overline{\Omega} \longrightarrow \overline{\Omega}$. Thus, the inverse problem is not uniquely solvable, but the non-uniqueness of the problem can be characterized. This makes it possible, e.g., to find the unique conductivity that is closest to isotropic ones [KLO]. Another related inverse problem is the Gel'fand problem, which uses boundary measurements at all frequencies, rather than at a fixed one. For this problem, uniqueness results are available; see, e.g., [BeK, KK], with a detailed exposition in [KKL].

We emphasize that for the above positive results for inverse problems it is assumed that the eigenvalues of the conductivity are bounded below and above by positive constants. Thus, a key point in the current works on invisibility that allows one to avoid the known uniqueness theorems is the lack of positive lower and upper bounds on the eigenvalues of these symmetric tensor fields.

For Maxwell's equations the inverse problem with the isotropic permittivity ε and permeability μ and the data given at one frequency was solved in [OPS]. The inverse problem with the anisotropic permittivity and permeability has been studied with data given at all frequencies (or in the time domain) when the permittivity and permeability tensors ε and μ are conformal to each other, i.e., multiples of each other by a positive scalar function; this condition has been studied in detail in [KLS]. For Maxwell's equations in the time domain, this condition corresponds to polarization-independent wave velocity. This seemingly special condition arises quite naturally also in the invisibility constructions, since the pushforward $(\widetilde{\varepsilon}, \widetilde{\mu})$ of an isotropic pair (ε, μ) by a diffeomorphism need not be isotropic but does satisfy this conformality.

Let us now return to the recent results on cloaking and the counterexamples for inverse problems. In 2006, several cloaking constructions were proposed. The constructions in [Le] are based on conformal mapping in two dimensions and are justified via change of variables on the exterior of the cloaked region. At the same time, [PSS1] proposed a cloaking construction for Maxwell's equations based on a singular transformation of the original space, again observing that, outside the cloaked region, the solutions of the homogeneous Maxwell equations in the original space become solutions of the transformed equations. The transformations used there are the same as used in [GLU2, GLU3] in the context of Calderón's inverse conductivity problem. The paper [PSS2] contained analysis of cloaking on the level of ray-tracing, full wave numerical simulations were discussed in [CPSSP], and the cloaking experiment at 8.5Ghz is in [Sc].

The electromagnetic material parameters used in cloaking constructions are degenerate and, due to the degeneracy of the equations at the surface of the cloaked region, it is important to consider rigorously (weak) solutions to Maxwell's equations on *all* of the domain, not just the exterior of the cloaked region. This analysis was carried out in [GKLU1]. There, various constructions for cloaking from observation are analyzed on the level of physically meaningful electromagnetic waves, i.e., finite energy distributional solutions of the equations. In the analysis of the problem, it turns out that the cloaking structure imposes hidden boundary conditions on such waves at the surface of the cloak. When these conditions are overdetermined, finite energy solutions typically do not exist. The time-domain physical interpretation of this was at first not entirely clear, but it now seems to be intimately related with blow-up of the fields, which would may compromise the desired cloaking effect [GKLU3]. We review the results here and give the possible remedies to restore invisibility.

We note that [GLU2, GLU3] gave, in dimensions $n \geq 3$, counterexamples to uniqueness for the inverse conductivity problem. Such counterexamples have now also been given and studied further in two dimensional case [KSVW, ALP2].

2 Basic constructions

The material parameters of electromagnetism, the electrical permittivity, $\varepsilon(x)$; magnetic permeability, $\mu(x)$; and the conductivity $\sigma(x)$ can be considered as coordinate invariant objects. If $F : \Omega_1 \longrightarrow \Omega_2, \quad y = F(x)$, is a diffeomorphism between domains in $\mathbb{R}^n$, then $\sigma(x) = [\sigma^{jk}(x)]_{j,k=1}^n$ on Ω_1 pushes forward to $(F_*\sigma)(y)$ on Ω_2, given by

$$(F_*\sigma)^{jk}(y) = \frac{1}{\det\left[\frac{\partial F}{\partial x}(x)\right]} \sum_{p,q=1}^{n} \frac{\partial F^j}{\partial x^p}(x) \, \frac{\partial F^k}{\partial x^q}(x) \sigma^{pq}(x) \Bigg|_{x=F^{-1}(y)} . \qquad (2)$$

The same transformation rule is valid for permittivity ε and permeability μ. It was observed by Luc Tartar (see [KV]) that it follows that if F is a diffeomorphism of a domain Ω fixing $\partial\Omega$, then the conductivity equations with the conductivities σ and $\widetilde{\sigma} := F_*\sigma$ have the same Dirichlet-to-Neumann map, producing infinite-dimensional families of indistinguishable conductivities. This can already be considered as a weak form of invisibility, with distinct conductivities being indistinguishable by external observations; however, nothing has been hidden yet.

On the other hand, a Riemannian metric $g = [g_{jk}(x)]_{j,k=1}^n$ is a covariant symmetric two-tensor. Remarkably, in dimension three or higher, a material parameter tensor and a Riemannian metric can be associated with each other by

$$\sigma^{jk} = |g|^{1/2} g^{jk}, \quad \text{or} \quad g^{jk} = |\sigma|^{2/(n-2)} \sigma^{jk}, \qquad (3)$$

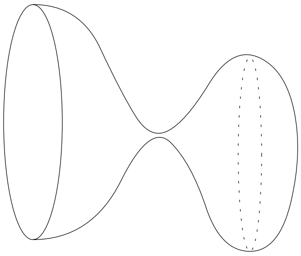


Figure 1: A family of manifolds that develops a singularity when the width of the neck connecting two parts goes to zero.

where $[g^{jk}] = [g_{jk}]^{-1}$ and $|g| = \det(g)$. Using this correspondence, examples of singular anisotropic conductivities in $\mathbb{R}^n, n \geq 3$, that are indistinguishable from a constant isotropic conductivity, in that they have the same Dirichlet-to-Neumann map, are given in [GLU3]. This construction is based on degenerations of Riemannian metrics, whose singular limits can be considered as coming from singular changes of variables. If one considers Figure 1, where the "neck" of the surface (or a manifold in the higher dimensional cases) is pinched, the manifold contains in the limit a pocket about which the boundary measurements do not give any information. If the collapsing of the manifold is done in an appropriate way, in the limit we have a Riemannian manifold which is indistinguishable from flat surface. This can be considered as a singular conductivity that appears the same as a constant conductivity to all boundary measurements.

To consider the above precisely, let $B(0,R) \subset \mathbb{R}^3$ be an open ball with center 0 and radius R. We use in the sequel the set $N = B(0,2)$, decomposed to two parts, $N_1 = B(0,2) \setminus \overline{B}(0,1)$ and $N_2 = B(0,1)$. Let $\Sigma = \partial N_2$ the the interface (or "cloaking surface") between N_1 and N_2.

We use also a "copy" of the ball $B(0,2)$, with the notation $M_1 = B(0,2)$. Let $g_{jk} = \delta_{jk}$ be the Euclidian metric in M_1 and let $\gamma = 1$ be the corresponding homogeneous conductivity. Define a singular transformation

$$F : M_1 \setminus \{0\} \to N_1, \qquad F(x) = (\frac{|x|}{2} + 1)\frac{x}{|x|}, \quad 0 < |x| \leq 2. \tag{4}$$

The pushforward $\widetilde{g} = F_*g$ of the metric g in F is the metric in N_1 given by

$$(F_*g)_{jk}(y) = \sum_{p,q=1}^{n} \frac{\partial F^p}{\partial x^j}(x) \, \frac{\partial F^q}{\partial x^k}(x) g_{pq}(x) \bigg|_{x=F^{-1}(y)} . \tag{5}$$

We use it to define a singular conductivity

$$\widetilde{\sigma} = \left\{ \begin{array}{ll} |\widetilde{g}|^{1/2}\widetilde{g}^{jk} & \text{for } x \in N_1, \\ \delta^{jk} & \text{for } x \in N_2 \end{array} \right. \tag{6}$$

in N. Then, denoting by $(r, \phi, \theta) \mapsto (r \sin\theta\cos\phi, r\sin\theta\sin\phi, r\cos\theta)$ the spherical coordinates, we have

$$\widetilde{\sigma} = \begin{pmatrix} 2(r-1)^2\sin\theta & 0 & 0 \\ 0 & 2\sin\theta & 0 \\ 0 & 0 & 2(\sin\theta)^{-1} \end{pmatrix}, \quad 1 < |x| \leq 2.$$

This means that in the Cartesian coordinates the conductivity $\widetilde{\sigma}$ is given by

$$\widetilde{\sigma}(x) = 2(I - P(x)) + 2(|x|-1)^2 P(x), \quad 1 < |x| < 2,$$

where I is the identity matrix and $P(x) = |x|^{-2}xx^t$ is the projection to the radial direction. We note that the anisotropic conductivity $\widetilde{\sigma}$ is singular on Σ in the sense that it is not bounded from below by any positive multiple of I. (See [KSVW] for a similar calculation for $n = 2$.)

Consider now the *Cauchy data* of all $H^1(N)$-solutions of the conductivity equation corresponding to $\widetilde{\sigma}$, that is,

$$C_1(\widetilde{\sigma}) = \{(u|_{\partial N}, \nu{\cdot}\, \widetilde{\sigma}\nabla u|_{\partial N}) \ : \ u \in H^1(n), \ \widetilde{\sigma}\nabla u \in \mathcal{D}'(N), \ \nabla{\cdot}\, \widetilde{\sigma}\nabla u = 0\},$$

where ν is the Euclidian unit normal vector of ∂N.

Theorem 2.1 *([GLU3]) The Cauchy data of H^1-solutions for all conductivities $\widetilde{\sigma}$ and γ on N coincide, that is, $C_1(\widetilde{\sigma}) = C_1(\gamma)$.*

This means that all boundary measurements for the homogeneous conductivity $\gamma = 1$ and the degenerated conductivity $\widetilde{\sigma}$ are the same. In the figure below there are analytically obtained solutions on a disc with metric $\widetilde{\sigma}$

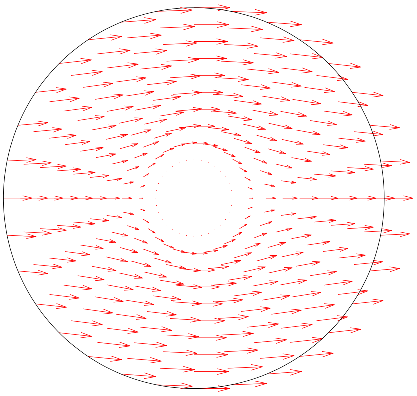


Figure 2: Analytic solutions for the currents

As seen in the figure, no currents appear near the center of the disc, so that if the conductivity is changed near the center, the measurements on the boundary ∂N do not change.

We note that a similar type of theorem is valid also for a more general class of solutions. Consider an unbounded quadratic form, A in $L^2(N)$,

$$A_{\widetilde{\sigma}}[u,v] = \int_N \widetilde{\sigma} \nabla u \cdot \nabla v \, dx$$

defined for $u, v \in \mathcal{D}(A_{\widetilde{\sigma}}) = C_0^\infty(N)$. Note that here, in the three dimensional case that we now consider, $\widetilde{\sigma}$ is bounded from above, but not from below. Let $\overline{A}_{\widetilde{\sigma}}$ be the closure of this quadratic form and say that

$$\nabla \cdot \widetilde{\sigma} \nabla u = 0 \quad \text{in } N$$

is satisfied in the finite energy sense if there is $u_0 \in H^1(N)$ supported in N_1 such that $u - u_0 \in \mathcal{D}(A_{\widetilde{\sigma}})$ and

$$\overline{A}_{\widetilde{\sigma}}[u - u_0, v] = -\int_N \widetilde{\sigma} \nabla u_0 \cdot \nabla v \, dx, \quad \text{for all } v \in \mathcal{D}(\overline{A}_{\widetilde{\sigma}}).$$

Then Cauchy data set of the finite energy solutions, denoted

$$C_f(\widetilde{\sigma}) = \{(u|_{\partial N}, \nu \cdot \widetilde{\sigma} \nabla u|_{\partial N}) \ : \ u \text{ is finite energy solution of } \nabla \cdot \widetilde{\sigma} \nabla u = 0\}$$

coincides with $C_f(\gamma)$. In three dimensions, this and similar results for the Helmholtz equation follow directly from [GKLU1].

3 Maxwell's equations

In what follows, we treat Maxwell's equations in non-conducting media, that is, for which $\sigma = 0$. We consider the electric and magnetic fields, E and H, as differential 1-forms, given in some local coordinates by

$$E = E_j(x)dx^j, \quad H = H_j(x)dx^j.$$

For 1-form $E(x) = E_1(x)dx^1 + E_2(x)dx^2 + E_3(x)dx^3$ we define the push-forward of E in F, denoted $\widetilde{E} = F_*E$, by

$$\begin{aligned}\widetilde{E}(\widetilde{x}) &= \widetilde{E}_1(\widetilde{x})d\widetilde{x}^1 + \widetilde{E}_2(\widetilde{x})d\widetilde{x}^2 + \widetilde{E}_3(\widetilde{x})d\widetilde{x}^3 \\ &= \sum_{j=1}^{3}\Big(\sum_{k=1}^{3}(DF^{-1})_j^k(\widetilde{x})\,E_k(F^{-1}(\widetilde{x}))\Big)d\widetilde{x}^j, \quad \widetilde{x} = F(x).\end{aligned}$$

A similar kind of transformation law is valid for 2-forms. We interpret the curl operator for 1-forms in $\mathbb{R}^3$ as being the exterior derivative, d. Maxwell's equations then have the form

$$\operatorname{curl} H = -ikD + J, \quad \operatorname{curl} E = ikB$$

where we consider the D and B fields the external current J (if present) as 2-forms. The constitutive relations are

$$D = \varepsilon E, \quad B = \mu H,$$

where the material parameters ε and μ are linear maps mapping 1-forms to 2-forms.

Let g be a Riemannian metric in $\Omega \subset \mathbb{R}^3$. Using the metric g, we define a specific permittivity and permeability by setting

$$\varepsilon^{jk} = \mu^{jk} = |g|^{1/2}g^{jk}.$$

To introduce the material parameters $\widetilde{\varepsilon}(x)$ and $\widetilde{\mu}(x)$ that make cloaking possible, we consider the map F given by (4), the Euclidean metric g in M_1 and $\widetilde{g} = F_*g$ in N_1 as before, and define the singular permittivity and permeability by the formula analogous to (6),

$$\varepsilon^{jk} = \mu^{jk} = \begin{cases} |\widetilde{g}|^{1/2}\widetilde{g}^{jk} & \text{for } x \in N_1, \\ \delta^{jk} & \text{for } x \in N_2. \end{cases} \tag{7}$$

These material parameters are singular on Σ, requiring that what it means for fields $(\widetilde{E}, \widetilde{H})$ to form a solution to Maxwell's equations must be defined carefully.

3.1 Definition of solutions of Maxwell equations

Since the material parameters $\widetilde{\varepsilon}$ and $\widetilde{\mu}$ are again singular, we need to define solutions carefully.

Definition 3.1 *We say that $(\widetilde{E},\widetilde{H})$ is a finite energy solution to Maxwell's equations on N,*

$$\nabla\times\widetilde{E}=ik\widetilde{\mu}(x)\widetilde{H},\quad \nabla\times\widetilde{H}=-ik\widetilde{\varepsilon}(x)\widetilde{E}+\widetilde{J}\quad \text{on } N, \tag{8}$$

if $\widetilde{E}$, $\widetilde{H}$ are one-forms and $\widetilde{D}:=\widetilde{\varepsilon}\,\widetilde{E}$ and $\widetilde{B}:=\widetilde{\mu}\,\widetilde{H}$ two-forms in N with $L^1(N,dx)$-coefficients satisfying

$$\|\widetilde{E}\|^2_{L^2(N,|\widetilde{g}|^{1/2}dV_0(x))}=\int_N \widetilde{\varepsilon}^{jk}\,\widetilde{E}_j\,\overline{\widetilde{E}_k}\,dV_0(x)<\infty, \tag{9}$$

$$\|\widetilde{H}\|^2_{L^2(N,|\widetilde{g}|^{1/2}dV_0(x))}=\int_N \widetilde{\mu}^{jk}\,\widetilde{H}_j\,\overline{\widetilde{H}_k}\,dV_0(x)<\infty; \tag{10}$$

where dV_0 is the standard Euclidean volume, $(\widetilde{E},\widetilde{H})$ is a classical solution of Maxwell's equations on a neighborhood $U\subset\overline{N}$ of ∂N:

$$\nabla\times\widetilde{E}=ik\widetilde{\mu}(x)\widetilde{H},\quad \nabla\times\widetilde{H}=-ik\varepsilon(x)\widetilde{E}+\widetilde{J}\quad \text{in } U,$$

and finally,

$$\int_N((\nabla\times\widetilde{h})\cdot\widetilde{E}-ik\widetilde{h}\cdot\widetilde{\mu}(x)\widetilde{H})\,dV_0(x)=0,$$

$$\int_N((\nabla\times\widetilde{e})\cdot\widetilde{H}+\widetilde{e}\cdot(ik\widetilde{\varepsilon}(x)\widetilde{E}-\widetilde{J}))\,dV_0(x)=0$$

for all $\widetilde{e},\widetilde{h}\in C_0^\infty(\Omega^1N)$.

Here, $C_0^\infty(\Omega^1N)$ denotes smooth 1-forms on N whose supports do not intersect ∂N, and the inner product "·" denotes the Euclidean inner product.

Surprisingly, the finite energy solutions do not exists for generic currents. To consider this, let M be the disjoint union of a ball $M_1=B(0,2)$ and a ball $M_2=B(0,1)$. These will correspond to sets N,N_1,N_2 after an appropriate changes of coordinates. We thus consider a map $F:M\setminus\{0\}=(M_1\setminus\{0\})\cup M_2\to N\setminus\Sigma$, where F mapping $M_1\setminus\{0\}$ to N_1 is the the map defined by formula (4) and F mapping M_2 to N_2 as the identity map.

Theorem 3.2 *([GKLU1]) Let E and H be 1-forms with measurable coefficients on $M \setminus \{0\}$ and $\widetilde E$ and $\widetilde H$ be 1-forms with measurable coefficients on $N \setminus \Sigma$ such that $E = F^*\widetilde E$, $H = F^*\widetilde H$. Let J and $\widetilde J$ be 2-forms with smooth coefficients on $M \setminus \{0\}$ and $N \setminus \Sigma$, that are supported away from $\{0\}$ and Σ. Then the following are equivalent:*

1. *The 1-forms $\widetilde E$ and $\widetilde H$ on N satisfy Maxwell's equations*

$$\begin{aligned} &\nabla \times \widetilde E = ik\widetilde\mu(x)\widetilde H, \quad \nabla \times \widetilde H = -ik\widetilde\varepsilon(x)\widetilde E + \widetilde J \quad \text{on } N, \\ &\nu \times \widetilde E|_{\partial N} = f \end{aligned} \tag{11}$$

   in the sense of Definition 3.1.

2. *The forms E and H satisfy Maxwell's equations on M,*

$$\begin{aligned} &\nabla \times E = ik\mu(x)H, \quad \nabla \times H = -ik\varepsilon(x)E + J \quad \text{on } M_1, \\ &\nu \times E|_{\partial M_1} = f \end{aligned} \tag{12}$$

   and

$$\nabla \times E = ik\mu(x)H, \quad \nabla \times H = -ik\varepsilon(x)E + J \quad \text{on } M_2 \tag{13}$$

   with Cauchy data

$$\nu \times E|_{\partial M_2} = b^e, \quad \nu \times H|_{\partial M_2} = b^h \tag{14}$$

   that satisfies $b^e = b^h = 0$.

Moreover, if E and H solve (12), (13), and (14) with non-zero b^e or b^h, then the fields $\widetilde E$ and $\widetilde H$ are not solutions of Maxwell equations on N in the sense of Definition 3.1.

The above theorem can be interpreted by saying that the cloaking of active objects is difficult, as the idealized model with non-zero currents present within the region to be cloaked, leads to non-existence of finite energy distributional solutions. We find two ways of dealing with this difficulty. One is to simply augment the above coating construction around a ball by adding a perfect electrical conductor (PEC) lining at Σ, in order to make the object inside the coating material to appear like a passive object. Alternatively, one can introduce a more elaborate construction, which we refer to as the *double coating.* Mathematically, this corresponds to a singular Riemannian metric which degenerates in the same way as one approaches Σ from both sides; physically it would correspond to surrounding both the inner and outer surfaces of Σ with appropriately matched metamaterials.

4 Cloaking an infinite cylindrical domain

In the following we change the geometrical situation where we do our considerations, and redefine the meaning of the used notations.

We consider next an infinite cylindrical domain. Below, $B_2(0,r) \subset \mathbb{R}^2$ is Euclidian disc with center 0 and radius r. Let us use in the following the notations $N = B_2(0,2) \times \mathbb{R}$, $N_1 = (B_2(0,2) \setminus B_2(0,1)) \times \mathbb{R}$, and $N_2 = B_2(0,1) \times \mathbb{R}$. Moreover, let M be the disjoint union of $M_1 = B_2(0,2) \times \mathbb{R}$ and $M_2 = B_2(0,1) \times \mathbb{R}$. Finally, let us denote in this section $\Sigma = \partial B_2(0,1) \times \mathbb{R}$, $L = \{(0,0)\} \times \mathbb{R} \subset M_1$. We define the map $F : M \setminus L \to N \setminus \Sigma$ in cylindrical coordinates by

$$F(r,\theta,z) = (1 + \frac{r}{2}, \theta, z)$$

Again, let g be the Euclidian metric on M, and $\varepsilon = 1$ and $\mu = 1$ be homogeneous material parameters in M. Using map F we define $\widetilde{g} = F_* g$ in $N \setminus \Sigma$ and define $\widetilde{\varepsilon}$ and $\widetilde{\mu}$ as in formula (7). By finite energy solutions of Maxwell's equations on N we will mean one-forms $\widetilde{E}$ and $\widetilde{H}$ satisfying the the conditions analogous to Definition 3.1.

Theorem 4.1 *([GKLU1]) Let E and H be 1-forms with measurable coefficients on $M \setminus L$ and $\widetilde{E}$ and $\widetilde{H}$ be 1-forms with measurable coefficients on $N \setminus \Sigma$ such that $E = F^*\widetilde{E}$, $H = F^*\widetilde{H}$. Let J and $\widetilde{J}$ be 2-forms with smooth coefficients on $M \setminus L$ and $N \setminus \Sigma$, that are supported away from L and Σ, respectively.*

Then the following are equivalent:

1. *On N, the 1-forms $\widetilde{E}$ and $\widetilde{H}$ satisfy Maxwell's equations*

$$\nabla \times \widetilde{E} = ik\widetilde{\mu}(x)\widetilde{H}, \quad \nabla \times \widetilde{H} = -ik\widetilde{\varepsilon}(x)\widetilde{E} + \widetilde{J} \quad \textit{in } N, \qquad (15)$$
$$\nu \times \widetilde{E}|_{\partial N} = f$$

and $\widetilde{E}$ and $\widetilde{H}$ are finite energy solutions.

2. *On M, the forms E and H are classical solutions to Maxwell's equations on M, with data*

$$b_1^e = \zeta\cdot E|_{\gamma_1},\quad b_2^e = \zeta\cdot E|_{\gamma_2},\quad b_1^h = \zeta\cdot H|_{\gamma_1},\quad b_2^h = \zeta\cdot H|_{\gamma_2}, \tag{16}$$

that satisfy

$$b_1^e(z) = b_2^e(z) \quad \textit{and} \quad b_1^h(z) = b_2^h(z), \quad z \in \mathbb{R}. \tag{17}$$

Moreover, if E and H solve Maxwell's equations on M with the boundary values (16) that do not satisfy (17), then then the fields $\widetilde{E}$ and $\widetilde{H}$ are not finite energy solutions of Maxwell equations on N.

Further analysis and numerical simulations, exploring the consequences of this non-existence result for cloaking, can be found in [GKLU3].

5 Cloaking a cylinder with the Soft-and-Hard boundary condition

Next, we consider N_2 as an obstacle, while the domain N_1 is equipped with a metric corresponding to the above coating in the cylindrical geometry. Motivated by the conditions at Σ in the previous section, we impose the soft-and-hard surface (SHS) boundary condition on the boundary of the obstacle. In classical terms, the SHS condition on a surface Σ [HLS, Ki1] is

$$\zeta \cdot E|_\Sigma = 0 \quad \text{and} \quad \zeta \cdot H|_\Sigma = 0,$$

where $\zeta = \zeta(x)$ is a tangential vector field on Σ, that is, $\zeta \times \nu = 0$. In other words, the part of the tangential component of the electric field E that is parallel to ζ vanishes, and the same is true for the magnetic field H. This was originally introduced in antenna design and can be physically realized by having a surface with thin parallel gratings filled with dielectric material [Ki1, Ki2, Li, HLS]. We consider this boundary condition when ζ is the vector field $\eta = \partial_\theta$, that is, the angular vector field that is tangential to Σ.

To this end, let us give still one more definition of weak solutions, appropriate for this construction. We consider only solutions on the set N_1; nevertheless, we continue to denote $\partial N = \partial N_1 \setminus \Sigma$.

Definition 5.1 *We say that the 1-forms $\widetilde{E}$ and $\widetilde{H}$ are* finite energy solutions *of Maxwell's equations on N_1 with the soft-and-hard (SH) boundary conditions on Σ,*

$$\begin{aligned}
&\nabla \times \widetilde{E} = ik\widetilde{\mu}(x)\widetilde{H}, \quad \nabla \times \widetilde{H} = -ik\widetilde{\varepsilon}(x)\widetilde{E} + \widetilde{J} \quad \text{on } N_1, && (18)\\
&\eta \cdot \widetilde{E}|_{\Sigma} = 0, \quad \eta \cdot \widetilde{H}|_{\Sigma} = 0, && (19)\\
&\nu \times \widetilde{E}|_{\partial N} = f,
\end{aligned}$$

if $\widetilde{E}$ and $\widetilde{H}$ are 1-forms on N_1 and $\widetilde{\varepsilon}\widetilde{E}$ and $\widetilde{\mu}\widetilde{H}$ are 2-forms with measurable coefficients satisfying

$$\|\widetilde{E}\|^2_{L^2(N_1,|\widetilde{g}|^{1/2}dV_0)} = \int_{N_1} \widetilde{\varepsilon}^{jk}\, \widetilde{E}_j\, \overline{\widetilde{E}_k}\, dV_0(x) < \infty, \qquad (20)$$

$$\|\widetilde{H}\|^2_{L^2(N_1,|\widetilde{g}|^{1/2}dV_0)} = \int_{N_1} \widetilde{\mu}^{jk}\, \widetilde{H}_j\, \overline{\widetilde{H}_k}\, dV_0(x) < \infty; \qquad (21)$$

Maxwell's equation are valid in the classical sense in a neighborhood U of ∂N:

$$\begin{aligned}
&\nabla \times \widetilde{E} = ik\widetilde{\mu}(x)\widetilde{H}, \quad \nabla \times \widetilde{H} = -ik\varepsilon(x)\widetilde{E} + \widetilde{J} \quad \text{in } U,\\
&\nu \times \widetilde{E}|_{\partial N} = f;
\end{aligned}$$

and finally,

$$\begin{aligned}
&\int_{N_1} ((\nabla \times \widetilde{h}) \cdot \widetilde{E} - ik\widetilde{h} \cdot \widetilde{\mu}(x)\widetilde{H})\, dV_0(x) = 0,\\
&\int_{N} ((\nabla \times \widetilde{e}) \cdot \widetilde{H} + \widetilde{e} \cdot (ik\widetilde{\varepsilon}(x)\widetilde{E} - \widetilde{J}))\, dV_0(x) = 0,
\end{aligned}$$

for all $\widetilde{e}, \widetilde{h} \in C_0^\infty(\Omega^1 N_1)$ satisfying

$$\eta \cdot \widetilde{e}|_{\Sigma} = 0, \quad \eta \cdot \widetilde{h}|_{\Sigma} = 0. \qquad (22)$$

We then have the following invisibility result.

Theorem 5.2 *([GKLU1]) Let E and H be 1-forms with measurable coefficients on $M_1 \setminus L$ and $\widetilde{E}$ and $\widetilde{H}$ be 1-forms with measurable coefficients on N_1 such that $E = F^*\widetilde{E}$, $H = F^*\widetilde{H}$. Let J and $\widetilde{J}$ be 2-forms with smooth coefficients on $M_1 \setminus L$ and $N_1 \setminus \Sigma$, that are supported away from L and Σ.*

Then the following are equivalent:

1. *On N_1, the 1-forms $\widetilde{E}$ and $\widetilde{H}$ satisfy Maxwell's equations with SH boundary conditions in the sense of Definition 5.1.*
2. *On M_1, the forms E and H are classical solutions of Maxwell's equations,*

$$\begin{aligned} &\nabla \times E = ik\mu(x)H, \quad \textit{in } M_1 \\ &\nabla \times H = -ik\varepsilon(x)E + J, \quad \textit{in } M_1, \\ &\nu \times E|_{\partial M_1} = f. \end{aligned} \tag{23}$$

This result implies that when the surface Σ is lined with a material implementing the SHS boundary condition, the finite energy distributional solutions exist for all incoming waves. Other boundary conditions making the problem solvable in some sense, using a different definition based on self-adjoint extensions of the operators, have been recently characterized in [W].

6 Artificial wormholes

Cloaking a ball or cylinder are particularly extreme examples of what has come to be known as *transformation optics* in the physics literature, and other interesting effects are possible. We sketch the construction of artificial electromagnetic wormholes, introduced in [GKLU2]. Consider first as in Fig. 3 a 3-dimensional wormhole manifold (or handlebody) $M = M_1 \# M_2$ where the components

$$\begin{aligned} &M_1 = \mathbb{R}^3 \setminus (B(O,1) \cup B(P,1)), \\ &M_2 = \mathbb{S}^2 \times [0,1] \end{aligned}$$

are glued together smoothly.

An optical device that acts as a wormhole for electromagnetic waves at a given frequency k can be constructed by starting with a two-dimensional finite cylinder

$$T = \mathbb{S}^1 \times [0,L] \subset \mathbb{R}^3$$

and taking its neighborhood $K = \{x \in \mathbb{R}^3 : \ \mathrm{dist}(x,T) < \rho\}$, where $\rho > 0$ is small enough and $N = \mathbb{R}^3 \setminus K$. Let us put on ∂K the SHS boundary condi-

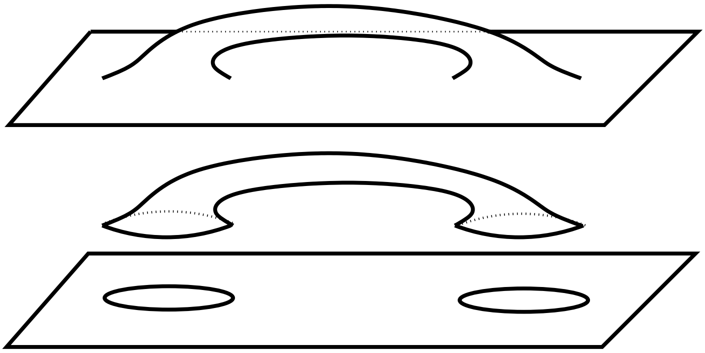


Figure 3: A schematic figure of two dimensional wormhole construction by gluing surfaces. Note that in the artificial wormhole construction components are three dimensional.

tion and cover K with "invisibility cloaking material", that in the boundary normal coordinates around K has the same representation as $\widetilde{\varepsilon}$ and $\widetilde{\mu}$ when cloaking an infinite cylinder. Finally, let

$$U = \{x : \ \mathrm{dist}(x, K) > 1\} \subset \mathbb{R}^3.$$

The set U can be considered both as a subset of $N \subset \mathbb{R}^3$ and the wormhole manifold M, $U \subset M_1$. Then all measurements of fields E and H in $U \subset M$ and $U \subset N$ coincide with currents that are supported in U, that is, thus $(N, \widetilde{\varepsilon}, \widetilde{\mu})$ behaves as the wormhole M in all external measurements.

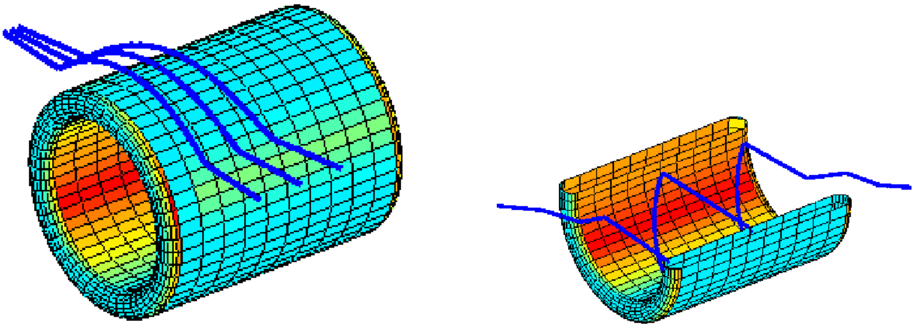


Fig. 4 (a) Rays travelling outside. (b) A ray travelling inside.

In Fig. 4, we give ray-tracing simulations in and near the wormhole. The obstacle in the figures is K, and the metamaterial corresponding to $\widetilde{\varepsilon}$ and $\widetilde{\mu}$ is not shown.

References

[AE] A. Alu and N. Engheta, Achieving transparency with plasmonic and metamaterial coatings, *Phys. Rev. E*, **72**, 016623 (2005)

[AP] K. Astala and L. Päivärinta: Calderón's inverse conductivity problem in the plane. *Annals of Math.*, **163** (2006), 265-299.

[ALP] K. Astala, M. Lassas, and L. Päiväirinta, Calderón's inverse problem for anisotropic conductivity in the plane, *Comm. Partial Diff. Eqns.* **30** (2005), 207–224.

[ALP2] K. Astala, M. Lassas, and L. Päivärinta, Limits of visibility and invisibility for Calderón's inverse problem in the plane, in preparation.

[BeK] M. Belishev, Y. Kurylev, To the reconstruction of a Riemannian manifold via its spectral data (B-method), *Comm. Part. Diff. Eqns.*, **17** (1992), 767–804.

[BT] R. Brown and R. Torres, Uniqueness in the inverse conductivity problem for conductivities with 3/2 derivatives in $L^p, p > 2n$, *J. Fourier Analysis Appl.*, **9**(2003), 1049-1056.

[BU] R. Brown and G. Uhlmann, Uniqueness in the inverse conductivity problem with less regular conductivities in two dimensions, *Comm. PDE*, **22**(1997), 1009-10027.

[C] A.P. Calderón, On an inverse boundary value problem, *Seminar on Numerical Analysis and its Applications to Continuum Physics (Rio de Janeiro, 1980)*, pp. 65–73, Soc. Brasil. Mat., Rio de Janeiro, 1980.

[CPSSP] S. Cummer, B.-I. Popa, D. Schurig, D. Smith, and J. Pendry, Full-wave simulations of electromagnetic cloaking structures, *Phys. Rev. E* **74**, 036621 (2006).

[GKLU1] A. Greenleaf, Y. Kurylev, M. Lassas, G. Uhlmann: Full-wave invisibility of active devices at all frequencies. *Comm. Math. Phys.* **275** (2007), 749-789.

[GKLU2] A. Greenleaf, Y. Kurylev, M. Lassas, G. Uhlmann: Electromagnetic wormholes and virtual magnetic monopoles from metamaterials. *Phys. Rev. Lett.* **99**, 183901 (2007).

[GKLU3] A. Greenleaf, Y. Kurylev, M. Lassas, G. Uhlmann: Effectiveness and improvement of cylindrical cloaking with the SHS lining. *Optics Express* **15** (2007), 12717-12734.

[GLU1] A. Greenleaf, M. Lassas, and G. Uhlmann, The Calderón problem for conormal potentials, I: Global uniqueness and reconstruction, *Comm. Pure Appl. Math* **56** (2003), no. 3, 328–352.

[GLU2] A. Greenleaf, M. Lassas, and G. Uhlmann, Anisotropic conductivities that cannot detected in EIT, *Physiolog. Meas.* (special issue on Impedance Tomography), **24** (2003), 413-420.

[GLU3] A. Greenleaf, M. Lassas, and G. Uhlmann, On nonuniqueness for Calderón's inverse problem, *Math. Res. Lett.* **10** (2003), no. 5-6, 685-693.

[HLS] I. Hänninen, I. Lindell, and A. Sihvola, Realization of Generalized Soft-and-Hard Boundary, *Prog. Electromag. Res.*, PIER **64**, 317 (2006).

[KK] A. Kachalov and Y. Kurylev, Multidimensional inverse problem with incomplete boundary spectral data, *Comm. Part. Diff. Eq.*, **23** (1998), 55-95.

[KKL] A. Kachalov, Y. Kurylev and M. Lassas, Inverse Boundary Spectral Problems, *Chapman and Hall/CRC Monogr. and Surv. in Pure and Appl. Math.,* **123**. Chapman and Hall/CRC, Boca Raton, 2001. xx+290 pp.

[Ki1] P.-S. Kildal, Definition of artificially soft and hard surfaces for electromagnetic waves, *Electron. Lett.* **24** (1988), 168–170.

[Ki2] P.-S. Kildal, Artificially soft and hard surfaces in electromagnetics, *IEEE Trans. Ant. and Prop.*, **38**, no. 10, 1537-1544 (1990).

[KKM] T. Kilpeläinen, J. Kinnunen, and O. Martio, Sobolev spaces with zero boundary values on metric spaces. *Potential Anal.* **12** (2000), no. 3, 233–247.

[KLO] V, Kolehmainen, M. Lassas, P. Ola, Inverse conductivity problem with an imperfectly known boundary, *SIAM J. Appl. Math.* **66** (2005), 365–383.

[KSVW] R. Kohn, H. Shen, M. Vogelius, and M. Weinstein, Cloaking via change of variables in Electrical Impedance Tomography, preprint (August, 2007).

[KV] R. Kohn and M. Vogelius, Identification of an unknown conductivity by means of measurements at the boundary, in *Inverse Problems, SIAM-AMS Proc.*, **14** (1984).

[Ku] Y. Kurylev, Multidimensional inverse boundary problems by the B-me: groups of transformations and uniqueness results, *Math. Comput. Modelling*, **18** (1993), 33-46.

[KLS] Y. Kurylev, M. Lassas, and E. Somersalo, Maxwell's equations with a polarization independent wave velocity: Direct and inverse problems, *Jour. de Mathématiques Pures et Appliqués*, **86** (2006), 237-270.

[LaU] M. Lassas and G. Uhlmann, Determining Riemannian manifold from boundary measurements, *Ann. Sci. École Norm. Sup.*, **34** (2001), no. 5, 771–787.

[LTU] M. Lassas, M. Taylor, and G. Uhlmann, The Dirichlet-to-Neumann map for complete Riemannian manifolds with boundary, *Comm. Geom. Anal.*, **11** (2003), 207-222.

[LeU] J. Lee and G. Uhlmann, Determining anisotropic real-analytic conductivities by boundary measurements, *Comm. Pure Appl. Math.*, **42** (1989), 1097–1112.

[Le] U. Leonhardt, Optical Conformal Mapping, *Science* **312** (23 June, 2006), 1777-1780.

[LeP] U. Leonhardt and T. Philbin, General relativity in electrical engineering, *New J. Phys.*, **8** (2006), 247, http://www.njp.org, doi: 10.1088/1367-2630/8/10/247.

[Li] I. Lindell, Generalized soft-and-hard surface, *IEEE Tran. Ant. and Propag.*, **50** (2002), 926–929.

[MBW] G. Milton, M. Briane, and J. Willis, On cloaking for elasticity and physical equations with a transformation invariant form, *New J. Phys.*, **8** (2006), 248, http://www.njp.org, doi:10.1088/1367-2630/8/10/248.

[MN] G. Milton and N.-A. Nicorovici, On the cloaking effects associated with anomalous localized resonance, *Proc. Royal Soc. A* **462** (2006), 3027–3059.

[N] A. Nachman, Reconstructions from boundary measurements, *Ann. of Math.* (2) **128** (1988), 531–576.

[N1] A. Nachman, Global uniqueness for a two-dimensional inverse boundary value problem, *Ann. of Math.* **143** (1996), 71-96.

[OPS] P. Ola, L. Päivärinta, E. Somersalo, An inverse boundary value problem in electrodynamics. *Duke Math. J.* **70** (1993), no. 3, 617–653.

[PPU] L. Päivärinta, A. Panchenko and G. Uhlmann, Complex geometrical optics for Lipschitz conductivities, *Revista Matematica Iberoamericana*, **19**(2003), 57-72.

[PSS1] J.B. Pendry, D. Schurig, and D.R. Smith, Controlling electromagnetic fields, *Science* **312** (23 June, 2006), 1780 - 1782.

[PSS2] J.B. Pendry, D. Schurig, and D.R. Smith, Calculation of material properties and ray tracing in transformation media, *Opt. Exp.* **14**, 9794 (2006).

[Sc] D. Schurig, J. Mock, B. Justice, S. Cummer, J. Pendry, A. Starr, and D. Smith, Metamaterial electromagnetic cloak at microwave frequencies, *Science* **314** (2006), no. 5801, pp. 977-980.

[Sh] V. Shalaev, W. Cai, U. Chettiar, H.-K. Yuan, A. Sarychev, V. Drachev, and A. Kildishev, Negative index of refraction in optical metamaterials *Optics Letters*, **30** (2005), 3356-3358

[SuU] Z. Sun and G. Uhlmann, Anisotropic inverse problems in two dimensions", *Inverse Problems*, **19**(2003), 1001-1010.

[S] J. Sylvester, An anisotropic inverse boundary value problem, *Comm. Pure Appl. Math.* **43** (1990), 201–232.

[SyU] J. Sylvester and G. Uhlmann, A global uniqueness theorem for an inverse boundary value problem. *Ann. of Math.* **125** (1987), 153–169.

[U] G. Uhlmann, Scattering by a metric, Chap. 6.1.5, in *Encyclopedia on Scattering*, Academic Pr., R. Pike and P. Sabatier, eds. (2002), 1668-1677.

[W] R. Weder, A Rigorous Time-Domain Analysis of Full–Wave Electromagnetic Cloaking (Invisibility), arXiv:0704.0248v1,2,3.

BIRATIONAL GEOMETRY OLD AND NEW

ANTONELLA GRASSI

Abstract. An important problem in many areas of geometry, is how to "complete" an open variety, and also, to understand the relations between the different ways to complete one given variety. In the language of algebraic geometry this becomes understanding birational equivalence. Two varieties are birationally equivalent if they have isomorphic Zariski open subsets, that is if they are isomorphic outside the complement of varieties of lower dimension. A classical problem in algebraic geometry is to describe quantities that are invariants under birational equivalence as well as to determine some convenient birational model for each given variety, a minimal model.

My talk will discuss some of the ideas involved, recent advances on the existence of minimal models and applications.

1. Plan

It is not possible to give here adequate credit to all the people who contributed to the results stated. The reference between brackets often merely indicates a paper where a proof could be found.

An important problem in many areas of geometry, is how to "complete" an open variety, and also, to understand the relations between the different ways to complete one given variety. In the language of algebraic geometry this becomes understanding birational equivalence. A classical problem in algebraic geometry is to describe quantities that are invariants under birational equivalence as well as to determine some convenient birational model ("minimal model") for each given variety. One such quantity is the ring of objects which transform like a tensor power of a differential of top degree, known as the canonical ring.

I will discuss some recent advances in birational geometry, in particular the proof of the existence of minimal models for varieties of general type and the (algebro-geometric proof) of the finite generation of the canonical ring. These results have been long standing conjectures in algebraic geometry.

In this discussion, I will attempt to assume no previous familiarity with algebraic geometry.

While the existence, and meaning, of minimal models turns out to be fairly clear in the case of curves, the finite generation is not always obvious. In the first part of the last century minimal models were defined and shown to exists for surfaces; a complete classification in birational classes was then later understood. The question of the finite generation of the canonical ring was settled much later, as a consequence of a the birational classification of surfaces.

2000 *Mathematics Subject Classification.* Primary 14E30; Secondary 14J99.

Key words and phrases. Algebraic geometry.

The definition of minimal models which works so well for surfaces does not make sense in higher dimension: an appropriate generalization was provided in the '80s, and the "minimal model program" was successfully carried out in dimension 3.

Since then many partial results have been obtained, and quite a few turn out to be essential to the proofs of the theorems which solve the conjectures mentioned at the beginning. These conjectures, and their resolutions, are discussed in Section 4. The preceding Sections are a background. Section 2 contains examples, facts, and results to introduce algebraic geometry, in particular birational algebraic geometry. Section 3 is an introduction to the minimal model program: we state results, examples, motivations and lingo. Some newer results and some general comments are in Section 5.

`Acknowledgments`: I would like to thank A. Collino, J. M^{c}Kernan, D. Morrison, M. Rossi, J. Shaneson and S. Shatz for helpful comments.

2. Old Birational Algebraic Geometry

2.1. **Some Algebraic Geometry.** The origin of algebraic geometry can perhaps be traced to René Descartes, who noticed that certain geometric objects, namely the graphs of equations, could be studied by combining techniques from algebra and geometry. The fundamental insight is that equations represent relations among different quantities and the representation of these equations in graph form results in either a curve, a surface or an object of higher dimension. When the equations are given in polynomial terms, the corresponding graphs are called *algebraic varieties*.

The polynomials are taken with *complex* coefficients, so our varieties will be *complex algebraic varieties*. The complex numbers are in fact in many ways easier than the real numbers: a complex polynomial of degree n in one variable always has n roots (counted with multiplicity) and differentiable functions of complex variables are better behaved than the real ones.

For example, if $f(z_1, z_2)$ is a complex polynomial of degree 2,

$$C = \{(z_{,1}\,, z_2) \in \mathbb{C}^2 \text{ such that } f(z_1, z_2) = 0\}$$

is an algebraic curve (and a plane conic). We will consider here *projective varieties*, that is varieties defined by *homogenous* polynomials in the projective space $\mathbb{P}^N$:

Definition 1. $\mathbb{P}^N = (\mathbb{C}^{N+1} \setminus 0)\ /\mathbb{C}^*$, where $[a_0, \cdots, a_N] = [\lambda a_0, \cdots, \lambda a_N]$, $a_j \in \mathbb{C}$, not all zero, and $\lambda \in \mathbb{C}^*$; recall that $\mathbb{C}^* \overset{def}{=} \mathbb{C} \setminus \{0\}$.

In the *Zariski topology* the closed sets are the algebraic subvarieties. For example the algebraic variety $Z = [0, z_1, z_2] \in \mathbb{P}^2$ defined by $z_0 = 0$ is a closed set, its complement $\mathcal{U} = \mathbb{P}^2 \setminus Z$ is an *open set* which can be identified with $\mathbb{C}^2$, via the map $[1, z_2, z_2] \mapsto (z_1, z_2) \in \mathbb{C}^2$. Note also that Z can be identified with $\mathbb{P}^1$.

The above identifications are examples of *isomorphisms*:
$f : X \dashrightarrow Y$ is a rational map if and only if it can be represented by well defined rational functions on open sets, satisfying obvious compatibility conditions. Note that the open sets in X do not need to cover the whole variety X.

Example 2. $f : \mathbb{P}^2 \dashrightarrow \mathbb{P}^1$ defined as $[z_0, z_1, z_2] \mapsto [z_0, z_1]$ is a rational map. In fact it is defined on the open set $\{\mathbb{P}^2 \setminus [z_0, z_1, 1]\} \subset \mathbb{P}^2$ and defined by a rational function (actually a polynomial).

f is birational if it is invertible on Zariski open sets and the inverse is also rational; f is called an isomorphism if f and its inverse are defined everywhere; in this case we write $X \simeq Y$.

Example 3. Let $C \subset \mathbb{P}^2$ be the cubic curve defined as $z_3 z_2^2 - z_1^3 = 0$; then $f : \mathbb{P}^1 \dashrightarrow C$ as $[1, t] \mapsto [t^2, t^3, 1]$ is a birational morphism, but it can be shown it is not an isomorphism.

Example 4. A ***blow up***. Let us consider $\mathbb{C}^2 \times \mathbb{P}^1$ with coordinates $((x_0, x_1), [z_0, z_1])$ and $B = \{((x_0, x_1), [z_0, z_1]) \in \mathbb{C}^2 \times \mathbb{P}^1 \text{ such that } x_j z_i - x_i z_j = 0\}$.
Then $f : B \to \mathbb{C}^2$ with $((x_0, x_1), [z_0, z_1]) \mapsto (x_0, x_1)$ is a birational morphism. Note also that $f_{|_{B \setminus f^{-1}(0,0)}} : B \setminus f^{-1}(0, 0) \to \mathbb{C}^2 \setminus (0, 0)$ is an isomorphism and that $f^{-1}(0, 0) \simeq \mathbb{P}^1$.

$E \overset{def}{=} f^{-1}(0, 0)$ is called the ***exceptional divisor*** of the birational morphism f (see also the introduction to Section 3).

Example 4 can be extended to a birational morphism between compact varieties. Let $\widehat{B}$ be the (Zariski) closure of B in $\mathbb{P}^2 \times \mathbb{P}^1$. Then f extends to: $\hat{f} : \widehat{B} \to \mathbb{P}^2$, which restricts to an isomorphism: $\hat{f}_{|_{\widehat{B} \setminus \hat{f}^{-1}([0,0,1])}} : \widehat{B} \setminus \hat{f}^{-1}([0, 0, 1]) \to \mathbb{P}^2 \setminus [0, 0, 1]$.

So the blow up construction gives two different ways to **"complete"** the open variety $\mathbb{P}^2 \setminus [0, 0, 1]$ to different projective varieties. Actually, since the above construction is local around $(0, 0) \in \mathbb{C}^2$, then it can be carried through for other varieties, and also, other dimensions.

An important problem in many areas of geometry is how to ***"complete"*** an open variety. One example is the compactification of the moduli space of smooth curves of genus g, $\mathcal{M}_g$ (see below for the definition of genus).

The search for the ***"minimal model"*** is in essence the search for the simplest way to birationally complete an open variety, and also, understand the relations between the different minimal models, if more than one model exists. As we will see below and in Section 3, in case of curve and surfaces, it was shown that every birational class has a *smooth* minimal model. [An algebraic variety is smooth (non singular) at a point P if the rank of the Jacobian of the defining polynomials (locally around P) is maximal. A theorem of Zariski shows that this definition does not depend on the embedding given by the defining polynomials.][1]

A related question is to describe quantities which are invariant under birational morphisms. Before defining some of these invariants (the Kodaira dimension $\kappa(X)$, the plurigenera, the canonical ring), let us see their significance in the case of complex curves:

Riemann showed that any compact connected orientable (real) surface can be embedded in a projective space $\mathbb{P}^N$, where the image, a complex curve C, is defined by a finite set of polynomial equations. The Riemann surfaces are determined up to homeomorphisms by the number g of holes, which is defined to be the ***genus*** of the complex curve C. The topological Euler characteristic is $\chi_{top}(C) = 2 - 2g$, which turns out to be the negative of the *degree* of the *canonical divisor* K_C. If

[1] We consider normal varieties, see for example [6]. If X is normal, then the singular locus has at most codimension 2. For example all normal curves are also smooth, normal surfaces are singular at most in a set of points.

the degree is negative, the curvature of the real surface (a sphere) is positive. If the degree is zero, the surface is flat (a torus); if the degree of the canonical divisor is positive, the curvature is negative. Equivalently the area of a circle of radius r is respectively smaller, equal or bigger than πr^2. The *Kodaira dimension* is the algebraic analogue of the curvature, and in the above case we say it is respectively negative, zero or positive.

2.1.1. *Canonical divisors, Kodaira dimensions, canonical rings.* An effective ***Weil*** divisor D on a smooth variety X is a codimension 1 subvariety of X, for example a collection of points on a curve or a collection of curves on a surface. A divisor is a finite sum $D = \sum_i a_i D_i$ with $a_i \in \mathbb{Z}$; $\sum_i a_i$ is the degree of D. A function f on X defines a divisor, namely the locus of zeroes and poles counted with multiplicity. We define an equivalence relation among divisors where the divisors of functions are the zero classes.

Locally on an open set $\mathcal{U}_j$ a divisor is defined by a function f_j. One could also define a divisor as the collection of these open sets and functions $\{(\mathcal{U}_j, f_j)\}$ with obvious compatibility conditions and equivalence relation. These are called ***Cartier divisors***, and coincide with the Weil divisors when X is smooth, see also Definition 13. One can also view Cartier divisors as invertible sheaves (or line bundles) [6]. Multiplication of functions induces tensor powers of the line bundles and sums of divisors. We will use the additive notation, say mD represents the m-th tensor power of the line bundle D.

Example 5. If $C \subset \mathbb{P}^2$ is a curve, and $H \subset \mathbb{P}^2$ a hyperplane, the intersection of H with C, denoted as $H \cdot C$, defines a divisor on C, which is called *very ample*. This generalizes to any subvariety $C \subset \mathbb{P}^N$ and a hyperplane $H \subset \mathbb{P}^N$.

The property of being very ample is intrinsic; a very ample divisor determines an embedding. It is a crucial problem to determine a priori which divisors are very ample and which divisors determine other morphisms (not necessarily embeddings); all the divisors which determine morphisms are called ***semiample***.

The set of global sections of a divisor D form vector space, denoted by $H^0(X, D)$. The ***canonical form*** on $\mathbb{C}^n$, with coordinates $(z_1, \cdots, z_n)$ is a top differential form, for example $dz_1 \wedge \cdots \wedge dz_n$. On a general smooth variety X, the canonical form is a collection of local volume forms on open sets with obvious compatibility conditions. The ***canonical divisor*** is the divisor naturally associated to it and denoted by K_X.

The **canonical ring** [17] is: $R(X, K_X) = \oplus_{m \geq 0} H^0(mK_X)$; we assume X to be connected and thus: $H^0(X, 0K_X) = \mathbb{C}$.

If $H^0(X, mK_X) = 0$, $\forall m > 0$, then the ***Kodaira dimension*** of X is defined to be negative (or $\kappa(X) = -1$). Otherwise, it is the growth rate of $dim H^0(X, mK_X)$ as m goes to infinity. Obviously $\kappa(X) \leq \dim(X)$; X is of ***general type*** if $\kappa(X) = \dim(X)$. The curves of general types are the curves with genus $g \geq 2$. It turns out that $\kappa(X) = trasc.deg.R(X, K_X) - 1$. There are many varieties of general type, for example any hypersurface of degree greater than $d+1$ in $\mathbb{P}^d$ (this follows from the adjunction formula, see for example [6]). $dim H^0(X, mK_X)$ is called the *m-plurigenus.*

Similarly, one defines $\kappa(X,D)$, the Kodaira dimension of a divisor D, as the growth rate of $dim H^0(X,mD)$ as m goes to infinity. If $\kappa(X,D)=\dim(X)$, D is called ***big*** and D is of ***general type***. If X is a variety of general type, K_X is big.

Example 6. (Genus $g=0$.) If $X=\mathbb{P}^1$, on the open set $\mathcal{U}_1 \simeq \{[s,1], s\in\mathbb{C}\}$ the local volume form is ds, if $\mathcal{U}_2\simeq\{[1,t]\}$, on $\mathcal{U}_1\cap\mathcal{U}_2$ $s=1/t$ and the local volume form is $ds=d(1/t)=-1/t^2dt$. The canonical divisor (Weil) corresponds to a pole with multiplicity 2; it is not an effective divisor, but its inverse is. In this case, $H^0(\mathbb{P}^1,K_X)=0$, $\kappa(X)<0$ and the canonical ring is trivial.

Example 7. (Genus $g=1$.) The torus $E=\mathbb{C}^2/\mathbb{Z}\oplus\imath\mathbb{Z}$ can be given the structure of an algebraic curve X, which can be identified with a smooth cubic in $\mathbb{P}^2$. It is easy to see that the volume form on $\mathbb{C}^2$, $dz_1\wedge dz_2$ induces a global volume form X. The divisor associated to it does not have poles nor zeroes, it is the trivial divisor. Here $H^0(E,K_X)=\mathbb{C}$, $\kappa(X)=0$, and the canonical ring is $R(X,K_X)=\oplus_m\mathbb{C}$.

Example 8. (*The tricanonical embedding.*) It turns out that if C is a curve of genus $g\geq 2$, then there is an embedding $C\hookrightarrow\mathbb{P}^N$ with the following property: if $H\subset P^N$ is an hyperplane, then $H\cdot C$ is $3K_C$. Then the canonical ring $R(C,K_C)$ can be reconstructed[2] from the coordinate ring of $\mathbb{P}^N$.

Remark 9. If X is a smooth curve

$$K_X=\Omega^1_X=T^*_X \text{ and } \chi_{top}(X)=degT_X=2-2g,$$

where T_X,T^*_X denote the tangent and cotangent bundle respectively. Note that $c_1(X)=-K_X$. Then K_X is non effective (we write $K_X<0$) if and only if $g=0$, K_X is trivial (we write $K_X=0$) if and only if $g=1$ and $K_X>0$ if and only if $g\geq 2$.

These examples show that in the case of curves, the canonical ring is always finitely generated. In the case of surfaces, it is much harder to show that the canonical ring is finitely generated; the proof relies on the existence of minimal models and the birational classification of surfaces.

In some cases the canonical divisor can be defined also when X is singular. For example, in the case of the surface below:

Example 10. (Kummer surface) Let $X=(E\times E)/<\imath,id>$, where E is the torus in Example 7 and $\imath:(z_1,z_2)\mapsto(-z_1,-z_2)$ with $(z_1,z_2)\in\mathbb{C}^2$. X has 16 singular points, and the canonical differential on $E\times E$ is preserved by the group action and induces a canonical divisor on X. The canonical divisor on X is still trivial: $H^0(X,K_X)\simeq\mathbb{C}$.

In some other cases the canonical divisor is not invariant under a group action:

Example 11. Let $X=\mathbb{C}^2/<e^{\pi i/3},e^{2\pi i/3},1>$, where each group element acts by multiplication on each factor. X is singular at $(0,0)$, but the canonical differential on $\mathbb{C}^2$ is not preserved by the group action. However, the third tensor power of the canonical differential is, and induces a Cartier divisor $3K_X$ on X. This is an example of a $\mathbb{Q}$-*Cartier divisor* (see Definition 13).

[2] In fact Serre's vanishing theorem for the first cohomology of the ideal sheaf of C in $\mathbb{P}^N$ (twisted by m) shows that there is a surjection $H^0(\mathbb{P}^N,mH)\to H^0(C,3mK_C)$.

3. Minimal models? An introduction.

We have seen that birational morphism between algebraic varieties X and Y is a map which is an isomorphism between Zariski open sets. Zariski open sets are dense in algebraic varieties, and the existence of a birational morphism between X and Y implies strong relations between them. For example, a birational morphism exists if and only if X and Y have isomorphic function fields.

In the case of curves, there is a unique smooth projective curve in each birational equivalence class. This is not true for surfaces: we have seen in Example 4 that $\mathbb{P}^2$ and $\widehat{B}$ are smooth and birational; we noted that any smooth point on a surface can be blown up to obtain another smooth, birationally equivalent, surface. The "Italian school" defined $\bar{S}$ to be ***minimal*** if any birational morphism from $\bar{S}$ was actually an isomorphism. It is a classical result that any surface has a (smooth) minimal model and that the model is *unique* if S is not birationally equivalent to $\mathbb{P}^1 \times C$, for some curve C, that is if S is not *ruled.* S is ruled if and only if $\kappa(S) < 0$. The Enriques-Severi-Kodaira classification then provides a finer structure theory for the non ruled surfaces, (see for example [1]). This classification is not discussed here.

For a long time it was believed that such a classification would be impossible in higher dimension. One of the major obstacles was the lack of a "good" analogue of the notion of minimal model. In 1982 Mori [15] constructed the first step of a contraction algorithm to build a minimal model for threefolds.

The starting point is the following observation: If $\kappa(S) \geq 0$, the classical definition of minimality is equivalent to the canonical divisor being ***nef*** i.e. $K_S \cdot C \geq 0$, for any curve C on S. The key idea of Mori's algorithm is to contract all curves C such that $K_X \cdot C < 0$, and define X to be *minimal* if $K_X \cdot E \geq 0$ for all curves E on X. In fact, it turns out that in the case of smooth surfaces, a birational morphism $f : S \to \bar{S}$ which is not an isomorphism contracts (to points) curves E on S such that $K_S \cdot E < 0$. Conversely, *Castelnuovo's contraction criterion* guarantees the existence of a morphism contracting any such curve.

To iterate the process in dimension 3 or higher one needs to allow certain mild singularities. If the exceptional locus of a contraction morphism is not a divisor, but say a curve C on a threefold X, then the singularities on the image variety are such that the algorithm cannot be applied again. To avoid the problem of these "small" contractions one would like to ***flip*** the curve C, that is to construct another threefold (with mild singularities) X^+, isomorphic to X outside C, but such that K_{X^+} intersects positively the "transform" of C ($K_{X^+} \cdot C^+ > 0$). Then one would need to show that a sequence of flips terminates. Mori's result on existence of flips in 1988 [16] completed the proof of the existence of such minimal models for threefolds with positive Kodaira dimension. The proof gives an algorithm for constructing minimal models.

Several groups of people have since been working on extending the same procedure to arbitrary algebraic varieties. If a minimal model $\bar{X}$ exists and $K_{\bar{X}}$ is also big, then $K_{\bar{X}}$ is semiample, by a result of Kamawata and Shokurov and the canonical ring is finitely generated, as in Example 8.

In the summer of 2005 Siu [19] announced the proof of the finite generation of the canonical ring for a smooth variety of general type. The proof is based on analysis, not on the existence on minimal models.

In the Fall of 2006, Birkhar, Cascini, Hacon and M^{c}Kernan [2] posted a paper on ArXiv on the existence of minimal models in all dimensions for varieties of general type. A corollary is the finite generation of the canonical ring. Their proof of the existence of minimal models for a n-dimensional variety of general type does not rely on the existence or termination of flips, but it is a clever round-about induction on n. The argument of the proof works in the ***log*** set up, which we will define in Section 3.1.3 below. We discuss their results in Section 4.

The existence of flips in all dimensions follows from finite generation.

3.1. **Minimal models: Mori's program.** In this section we summarize Mori's program, and some of the results obtained in this direction, especially for the case of threefolds.

3.1.1. *The Mori Cone; Ample, Nef, Pseudoeffective cones and Big divisors.*

A 1-*cycle* is a linear combination of subvarieties of dimension 1 of a variety X:

Definition 12. (The Mori Cone) Let X be a smooth variety. $NE(X)$ denotes the cone generated by classes of effective 1-cycles, with *real coefficients*, modulo numerical equivalence, and $\overline{NE}(X)$ denotes its closure. A ray R is defined as $R = \mathbb{R}^+[Z]$, for some $Z \in \overline{NE}(X)$. The ray R is ***extremal*** (in the sense of convex geometry) if $z_1 + z_2 \in R \Rightarrow z_1, z_2 \in R$ and $K_X \cdot R < 0$.

Let R be an extremal ray; we are interested in the *supporting divisors* of R, that is: the Cartier divisors D such that $D \cdot Z \geq 0$, for all $Z \in \overline{NE}(X)$ with equality if and only if $Z \in R$. If D is a supporting divisor for an extremal ray R, then mD is ***base point free*** for sufficiently large integers m, that is there exists a morphism

$$\phi_{|mD|} : X \to X_1$$

which contracts exactly the curves C such that $[C] \in R$ (see, for example [14]); we then say that D is semiample. Proving that every extremal ray has such a supporting divisor is a first step in generalizing Castelnuovo's contraction criterion for surfaces [1].

A divisor D is called **ample** if mD is very ample (see Example 5), for some $m \in \mathbf{N}$; Kleiman's criterion, see for example [14], says that D is ample if $D \cdot C > 0$, for every curve $C \in \overline{NE}(X)$. A divisor D is ***nef*** if $D \cdot C \geq 0$, for every curve $C \in \overline{NE}(X)$; the nef cone and the Mori cone are then dual cones (with respect to the intersection pairing).

It is also useful to consider the cone of effective divisors, with real coefficients, modulo numerical equivalence (which means: from the point of view of intersection with curves, they behave the same). Its closure is called the **pseudo-effective cone**; a divisor D in the closure is called ***pseudo-effective***.

A Cartier divisor D on a variety X is ***big*** if the dimension of the spaces of sections $H^0(X, mD)$ grows like $m^{\dim(X)}$.

Note that the following four inclusions hold:

$$\begin{tikzcd} Ample \arrow[r, hook] \arrow[d] & Eff \arrow[d] \\ Nef \arrow[r, hook] & Pseudoeff \end{tikzcd}$$

Definition 13. (See also Example 11.) A Weil divisor D on a normal variety is called $\mathbb{Q}$-***Cartier*** if mD is Cartier for some $m > 0$. X is $\mathbb{Q}$-***factorial*** if every Weil divisor is $\mathbb{Q}$-Cartier. Let D and E be $\mathbb{Q}$-Cartier divisors. $D =_{\mathbb{Q}} E$ means that there exists a m such that mD and mE are equivalent divisors. If C is a curve, we also define:

$$D \cdot C \overset{def}{=} \frac{1}{m}\,(mD \cdot C).$$

The above definitions can be applied also to $\mathbb{Q}$-divisors, and $\mathbb{R}$-divisors. In fact the proofs in [2] relies on $\mathbb{R}$-divisors.

3.1.2. *Singularities.* It turns out that if $dim(X) \geq 3$, contracting curves C on extremal rays may produce *singularities* on the image variety (see [15]). The nice thing is that the program also works with these singularities, and these are called *terminal.*

Remark 14. It turns out that if the canonical ring $R(X, K_X)$ of a variety of general type X is finitely generated, then $R(X, K_X)$ defines a, possibly singular, variety W; more precisely, $W = \text{Proj}(R(X, K_X))$. W is called the ***canonical model*** of X, and the singularities are called *canonical.* The canonical model is unique.

We have the following Definition/Theorem:

X has at most ***canonical*** singularities if K_X is $\mathbb{Q}$-Cartier and for every resolution $h : \widetilde{X} \to X$, we have $mK_{\widetilde{X}} = mh^*(K_X) + F$, for some effective divisor F supported on the exceptional locus. If every irreducible component of the exceptional locus appears in F with strictly positive multiplicity, then X is said to have ***terminal*** singularities.

Definition 15. X is a ***canonical*** (resp. ***minimal***) model if X has at worst canonical (resp. terminal) singularities and K_X is an ample (resp. nef) $\mathbb{Q}$-Cartier divisor.

If $dim(X) = 2$, then the canonical singularities are also known as rational double points, and they are locally of the form $\mathbb{C}^2/\Gamma$, where $\Gamma \subset \text{SL}(2, \mathbb{C})$ is a finite subgroup. The singularities in Example 10 are canonical. If X is a minimal model, then X is smooth and a minimal in the classical sense.

Theorem 16. [9]*(The cone theorem) Let X be a projective variety with at worst canonical singularities. Then every extremal ray has a supporting semiample divisor.*

Remark 17. If X is not a minimal model, there exists and extremal ray R and a supporting semiample divisor D; D determines the morphism $f : X \to Y$ (see Example 5):

Theorem 18. [16] *Let X be a projective variety with only $\mathbb{Q}$-factorial terminal singularities. If X is not a minimal model, then there exists a surjective morphism $f : X \to Y$ to a normal projective variety Y with connected fibers and one of the following holds:*

(1) *$dim(X) > dim(Y)$ (Mori fiber space).*
(2) *f is birational and contracts a divisor (divisorial contraction)*
(3) *f is birational and contracts no divisor (small contraction).*

Furthermore:

Case (1) holds if and only if $\kappa(X) < 0$.
In Case (2) Y has at most terminal singularities.
In Case (3) Y is not $\mathbb{Q}$-factorial.

Case (3) appears to be the "bad" case, because the image variety does not have terminal singularities and the above Theorems cannot be applied again. In the case of threefolds, Mori proved the following:

Theorem 19. [16] *Let $f : X \to Y$ be a birational morphism between normal projective threefolds as in Case (3) of the previous Theorem. Then there is a birational morphism $f^+ : X^+ \to Y$*

$$\begin{tikzcd}
X \arrow[rr, dashed, leftrightarrow] \arrow[dr, "f"'] & & X^+ \arrow[dl, "f^+"] \\
& Y &
\end{tikzcd}$$

where X^+ is a projective threefod with only $\mathbb{Q}$-factorial terminal singularities, f' contracts no divisors and $K_{X^+} \cdot C > 0$, for any curve C contracted by f^+.

The birational map between X and X^+ is an isomorphism in codimension 1, and is called a ***flip***.

This gives the *minimal model algorithm* for threefolds:

Let X be a threefold with $\mathbb{Q}$-factorial terminal singularities:

If K_X is nef, STOP: X is a minimal model.
If K_X is not nef: contract any extremal ray R (as in Theorem 18).
In case (1) X is a Mori fiber space: STOP.
In case (2): replace X by Y,
In case (3): replace X by X^+.
The process is then repeated until the canonical bundle is nef or Case (1) holds.

In fact neither (2) nor (3) can occur an infinite number of times. The termination of Step (2) follows from the fact that at every step the second betti number decreases. The termination of (3) is subtle.

This algorithm provides a minimal model for threefolds.

Remark 20. The choice of a particular extremal ray to contract is arbitrary and minimal models are not unique. There are examples of surfaces with infinitely many extremal rays.
The vareity of general type are somewhat special in this regard: in fact, it turns out (in the hypothesis of the cone theorem) that if a variety is of general type then the number of extremal rays is always finite.

3.1.3. *Log minimal pairs.* There exists also a ***log version*** of the minimal model program:
Let $(X, \mathbf{\Delta})$ be a pair consisting of a variety X and an effective $\mathbb{Q}(\mathbb{R})$-Cartier divisor in X. We might want to consider such a pair, for example, when we have an open variety M and a compactification $X = \bar{M}$, in this case $\mathbf{\Delta}$ is the boundary divisor.

Another case is given by the elliptic fibrations, which we discuss here in the case of smooth surfaces. An elliptic surface is morphism: $\pi : X \to C$, to a curve C where all the fibers, except a finite number, are elliptic curves; let also assume that X is minimal, for simplicity's sake. It is then a classical result [11] that $K_X = \pi^*(K_C + \mathbf{\Delta})$, where $\mathbf{\Delta}$ is a $\mathbb{Q}$-divisor supported on the image of the fibers of π which are not smooth elliptic curves. These singular curves determine also the coefficients of $\mathbf{\Delta}$.

We consider in particular ***kawamata log terminal pairs*** (klt for short), which appears to be the largest class of singularities for which the minimal model program has been shown to run.

Definition 21. Let $(X, \mathbf{\Delta})$ be a log pair such that $0 < a_i < 1$, where $\mathbf{\Delta} = \sum_i a_i \mathbf{\Delta}_i$. The pair is kawamata log terminal if and only if, for any resolution of X, $f : \widetilde{X} \to X$ such that the union of the strict transform of $\mathbf{\Delta}$ and the exceptional locus are supported on divisors with simple normal crossings, we have:
$K_{\widetilde{X}} + \widetilde{\mathbf{\Delta}} = f^*(K_X + \mathbf{\Delta})$, where if $\widetilde{\mathbf{\Delta}} = \sum d_j \widetilde{\mathbf{\Delta}}_j$, $d_j < 1, \forall j$.

For example, the pairs $(C, \mathbf{\Delta})$ in the base of the elliptic fibration described above are ktl: C is taken to be smooth, and $\mathbf{\Delta}$ consists of a finite set of points, so the only condition to verify is the one on the coefficients, which follows from [11].

A curve C on X is log-extremal if $(K_X + \mathbf{\Delta}) \cdot C < 0$ and C is extremal in the usual sense. For example, a version of the "Cone theorem" and contraction theorem still hold and the singularities of the image variety are ***kawamata log terminal***.

In the next Section we will see the log model at work, in the proof of Theorem 23. In some sense, the log pairs in dimension n work as a step towards dimension $n + 1$.

3.1.4. *Relative minimal model program; Finite generation and flips.* There is a "relative" version of the above algorithm, for a morphism f: $X \to Y$. This relative version is also crucial in the proofs of [2]. In these notes we restrict to the following:

Remark 22. (See [9]) Let $f : X \to Y$ be a small contraction.
The sheaf $R_f = \bigoplus_{m \in \mathbf{N}} f_*(mK_X)$ is a sheaf of finitely generated algebra if an only if the flip of f exists. The variety X^+ and the morphism f^+ are then unique (see Remark 14).

4. New results

Theorem 23. [2] *Let X be a smooth projective variety of general type of dim n. Then the minimal model of X exists.*

Theorem 24. [2, 5, 19] *Let X be a smooth variety, then the canonical ring $R(X, K_X)$ is finitely generated.*

4.1. **On the proof of Theorem 23.** The argument uses a log version of the minimal model program, and it relies on induction on n. Some features of varieties of general types have been discussed in Example 8 and Remark 20.

We have seen that the crucial points in completing the minimal model algorithm are the existence and termination of flips. The argument of [2] cleverly avoids both issues: First the problem is modified by considering a suitable log pair $K_X + D$ and showing that flips exist for these particular pairs, following ideas of Shokurov; these are called *pl flips.* The issue of termination is then avoided by considering a

minimal model with scaling and induction. To give a flavor of the argument used in [2] we follow the reasoning, presented in [2] for a very particular case:
Assume that K_X is *big* (see Section 3.1.1) and that $D \in |mK_X|$ is smooth and irreducible.

Consider the pair $K_X + D$. It is clear that K_X is nef if and only if $K_X + D$ is nef; conversely R is an extremal ray for K_X if and only if it is an extremal ray for $K_X + D$. If K_X is not nef, there exists an extremal ray R and a contraction $\phi : X \to Z$, as in Theorem 16.

Case 1: ϕ is a divisorial contraction: then K_Z is trivial and Z is a minimal model.

Case 2: ϕ is a small contraction: ϕ is a small contraction for the pair $K_X + D$, and the results on existence of flips for this particular pair applies. We return to this point below in Remark 25.

Let $\varphi : X \dashrightarrow X'$ be the flip. One still has to show that any sequence of such flips terminates. Now the authors of [2], consider an ample divisor H on X, a positive number $t \in \mathbb{R}_{\geq 0}$ and the pair: $K_X + D + tH$, and restrict to D. Then $(K_X + D)_{|}D = K_D$ (by adjunction, see [6]); let $G = H_{|_D}$. We have then the pair $K_D + tG$ on D, and $\dim(D) = n - 1$; one can apply induction to show that the flips terminate for these pairs. We return to this point below in Remark 26.

Remark 25. The existence of such flips is also proved by induction, following ideas of Kawamata, Shokurov, Siu.

Remark 26. Here the result is proved by showing the finiteness of the possible models of pairs that one would get by flipping the pair $K_D + G$. These are not necessarily log terminal, so a result for broader singularities is needed. Here the hypothesis of general type is used. (See also Remark 20.)

In both cases the induction steps use log pairs $K_X + \Gamma$, for some suitable Γ and the relative setting. In the general case, the strategy is to add a suitable divisor H and consider pair $K_X + \mathbf{\Delta} + tH$. Quite a few subtleties and difficulties occur here and the relative case of the minimal program needs to be considered.

In [2] the following theorem is proved, and Theorem 23 follows as a Corollary.

Theorem 27. *Let X be a projective variety. If $K_X + \mathbf{\Delta}$ is big and $K_X + \mathbf{\Delta}$ is kawamata log terminal, then there exists a log terminal model of $K_X + \mathbf{\Delta}$.*

[A even more general relative version of the above Theorem holds.]

4.2. **On a proof of Theorem 24.** If $\kappa(X) \leq 0$ then $R(X, K_X)$ is obviously finitely generated, as in Examples 6 and 7. It turns out that if $\kappa(X) \geq 1$, then there exists a variety Y and a klt and big log pair: $K_Y + \mathbf{\Delta}$ such that $R(X, K_X) = R(Y, K_Y + \mathbf{\Delta})$ [5].
A model for this result is given by the minimal elliptic surfaces of $\kappa(X) \geq 1$ (see Section 3.1.3). It is again a classical result that $K_X = \pi^*(K_C + \mathbf{\Delta})$, where $K_C + \mathbf{\Delta}$ is big on C.
By Theorem 27, a log terminal model $(\bar{X}, K_{\bar{X}} + \bar{\mathbf{\Delta}})$ exists for the pair $K_Y + \mathbf{\Delta}$; since the log canonical ring is a birational invariant we have $R(Y, K_Y + \mathbf{\Delta}) = R(\bar{X}, K_{\bar{X}} + \bar{\mathbf{\Delta}})$. The base point free theorem (see for example [9]) then shows that $K_{\bar{X}} + \bar{\mathbf{\Delta}}$ is semiample. Then $R(Y, K_{\bar{X}} + \bar{\mathbf{\Delta}})$ is finitely generated (see Example 8), and so is $R(X, K_X)$.

The more general statement of the finite generation is:

Theorem 28. *Let X be a projective variety. If $K_X+\mathbf{\Delta}$ is is kawamata log terminal, and $K_X+\mathbf{\Delta}$ is $\mathbb{Q}$-Cartier. Then the ring*

$$R(X, K_X+\mathbf{\Delta}) = \bigoplus_{m\in\mathbf{N}} H^0((\llcorner m(K_X+\mathbf{\Delta})\lrcorner)$$

is finitely generated. Here $\lfloor k\rfloor$ means the round down to an integer.

Theorems 24 and 27 hold also in a relative setting. In particular the relative finite generation (see also Remark 22):

Theorem 29. *Klt flips exist.*

Theorem 30. *Let X be a smooth variety of general type, then $\operatorname{Proj}(R(X,K_X) = W$ is the canonical model of X (see Remark 14). The canonical model is unique.*

5. Newer results and open problems: Beyond general type?

An obvious question is if the several results stated here for varieties of general type can be shown for all varieties, without the assumption of general type.

For example, X is a minimal model and K_X is ample (then X is of general type), then the minimal model is unique. This is not true in general, and the question is how are the different minimal models related. If X and X' are two different minimal models in the same birational class, then they are isomorphic outside a set of codimension at least 2 [10], as in the following (local) example:

Example 31. Let $\bar{X} = \{xy+zw=0\} \subset \mathbb{C}^4$; $\bar{X}$ is singular at the origin. Now consider the threefolds $X \subset \mathbb{C}^4 \times \mathbb{P}^1$ and $X' \subset \mathbb{C}^4 \times \mathbb{P}^1$ defined by the equations: $\{y_0w+y_1y=0, y_0x+y_1z=0\}$ and $\{y_0w+y_1x=0, y_0y+y_1z=0\}$ respectively, with $[y_0,y_1] \in \mathbb{P}^1$. X and X' are smooth and the natural projection morphisms $\phi : X \longrightarrow \mathbb{C}^4$ and $\phi' : X' \longrightarrow \mathbb{C}^4$ are a isomorphism outside the origin of $\mathbb{C}^4$. The composition $\varphi = \phi \cdot {\phi'}^{-1} : X \dashrightarrow X'$ is an isomorphism outside $L = \phi^{-1}(0,0,0,0)$ and $L' = \phi'^{-1}(0,0,0,0)$. It can be shown that $K_X \cdot L = K_{X'} \cdot L' = 0$. [OK, this can be an exercise!] This transformation is called a ***flop***.

We have the following:

Theorem 32. [8] *Let X and X' two minimal models and $\phi : X \dashrightarrow X'$ a birational map. Then ϕ is a composition of flops.*

5.1. **On the proof.** The statement is obtained by the considering a klt log pair $(X, \epsilon D)$, where D is an ample divisor in X and then decomposing ϕ as a composition of *klt flips* for the pair. Each klt flip for the pair (X_i, D_i) is also a flop for the terminal variety X. The argument uses existence of klt flips, a MMP with scaling and the termination of certain kind of flips, also shown in [2]. Note that the result was first proved by [2], in the case of general type.

The existence of minimal models for varieties of non general type is still unresolved, as I am typing. Once the existence of minimal model is established, one could also ask if there exists a classification in birational classes like the famous Severi-Enriques-Kodaira classification for surfaces of non negative Kodaira dimension. And also construct a finer classifications for Mori fiber spaces, when the Kodaira dimension is negative.

Another unresolved question is whether on a minimal model the canonical divisor is semiample: the theorem of Shokurov and Kawamata (quoted in the comments on the proof of Theorem 24) holds if the canonical divisor is not big.

References

1. A. Beauville, *Algebraic Surfaces*, Cambrige University Press, 1983.
2. C. Birhkar, P. Cascini, C. Hacon, J. M^{c}Kernan, *Existence of minimal model for varieties of log general type*, http://math.mit.edu/ mckernan/Papers/papers.html, May 3, 2007.
3. O. Debarre, *Higher-dimensional algebraic geometry*, Universitext, Springer-Verlag, New York, 2001
4. A. Corti, P. Hacking, J, Kollár, R. Lazarsfeld, *Lectures on Flips and Minimal Models*, ArXiv:math.AG/0706.0494, 1–28, 2007.
5. O. Fujino, S. Mori *A canonical bundle formula*, J. Differential Geom. **56**, no. 1, (2000), 167–188.
6. R. Hartshorne, *Algebraic Geometry*, Springer–Verlag, 1977.
7. C. Hacon,J. M^{c}Kernan, *On the existence of flips* ArXiv:math.AG/0507597, 2005.
8. Y. Kawamata *Flops connect minimal models* ArXiv:math.AG/0704.1013 2007, 1–5.
9. J. Kollár, S. Mori, Birational geometry of algebraic varieties. With the collaboration of C. H. Clemens and A. Corti. textitCambridge Tracts in Mathematics, **134**, Cambridge University Press, Cambridge, 1998.
10. J. Kollár, *Flops*, Nagoya Math. J. **113** (1989), 15–36.
11. K. Kodaira *Collected Works*, **vol. III**, Princeton University Press, 1975.
12. R. Lazarsfeld, *Positivity in algebraic geometry II, Classical setting: line bundles and linear series*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics, **49**, Springer-Verlag, Berlin, 2004
13. R. Lazarsfeld, *Positivity in algebraic geometry II, Positivity for vector bundles, and multiplier ideals*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics, **48**, Springer-Verlag, Berlin, 2004
14. K. Matsuki *Introduction to the Mori program*, Universitext. Springer-Verlag, New York, 2002. xxiv+478 pp
15. S. Mori Threefolds whose canonical bundles are not numerically effective, **116** (1982), 133–176.
16. S. Mori *Flip theorem and the existence of minimal models for 3-folds*, Jour. Amer. Math. Soc. **1** (1988) 117–253.
17. D. Mumford *The canonical ring of an algebraic variety, Appendix to Zariski's paper "The theorem of Riemann-Roch for high multiples of an effective divisor on an algebraic surface"*, Ann. of Math. **76** (2) 1962 612–615
18. M. Reid, *Canonical 3-folds*, Journs de Gmetrie Algrique d'Angers, Juillet 1979/Algebraic Geometry, Angers, 1979, Sijthoff & Noordhoff, Alphen aan den Rijn—Germantown, Md., 273–310, 1980.
19. Y.-T. Siu *A general non-vanishing theorem and an analytic proof of the finite generation of the canonical ring*, arXiv:math.AG/0610740
20. V.V. Shokurov, *Numerical geometry of algebraic varieties*, Proceedings of the International Congress of Mathematicians, **Vol. 1**, 2 (Berkeley, Calif., 1986), Amer. Math. Soc., Providence, RI, 672–681, 1987.

Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104
E-mail address: grassi@math.upenn.edu

CONFORMAL INVARIANCE AND $2-d$ STATISTICAL PHYSICS

GREGORY F. LAWLER

Abstract. A number of two-dimensional models in statistical physics are conjectured to have scaling limits at criticality that are in some sense conformally invariant. In the last ten years, the rigorous understanding of such limits has increased significantly. I give an introduction to the models and one of the major new mathematical structures, the Schramm-Loewner Evolution (SLE).

1. Critical Phenomena

Critical phenomena in statistical physics refers to the study of systems at or near the point at which a phase transition occurs. There are many models of such phenomena. We will discuss some discrete equilibrium models that are defined on a lattice. There are essentially measures on paths or configurations where configurations are weighted by their energy $\mathcal{E}$ with a preference for paths of smaller energy. These measures depend on at least one parameter. A standard parameter in physics $\beta = c/T$ where c is some fixed constant and T stands for temperature and the measure given to a configuration is $e^{-\beta\mathcal{E}}$. Phase transitions occur at critical values of the temperature corresponding to, e.g., the transition from a gaseous to liquid state. We use β for the parameter although for understanding the literature it is useful to know that large values of β correspond to "low temperature" and small values of β correspond to "high temperature". Small β (high temperature) systems have weaker correlations than large β (low temperature) systems. In a number of models, there is a critical β_c such that qualitatively the system has three regimes $\beta < \beta_c$ (high temperature), $\beta = \beta_c$ (critical) and $\beta > \beta_c$ (low temperature).

A standard procedure is to define a model on a finite subset of a lattice, and then let the lattice size grow. Equivalently, one can consider a bounded region and consider finer and finer lattices inside the region. In either case, one would like to know the scaling or continuum limit of the system. For many models, it can be difficult to even describe what kind of object one has in the limit, and then it can be much more difficult to prove such a limit exists.

The behavior of the models I will describe varies significantly in different dimensions. In two dimensions, Belavin, Polyakov, and Zamolodchikov predicted[1] [3, 4] that many systems at criticality (at β_c) had scaling limits that were in some sense conformally invariant. This assumption and nonrigorous techniques of conformal field theory allowed for exact calculation of a number of "critical exponents" and other quantities in the limit. For mathematicians, many of the arguments were

2000 *Mathematics Subject Classification.* Primary 82B27 ; Secondary 30C35, 60J65, 82B27.
Research supported by National Science Foundation grant DMS-0734151.

[1] I use the word "predicted" to mean that the result was mathematically nonrigorous, but had significant theoretical argument behind it. There is much nontrivial, deep mathematics in the conformal field theory arguments and other theories in mathematical physics. I use the word "predict" rather then "conjecture" to acknowledge this.

unsatisfactory, not only because there was incomplete proof, but also because in many cases there was no precise statement of what kind of limit was taken or what the limiting object was. However, the predictions made from these arguments were very consistent with numerical simulations, so it was clear that there was something essentially correct about the arguments.

There are a number of different approaches to studying two-dimensional critical phenomena. This paper will focus on the progress in understanding the geometric and fractal properties of the scaling limit. The big step was the introduction of the Schramm-Loewner evolution (or stochastic Loewner evolution as Schramm called it) which is a measure on continuous curves. The definition combines work in classical function theory by Loewner with the fundamental quantity in stochastic analysis, Brownian motion.

2. Lattice models

2.1. **Self-avoiding walk.** A self-avoiding walk (SAW) in the lattice $\mathbb{Z}^2$ is a nearest neighbor path that has no self-intersections. SAWs arose as a model of polymer chains (in a dilute solution). Roughly speaking, a single polymer chain consists of a number of monomers that take a random shape with the only constraint being that the polymer cannot cross itself. We put a measure on SAWs for which all walks of the same length have the same measure. More precisely we assign measure $e^{-\beta n}$ to each walk

$$\omega = [\omega_0, \ldots, \omega_n],$$

with $|\omega_j - \omega_{j-1}| = 1$ for each j and $\omega_j \neq \omega_k, j < k$. Trying to understand the SAW problem leads to the following (easily stated but still open) problems:

- How many SAWs are there of length n with $\omega_0 = 0$?
- If one chooses a SAW at random from the set of SAWs of length n, what is the typical end-to-end (Euclidean) distance of the chain?

It is conjectured that there exist c, β_c, γ such that the number of walks of length n, C_n satisfies

$$C_n \sim c\, e^{\beta_c n}\, n^{\gamma - 1}.$$

A simple subadditivity argument shows the existence of a β_c such that $\log C_n \sim \beta_c n$, but the more precise asymptotics are still open questions. The number γ is one of the "critical exponents" for the problem. Another critical exponent, usually denoted ν, states that the average end-to-end distance grows like n^ν.

Let us consider SAWs with boundary conditions. In Figure 1, we have an $N \times N$ square in $\mathbb{Z}^2$ and let z, w be boundary points on opposites sides. Consider the set of SAWs starting at z ending at w and otherwise staying in the box. We give each such walk ω measure $e^{-\beta|\omega|}$ where $|\cdot|$ denotes the length (number of edges) in the walk ω. The total mass of this measure

$$Z_N = Z_{N,\beta} = \sum_{\omega : z \to w} e^{-\beta|\omega|}$$

is often called the *partition function.* For any N, β, we get a probability measure on paths by dividing by Z_N. For large N, the behavior of this probability measure varies depending on β:

- If $\beta < \beta_c$, then Z_N grows exponentially in N. The penalty for having many bonds is not high, and a typical path tends to fill up the square.

- If $\beta > \beta_c$, then Z_N decays exponentially in N. The penalty for having many bonds is high, and a typical path goes from z to w without visiting many more sites than necessary.
- If $\beta = \beta_c$, then Z_N neither grows nor decays exponentially. It is expected that it decays like a power of N. The typical path is a typical SAW path of $N^{1/\nu}$ steps and roughly looks $(1/\nu)$-dimensional.

Let us focus on the critical value $\beta = \beta_c$. If we scale space and time, then we might hope to get a probability measure on continuous paths connecting two boundary points of a square. This measure would be supported on curves whose fractal dimension is $1/\nu$.

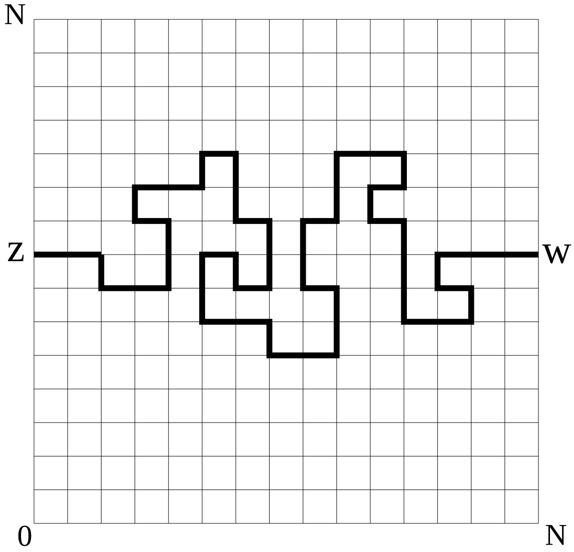


FIGURE 1. Self-avoiding walk in a domain

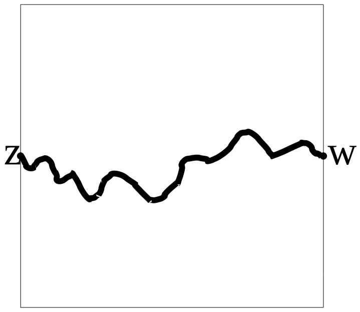


FIGURE 2. Scaling limit of SAW

This existence of this limit for SAW has not been proved. However, let us suppose that such a limit exists. Suppose we took a different domain with two boundary points and considered a similar limit measure at $\beta = \beta_c$. This gives a measure on curves in the new domain. If the new domain is simply connected, we can also get a probability measure on curves by starting with the measure on the square and mapping these curves to the new domain. (There are two issues we are not dealing with. One is the local lattice effect at the boundary if the boundary does not match up nicely with the lattice as in the case of the square above. The other is the parametrization of the curves. We will consider two curves the same if one is an increasing reparametrization of the other). Conformal invariance would imply that one can obtain one measure from the other by means of a conformal transformation.

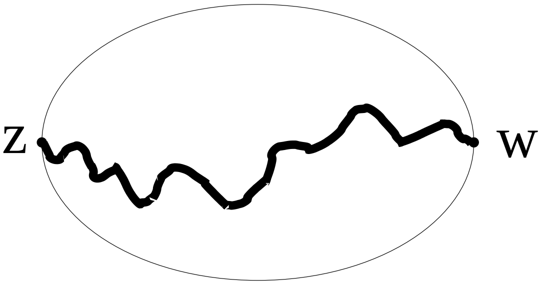


FIGURE 3. Scaling limit of SAW in a different domain

2.2. **Simple random walk.** There is a variation of the last model that is much better understood by probabilists. Suppose we do not put on a self-avoidance constraint and consider all nearest-neighbor random walks. Then the number of walks of length n is 4^n and hence $e^{\beta_c} = 4$. The typical end-to-end distance of a simple random walk is $O(n^{1/2})$, i.e., $\nu = 1/2$. For simple random walk, one can show that the partition function for the square in Figure 4 with $\beta = \beta_c$ is comparable to N^{-2}. When we scale the paths as above, we get a well-known limit, known as *Brownian motion.* (More precisely, we get a Brownian "excursion" from z to w — this is a modification of Brownian motion for which the path starting at z goes immediately into the domain and then exits the domain at w.)

It is known that the scaling limit of Brownian excursion is conformally invariant. The result for Brownian motion goes back to Paul Lévy [30]. In fact, it is implicit in earlier work — the basic fact is that harmonic functions in two dimensions are invariant under conformal transformations.

2.3. **Loop-erased random walk.** The loop-erased random walk (LERW) is the measure obtained from simple random walk by erasing loops. In other words, we start with the simple random walk measure and for each walk ω we obtain a walk without self-intersections by erasing the loops. (The walk that one obtains from loop-erasure depends on the order in which the points are erased — in this case, we specify that the loops are erased chronologically.) The LERW arises in a number of situations; e.g., as the distribution of a typical geodesic in a spanning tree of a graph where the tree is chosen from the uniform distribution of all spanning trees.

Since the partition function (the sum of the weights of all the paths) is the same as for simple random walk, it has the same β_c and partition function. However,

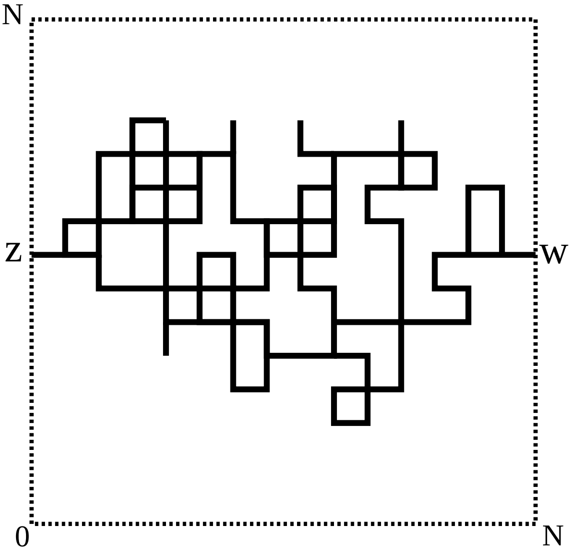


FIGURE 4. Simple random walk in D

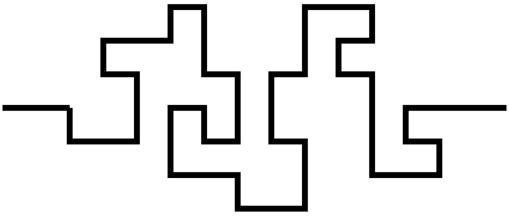


FIGURE 5. The walk obtained from erasing loops chronologically from the simple walk above

the number of bonds in a typical LERW is smaller than that for the simple random walk. It was predicted by physicists (and a version proved by Rick Kenyon [17]) that the typical LERW has on the order of $N^{5/4}$ bonds. If we take a scaling limit, than we would expect that the paths in the scaling limit would have fractal dimension 5/4.

We also would conjecture that the LERW gives a conformally invariant limit. For example, one could use the fact that simple random walk has a conformally invariant limit and note that the loop-erasing procedure which only looks at the order of the points should also be conformally invariant. There are other relations between LERW and simple random walk that would lead us to conjecture this.

It is not so easy to give a formula for the weight of any particular path ω under this measure. For each self-avoiding path, the LERW measure of the path is the measure of all the simple random walks whose loop-erasure is ω. It turns out that one can write this weight as

$$4^{-|\omega|}\, e^{\Lambda(\omega)}$$

where $\Lambda(\omega)$ is a measure of the number of loops in the domain that intersect ω. The scaling limit of this loop measure arises in studying the scaling limit of a number of models.

2.4. **Percolation.** Percolation can be considered a model of permeability of a material. We will describe a lattice model for percolation on the triangular lattice. Suppose that every point in the triangular lattice in the upper half plane is colored black or white independently with white having probability p. A typical realization with $p = 1/2$ is illustrated in Figure 6 (if one ignores the bottom row). We think of a white site as being "open" through which liquid can flow. The general question is whether or not there is an infinite collection of open sites that are connected. The value $p = 1/2$ is "critical" for the triangular lattice in that for $p > 1/2$, there will be an infinite connected cluster of white sites while for $p < 1/2$, this will not be true. We will consider critical percolation, $p = 1/2$.

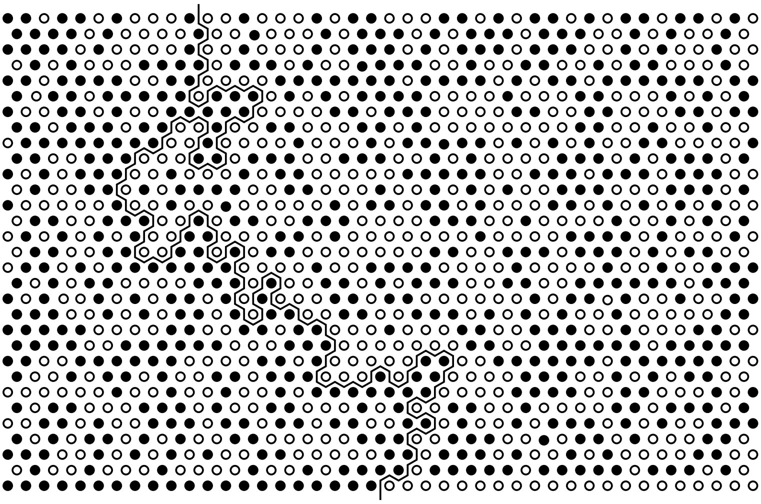


Figure 6. The percolation exploration process.

We now put a boundary condition on the bottom row as illustrated — all black on one side of the origin and all white on the other side. Once all the other colors in the upper half plane have been chosen, there is a unique curve starting at the bottom row that has all white vertices on one side and all black vertices on the other side. This is called the *percolation exploration process*. Similarly we could start with a domain D and two boundary points z, w; give a boundary condition of black on one of the arcs and white on the other arc; put a fine triangular lattice inside D; color vertices black or white independently with probability $1/2$ for each; and then consider the path connecting z and w. In the limit, one might hope for a continuous interface.

There is another conformal invariant for percolation first predicted by John Cardy [7, 8]. Suppose D is a simply connected domain and the boundary is divided into four arcs, A_1, A_2, A_3, A_4 in counterclockwise order. Let $P_D(A_1, A_3)$ be the limit as the lattice spacing goes to zero of the probability that in a percolation cluster as above there is a connected cluster of white vertices connecting A_1 to A_3.

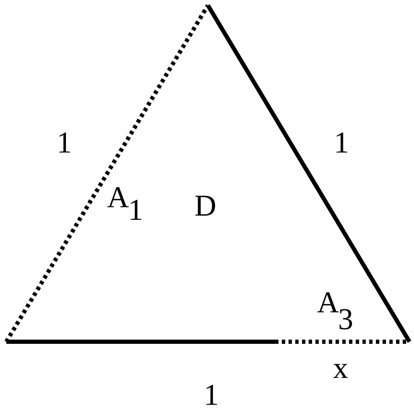


FIGURE 7. Cardy's formula: $P_D(A_1, A_3) = x$.

This should be a conformal invariant. It turns out that the nicest domain to give the formula is an equilateral triangle, see Figure 7.

2.5. **Ising model.** The Ising model is a simple model of a ferromagnet. Consider the triangular lattice as in Figure 6. Again we color the vertices black or white although we now think of the colors as spins —black is a spin up and white is a spin down. If x is a vertex, we let $\sigma(x) = 1$ if x is colored black and $\sigma(x) = -1$ if x is colored white. The measure on configurations is such that neighboring spins like to be aligned with each other.

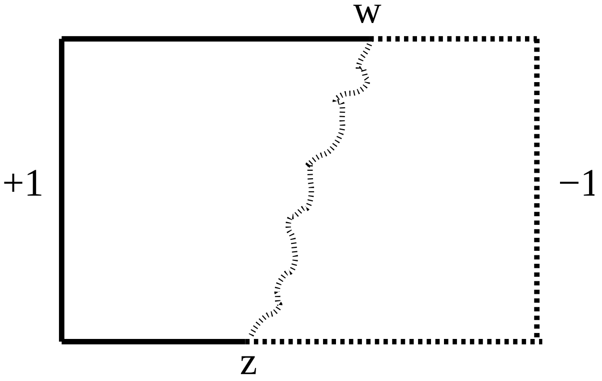


FIGURE 8. Ising interface.

It is easiest to define the measure for a finite collection of spins. Suppose D is a bounded domain in $\mathbb{C}$ with two marked boundary points z, w which give us two boundary arcs. We consider a fine lattice in D and fix boundary conditions $+1$ and -1 respectively on the two boundary arcs. Each configuration of spins is given energy

$$\mathcal{E} = -\sum_{x \sim y} \sigma(x)\,\sigma(y),$$

where $x \sim y$ means that x, y are nearest neighbors. We then give measure $e^{-\beta\mathcal{E}}$ to a configuration of spins. If β is small, then the correlations are localized and spins separated by a large distance are almost independent. If β is large, there is long-range correlation. There is a critical β_c that separates these two phases. At this critical value the interface is a random fractal.

3. Assumptions on the limit

In trying to find possible scaling limits for models as above, Schramm started by giving assumptions that a scaling limit of the curves arising in critical systems should satisfy. Suppose there exists a family of probability measures $\mu(D; z, w)$, indexed by a collection of domains D and distinct boundary points $z, w \in \partial D$, supported on curves connecting z and w in D. We assume that the family satisfies conformal invariance.

- **Conformal Invariance.** If $f : D \to f(D)$ is a conformal transformation, then
$$f \circ \mu(D; z, w) = \mu(f(D); f(z), f(w)).$$

To be more precise, if γ is a curve in D from z to w, then $f \circ \gamma(t) = f(\gamma(t))$ gives a curve in $f(D)$ from $f(z)$ to $f(w)$. We adopt the convention that two curves are equivalent if one is an (increasing) reparametrization of the other. For this reason, we do not worry about the parametrization of $f \circ \gamma$. If we have a measure on curves γ in D, then f induces a measure on curves $\tilde{\gamma}$ in $f(D)$ by the map $\gamma \mapsto f \circ \gamma$.

We now consider all the examples in the previous section, except the simple random walk (for which we know the scaling limit, Brownian motion). For all the other systems, the discrete models satisfy a property called the *domain Markov property* which we would expect to be satisfied by the scaling limit. We will state this property under the assumption that the curves γ are simple (non-self-intersecting). Suppose we are interested in the distribution of the curve γ and we observe an initial part of the curve $\gamma(0, t]$. Let D_t denote the slit domain $D \setminus \gamma(0, t]$. (More generally, if γ can have self-intersections we let D_t denote the connected component of $D \setminus \gamma(0, t]$ that contains w on the boundary. We require our limit to satisfy the "non-crossing" condition that $\gamma(t) \in \partial D_t$. This is not satisfied for the Brownian excursion.)

- **Domain Markov Property.** Given $\gamma[0, t]$ the distribution of the remainder of the path is given by $\mu(D_t; \gamma(t), w)$.

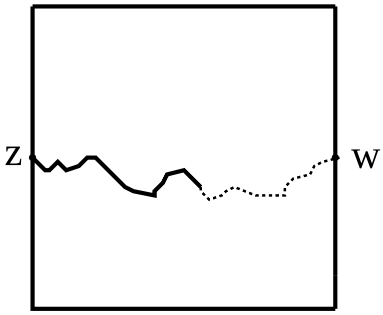


Figure 9. Domain Markov property

Schramm showed that if we restrict to simply connected domains D, then there is only a one-parameter family of probability measures that satisfy both conformal invariance and the domain Markov property. We will explain why in the next two sections. The Riemann mapping theorem tells us that all simply connected domains in $\mathbb{C}$ with nontrivial boundary are conformally equivalent. In particular, if D is a simply connected domain with distinct boundary points z, w, then there exists a conformal transformation

$$F : \mathbb{H} = \{x + iy : y > 0\} \to D, \quad F(0) = z, \quad F(\infty) = w.$$

The map is not unique, but if $\tilde{F}$ is another such transformation, then $\tilde{F}(z) = F(rz)$ for some $r > 0$. If we can define the measure $\mu(\mathbb{H}; 0, \infty)$, then the measure $\mu(D; z, w)$ for all simply connected D and distinct $z, w \in \partial D$ will be determined.

4. Loewner differential equation

One of the biggest problems of classical function theory in the twentieth century was the Bieberbach conjecture. The Riemann mapping theorem showed that the study of simply connected domains reduces to the study of univalent, i.e., one-to-one and analytic, functions on the unit disk $\mathbb{D}$. By translation, dilation, and rotation, it suffices to consider the set $\mathcal{S}$ of such functions with $f(0) = 0, f'(0) = 1$. Each such function can be written as

$$f(z) = z + \sum_{n=2}^{\infty} a_n\, z^n.$$

Ludwig Bieberbach proved that $|a_2| \leq 2$ and conjectured that $|a_n| \leq n$ for all n. (There exists a particular f that obtains the maximal values.) Compactness arguments show that it suffices to prove the inequality for slit domains of the form $\mathbb{C}\setminus\gamma[t, \infty)$ where γ is a simple curve with $\gamma(z) \to \infty$, $\gamma(-\infty) = 0$. Charles Loewner[2] [31] derived a differential equation which described the dynamics of the coefficients $a_n(t)$ in t. He used this to prove $|a_3| \leq 3$, and the general technique became an important tool in the study of conformal maps. In fact, the proof of the Bieberbach conjecture by Louis de Branges [5] uses such chains.

There are a number of versions of the Loewner differential equation. Schramm found it most convenient to consider a version in the upper half plane $\mathbb{H}$. Suppose $\gamma : (0, \infty) \longrightarrow \mathbb{H}$ is a simple curve with $\gamma(0+) = 0$. For each t, let H_t denote the slit domain $\mathbb{H} \setminus \gamma(0, t]$. Using the Riemann mapping theorem, one can show that there is a unique conformal transformation g_t of H_t onto $\mathbb{H}$ satisfying $g_t(z) = z + o(1)$ as $z \to \infty$. This can be expanded at infinity as

$$g(z) = z + \frac{a(t)}{z} + O(|z|^{-2}),$$

where $a(t) \geq 0$ depends on the curve $\gamma(0, t]$. The quantity $a(t)$ is a "half-plane capacity". It can be shown that $t \mapsto a(t)$ is a continuous, strictly increasing function. By making a slightly stronger assumption on the curve γ, we can also assume that $a(t) \to \infty$. Since a is strictly increasing we can reparametrize the curve γ so that

[2] He spelled his name Karel (in Czech) or Karl (in German) Lówner in Europe but adopted the spelling Charles Loewner when he moved to the United States.

$a(t)$ grows linearly, say $a(t) = 2t$. If we do this, then the (chordal or half-plane) Loewner differential equation states that the function $t \mapsto g_t(z)$ satisfies

$$\partial_t g_t(z) = \frac{2}{g_t(z) - U_t}, \quad g_0(z) = z, \tag{4.1}$$

where $U_t = g_t(\gamma(t))$. (Although $\gamma(t) \in \partial H_t$, one can show that there is a unique continuous extension of g_t to the boundary at $\gamma(t)$.) Moreover, the function $t \mapsto U_t$ is a continuous function from $[0,\infty)$ to $\mathbb{R}$. For $z \in \mathbb{H}$, the solution exists up to time $T_z \in (0,\infty]$. In fact, for a simple curve γ, $T_z = \infty$ if $z \notin \gamma(0,\infty)$ and $T_{\gamma(t)} = t$. To understand this equation, let us consider $t = 0$,

$$\partial_t g_t(z)\Big|_{t=0} = \frac{2}{z}.$$

The function $z \mapsto 1/z$ is (a multiple of) the complex form of the Poisson kernel in the upper half plane. The Loewner equation states that if one parametrizes the curve by capacity, then the change in the conformal map g_t is determined by the Poisson kernel.

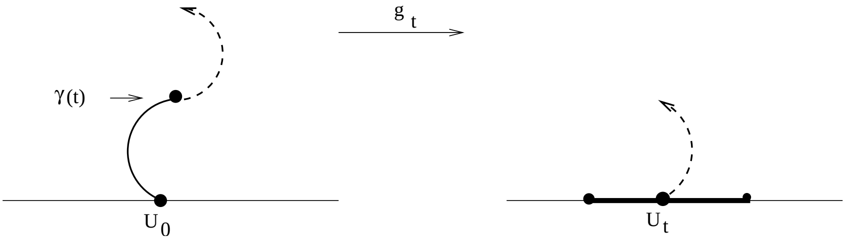


FIGURE 10. the conformal map g_t

If we start with a continuous function $t \mapsto U_t$ on the real line, then for each $z \in \mathbb{H}$, we can find a solution to the equation (4.1) that exists up to a time $T_z \in (0,\infty]$. In fact, for fixed t, the function g_t is the unique conformal transformation of $H_t := \{z : T_z > t\}$ onto $\mathbb{H}$ that satisfies $g_t(z) - z = o(1)$ as $z \to \infty$. However, it is not always the case that the domain H_t will be a simple slit domain as above.

5. Schramm-Loewner evolution

Suppose we have a random simple curve $\gamma(t)$ in $\mathbb{H}$ that arises from a family of curves satisfying conformal invariance and the domain Markov property. The random curves $\gamma(t)$ generate random one-dimensional continuous functions U_t by (4.1). Conformal invariance and the domain Markov property translate into assumptions on U_t:

- For each $s < t$, the random variable $U_t - U_s$ is independent of $U_{s'}, 0 \leq s' \leq s$ and has the same distribution as U_{t-s}.

Since the measure $\mu(\mathbb{H}; 0, \infty)$ should be invariant under dilations $z \mapsto rz$, the function $r\gamma(t)$ should have the same distribution (modulo reparametrization) as $\gamma(t)$. Using properties of capacity, we can see that if we set $\tilde{\gamma}(t) = r\gamma(t/r^2)$, then $\tilde{\gamma}(t)$ is parametrized by capacity. Using this and doing the appropriate change of variables, we also get

- The distribution of rU_{t/r^2} should be the same as the distribution of U_t.

It is well known to probabilists that there is only a one-parameter family of processes U_t that satisfy these conditions,

$$U_t = \sqrt{\kappa}\, W_t,$$

where W_t is a standard Brownian motion. One can choose any $\kappa > 0$. This leads to Schramm's definition.

Definition 5.1. *(Chordal) SLE_κ (from 0 to ∞ in $\mathbb{H}$)* is the random collection of conformal maps g_t obtained by solving (4.1) with $U_t = \sqrt{\kappa}\, W_t$.

Chordal SLE_κ in other simply connected domains is defined by conformal transformation. The invariance of SLE_κ in $\mathbb{H}$ under dilations shows that this definition is independent of the choice of map. It is not obvious, but was proved [33] that there is a random function $\gamma : [0,\infty) \to \overline{\mathbb{H}}$ such that for each t, g_t is the conformal transformation of the unbounded component of $\mathbb{H} \setminus \gamma(0,t]$ onto $\mathbb{H}$. This curve is called the SLE_κ curve or SLE_κ trace. If D is a simply connected domain and z, w are distinct boundary points, then SLE_κ in D from z to w is obtained by conformal transformation.

It turns out that the qualitative behavior of the curve varies greatly as the parameter κ varies.

- If $0 < \kappa \leq 4$, the curve γ is simple and $\gamma(0,\infty) \subset \mathbb{H}$.
- If $4 < \kappa < 8$, the curve γ has self-intersections but is "non-crossing" and does not fill up the plane. Non-crossing implies that for all t, $\gamma(t)$ is on the boundary of the unbounded connected component of $\mathbb{H} \setminus \gamma(0,t]$. γ also intersects $\mathbb{R} \setminus \{0\}$.
- If $\kappa \geq 8$, the curve is non-crossing and plane-filling! For each $z \in \overline{\mathbb{H}}$ there is a t such that $\gamma(t) = z$.

Moreover [2, 33], the Hausdorff dimension of the paths is given by a simple formula

$$d_\kappa := \dim(\gamma[0,t]) = 1 + \frac{\kappa}{8}, \qquad t > 0, \quad \kappa \leq 8. \tag{5.1}$$

For $\kappa \geq 8$, the dimension is two since the curve is plane-filling. The relation $\kappa \leftrightarrow d_\kappa$ is a bijection between $[0,8]$ and $[1,2]$. ($\kappa = 0$ corresponds to a straight line, i.e., the curve $\gamma(t) = cti$. It is the solution of the Loewner equation with $U_t \equiv 0$.)

The basic tool for studying SLE is stochastic calculus. Fix $z \in \overline{\mathbb{H}} \setminus \{0\}$ and let $\tilde{Z}_t = \tilde{Z}_t(z) = g_t(z) - U_t$. Then the Loewner equation (4.1) becomes the stochastic differential equation (SDE)

$$d\tilde{Z}_t = \frac{2}{\tilde{Z}_t}\, dt + \sqrt{\kappa}\, dB_t,$$

where $B_t = -W_t$ is a standard Brownian motion. If we do the time change $Z_t = \tilde{Z}_{t/\kappa}$, this equation becomes

$$dZ_t = \frac{a}{Z_t}\, dt + dW_t, \tag{5.2}$$

where W_t is a standard Brownian motion and $a = 2/\kappa$. This SDE is called the Bessel equation. It is well known that for real Z_t, if $a \geq 1/2$ ($\kappa \leq 4$), the solutions to this equation never reach the origin, i.e., the push away from the origin of a/Z_t is big enough to withstand the randomness in the Brownian motion W_t. On the other hand, if $a < 1/2$ ($\kappa > 4$), the solutions eventually hit the origin. Geometrically, this corresponds to the fact that for $\kappa \leq 4$, the point 1 stays on the boundary of H_t which is the same as saying that the curve $\gamma(0, \infty)$ never hits $(1, \infty)$.

We give one example of how standard tools from stochastic calculus are used to compute a probability for SLE. Let $E = E_z$ be the event that z lies on the "left side" of the SLE path $\gamma(0, \infty)$. By scaling $\mathbf{P}(E_z)$ depends only on $\arg(z)$ and hence we can write $\mathbf{P}(E) = f(\arg z)$ for some function f with $f(0) = 0, f(\pi) = 1$. If

$$J_t = \arg(Z_t),$$

then $f(J_t)$ denotes the conditional probability of the event E given $\gamma(0, t]$. In probabilistic notation

$$f(J_t) = \mathbf{P}[E_z \mid \mathcal{F}_t],$$

where $\mathcal{F}_t$ denotes the "information" contained in $\gamma(0, t]$. The process $M_t = f(J_t)$ is an example of a *martingale*, a process which satisfies $\mathbf{E}[M_t \mid \mathcal{F}_s] = M_s$ for $s < t$. Itø's formula (the fundamental theorem of stochastic caculus) and (5.2) give

$$dJ_t = (1 - 2a) \cos J_t \, \sin J_t \, dt - \sin J_t \, dB_t.$$

and another application of Itô's formula gives

$$df(J_t) = \left[(1 - 2a) \cos J_t \, \sin J_t \, f'(J_t) + \frac{1}{2} \sin^2 J_t f''(J_t)\right] dt + [\cdots] \, dB_t.$$

If we have a process such as $f(J_t)$ written in terms of a stochastic differential equation as above and we know that $f(J_t)$ is a martingale then it must follow that the dt term is zero, i.e., that $f(\theta)$ satisfies

$$2(1 - 2a) \, f'(\theta) \, \cot \theta + f''(\theta) = 0,$$

which yields

$$f(\theta) = \int_0^{\pi} \frac{c \, d\theta}{\sin^{2-4a} \theta},$$

where c is chosen so that $f(\pi) = 1$. Note that the integral is finite only if $2 - 4a < 1$, i.e., if $\kappa < 8$. In the plane-filling regime $\kappa \geq 8$, the point z is hit and hence is not on the right or left side.

6. Central charge and other parameters

Families of curves on simply connected domains that satisfy conformal invariance and the domain Markov property must be SLE curves. However, there is a one-parameter family of such curves. For a particular model, one must determine what is the value of κ that corresponds to that model. Sometimes, this can be determined by computing some quantity for that model or demonstrating some property that would need to be satisfied in the limit. For example, if we knew the Hausdorff dimension of the curves in the limit then if this dimension is less than two, we would know κ by (5.1). In practice, the dimension is generally one of the properties that we are trying to determine about the limiting curves, so we could not use this to find κ.

Another quantity is the *(boundary) scaling exponent* which is sometimes called the *conformal weight* or a *scaling dimension.* This exponent can be defined directly in terms of SLE but it corresponds to an exponent we have seen before. Recall the self-avoiding walk and the loop-erased random walk. The partition function at $\beta = \beta_c$ is conjectured to satisfy a power law, $Z_N \sim N^{-2b}$ for an exponent b. For

SLE_κ, this exponent corresponds to the the boundary scaling exponent that can be computed

$$b = \frac{6-\kappa}{2\kappa}.$$

As mentioned before, properties of simple random walk determine the behavior of its (and hence also the LERW) partition function, which lead us to identify $b = 1$ or $\kappa = 2$ for loop-erased walk. In the case of critical percolation, the partition function is actually equal to 1 because the weights of configurations are determined by a probability measure. This would lead one to guess $b = 0$ or $\kappa = 6$. For other examples, such as SAW, the scaling exponent is something one would like to compute for the model.

In conformal field theory there is a constant $\mathbf{c}$, called the *central charge*, which is used to describe the theory. (The term central comes from its relationship to central extensions of Lie algebra which we will not discuss here.) Bertrand Duplantier was the first to conjecture the following relationship between $\mathbf{c}$ and κ,

$$\mathbf{c} = \frac{(3\kappa-8)\,(6-\kappa)}{2\kappa}, \quad \kappa = \frac{(13-\mathbf{c}) \pm \sqrt{(13-\mathbf{c})^2 - 144}}{3}.$$

Note that the relationship $\kappa \leftrightarrow \mathbf{c}$ is two-to-one except at the double root $\mathbf{c} = 1, \kappa = 4$. Also, for $\kappa > 0$, $c \leq 1$. We define $\kappa' = 16/\kappa$ to be the "dual value" of the SLE parameter; κ and κ' have the same central charge $\mathbf{c}$. If $\kappa < 4$, then the SLE paths are simple, but the curves for the dual value $\kappa' > 4$ are not simple. There is a duality conjecture that has been proven only for certain values of κ, that states that if $\kappa \leq 4$ and γ is an $SLE_{\kappa'}$ curve, then for each t the "outer boundary" of the curve (which can be defined as $\mathbb{H} \cap \partial H_t$ where H_t is as above) looks like an SLE_κ curve.

One can give an interpretation of the central charge in terms of how the measure on configurations changes when a domain is perturbed. Suppose $\kappa \leq 4$, D is a domain with smooth boundary and we try to define the limiting measures as in the first section. However, instead of dividing by Z_N we will multiply by N^{2b}. If $Z_N \sim cN^{-2b}$ as we expect (where the constant c depends on the domain and boundary points, and we ignore the serious lattice effects at the boundary), then the limiting measure will be a nonzero finite measure on paths that is not necessarily a probability measure. Let us call this measure $m(D; z, w)$. Now suppose $D_1 \subset D$ with the perturbation being away from z, w. Then the measure $m(D_1; z, w)$ is absolutely continuous with respect to $m(D; z, w)$. In fact, we can give the Radon-Nikodym derivative

$$\frac{dm(D_1; z, w)}{dm(D; z, w)}\,(\gamma) = e^{(\mathbf{c}/2)\,\Lambda(D; D_1, \gamma)}\, 1\{\gamma \subset D_1\},$$

where $\Lambda(D; D_1, \gamma)$ is a conformal invariant (independent of κ) that roughly gives the measure of the set of Brownian loops in D that intersect both γ and D_1. This measure was introduced in [29].

Examples

- The discrete measure on self-avoiding walks gives all walks of the same length the same measure. Hence in a scaling limit, one would expect that boundary perturbation would not affect the measure, i.e., that the limit of SAWs should have central charge $\mathbf{c} = 0$. This property is called the

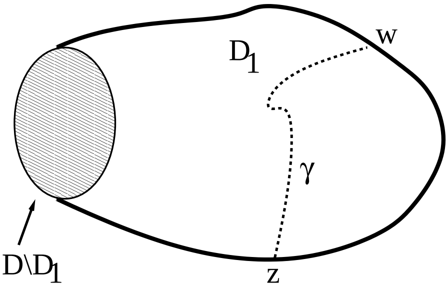


Figure 11. Boundary perturbation

restriction property. Assuming that the limit lies on simple curves, we see that the only possible value is $\kappa = 8/3$.

- In the discrete measure on loop-erased random walk, perturbing the domain *does* affect the measure of a path. The measure of a particular walk in the LERW is the measure of all the simple random walks whose loop-erasure gives that walk. If we shrink the domain we lose some of the simple walks that produce our walk and hence the measure decreases. Hence we would expect that $\mathbf{c} < 0$. We have already given a reason why $\kappa = 2$ should correspond to LERW which would give $\mathbf{c} = -2$. On the discrete level one can give a precise description of how the measure changes in terms of random walk loops that hit both the path and the part of the domain being removed. The Brownian loop measure is a continuous limit of the corresponding measure on random walks.

7. Particular Models

7.1. **Schramm's original paper.** In [34], Oded Schramm considered two of the models we discussed, LERW and percolation. Assuming conformal invariance and the domain Markov property, he deduced that *if the processes had a conformally invariant limit*, then they must be SLE_κ for some κ. He also did some computations for SLE to determine what the values must be. For LERW he determined $\kappa = 2$ by some winding number calculations and for percolation he determined $\kappa = 6$ by proving an analogue of Cardy's formula for SLE and showing which one matched up appropriately.

7.2. **Brownian paths.** There are a number of problems about the geometric and fractal properties of planar Brownian motion which have been solved using SLE. Benoit Mandelbrot [32] looked at "Brownian islands" which are formed by taking planar Brownian motions, conditioning them to begin and end at the same point, and then filling in the set surrounded by the path. The "coastline" or "outer boundary" of this set is an interesting fractal that appeared to him (and was tested by numerical simulation) to be of fractal dimension 4/3. Since this was also the conjectured dimension for the scaling limit of self-avoiding walk, he proposed that the outer boundary of Brownian motion might give a model of the limit of SAWs. The dimension of this coastline can be shown to be given in terms of a particular value of the *intersection exponents* for Brownian motion. Wendelin Werner and I [27, 28] were trying to compute the intersection exponents and had discovered at least heuristically that these exponents were related to those for SAW and critical percolation. This work was going on at the same time that Schramm was introducing his new process as a model for scaling limits of LERW and percolation. The three us joined forces to see if SLE could be applied to the intersection exponent problem. This program turned out to be successful [21, 22, 23] and we proved Mandelbrot's conjecture. In the process, we discovered an important property *locality* that distinguishes $\kappa = 6$ from all values. Although SLE_6 and Brownian motion are very different processes, their outer boundaries are the same and are versions of $SLE_{8/3}$. This is the duality relation for $\mathbf{c} = 0$.

7.3. **Percolation.** When Cardy [7, 8] gave his original prediction for the crossing probabilities for critical percolation, he gave it in the upper half plane $\mathbb{H}$ for which it is a hypergeometric function. Lennart Carleson made the observation that the formula was nicer if one maps on the equilateral triangle and he also suggested that it might be easier to prove limit theorems for the discrete model on the triangular lattice rather than the square lattice. Indeed, Schramm's percolation exploration process is much easier to define on the triangular lattice. Stas Smirnov [38] proved that the scaling limit of percolation satisfies Cardy's formula and in the process established that the percolation exploration process approaches SLE_6. The scaling limit of percolation is much richer than just one boundary curve; however, see [6], one can use an infinite collection of boundary curves to construct the scaling limit. Smirnov's proof unfortunately holds only for the triangular lattice and it is still open to prove Cardy's formula for other lattices such as the square lattice. The percolation exploration process satisfies a locality property inasmuch as to determine an initial segment of the path one needs only examine the colors of the sites adjacent to the path. In particular, if the domain were perturbed away from the path, this would not affect the distribution. The SLE_6 locality property is a continuous analogue of this. Although the discrete percolation exploration process has no self-intersections, its scaling limit, SLE_6, does not have simple paths.

7.4. **Loop-erased random walk.** The loop-erased random walk is related to a number of models, in particular the problem of choosing a spanning tree at random from the collection of all spanning trees on a graph (or, similarly, counting the number of spanning trees). See, e.g., [41]. Roughly speaking, the path connecting two points on a uniform spanning tree has the distribution of a LERW. The spanning tree problem is also related to dimer configurations. Kenyon [17] used this relation to show that the average number of steps in a LERW on an $N \times N$ box is $N^{5/4}$. The process is also a determinantal process and this combined with conformal invariance of Brownian motion can establish conformal invariance for some quantities without discussing SLE [13, 18]. In [24], LSW established that the scaling limit of loop-erased random walk is SLE_2 and also the scaling limit of uniform spanning trees (appropriately defined) gives SLE_8. This is another example of the duality relation, this time for $\mathbf{c} = -2$. While there are technical difficulties in the limit, one reason why the problem is tractable is that the LERW and the uniform spanning tree are constructed from simple random walk and it is well known that the limit of simple random walk is Brownian motion. The determination of κ comes from the calculation.

7.5. **Gaussian free field.** The Gaussian free field in two dimensions is another fundamental conformally invariant object. We will not define it here but rather recommend [37] for an introduction. Schramm and Sheffield [35, 36] have shown that scaling limits for interfaces for the Gaussian free field (as well as for a somewhat simpler model, the harmonic explorer) are SLE_4 curves. This agrees with what was already known that the free field was a $\mathbf{c} = 1$ model. The proofs of the scaling limit are quite involved. However, as in the case of LERW, the discrete model can be described in terms of random walks and one uses the convergence of random walk to Brownian motion to establish the result.

7.6. **Ising model and related models.** The interfaces for the Ising model are predicted to converge to SLE_3 curves. There is current exciting work by Smirnov establishing this limit as well as giving a general procedure for determining limits for Potts models and random cluster models (for which the Ising model is an example). An introduction to this work can be found in his expository paper [39].

7.7. **Self-avoiding walk.** If a scaling limit for self-avoiding walk exists, then it must satisfy the restriction property. This leads us to predict that the scaling limit is $SLE_{8/3}$, and numerical simulations [15, 16] give strong evidence that the conjecture is correct. However, there is no proof that the scaling limit exists.

8. Comments

I will end by making a number of comments about SLE and current and future research trends.

- The relation between conformal invariance and two-dimensional critical phenomena has been studied extensively in the physics community, see, e.g. [9, 14, 11]. For many mathematicians (this author included), this research has been hard to read not just because mathematical details were missing but because the basic notions were not defined precisely enough (for us mortal mathematicians). Many of these ideas can be made precise with SLE — in fact, there have been a number of papers (see, e.g., [1, 10]) that have used SLE to bridge the communication gap between the mathematics and physics communities. There is still much work to be done but one could hope that much of the work of the physics community will be able to be made rigorous using SLE. I should point out that much of the work in the nonrigorous treatments of conformal field theory and relation to critical phenomena is either rigorous or relatively easily "rigorizable".
- Understanding SLE in non-simply connected domains and more general Riemann surfaces is an ongoing project. The beauty of Schramm's construction is that conformal invariance and the domain Markov property are sufficient to characterize SLE in all simply connected domains. Non-simply connected domains D have the property that if one slits the domain (say by taking $D \setminus \gamma(0,t]$), then the slit domain is not conformally equivalent to the original domain. To understand such processes one probably needs to consider measures on paths that are not probability measures. In most cases one expects these measures to be (at least locally) absolutely continuous with respect to SLE in simply connected domains.
- SLE by definition is a random process that moves in a domain with moving boundary. However, the processes which are modeled by SLE are not naturally built in time by the formation of an SLE path. Instead, the configurations are weighted by an equilibrium measure and the "dynamics"

of an SLE are really just a description of how a conditional probability or conditional expectation changes as we observe more of the path. The fact that SLE gives a dynamic rather than equilibrium view of systems has made some things harder to prove than they really should be. Dapeng Zhan [42] recently gave a nice proof of the reversibility of SLE paths for $\kappa \leq 4$, i.e., if γ is a chordal SLE_κ path from z to w in D, then the time reversal of this path has the same distribution as an SLE_κ from w to z in D.

- This paper has focused on only one major aspect of recent work in conformal invariance and two-dimensional statistical mechanics: the geometric limiting curves given by SLE and random walk loops. There is also a large amount of exciting work in related areas: combinatorics, random matrices, integrable systems, complex and algebraic geometry. To even touch on these areas briefly would make this paper too long. The theory of two-dimensional critical phenomena is a very rich area that combines many different areas of mathematics.
- One of the major limitations of the theory of conformally invariant systems and critical phenomena is that it works only in two dimensions. There are many open, and probably significantly harder, problems understanding processes in three dimensions. In fact, numerical simulations suggest that the answers may not be nice. For example, the critical exponents in two dimensions turn out to be rational numbers but there is no reason to believe that the corresponding exponents (e.g., γ and ν for self-avoiding walk) are rational numbers.
- For a more extensive survey of SLE and its relation to discrete models see [40]. For a mathematical treatment of the Loewner equation and SLE with less discussion of the relavant discrete models , see [19].

References

1. M. Bauer, D. Bernard (2003), Conformal field theories of stochastic Loewner evolutions. Comm. Math. Phys. **239**, 493–521.
2. V. Beffara (2007), The dimension of the SLE curves, to appear in Ann. Probab.
3. A. Belavin, A. Polyakov, A. Zamolodchikov (1984), Infinite conformal symmetry of critical fluctuations in two dimensions, J. Stat. Phys. **34**, 763–774.
4. A. Belavin, A. Polyakov, A. Zamolodchikov (1984), Infinite conformal symmetry in two-dimensional quantum field theory. Nuclear Phys. B **241**, 333–380.
5. L. de Branges(1985), A proof of the Bieberbach conjecture, Acta Math. **154**, 137–152.
6. F. Camia, C. Newman (2006). Two-dimensional critical percolation: the full scaling limit. Comm. Math. Phys. **268**, 1–38.
7. J. Cardy (1984), Conformal invariance and surface critical behavior, Nucl. Phys. B **240** (FS12), 514–532.
8. J. Cardy (1992), Critical percolation in finite geometries, J. Phys. A **25**, L201–L206.
9. J. Cardy (1996), *Scaling and Renormalization in Statistical Physics*, Cambridge.
10. J. Cardy (2005), SLE for theoretical physicists, Annals of Phys. **318**,81-118
11. P. Di Francesco, P, Mathieu, D, Sénéchal (1997), *Conformal Field Theory*, Springer.
12. B. Doyon, V. Riva, J. Cardy (2006), Identification of the stress-energy tensor through conformal restriction in SLE and related processes. Comm. Math. Phys. **268**, 687–716.
13. S. Fomin (2001), Loop-erased walks and total positivity, Trans. Amer. math. Scot. **353**, 3563–3583.
14. M. Henkel (1999), *Conformal Invariance and Critical Phenomena*, Springer.
15. T. Kennedy (2003), Monte Carlo tests of SLE predictions for $2D$ self-avoiding walks, Phys. Rev. Lett. **88**, 130601.

16. T. Kennedy (2004), Conformal invariance and stochastic Loewner evolution predictions for the 2D self-avoiding walk–Monte Carlo tests, J. Statist. Phys. **114**, 51–78.
17. R. Kenyon (2000), The asymptotic determinant of the discrete Laplacian, Acta Math. **185**, 239–286.
18. M. Kozdron, G. Lawler (2005), Estimates of random walk exit probabilities and application to loop-erased random walk, Electron. J. probab., **10**, 1442–1467.
19. G. Lawler (2005). *Conformally Invariant Processes in the Plane*, American Mathematical Society.
20. G. Lawler (2006), The Laplacian-b random walk and the Schramm-Loewner evolution, Illinois J. Math. **50**, no. 1-4, 701–746.
21. G. Lawler, O. Schramm, W. Werner (2001), Values of Brownian intersection exponents I: Half-plane exponents, Acta Math. **187**, 237–273.
22. G. Lawler, O. Schramm, W. Werner (2001), Values of Brownian intersection exponents II: Plane exponents, Acta Math. **187**, 275–308.
23. G. Lawler, O. Schramm, W. Werner (2002), Analyticity of intersection exponents for planar Brownian motion, Acta Math. **189**, 179–201.
24. G. Lawler, O. Schramm, W. Werner (2004), Conformal invariance of planar loop-erased random walks and uniform spanning trees, Annals of Probab. **32**, 939–995.
25. G. Lawler, O. Schramm, W. Werner (2004), On the scaling limit of planar self-avoiding walk, in *Fractal Geometry and Applications: A Jubilee of Benoit Mandelbrot*, Vol II., M. Lapidus, M. van Frankenhuijsen, ed., Amer. Math. Soc., 339–364.
26. G. Lawler, O. Schramm, W. Werner (2003), Conformal restriction: the chordal case, J. Amer. Math. Soc. **16**, 917–955.
27. G. Lawler, W. Werner (1999), Intersection exponents for planar Brownian motion, Annals of Probability **27**, 1601–1642.
28. G. Lawler, W. Werner (2000), Universality for conformally invariant intersection exponents, J. Europ. Math. Soc. **2**, 291-328.
29. G. Lawler, W. Werner (2004), The Brownian loop soup, Probab. Theory Related Fields **128**, 565–588.
30. P. Lëvy (1946), *Processus Stochasticques et Mouvement Brownien*, Gauthier-Villars.
31. K. Löwner (1923), Untersuchungen über schlichte konforme Abildugnen des Einheitskreises I, Math. Ann. **89**, 103–121.
32. B. Mandelbrot (1982), *The Fractal Geometry of Nature*, Freeman.
33. S. Rohde, O. Schramm (2005), Basic properties of SLE. Ann. of Math. **161**, 883–924..
34. O. Schramm (2000), Scaling limits of loop-erased random walks and uniform spanning trees, Israel J. Math. **118**, 221–288.
35. O. Schramm, S. Sheffield (2005), Harmonic explorer and its convergence to SLE$_4$. Ann. Probab. **33**, 2127–2148.
36. O. Schramm, S. Sheffield, Contour lines of the discrete Gaussian free field, preprint.
37. S. Sheffield, Gaussian free fields for mathematicians, preprint.
38. S. Smirnov (2001), Critical percolation in the plane: Conformal invariance, Cardy's formula, scaling limits, C. R. Acad. Sci. Paris Sér. I Math. **333** no. 3, 239–244.
39. S. Smirnov (2006) Towards conformal invariance of $2D$ lattice models, *International Conngress of Mathematicians, Madrid 2006*, Eur. Math. Soc. 1421–1451.
40. W. Werner (2004), Random planar curves and Schramm-Loewner evolutions, Ecole d'Eté de Probabilités de Saint-Flour XXXII - 2002, Lecture Notes in Mathematics **1840**, Springer-Verlag, 113–195.
41. D. Wilson (1996), Generating spanning trees more quickly than the cover time in *Proceedings of the Twenty-Eighth Symposium on the Theory of Computing*, ACM, 296–303.
42. D. Zhan (2007), Reversibility of chordal SLE, to appear in Annals of Prob.

Department of Mathematics, University of Chicago, 5734 S. University Ave., Chicago, IL 60637-1546

E-mail address: lawler@math.uchicago.edu

WHY ARE SOLITONS STABLE?

TERENCE TAO

Abstract. The theory of linear dispersive equations predicts that waves should spread out and disperse over time. However, it is a remarkable phenomenon, observed both in theory and practice, that once nonlinear effects are taken into account, *solitary wave* or *soliton* solutions can be created, which can be stable enough to persist indefinitely. The construction of such solutions is relatively straightforward, but the fact that they are *stable* requires some significant amounts of analysis to establish, in part due to symmetries in the equation (such as translation invariance) which create degeneracy in the stability analysis. The theory is particularly difficult in the *critical* case in which the nonlinearity is at exactly the right power to potentially allow for a self-similar blowup. In this article we survey some of the highlights of this theory, from the more classical orbital stability analysis of Weinstein and Grillakis-Shatah-Strauss, to the more recent asymptotic stability and blowup analysis of Martel-Merle and Merle-Raphael, as well as current developments in using this theory to rigorously demonstrate controlled blowup for several key equations.

1. Introduction

In these notes we shall eventually describe recent developments in the stability of solitons. Before we discuss solitons, however, we need to first describe the wider context of dispersive equations, and why even the very existence of solitons were initially such a surprising phenomenon. Due to lack of time, a bibliography was not prepared for these notes; this will be rectified in the proceedings article version of these notes. We will also suppress technical details and computations, and on the most general formulation of variosu results, and try instead to just give a flavour of the phenomena that we are trying to understand, and a glimpse of some of the tools we use to try to explain such phenomena.

In classical physics, it has been realised for centuries that the behaviour of idealised vibrating media (such as waves on string, on the surface of a body of water, or in air), in the absence of friction or other dissipative forces, can be modeled by a number of partial differential equations, known collectively as *dispersive equations*. Model examples of such equations include the following:

- The *free wave equation*
$$u_{tt} - c^2 \Delta u = 0,$$
where $u : \mathbf{R} \times \mathbf{R}^d \to \mathbf{R}$ represents the amplitude $u(t, x)$ of a wave at a point in a spacetime with d spatial dimensions, $\Delta := \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2}$ is the spatial Laplacian on $\mathbf{R}^d$, u_{tt} is short for $\frac{\partial^2 u}{\partial t^2}$, and $c > 0$ is a fixed constant (which can be rescaled to equal 1 if one wishes). This equation models the

2000 *Mathematics Subject Classification.* Primary .

evolution of waves in a medium which has a fixed speed c of propagation in all directions.

- The *linear (time-dependent) Schrödinger equation*

$$i\hbar u_t + \frac{\hbar^2}{2m}\Delta u = Vu \tag{1.1}$$

where $u : \mathbf{R} \times \mathbf{R}^d \to \mathbf{C}$ is the wave function of a quantum particle, $\hbar, m > 0$ are physical constants (which can be rescaled to equal 1 if one wishes), and $V : \mathbf{R}^d \to \mathbf{R}$ is a potential function, which we assume to depend only on the spatial variable x. This equation models the evolution of a quantum particle in space in the presence of a classical potential well V.

- The *nonlinear Schrödinger (NLS) equation*

$$iu_t + \Delta u = \mu|u|^{p-1}u \tag{1.2}$$

where $p > 1$ is an exponent and $\mu = \pm 1$ is a sign (the case $\mu = +1$ is known as the *defocusing* case, and $\mu = -1$ as the *focusing* case). This equation can be viewed as a variant of the linear Schrödinger equation (with the constants $\hbar$ and m normalised away), in which the potential V now depends in a nonlinear fashion on the solution itself. This equation no longer has a physical interpretation as the evolution of a quantum particle, but can be derived as a model for quantum media such as Bose-Einstein condensates.

- The *(time-dependent) Airy equation*

$$u_t + u_{xxx} = 0 \tag{1.3}$$

where $u : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ is a scalar function. This equation can be derived as a very simplified model for propagation of low amplitude water waves in a shallow canal, by starting with the Euler equations, making a number of simplifying assumptions to discard nonlinear terms, and the normalising all constants to equal 1.

- The *Korteweg-de Vries (KdV) equation*

$$u_t + u_{xxx} + 6uu_x = 0 \tag{1.4}$$

which is a more refined version of the Airy equation in which the first nonlinear term is retained. The constant 6 that appears here is not essential, but turns out to be convenient when connecting this equation to the theory of inverse scattering (of which more will be said later).

- The *generalised Korteweg-de Vries (gKdV) equation*

$$u_t + u_{xxx} + (u^p)_x = 0 \tag{1.5}$$

for $p > 1$ an integer; the case $p = 2$ is essentially the KdV equation, and the case $p = 3$ is known as the *modified Korteweg-de Vries (mKdV) equation.* The case $p = 5$ is particularly interesting due to its *mass-critical* nature, which we will discuss later.

These are all evolution equations; if we specify the initial position $u(0, x) = u_0(x)$ of a wave at time $t = 0$, we expect these equations to have a unique solution with that initial data[1] for all future times $t > 0$. Actually, all of the above equations are

[1] For the wave equation, which is second-order in time, we also need to specify the initial velocity $\partial_t u(0, x) = u_1(x)$.

time-reversible (for instance, if $(t,x) \mapsto u(t,x)$ solves (1.4), then $(t,x) \mapsto u(-t,-x)$ also solves (1.4)) so we also expect the initial data at time $t = 0$ to determine the solution at all past times $t < 0$. (This is in sharp contrast to *dissipative* equations such as the heat equation $u_t = \Delta u$, which are solvable forward in time but are not solvable backwards in time, at least in the category of smooth functions, due to an irreversible loss of energy and information inherent in this equation as time moves forward.)

Solutions to the above equations have two properties, which may seem at first to contradict each other. The first property is that all of these equations are *conservative*; there exists a *Hamiltonian* $v \mapsto H(v)$, which is a functional that assigns a real number $H(v)$ to any (sufficiently smooth and decaying[2]) function v on the spatial domain $\mathbf{R}^d$, such that the Hamiltonian[3] $H(u(t))$ of a (sufficiently smooth and decaying) solution $u(t)$ to the above equation is conserved in time:

$$H(u(t)) = H(u(0)), \text{ or equivalently } \partial_t H(u(t)) = 0.$$

More specifically, the Hamiltonian is given by

$$H(u(t), u_t(t)) := \int_{\mathbf{R}^d} \frac{1}{2}|u_t(t,x)|^2 + \frac{1}{2}|\nabla_x u(t,x)|^2 \, dx$$

for the wave equation,

$$H(u(t)) := \int_{\mathbf{R}^d} \frac{\hbar^2}{2m}|\nabla_x u(t,x)|^2 + V(x)|u(t,x)|^2 \, dx$$

for the linear Schrödinger equation,

$$H(u(t)) := \int_{\mathbf{R}^d} \frac{1}{2}|\nabla_x u(t,x)|^2 + \frac{\mu}{p+1}|u(t,x)|^{p+1} \, dx$$

for the nonlinear Schrödinger equation,

$$H(u(t)) := \int_{\mathbf{R}} u_x(t,x)^2 \, dx$$

for the Airy equation,

$$H(u(t)) := \int_{\mathbf{R}} u_x(t,x)^2 - 2u(t,x)^3 \, dx \tag{1.6}$$

for the Korteweg-de Vries equation, and

$$H(u(t)) := \int_{\mathbf{R}} u_x(t,x)^2 - \frac{2}{p+1} u(t,x)^{p+1} \, dx \tag{1.7}$$

for the generalised Korteweg-de Vries equation. In all of these cases, the conservation of the Hamiltonian can be formally verified by computing $\partial_t H(u(t))$ via

[2] To simplify the exposition, we shall largely ignore the important, but technical, analytic issues of exactly how much regularity and decay one needs in order to justify all the computations and assertions given here. In practice, one usually first works in the category of *classical solutions* - solutions that are smooth and rapidly decreasing - and then uses rigorous limiting arguments (and in particular, exploiting the *low-regularity well-posedness theory* of these equations) to extend all results to more general classes of solutions, such as solutions in the energy space $H^1(\mathbf{R}^d)$.

[3] Again, with the wave equation, the Hamiltonian depends on the instantaneous velocity $u_t(t)$ of the solution at time t as well as the instantaneous position $u(t)$.

differentiation under the integral sign, substituting the evolution equation for u, and then integrating by parts; we leave this to the reader as an exercise[4].

Actually, the Hamiltonian is not the only conserved quantity available for these equations; each of these equations also enjoy a number of symmetries (e.g. translation invariance), which (by Noether's theorem) leads to a number of important additional conserved quantities, which we will discuss later. It is often helpful to interpret the Hamiltonian as describing the total *energy* of the wave.

The conservative nature of these equations means that even for very late times t, the state $u(t)$ of solution is still[5] "similar" to the initial state $u(0)$, in the sense that they have the same energy. This may lead one to conclude that solutions to these evolution equations should evolve in a fairly static, or fairly periodic manner; after all, this is what happens to solutions to finite-dimensional systems of ordinary differential equations which have a conserved energy H which is *coercive* in the sense that the *energy surfaces* $\{H = \text{const}\}$ are always bounded.

However, this intuition turns out to not be correct in the realm of dispersive equations, even though such equations can be thought of as infinite-dimensional systems of ODE with a conserved energy, and even though this energy usually exhibits coercive properties. This is ultimately because of a second property of all of these equations, namely *dispersion.* Informally, dispersion means that different components of a solution u to any of these equations travel at different velocities, with the velocity of each component determined by the velocity. As a consequence, even though the state of solution at late times has the same energy as the initial state, the different components of the solution are often so dispersed that the solution at late times tends to have much smaller amplitudes than at early times[6]. Thus, for instance, it is perfectly possible for the solution $u(t)$ to go to zero in $L^\infty(\mathbf{R}^d)$ norm as $t \to \pm\infty$, even as its energy stays constant (and non-zero). The ability to go to zero as measured in one norm, while staying bounded away from zero in another, is a feature of systems with infinitely many degrees of freedom, which is not present when considering systems of ODE with only boundedly many degrees of freedom.

One can see this dispersive effect in a number of ways. One (somewhat informal) way is to analyse *plane wave* solutions

$$u(t,x) = Ae^{it\tau + x\cdot\xi} \tag{1.8}$$

for some non-zero amplitude A, some temporal frequency $\tau \in \mathbf{R}$, and some spatial frequency $\xi \in \mathbf{R}^d$. For instance, for the Airy equation (1.3), one easily verifies that (1.8) solves[7] (1.3) exactly when $\tau = \xi^3$; this equation is known as the *dispersion relation* for the Airy equation. If we rewrite the right-hand side of (1.8) in this case

[4] One can also formally establish conservation of the Hamiltonian by interpreting each of the above dispersive equations in turn as an infinite-dimensional Hamiltonian system, but we will not adopt this (important) perspective here.

[5] This is assuming that the solution exists all the way up to this time t, which can be a difficult task to establish rigorously, especially if the initial data was rough. Again, we suppress these important technical issues for simplicity.

[6] This phenomenon may seem to be inconsistent with time reversal symmetry. However, this dispersive effect only occurs when the initial data is spatially localised; dispersion sends localised high-amplitude states to broadly dispersed, low-amplitude states, but (by time reversal) can also have the reverse effect.

[7] Strictly speaking, one needs to allow solutions u to (1.3) to be complex-valued here rather than real-valued, but this is of course a minor change.

as $Ae^{i\xi(x-(-\xi^2)t)}$, this asserts that a plane wave solution to (1.3) has a *phase velocity* $-\xi^2$ which is the negative square of its spatial frequency ξ. Thus we see that for this equation, higher frequency plane waves have a much faster phase velocity than lower frequency ones, and the velocity is always in a leftward direction. Similar analyses can be carried out for the other equations given above, though in those equations involving a nonlinearity or a potential, one has to restrict attention to small amplitude or high frequency solutions so that one can (non-rigorously) neglect the effect of these terms. For instance, for the Schrödinger equation (1.1) (at least with $V = 0$) one has the dispersion relation

$$\tau = -\frac{\hbar^2}{2m}|\xi|^2. \tag{1.9}$$

However, as is well known in physics, the phase velocity does not determine the speed of propagation of information in a system; that quality is instead controlled by the *group velocity*, which typically is slightly different from the phase velocity. To explain this quantity, let us modify the ansatz (1.8) by allowing the amplitude A to vary slowly in space, and propagate in time at some velocity $v \in \mathbf{R}^d$. More precisely, we consider solutions of the form

$$u(t,x) \approx A(\varepsilon(x-vt))e^{it\tau + x\cdot\xi}$$

where A is a smooth function and $\varepsilon > 0$ is a small parameter, and we shall be vague about what the symbol "$\approx$" means. If we have $\tau = \xi^3$ as before, then a short (and slightly non-rigorous) computation shows that

$$u_t + u_{xxx} \approx \varepsilon(v + 3\xi^2)A'(\varepsilon(x-vt))e^{it\tau + x\cdot\xi} + O(\varepsilon^2).$$

Thus we see that in order for u to (approximately) solve (1.3) up to errors of $O(\varepsilon^2)$, the group velocity v must be equal to $-3\xi^2$, which is three times the phase velocity $-\xi^2$. Thus, at a qualitative level at least, we still have the same predicted behaviour as before; all frequencies propagate leftward, and higher frequencies propagate faster than lower ones. In particular we expect localised high-amplitude states, which can be viewed (via the Fourier inversion formula) as linear superpositions of plane waves of many different frequencies, to disperse leftwards over time into broader, lower-amplitude states (but still with the same energy as the original state, of course).

One can perform similar analyses for other equations. For instance, for the linear Schrödinger equation, and assuming either high frequencies or small potential, one expects waves to propagate at a velocity proportional to their frequency, according to *de Broglie's law* $mv = \hbar\xi$; similarly for the nonlinear Schrödinger equation when one assumes either high frequencies or small amplitude. In contrast, for wave equation, this analysis suggests that waves of (non-zero) frequency ξ should propagate at velocities $\frac{\xi}{|\xi|}c$; thus the propagation speed c is constant but the propagation direction $\frac{\xi}{|\xi|}$ varies with frequency, leading to a weak dispersive effect. For more general dispersive equations, the group velocity can be read off of the dispersion relation $\tau = \tau(\xi)$ by the formula $v = -\nabla_\xi \tau$ (whereas in contrast, the phase velocity is $-\frac{\xi}{|\xi|^2}\tau$).

In the case of the Schrödinger equation with $V = 0$, one can see the dispersive effort more directly by using the explicit solution formula

$$u(t,x) = \frac{1}{(2\pi\hbar t/m)^{d/2}} \int_{\mathbf{R}^d} e^{i\frac{m|x-y|^2}{2\hbar^2 t}} u_0(y)\, dy$$

for $t \neq 0$ and all sufficiently smooth and decaying initial data u_0. Indeed, we immediately conclude from this formula (formally, at least) that if u_0 is absolutely integrable, then $\|u(t)\|_{L^\infty(\mathbf{R}^d)}$ decays at the rate $O(|t|^{-d/2})$.

In the case of linear equations such as the Airy equation (1.3), there is a similar explicit formula (involving the Airy function $\mathrm{Ai}(x)$ instead of complex exponentials), but one can avoid the use of special functions by instead proceeding using the Fourier transform and the principle of stationary phase. Indeed, by starting with the Fourier inversion formula

$$u_0(x) = \int_{\mathbf{R}} \hat{u}_0(\xi) e^{ix\xi}\ d\xi$$

where $\hat{u}_0(\xi) := \frac{1}{2\pi}\int_{\mathbf{R}} u_0(x) e^{-ix\xi}\ dx$ is the Fourier transform of u_0, and noting as before that $e^{it\xi^3+ix\xi}$ is the solution of the Airy equation with initial data $e^{ix\xi}$, we see from the principle of superposition (and ignoring issues of interchanging derivatives and integrals, etc.) that the solution u is given by the formula

$$u(t,x) = \int_{\mathbf{R}} \hat{u}_0(x) e^{it\xi^3+ix\xi}\ d\xi. \tag{1.10}$$

If u_0 is a Schwartz function (infinitely smooth and decreasing faster than any polynomial), then its Fourier transform is also Schwartz and thus slowly varying. On the other hand, as t increases, the phase $e^{it\xi^3+i\xi}$ oscillates more and more rapidly (for non-zero ξ), and so we expect an increasing amount of cancellation in the integral in (1.10), leading to decay of u as $t \to \infty$. This intuition can be formalised using the methods of stationary phase (which can be viewed as advanced applications of the undergraduate calculus tools of integration by parts and changes of variable), and can for instance be used to show that $\|u(t)\|_{L^\infty(\mathbf{R})}$ decays at a rate $O(t^{-1/3})$ in general.

This technique of representing a solution as a superposition of plane waves also works (with a twist) for the linear Schrödinger equation (1.1) in the presence of a potential V, provided that the potential is sufficiently smooth and decaying. The basic idea is to replace the plane waves (1.8) by *distorted plane waves* $\Phi(\tau, x)e^{it\tau}$, where (in order to solve (1.1)) Φ has to solve the *time-independent Schrödinger equation*

$$-\tau\hbar\Phi + \frac{\hbar^2}{2m}\Delta\Phi = V\Phi, \tag{1.11}$$

and then to try to represent solutions u to (1.1) as superpositions

$$u(t,x) = \int a(\tau)\Phi(\tau,x)e^{it\tau}\ d\tau$$

where we are being intentionally vague as to what the range of integration is. If we restrict attention to negative values of τ, then it turns out (by use of scattering theory) that we can construct distorted plane waves $\Phi(\tau, x)$ which asymptotically resemble the standard plane waves $e^{ix\cdot\xi}$ as $|x| \to \infty$, where ξ is a frequency obeying the dispersion relation (1.9). If u is composed entirely of these waves, then one has a similar dispersive behaviour to the free Schrödinger equation (for instance, under suitable regularity and decay hypotheses on V and u_0, $\|u(t)\|_{L^\infty(\mathbf{R}^d)}$ will continue to decay like $O(t^{-d/2})$). In such cases we say that u is in a *radiating state.* In many important cases (such as when the potential V is non-negative, or is small in certain function space norms), all states (with suitable regularity and decay hypotheses) are radiating states. However, when V is large and allowed to be negative, it is also possible[8] to contain *bound states*, in which τ is positive, and the distorted plane wave $\Phi(\tau, x)$ is replaced by an *eigenfunction* Φ, which continues to solve the equation (1.11), but now Φ decays exponentially to zero as $|x| \to \infty$, instead of oscillating like a plane wave as before. (Informally, this is because once τ is positive, the dispersion relation (1.9) is forcing ξ to be imaginary rather than real.) In particular, Φ lies in $L^2(\mathbf{R}^d)$, and so $-\tau$ becomes an eigenvalue of the Schrödinger operator[9] $H := -\frac{\hbar^2}{2m}\Delta + V$. Because multiplication V is a compact operator relative to $-\frac{\hbar^2}{2m}\Delta$, standard spectral theory shows that the set of eigenvalues $-\tau$ is discrete (except possibly at the origin $-\tau = 0$). Note that it is necessary for V to take on negative values in order to obtain negative eigenvalues, since otherwise the operator H is positive semi-definite.

If u_0 consists of a superposition of one or more of these eigenfunctions, e.g.

$$u_0 = \sum_k c_k \Phi(\tau_k, x)$$

where $-\tau_k$ ranges over finitely many of the eigenvalues of H, then we formally have

$$u(t) = \sum_k c_k e^{it\tau_k} \Phi(\tau_k, x),$$

and so we see that $u(t)$ oscillates in time but does not disperse in space. In this case we say that u is a *bound state*. Indeed, the evolution is instead *almost periodic*, in the sense that $\liminf_{t\to\infty} \|u(t) - u_0\|_{L^2(\mathbf{R}^d)} = 0$, or equivalently that the orbit $\{u(t) : t \in \mathbf{R}\}$ is a precompact subset of $L^2(\mathbf{R}^d)$.

By further application of spectral theory, one can show that an arbitrary state u_0 (in, say, $L^2(\mathbf{R}^d)$) can be decomposed as the orthogonal sum of a radiating state, which disperses as $t \to \infty$, and a bound state, which evolves in an almost periodic manner. Indeed this decomposition corresponds to the decomposition of the spectral measure of H into absolutely continuous and pure point components.

2. Solitons

We have seen how solutions to linear dispersive equations either disperse completely as $t \to \infty$, or else (in the presence of an external potential) decompose into a superposition of a radiative state that disperses to zero, plus a bound state that exhibits phase oscillation but is otherwise stationary.

In everyday physical experience with water waves, we of course see that such waves disperse to zero over time; once a rock is thrown into a pond, for instance, the amplitude of the resulting waves diminish over time. However, one does see in nature water waves which refuse to disperse for astonishingly long periods of time, instead moving at a constant speed without change in shape. Such *solitary waves*

[8] When V does not decay rapidly, then there can also be some intermediate states involving the singular continuous spectrum of the Schrödinger operator $-\frac{\hbar^2}{2m}\Delta + V$, which disperse over time slower than the radiating states but faster than the bound states. One can also occasionally have *resonances* corresponding to the boundary case $\tau = 0$, which exhibit somewhat similar behaviour. For simplicity of exposition, we will not discuss these (important) phenomena.

[9] This operator H is related to the Hamiltonian $H(u)$ discussed earlier by the formula $H(u) = \langle Hu, u\rangle$, where $\langle u, v\rangle := \int_{\mathbf{R}^d} u\overline{v}$ is the usual inner product on $L^2(\mathbf{R}^d)$.

or *solitons*[10] were first observed by John Scott Russell, who followed such a wave in a shallow canal on horseback for over a mile, and then reproduced such a travelling wave in a wave tank.

This phenomenon was first explained mathematically by Korteweg and de Vries in 1895, using the equation (1.4) that now bears their name (although this equation was first proposed as a model for shallow wave propagation by Boussinesq a few decades earlier). Indeed, if one considers travelling wave solutions to (1.4) of the form

$$u(t,x) = f(x-ct)$$

for some velocity c, then this will be a solution to (1.4) as long as f solves the ODE

$$-cf' + f''' + 6ff' = 0.$$

If we assume that f decays at infinity, then we can integrate this third-order ODE to obtain a first-order ODE

$$-cf + f'' + 3f^2 = 0.$$

For $c > 0$, this ODE admits the localised explicit solutions $f(x) = cQ(c^{1/2}(x-x_0))$ for any $x_0 \in \mathbf{R}$, where Q is the explicit Schwartz function $Q(x) := \frac{1}{2}\operatorname{sech}^2(\frac{x}{2})$. For $c \leq 0$, one can show that there are no localised solutions other than the trivial solution $f \equiv 0$. Thus we obtain a family of explicit travelling wave solutions

$$u(t,x) = cQ(c^{1/2}(x-ct-x_0)) \tag{2.1}$$

to the KdV equation; the parameter c thus controls the speed, amplitude, and width of the wave, while x_0 determines the initial location.

Interestingly, all the solutions (2.1) move to the *right*, while radiating states move to the left. This phenomenon is somewhat analogous to the situation with the linear Schrödinger equation, in which the temporal frequency τ (which is somewhat like the propagation speed c in KdV) was negative for radiating states and positive for bound states. Similar travelling wave solutions can also be found for gKdV and NLS equations, though in higher dimensions $d > 1$ one cannot hope to obtain such explicit formulae for these solutions, and instead one needs to use more modern PDE tools, such as calculus of variations and other elliptic theory methods, in order to build such solutions. There are also larger and more oscillatory "excited" travelling wave solutions which, unlike the "ground state" travelling wave solutions described above, exhibit changes of sign, but we will not discuss them here.

Early numerical analyses of the KdV equation revealed that these soliton solutions (2.1) were remarkably stable. Firstly, if one perturbed a soliton by adding a small amount of noise, then the noise would soon radiate away from the soliton, leaving the soliton largely unchanged (other than some slight perturbation in the c and x_0 parameters); these phenomena are described mathematically by results on the *orbital stability* and *asymptotic stability* of solitons, of which more will be said later. This is perhaps unsurprising, given that solitons move rightwards and radiation moves leftwards, but one has to bear in mind that equations such as (1.4)

[10] Strictly speaking, a wave which is localised and maintains its form for long periods of time is merely a *travelling wave*. A soliton is a travelling wave with the additional property that solitons and other radiation can pass through it without destroying its form. The question as to whether the travelling waves presented here are true solitons in that sense will in fact be the central topic of discussion of these notes.

are not linear, and in particular one cannot obviously superimpose a soliton and a radiative state to create a new solution to the KdV equation.

What was even more surprising was what happened if one considered collisions between two solitons, for instance imagining initial data of the form

$$u(0,x) = c_1 Q(c_1^{1/2}(x - x_1)) + c_2 Q(c_2^{1/2}(x - x_2))$$

with $0 < c_2 < c_1$ and x_1 far to the left of x_2; thus initially we have a larger, fast-moving soliton to the left of a shallower, slow-moving soliton. If the KdV equation were linear, the solution would now take the form

$$u(t,x) = c_1 Q(c_1^{1/2}(x - c_1 t - x_1)) + c_2 Q(c_2^{1/2}(x - c_2 t - x_2))$$

and so the faster solitons would simply overtake the slower one, with no interaction between the two. At the other extreme, with a strongly nonlinear equation, one could imagine all sorts of scenarios when two solitons collide, for instance that they scatter into radiation or into many smaller solitons, combine into a large soliton, and so forth. However, the KdV equation exhibits an interesting intermediate behaviour: the solitons do interact nonlinearly with each other during collision, but then emerge from that collision almost unchanged, except that the solitons have been shifted slightly by their collision. In other words, for very late times t, the solution approximately takes the form

$$u(t,x) \approx c_1 Q(c_1^{1/2}(x - c_1 t - x_1 - \theta_1)) + c_2 Q(c_2^{1/2}(x - c_2 t - x_2 - \theta_2))$$

for some additional shift parameters $\theta_1, \theta_2 \in \mathbf{R}$.

More generally, if one starts with *arbitrary* (but smooth and decaying) initial data, what usually happens (numerically, at least) with evolutions of equations such as (1.4) is that some non-linear (and chaotic-seeming) behaviour happens for a while, but eventually most of the solution radiates away to infinity and a finite number of solitons emerge, moving away from each other at different rates. Quite remarkably, this behaviour can in fact be justified rigorously for the KdV equation and a handful of other equations (such as the NLS equation in the cubic one-dimensional case $d = 1, p = 3$) due to the inverse scattering method, which we shall discuss shortly, although even in those cases, there are some exotic solutions, such as "breather" solutions, which occasionally arise and which do not evolve to a superposition of solitons and radiation, but instead exhibit periodic or almost periodic behaviour in time. Nevertheless, it is widely believed (and supported by extensive numerics) that for many other dispersive equations (roughly speaking, those equations whose nonlinearity is not strong enough to cause finite time blowup, and more precisely for the *subcritical* equations), solutions with "generic" initial data should eventually resolve into a finite number of solitons, moving at different speeds, plus a radiative term which goes to zero. This (rather vaguely defined) conjecture goes by the name of the *soliton resolution conjecture*. Except for those few equations which admit exact solutions (for instance, by inverse scattering methods), the conjecture remains unsolved in general, in part because we have very few tools available that can say anything meaningful about *generic* data in a certain class (e.g. with some function norm bounds) without also being applicable to *all* data in that class; thus the presence of a few exotic solutions that do not resolve into solitons and radiation seems to prevent us from tackling all the other cases. Nevertheless, there are certain important regimes in which we do have a good understanding. One of these is the perturbative regime near a single soliton, in which the initial state u_0

is close to that of a soliton such as (2.1); this case will be the main topic of our discussion. More recently, results have begun to emerge on *multisoliton* states, in which the solution is close to the superposition of many widely separated solitons, and even more recently still, there has been some results on the collision between a very fast narrow soliton and a very slow broad one. However, it seems that truly non-perturbative regimes, such as the collisions between two solitons of comparable size, remain beyond the reach of current tools (perhaps requiring a new advance in our understanding of dynamical systems in general).

3. The inverse scattering approach

We now briefly mention the technique of *inverse scattering*, which is a non-perturbative approach which allows one to control the evolution of solutions to completely integrable equations such as (1.4). This is a vast subject, which can be viewed from many different algebraic and geometric perspectives; we shall content ourselves with describing the approach based on *Lax pairs*, which has the advantage of simplicity, provided that one is willing to accept a rather miraculous algebraic identity.

The identity in question is as follows. Suppose that u solves the KdV equation (1.4). As always we assume enough smoothness and decay to justify the computations that follow. For every time t, we consider the time-dependent differential operators $L(t)$, $P(t)$ acting on functions on the real line $\mathbf{R}$, defined by

$$\begin{aligned} L(t) &:= \frac{d^2}{dx^2} + u(t) \\ P(t) &:= \frac{d^3}{dx^3} + \frac{3}{4}(\frac{d}{dx}u(t) + u(t)\frac{d}{dx}) \end{aligned}$$

where we view $u(t)$ as a multiplication operator, $f \mapsto u(t)f$. One can view $P(t)$ as a truncated (non-commutative) Taylor expansion of $L(t)^{3/2}$. In view with this interpretation, it is perhaps not so surprising that $L(t)$ and $P(t)$ "almost commute"; the commutator $[P(t), L(t)] := P(t)L(t) - L(t)P(t)$ of the third order operator $P(t)$ and the second order operator $L(t)$ would normally be expected to be fourth order, but in fact things collapse to just be zeroth order. Indeed, after some computation, one eventually obtains

$$[P(t), L(t)] = \frac{1}{4}(u_{xxx}(t) + 6u(t)u_x(t)).$$

In particular, if we substitute in (1.4), we obtain the remarkable *Lax pair equation*

$$\frac{d}{dt}L(t) = 4[P(t), L(t)]. \tag{3.1}$$

If we non-rigorously treat the operators $L(t)$, $P(t)$ as if they were matrices, we can interpret this equation as follows. Using the Newton approximation

$$L(t+dt) \approx L(t) + dt\frac{d}{dt}L(t); \quad \exp(\pm P(t)dt) \approx (1 \pm P(t)dt)$$

for infinitesimal dt, we see from (3.1) that

$$L(t+dt) \approx \exp(4P(t)dt)L(t)\exp(-4P(t)dt). \tag{3.2}$$

This informal analysis suggests that $L(t+dt)$ is a conjugate of $L(t)$, and so on iterating this we expect $L(t)$ to be a conjugate of $L(0)$. In particular, the *spectrum* of $L(t)$ should be time-invariant. Since $L(t)$ is determined by $u(t)$, this leads to a rich source of invariants for $u(t)$.

The above analysis can be made more rigorous. For instance, one can show that the traces[11] $\mathrm{tr}(e^{sL(t)})$ of heat kernels are independent of t for any fixed $s > 0$; expanding those traces in powers of s one can recover an infinite number of conservation laws, which includes the conservation of the Hamiltonian (1.6) as a special case. We will not pursue this approach further here. Another way to proceed is to consider solutions to the generalised eigenfunction equation

$$L(t)\phi(t,x) = \tau\phi(t,x) \tag{3.3}$$

for some $\tau \in \mathbf{R}$ and some smooth function $\phi(t,x) = \phi(t,x;\tau)$ (not necessarily decaying at infinity). If the equation (3.3) holds for a single time t (e.g. $t = 0$), and if ϕ then evolves by the equation

$$\phi_t(t,x) = 4P(t)\phi(t,x) \tag{3.4}$$

for all t, one can verify (formally, at least) that (3.3) persists for all t, by differentiating (3.3) in time and substituting in (3.1) and (3.4). (The astute reader will note that these manipulations are equivalent to those used to produce (3.2).

This now suggests an strategy to solve the KdV equation exactly from an arbitrary choice of initial data $u(0) = u_0$.

(1) Use the initial data u_0 to form the operator $L(0)$, and then locate the generalised eigenfunctions $\phi(0,x;\lambda)$ for each choice of spectral parameter τ.
(2) Evolve each generalised eigenfunction ϕ in time by the equation (3.4).
(3) Use the generalised eigenfunctions $\phi(t,x;\tau)$ to recover $L(t)$ and $u(t)$.

This strategy looks very difficult to execute, because the operator $P(t)$ itself depends on $u(t)$, and so (3.4) cannot be solved exactly without knowing what $u(t)$ is - which is exactly what we are trying to find in the first place! But we can break this circularity by only seeking to solve (3.4) *at spatial infinity* $x = \pm\infty$. Indeed, if $u(t)$ is decaying, and $\tau = -\xi^2$ for some real number ξ, then we see that solutions $\phi(t,x)$ to (3.3) must take the form

$$\phi(t,x) \approx a_\pm(\xi;t)e^{i\xi x} + b_\pm(\xi;t)e^{-i\xi x}$$

as $x \to \pm\infty$, for some quantities $a_\pm(\xi;t), b_\pm(\xi;t) \in \mathbf{C}$, which we shall refer to as the *scattering data* of $L(t)$. (One can normalise, say, $a_-(0) = 1$ and $b_-(\xi;0) = 0$, and focus primarily on $a_+(\xi;t)$ and $b_+(\xi;t)$, if desired.) Applying (3.4) and using the decay of $u(t)$ once again, we are then led (formally, at least) to the asymptotic equations

$$\partial_t a_\pm(\xi;t) = 4(i\xi)^3 a_\pm(\xi;t); \quad \partial_t b_\pm(\xi;t) = 4(-i\xi)^3 b_\pm(\xi;t)$$

which can be explicitly solved[12];

$$a_\pm(\xi;t) = e^{-4i\xi^3 t}a_\pm(\xi;0); \quad b_\pm(\xi;t) = e^{-4i\xi^3 t}b_\pm(\xi;0). \tag{3.5}$$

[11] Actually, to avoid divergences we need to consider normalised traces, such as $\mathrm{tr}(e^{sL(t)} - e^{s\frac{d^2}{dx^2}})$.

[12] Note the resemblance of the phases here to those in (1.10). This is not a co-incidence, and indeed the scattering and inverse scattering transforms can be viewed as nonlinear versions of the Fourier and inverse Fourier transform.

This only handles the case of negative energies $\lambda < 0$. For positive energies, say $\lambda = +\xi^2$ for some $\xi > 0$, the situation is somewhat similar; in this case, we have a discrete set of ξ for which we have a decaying solution $\phi(t,x)$, with $\phi(t,x) \approx c_\pm(\xi;t)e^{\mp\xi x}$ for $x \to \pm\infty$, where

$$c_\pm(\xi;t) = e^{\mp 4\xi^3 t} c_\pm(\xi;0). \tag{3.6}$$

This suggests a revised strategy to solve the KdV equation exactly:

(1) Use the initial data u_0 to form the operator $L(0)$, and then locate the scattering data $a_\pm(\xi;0)$, $b_\pm(\xi;0)$, $c_\pm(\xi;0)$.
(2) Evolve the scattering data by the equations (3.5), (3.6).
(3) Use the scattering data at t to recover $L(t)$ and $u(t)$.

The main difficulty in this strategy is now the third step, in which one needs to solve the *inverse scattering problem* to recover $u(t)$ from the scattering data. This is a vast and interesting topic in its own right, and involves complex-analytic problems such as the Riemann-Hilbert problem; we will not discuss it further here. Suffice to say, though, that after some work, it is possible to execute the above strategy for sufficiently smooth and decaying initial data u to obtain what is essentially[13] an explicit formula for u.

The relationship of all this to solitons is as follows. Recall from our discussion of the linear Schrödinger equation (1.1) that the operator $L(0) = \frac{d^2}{dx^2} + u_0$ is going to have radiating states (or absolutely continuous spectrum) corresponding to negative energies $\tau = -\xi^2 < 0$, and a discrete set of positive eigenfunctions corresponding to positive energies $\tau = +\xi^2 > 0$. Generically, the eigenvalues are simple. In that case, it turns out that the inverse scattering procedure relates each eigenvalue $+\xi^2$ of $L(0)$ to a soliton present inside u_0; the value of ξ determines the scaling parameter c of the soliton, and the scattering data $c_\pm(\xi;0)$ determines (in a slightly complicated fashion, depending on the rest of the spectrum) the location of the solitons. The remaining scattering data $a_\pm(\xi;0)$, $b_\pm(\xi;0)$ determines the radiative portion of the solution. As the solution evolves, the spectrum stays constant, but the data $a_\pm, b_\pm, c_\pm$ changes in a controlled manner; this is what causes the solitons to move and the radiation to scatter. It turns out that the exact location of each soliton depends to some extent on the relative sizes of the constants $c_\pm$, which are growing or decaying exponentially at differing rates; it is because of this that as one soliton overtakes another, the location of each soliton gets shifted slightly.

4. The analytic approach

The inverse scattering method gives extremely powerful and precise information on very general (and in particular, non-perturbative) solutions to equations such as the Korteweg-de Vries equation. However, it does not seem to be directly applicable to more general equations, such as the gKdV equation (1.5) for[14] $p \neq 2, 3$. For instance, no reasonable Lax pair formulation exists for these equations. We now

[13] The solution is not quite expressible as a closed-form integral, as in (1.10), but can be built out of solving a number of ordinary differential-integral equations (such as the Gelfand-Levitan-Marchenko equation), which turns out to suffice for the purposes of analysing the asymptotic behaviour of the solution.

[14] The modified KdV equation with $p = 3$ turns out to also be completely integrable; in fact, it can even be transformed directly into the KdV equation by a simple operation known as the *Miura transform*, which we will not discuss further here.

turn to more analytic techniques, which are less sensitive to the fine algebraic structure of the equation, although they still do rely very heavily on conservation laws and their relatives, such as monotonicity formulae.

We shall mostly restrict attention to the gKdV equation (1.5). We have already identified one conserved quantity for this equation, namely the energy (1.7). Another such conserved quantity is the mass

$$M(u(t)) := \int_{\mathbf{R}} u(t,x)^2\ dx.$$

Together, the mass and energy can (in some cases) control the H^1 norm

$$\|u(t)\|_{H^1_x(\mathbf{R})}^2 := \int_{\mathbf{R}} u(t,x)^2 + u_x(t,x)^2\ dx.$$

Indeed, if we are in the *mass-subcritical* case $p < 5$, then the *Gagliardo-Nirenberg inequality*

$$\int_{\mathbf{R}} v^{p+1} \leq C(p)(\int_{\mathbf{R}} v^2)^{\frac{p+3}{4}}(\int_{\mathbf{R}} v_x^2)^{\frac{p-1}{4}}, \tag{4.1}$$

valid for any v with suitable decay and regularity, gives us the *a priori* bound

$$\|u(t)\|_{H^1_x(\mathbf{R})} \leq C(M(u(t)), H(u(t))) = C(M(u_0), H(u_0)) \tag{4.2}$$

for some quantity $C(M(u_0), H(u_0))$ depending on the initial mass or energy. The condition $p < 5$ is necessary to ensure that the exponent $\frac{p-1}{4}$ in (4.1) is strictly less than 1. This condition can also be deduced from scaling heuristics, by investigating how the mass and energy transform under the scale invariance

$$u(t,x) \mapsto \lambda^{-\frac{2}{p-1}} u(\frac{t}{\lambda^3}, \frac{x}{\lambda}). \tag{4.3}$$

It is possible to use the *a priori* bound (4.2), combined with the Picard iteration method for constructing solutions, and some moderately advanced estimates from harmonic analysis, to show that the equation (1.5) in the mass-subcritical case admits a unique global smooth solutions from arbitrary smooth, decaying data. Thus there is no problem with existence, uniqueness, or regularity when it comes to these equations; the only remaining analytic issue (albeit a difficult one) is to understand the asymptotic behaviour of solutions.

The above analysis for $p < 5$ is valid no matter how large the mass $M(u) = M(u_0)$ of the solution is. If we then turn to the *mass-critical* case $p = 5$, the situation changes; the *a priori* bound (4.2) is now only valid when the mass $M(u_0)$ is sufficiently small. In fact, by using the sharp Gagliardo-Nirenberg inequality of Weinstein, one can be more precise as follows. Given any $p > 1$, the equation (1.5) admits a family of soliton (or travelling wave) solutions similar to (2.1), namely

$$u(t,x) = c^{1/(p-1)} Q(c^{1/2}(x - x_0 - ct)) \tag{4.4}$$

where

$$Q(x) := \left(\frac{p+1}{2\cosh^2(\frac{p-1}{2}x)} \right)^{1/(p-1)}$$

is a positive, smooth, rapidly decreasing solution to the ODE

$$Q_x x + Q^p = Q. \tag{4.5}$$

(In fact, up to translation, Q is the only such solution.) In particular, we have the *standard soliton* solution to gKdV

$$u(t,x) = Q(x-t); \tag{4.6}$$

all other solitons differ from the standard soliton only by the scaling (4.3) and the translation invariance.

In the mass-critical case $p = 5$, all of the solitons have the same mass, namely

$$M(u) = M(Q) = \int_{\mathbf{R}} Q^2.$$

The Gagliardo-Nirenberg inequality of Weinstein can then be used to show that one has the a priori bound (4.2) in the $p = 5$ case as long as one only considers solutions with mass strictly less than that of the ground state, and so long as the solution exists. The latter caveat is a substantial one; it is conjectured that one has global existence of solutions for gKdV of the $p = 5$ from smooth decaying initial data whenever the mass is less than that of the ground state, but this conjecture is still open. (However, global existence is known if the mass is sufficiently small, by a perturbative argument based on the contraction mapping principle and some harmonic analysis estimates.) An important recent work of Martel and Merle, though, shows in this case that singularities can form for data arbitrarily close to the ground state (but of slightly larger mass).

In the *mass-supercritical case* $p > 5$ the situation is very unclear, due to the lack of good *a priori* etimates in this case; it is likely that singularities do form in this case for large initial data, but this has not been rigorously established. It is however known that solitons are unstable in this setting.

We return now to the mass-subcritical case $p < 5$, in which global existence and regularity are assured, and now consider the problem of *stability* of one of the soliton solutions (4.4). By taking advantage of the scaling and translation invariance in the problem, we can reduce matters to considering the stability of the standard soliton $Q(x-t)$. For instance, if we are given a solution u which is close to this soliton at time zero, for instance in the sense that

$$\|u(0) - Q\|_{H^1(\mathbf{R})} \leq \sigma \tag{4.7}$$

for some sufficiently small σ, is this enough to guarantee that u stays close to (2.1) for much later times, thus

$$\|u(t) - Q(\cdot - t))\|_{H^1(\mathbf{R})} \leq \sigma' \tag{4.8}$$

for some other small σ' depending on σ, and for large times t? (For small times t, the local well-posedness theory allows one to obtain bounds of this form, but with σ' replaced by $C\sigma \exp(Ct)$ for some constant C depending only on p.) One can phrase this question in other norms than the H^1 norm, of course, but this norm turns out to be rather natural due to its connection with the Hamiltonian (which we have already seen in (4.2)).

This type of *absolute stability* of the soliton is too strong a property to hold, basically because it is not compatible with the scale invariance (4.3). Indeed, consider the soliton solution (4.4) with $x_0 = 0$ and $c = 1 + O(\sigma)$ very close to 1. Then (4.7) holds, but (4.8) will fail for sufficiently large times t, because u has most of its mass (and H^1 norm) near ct, whereas (2.1) has most of its mass near ct. The point is that by rescaling the soliton very slightly, one can adjust the speed of that soliton, which will eventually over time cause the perturbed soliton to diverge from the original soliton. Note that this conclusion has nothing to do with the H^1 norm, and would work for basically any reasonable function space norm.

However, even though this perturbed soliton is far from the original soliton at late times t, it is still close to a *translation* of that original soliton (by $ct - t$). Equivalently, if we define the *ground state curve*

$$\Sigma = \{Q(\cdot - x_0) : x_0 \in \mathbf{R}\} \subset H^1(\mathbf{R})$$

consisting of all translates of the ground state[15] Q, then we see that $u(t)$ stays close to Σ for all t. To put it another way, while $u(t)$ does not stay close to $Q(\cdot - t)$ for each time t, the *orbit* $\{u(t) : t \in \mathbf{R}\}$ stays close to the orbit $\{Q(\cdot - t) : t \in \mathbf{R}\} = \Sigma$. Indeed, this is a general phenomenon (established from the work of Benjamin, Bona, and Weinstein):

Theorem 4.1 (Orbital stability of sub-critical gKdV)**.** *Let* $1 < p < 5$*. If* $u_0 \in H^1(\mathbf{R})$ *is such that* $\mathrm{dist}_{H^1}(u_0, \Sigma)$ *is sufficiently small (say less than* σ *for some small constant* $\sigma > 0$*), and* u *is the solution to* (1.5) *with initial data* u_0*, then we have*

$$\mathrm{dist}_{H^1}(u(t), \Sigma) \lesssim \mathrm{dist}_{H^1}(u_0, \Sigma)$$

for all t*. Here we use* $X \lesssim Y$ *or* $X = O(Y)$ *to denote the estimate* $|X| \leq CY$ *for some* C *that depends only on* p*, and* $X \sim Y$ *as shorthand for* $X \lesssim Y \lesssim X$*.*

This theorem is proven by a variant of the classical *Lyapunov functional* method for establishing absolute stability. Let us briefly recall how that method works. Suppose we were able to find a functional $u \mapsto L(u)$ on H^1 with the following properties:

1. If u is an H^1 solution to (1.5), then $L(u(t))$ is non-increasing in t.
2. Q is a local minimiser of L, thus $L(u) - L(Q) \geq 0$ for all u sufficiently close to Q in H^1.
3. Furthermore, the minimum is non-degenerate in the sense that $L(u) - L(Q) \sim \|u - Q\|_{H^1}^2$ for all u sufficiently close to Q in H^1.

These three facts would then easily imply that Q is absolutely stable. Indeed, if u_0 is close to Q, then $L(u_0)$ is close to (but not smaller than) $L(Q)$, which implies that $L(u(t))$ is also close to but not smaller than $L(Q)$ for all $t > 0$, which implies (by a continuity argument) that $u(t)$ is close to Q for all $t > 0$. (The case $t < 0$ can then be handled by time reversal symmetry $u(t, x) \mapsto u(-t, -x)$, or equivalently by considering $L(u(-\cdot))$ instead of $L(u(\cdot))$.)

We already saw, though, that Q is not absolutely stable, and so such a Lyapunov functional cannot exist. However, we can still hope to obtain a modified Lyapunov functional $u \mapsto L(u)$ which implies orbital stability instead of absolute stability. More precisely, we require L to be such that

1. If u is an H^1 solution to (1.5), then $L(u(t))$ is non-increasing in t.
2. $L(u) = L(Q)$ for all $u \in \Sigma$.
3. Σ is a local minimiser of L, thus $L(u) - L(Q) \geq 0$ for all u sufficiently close to Σ in H^1.

[15] More suggestively, one should think of Σ as the space of all possible soliton states whose conserved statistics (in particular, mass and energy) agree with that of the ground state. In the case of the sub-critical NLS equation (1.2), Σ then becomes a cylinder, formed by considering the action of both translation $Q(x) \mapsto Q(x - x_0)$ and phase rotation $Q(x) \mapsto e^{i\theta}Q(x)$ on Q. In the critical case, the dimension of Σ increases due to the additional symmetry of scale invariance, as we shall shortly see.

(4) Furthermore, the minimum is non-degenerate in the sense that $L(u) - L(Q) \sim \operatorname{dist}_{H^1}(u, \Sigma)^2$ for all u sufficiently close to Σ in H^1.

It is not hard to see that this would imply Theorem 4.1. The task then reduces to locating the functional L with the stated properties. From properties 2,4 it seems reasonable to look for an L which is translation invariant. From property 1 and time reversal symmetry it seems reasonable to look for an L which is conserved, such as a combination of the mass $M(u)$ and the energy $H(u)$. It also has to be a functional for which Q is a local minimum, thus (4.5) should essentially be the Euler-Lagrange equation for L. With all these heuristics, one is soon led to the candidate

$$L(u) := H(u) + M(u).$$

It is then not hard to verify most of the required properties for L, especially if we *define* Q to be the minimiser of L. The one tricky thing is to show the strict non-degeneracy $L(u) - L(Q) \gtrsim \operatorname{dist}_{H^1}(u, \Sigma)^2$ when u is close to Σ. One difficulty here is the translation invariance of the estimate; if we do not break this symmetry, then we are forced to only use translation-invariant methods to establish the estimate, which greatly reduces the range of tools available. Hence we shall *break* the symmetry by decomposing

$$u = Q(\cdot - x_0) + \varepsilon$$

for some small function[16] $\varepsilon \in H^1(\mathbf{R})$ and some $x_0 \in \mathbf{R}$. There are a number of ways we can choose the parameter x_0. The most obvious approach is to pick $Q(\cdot - x_0)$ to be the translated ground state which is closest to u in H^1 norm, thus minimising $\|\varepsilon\|_{H^1}$. By elementary calculus, this allows us to obtain the orthogonality condition

$$\langle \varepsilon, Q'(\cdot - x_0) \rangle_{H^1(\mathbf{R})} = 0$$

where $\langle u, v \rangle_{H^1(\mathbf{R})} := \int_{\mathbf{R}} uv + u_x v_x$ is the standard inner product on H^1. Other choices of x_0 will lead to a slightly different orthogonality condition; some orthogonality conditions are more suitable for some applications than others, but we will not explore this technical issue further here.

We can then break (or "spend") the translation invariance by normalising x_0 to be zero, thus $u = Q + \varepsilon$ and ε is orthogonal to Q'. (Note that Q' represents the infinitesimal action of the translation group at Q.) Now, since Q is a minimiser of L we have (formally, at least) the Taylor expansion

$$L(Q + \varepsilon) = L(Q) + \frac{1}{2} L''(Q)(\varepsilon, \varepsilon) + O(\varepsilon^3)$$

where $L''(Q) : H^1(\mathbf{R}) \times H^1(\mathbf{R}) \to \mathbf{R}$ is some explicit positive semi-definite symmetric bilinear form. The task is then to show that

$$L''(Q)(\varepsilon, \varepsilon) \gtrsim \|\varepsilon\|_{H^1}^2$$

when ε is orthogonal to Q'. (An orthogonality condition of this sort is necessary; since L is translation invariant, we easily verify that $L''(Q)$ must annihilate Q'.) This is a spectral gap condition on $L''(Q)$, which can be viewed as a positive-definite self-adjoint operator, and can be established by spectral methods; the key ingredient needed is a *uniqueness* result that asserts that Q and its translates are the *only* minimisers of L. Details can be found in the paper of Weinstein.

[16] Note here that ε is denoting a *function* rather than a number! This notation is traditional in the literature.

Another way to state the above results is that if a global solution u starts off close to the ground state curve Σ, then at later times one has the decomposition

$$u(t,x) = Q(x-x(t)) + \varepsilon(t,x-x(t)) \tag{4.9}$$

for some function $x : \mathbf{R} \to \mathbf{R}$ (which tracks the position of the soliton component of u) and some error term ε, which is small in H^1. We have the freedom to impose one (non-degenerate) orthogonality condition of our choice on ε, such as $\langle \varepsilon, Q' \rangle_{H^1} = 0$, by choosing $x(t)$ appropriately.

The question then arises as to what happens to the error ε over time, or to the position $x(t)$. We return to the model example of the rescaled soliton (4.4). In this case we can take $x(t) = x_0 + ct$ and $\varepsilon(t,x) = c^{1/(p-1)}Q(c^{1/2}x) - Q(x)$ (this ε does not quite obey the above orthogonality condition, but this will not concern us). Thus we see in this case that ε does not disperse to zero in any sense. However, we can hope to "quotient out" this scaling and obtain a decomposition (4.9) which has a better error term ε. Indeed, if we replace the ground state curve Σ by the *ground state surface*

$$\Sigma' := \{c^{1/(p-1)}Q(c^{1/2}(\cdot - x_0)) : c > 0, x_0 \in \mathbf{R}\} \supset \Sigma$$

and approximate $u(t)$ by an element of Σ', we can obtain a more refined decomposition

$$u(t,x) = R(t,x) + \varepsilon(t,x) \tag{4.10}$$

where R is a soliton-like state

$$R(t,x) := c(t)^{1/(p-1)}Q(c(t)^{1/2}(x-x(t))) \tag{4.11}$$

for some *modulation functions* $x : \mathbf{R} \to \mathbf{R}$ and $c : \mathbf{R} \to \mathbf{R}^+$ (with c always close to 1), and ε is small in H^1 (one can take $\|\varepsilon\|_{H^1} = O(\sigma)$ if $\mathrm{dist}_{H^1}(u_0, \Sigma) \leq \sigma$). The point is now that by enlarging the dimension of the approximating surface from 1 to 2, the error ε is now allowed to enjoy *two* orthogonality conditions rather than just one. There are again several choices of which orthogonality conditions to pick (anything which is suitably "transverse" to Σ' will do); a typical set of choices is

$$\int_{\mathbf{R}} R(t,x)\varepsilon(t,x)\ dx = \int_{\mathbf{R}} (x-x(t))R(t,x)\varepsilon(t,x)\ dx = 0. \tag{4.12}$$

Now we can hope that with such a refined decomposition that the error ε will disperse, especially in the neighbourhood of $x(t)$, so that in the vicinity of the solition $R(t,x)$, the solution u converges locally to $R(t,x)$. Let us informally say that the soliton is *asymptotically stable* if we have a result of this form. Such stability results can be obtained for the completely integrable cases $p = 2, 3$ by using the inverse scattering methods of the previous section. For general sub-critical p, such results were first obtained by Pego and Weinstein and by Mizumachi, for perturbations of the ground state which were strongly localised (e.g. assuming exponential decay at infinity). More recently, Martel and Merle were able to consider more general perturbations which were only assumed to be small in the energy norm H^1; this generalisation is important for the purposes of understanding how a soliton will collide with another shallow broad soliton, which may have small energy but will not have strong localisation properties. In particular, they showed

Theorem 4.2 (Asymptotic stability for sub-critical gKdV). *Let the notation and assumptions be as in Theorem 4.1. Then we have a decomposition of the form* (4.10), *with* $c(t) = c_+$ *constant and close to* 1, $x(t)$ *differentiable with*

$$\lim_{t \to +\infty} x'(t) = c_+, \tag{4.13}$$

and the error term ε *obeying the local decay* $\|\varepsilon(t)\|_{H^1(x>\beta t)} \to 0$ *as* $t \to +\infty$ *for any* $\beta > 0$.

Roughly speaking, this asserts that as $t \to \infty$, the solution resolves into a soliton moving to the right at an asymptotically constant speed c_+, plus an error term which is radiating to the left; this is of course consistent with the soliton resolution conjecture. (In the case $p = 4$, and with the additional scale-invariant assumption that $u - Q$ is small in $\dot H^{1/6}$, a refinement of this result was given by the author, asserting that ε in fact converges asymptotically to a solution of the Airy equation (1.3).)

The estimate (4.13) implies the asymptotic $x(t) = c_+ t + o(t)$. It is not entirely clear what the nature of the $o(t)$ error is; one might naively expect to obtain a refined asymptotic of the form $x(t) = c_+ t + x_+ + o(1)$, but it turns out that by inverse scattering methods one can give an example in the $p = 2$ case in which one has the asymptotic $x(t) = t + \kappa\sqrt{\log t} + o(\sqrt{\log t})$ for some $\kappa > 0$.

We now sketch the ideas used to prove Theorem 4.2. The first step is to pass from (1.5), which is an equation describing the dynamics of u, to equations describing the dynamics of ε, $x(t)$, and $c(t)$. This can be done by algebraic manipulations[17]. Indeed, if one substitutes (4.10), (4.11) into (1.5), one eventually obtains the equation

$$\varepsilon_t + \varepsilon_{xxx} + (pR^{p-1}\varepsilon)_x = F_{x(t),c(t),x'(t),c'(t)} + N(\varepsilon, R) \tag{4.14}$$

for ε where the forcing term $F_{x(t),c(t),x'(t),c'(t)}$ (caused by changes in the modulation parameters) is the explicit smooth function

$$F_{x(t),c(t),x'(t),c'(t)} := -\frac{c'(t)}{c(t)}(\frac{2R}{p-1} + (x - x(t))R_x) + (x'(t) - c(t))R_x \tag{4.15}$$

and $N(\varepsilon, R)$ (caused by self-interactions of the radiation term ε) is the nonlinearity

$$N(\varepsilon, R) := ((R+\varepsilon)^p - R^p - pR^{p-1}\varepsilon)_x.$$

As for the evolution of $x(t)$ and $c(t)$, one can differentiate (4.12) to obtain a 2×2 linear system of equations (known as the *modulation equations*) expressing the evolution $x'(t)$ and $c'(t)$ of the modulation parameters in terms of various integrals involving R, ε and its derivatives. The exact form of these modulation equations is not important for our purposes; the only thing which matters is the type of control that one gets on $x'(t)$ and $c'(t)$. By comparison with the soliton solutions (4.4) one expects $x'(t)$ to be close to 1, and $c'(t)$ to be close to 0. For most choices of orthogonality conditions, the degree of this closeness will only be linear in ε. But if one uses the specific orthogonality conditions (4.12), it turns out that there are particular cancellations which allow the error here to be *quadratic* in ε, at least as regards the variation of the scale parameter $c(t)$. Indeed, one can show after some

[17] As always, we ignore the analytic issues of how to justify all the formal computations in the case when u is low regularity; this can be done by standard (and boring) regularisation and limiting arguments.

computation (exploiting the exponential decay of R and its derivatives away from $x(t)$) that

$$|c'(t)| + |x'(t) - 1|^2 \lesssim \int_{\mathbf{R}} \varepsilon^2(t,x) e^{-|x-x(t)|}\, dx. \tag{4.16}$$

This is a rather strong estimate; it asserts that the error term ε only has a linear influence on the velocity, and a quadratic influence on the change in scale, and only when a significant portion of the mass of ε is stationed near the soliton. These bounds are particularly useful in controlling the size of the forcing term $F_{x(t),c(t),x'(t),c'(t)}$.

The right-hand side of (4.14) now consists primarily of terms which behave quadratically or higher in ε. This raises the hope that one can use perturbation theory to approximate the evolution here by that of the linearised equation $\varepsilon_t + \varepsilon_{xxx} + (pR^{p-1}\varepsilon)_x = 0$. (There is still one term, namely the drift term $(x'(t) - c(t))R_x$in (4.15), in the right-hand side which exhibits linear behaviour, but this term only causes a translation in ε and is thus be manageable.) To do this, we need[18] to somehow exploit the fact that the linearised equation is trying to propagate ε to the left, while the soliton is moving to the right.

One particularly elegant way to achieve this is via *virial identities.* Let us motivate these identities in the simple model case of the Airy equation (1.3). This equation has a conserved mass, indeed one quickly computes using (1.3) and integration by parts that

$$\partial_t \int_{\mathbf{R}} u^2 = \int_{\mathbf{R}} 2uu_t = -\int_{\mathbf{R}} 2uu_{xxx} = 0.$$

To affirm the intuition that the mass of u should be propagating leftward, let us now introduce the virial quantity $\int_{\mathbf{R}} xu^2$, which one can think of as the mean position of u. We compute

$$\begin{aligned}
\partial_t \int_{\mathbf{R}} xu^2 &= \int_{\mathbf{R}} 2xuu_t \\
&= -\int_{\mathbf{R}} 2xuu_{xxx} \\
&= \int_{\mathbf{R}} 2uu_{xx} + 2xu_x u_{xx} \\
&= -\int_{\mathbf{R}} 3u_x^2.
\end{aligned}$$

In particular, we see that $\int_{\mathbf{R}} xu^2$ is a decreasing function of time, which is a quantitative realisation of the intuition of leftwards propagation. If we instead replace x by $x - x(t)$, we get even faster decay:

$$\partial_t \int_{\mathbf{R}} (x - x(t))u^2 = -\int_{\mathbf{R}} 3u_x^2 - x'(t) \int_{\mathbf{R}} u^2.$$

In particular, if $x'(t) \gtrsim 1$ (which is the situation we are in above), we have

$$\partial_t \int_{\mathbf{R}} (x - x(t))u^2 \leq -c\|u\|_{H^1(\mathbf{R})}^2$$

[18] If we do not exploit this fact, then our control on the dispersive effects of the linearised equation is too weak; we can only hope to obtain decay of $O(t^{-1/3})$ on ε at best, which is insufficient to allow us to neglect the quadratic nonlinearity terms.

for some $c > 0$.

It turns out that one can do the same sort of thing for (4.14). Indeed, one can show (after lengthy computations) that (formally, at least) we have

$$\partial_t \int_{\mathbf{R}} (x - x(t))\varepsilon^2 \leq -c\|\varepsilon\|^2_{H^1(\mathbf{R})} \tag{4.17}$$

for some $c > 0$. This estimate strongly suggests that ε will move to the left of the soliton over time. Unfortunately, this "global" virial identity cannot be used directly in the above analysis, because the integral on the left-hand side may be divergent due to lack of spatial decay on ε. However, this can be rectified by the usual trick of localising the weight $x - x(t)$. Indeed, one can show that for sufficiently large $A > 1$, we have the "local" virial identity

$$\partial_t \int_{\mathbf{R}} \Psi_A(x - x(t))\varepsilon^2 \leq -c \int_{\mathbf{R}} (\varepsilon^2 + \varepsilon_x^2) e^{-|x-x(t)|/A}$$

for some bounded increasing function $\Psi(x - x(t))$ which equals $x - x(t)$ for $|x - x(t)| \leq A$, and is of magnitude $O(A)$ throughout. Thus the quantity $\int_{\mathbf{R}} \Psi_A(x - x(t))\varepsilon^2$ is monotone decreasing, while also being controlled by A times the mass. If we then integrate this in time we obtain an important spacetime bound

$$\int_{\mathbf{R}} \int_{\mathbf{R}} (\varepsilon^2 + \varepsilon_x^2) e^{-|x-x(t)|/A}\ dxdt \lesssim A\sigma.$$

This is the first indication of dispersion away from the soliton; it asserts that the radiation term ε cannot linger near the soliton $x(t)$ for extended periods of time.

This estimate, combined with (4.16), is already enough to demonstrate convergence of the scale parameter $c(t)$ to an asymptotic limit c_+. It also shows that the forcing term in (4.14) decays quite quickly in time; in particular, the quadratic nature of the nonlinearity shows that it decays integrably in time, with the exception of the drift term $(x'(t) - c(t))R_x$ which can be dealt with by hand. Because of this, it is possible to use energy estimates to conclude the full strength of Theorem 4.2.

We now describe an alternate approach to asymptotic stability, also due to Martel and Merle, which uses a more sophisticated and general strategy which has since shown to be useful for many other equations, including critical equations. The basic strategy is to use the *compactness-and-contradiction method*, which we informally summarise as follows.

(1) Suppose we wish to show some asymptotic property of a solution $u(t)$ as $t \to +\infty$. We assume for contradiction that this property does not occur.

(2) By using weak compactness, we then extract a sequence t_n of times going to infinity in which (suitably normalised versions) of the state $u(t_n)$ are weakly convergent in some sense, but which violate the property in some quantitative manner. In particular, (suitably normalised versions of) $u(t + t_n)$ should converge weakly to some asymptotic solution $u_\infty(t)$ of the original equation (now defined for all times $t \in \mathbf{R}$), which continues to violate the desired property.

(3) By using the dispersive properties of the equation, show that the asymptotic solution $u_\infty(t)$ obeys some *strong* compactness properties, or equivalently that the evolution $t \mapsto u_\infty(t)$ is *almost periodic* in some strong topology. (At this point u_∞ is behaving somewhat like the dispersive analogue of a solution to an elliptic PDE or variational problem.)

(4) Using more dispersive properties of the equation, upgrade the strong compactness to obtain further regularity and decay of the solution. (This step is roughly analogous to the exploitation of *elliptic regularity* in the theory of elliptic PDE.)

(5) Establish a *Liouville theorem* or *rigidity theorem*, that the only solutions close to solitons which exhibit strong compactness, regularity, and decay properties are the solitons itself. This is the most difficult step, and often requires full use of the conservation laws and monotonicity formulae of the equation. (This is analogous to Liouville theorems in elliptic PDE, the most famous of which is the assertion that the only bounded holomorphic or harmonic functions on $\mathbf{C}$ or $\mathbf{R}^d$ are the constants.)

(6) We conclude that u_∞ is itself a soliton, which we then combine with the fact that it violates the required property to obtain a contradiction.

The compactness-and-contradiction method is extremely powerful in analysing many nonlinear parabolic and dispersive equations, for instance a variant of this method for Ricci flow also plays a crucial role in Perelman's recent proof of the Poincaré conjecture. Another variant of this method is also very useful in establishing large data global well-posedness results for critical equations, though we will not discuss this topic further here. The one drawback of the method is that, by being indirect and relying so strongly on compactness methods, it does not easily provide any sort of quantitative bound in its conclusions, in contrast to the previous arguments used to prove Theorem 4.2, which were direct and easily provide explicit bounds.

Let us now sketch how this method is applied to give a new proof of Theorem 4.2. Actually we will just prove the slightly weaker claim that the translated radiation terms $\varepsilon(t, x - x(t))$ converges weakly in $H^1(\mathbf{R})$ to zero as $t \to +\infty$; note this weak convergence implies for instance that $\varepsilon(t, x - x(t))$ converges locally uniformly to zero, and so the radiation term eventually vacates the neighbourhood of the soliton. One can upgrade this convergence to obtain results closer in strength in Theorem 4.2, but we will not do so here.

To prove this weak convergence claim, we use the compactness-and-contradiction method. Suppose for contradiction that $\varepsilon(t, x - x(t))$ does not converge weakly to $H^1(\mathbf{R})$ as $t \to +\infty$. Since ε is bounded in H^1, weak compactness then shows that there exists a sequence of times $t_n \to \infty$ such that $\varepsilon(t_n, x - x(t_n))$ converges weakly in H^1 to some non-trivial limit $\varepsilon_\infty(0, x)$; one can also assume that $c(t_n)$ converges to some limit c_+. Due to some weak continuity properties of the gKdV flow (which can be proven by harmonic analysis methods) one can then show that $u(t + t_n, x + x(t_n))$ converges weakly (and locally in time) to some limiting solution $u_\infty(t, x) = R_\infty(t, x) + \varepsilon_\infty(t, x)$, where R_∞ and ε_∞ obey similar estimates to R and ε, and R_∞ is defined using some modulation parameters $c_\infty(t)$ and $x_\infty(t)$.

The normalised radiation terms $\varepsilon_\infty(t, x - x(t))$ stay bounded in H^1. By the Rellich compactness theorem, this means that they are locally precompact in L^2, i.e. their restriction to any compact spatial interval I lies in a compact subset of $L^2(I)$. We now assert that these terms are in fact *globally* precompact in L^2. This is equivalent to asserting that for any $\delta > 0$, we must have some radius R such that we have very little mass on the left,

$$\int_{x < x(t) - R} |\varepsilon_\infty(t, x)|^2 \, dx < \delta \tag{4.18}$$

and very little mass on the right

$$\int_{x>x(t)+R} |\varepsilon_\infty(t,x)|^2\, dx < \delta. \tag{4.19}$$

We briefly sketch why one would expect these claims to be true. Suppose that (4.19) failed, then a non-zero portion of the mass of ε_∞ at some time would be far to the right of the soliton. Returning to the original solution, we see that there exist arbitrarily large times t for which a significant portion of the mass of $\varepsilon(t)$ is to the right of $x(t)$. Now we evolve backwards in time, back to time 0. Away from the soliton, mass has a tendency to move leftwards as one goes forwards in time, and thus rightwards as one goes backwards in time. One can make this precise (by using crude forms of the local virial identity alluded to before), and conclude that at time 0, a significant portion of the mass of $\varepsilon(0)$ is to the right of $x(t)$. But t can be arbitrarily large, and so $x(t)$ can be arbitrarily large also (recall that $x'(t)$ stays close to 1). This contradicts the monotone convergence theorem, and so (4.19).

The proof of (4.18) is similar; if (4.18) failed, then at some point a non-zero portion of the mass of ε_∞ lies far to the left of the soliton, and thus we have strictly less than $M(\varepsilon_\infty)$ of the mass of ε_∞ near or to the right of the soliton. By the above discussion, we see that we have strictly less than $M(\varepsilon_\infty)$ of the mass of $\varepsilon(t)$ near or to the right of $x(t)$ for a sequence of arbitrarily large times t. But by using local virial-type identities to control the propagation of mass, this loss of mass to the left is irreversible, and in fact we have strictly less than $M(\varepsilon_\infty)$ of the mass of $\varepsilon(t)$ near or to the right of $x(t)$ for *all* sufficiently large times t. But then it is not possible for $\varepsilon(t+t_n, x+x(t_n))$ to converge weakly to $\varepsilon_\infty(t,x)$, a contradiction.

It turns out that one can upgrade the bounds (4.18), (4.19) significantly, to obtain a pointwise uniform exponential decay estimate of the form

$$|\varepsilon_\infty(t,x)| \lesssim \|\varepsilon_\infty(t)\|_{H^1} e^{-c|x-x(t)|} \tag{4.20}$$

for some $c > 0$. This is established by a long-time analysis of (4.14) and exploits the fact that the fundamental solution to the Airy equation (1.3) decays exponentially fast in the rightwards direction. This uniformity estimate is crucial in what follows.

It then remains to establish the *Liouville theorem* that if ε_∞ is a sufficiently small (in H^1) solution to a nonlinear equation (4.14) with $\varepsilon_\infty(t, x - x(t))$ compact in L^2, then ε_∞ must vanish. To prove this, we first use another compactness-and-contradiction argument in order to eliminate the nonlinear terms in (4.14). If the claim failed, then we could find a sequence of solutions ε_n to (4.14) which converged to zero in H^1 norm as $n \to \infty$, and were each compact in L^2, but were non-zero. Normalising each ε_n by its H^1 norm, we see from from the uniform estimate (4.20) that the resulting sequence is still compact in L^2. Thus we can take a limit and obtain a nontrivial solution ε to the *linearised* equation

$$\varepsilon_t + \varepsilon_{xxx} + (pR^{p-1}\varepsilon)_x = \alpha(t)R_x$$

for some scalar quantity $\alpha(t)$, which stays compact in L^2 and bounded in H^1 (and obeys the orthogonality conditions (4.12)). The task is thus to show that there is no such solution other than the trivial solution $\varepsilon = 0$, which will establish the Liouville theorem and thus Theorem 4.2.

At this point, one can now use the global virial estimate (4.17), which is valid here due to the exponential decay of ε. If ε is non-trivial, it has an H^1 norm bounded away from zero, which in conjunction with the L^2 compactness shows that the right-hand side of (4.17) is negative and bounded away from zero. But this forces $\int_{\mathbf{R}}(x - x(t))\varepsilon^2$ to go to $-\infty$, which contradicts the exponential decay of ε. This finally finishes the argument.

This proof was significantly more complicated than the direct proof, but the underlying strategy is much more powerful: it uses compactness methods to strip away all the inessential portions of the dynamics, leaving a very smooth and localised solution to which global estimates can be applied. This dispenses with the need for any cutoffs in space or frequency, which can significantly complicate the analysis.

5. The critical case

In the full version of these lecture notes, we discuss the more difficult mass-critical case $p = 5$, in which the scale invariance now plays a much more delicate role. We will also discuss more recent developments regarding stability of multisolitons, and on collisions between solitons.

UCLA Department of Mathematics, Los Angeles, CA 90095-1596.
E-mail address: tao@@math.ucla.edu

CURRENT EVENTS BULLETIN
Previous speakers and titles

For PDF files of talks, and links to *Bulletin of the AMS* articles, see http://www.ams.org/ams/current-events-bulletin.html.

January 7, 2007 (New Orleans, Louisiana)

Robert Ghrist, University of Illinois, Urbana-Champaign
Barcodes: The persistent topology of data

Akshay Venkatesh, Courant Institute, New York University
Flows on the space of lattices: work of Einsiedler, Katok and Lindenstrauss

Izabella Laba, University of British Columbia
From harmonic analysis to arithmetic combinatorics

Barry Mazur, Harvard University
The structure of error terms in number theory and an introduction to the Sato-Tate Conjecture

January 14, 2006 (San Antonio, Texas)

Lauren Ancel Myers, University of Texas at Austin
Contact network epidemiology: Bond percolation applied to infectious disease prediction and control

Kannan Soundararajan, University of Michigan, Ann Arbor
Small gaps between prime numbers

Madhu Sudan, MIT
Probabilistically checkable proofs

Martin Golubitsky, University of Houston
Symmetry in neuroscience

January 7, 2005 (Atlanta, Georgia)

Bryna Kra, Northwestern University
The Green-Tao Theorem on primes in arithmetic progression:
A dynamical point of view

Robert McEliece, California Institute of Technology
Achieving the Shannon Limit: A progress report

Dusa McDuff, SUNY at Stony Brook
Floer theory and low dimensional topology

Jerrold Marsden, Shane Ross, California Institute of Technology
New methods in celestial mechanics and mission design

László Lovász, Microsoft Corporation
Graph minors and the proof of Wagner's Conjecture

January 9, 2004 (Phoenix, Arizona)

Margaret H. Wright, Courant Institute of Mathematical Sciences, New York University
The interior-point revolution in optimization: History, recent developments and lasting consequences

Thomas C. Hales, University of Pittsburgh
What is motivic integration?

Andrew Granville, Université de Montréal
It is easy to determine whether or not a given integer is prime

John W. Morgan, Columbia University
Perelman's recent work on the classification of 3-manifolds

January 17, 2003 (Baltimore, Maryland)

Michael J. Hopkins, MIT
Homotopy theory of schemes

Ingrid Daubechies, Princeton University
Sublinear algorithms for sparse approximations with excellent odds

Edward Frenkel, University of California, Berkeley
Recent advances in the Langlands Program

Daniel Tataru, University of California, Berkeley
The wave maps equation

2008 CURRENT EVENTS BULLETIN *Committee*

Alejandro Adem, *University of British Columbia*

David Aldous, *University of California, Berkeley*

Peter Bates, *Michigan State University*

David Eisenbud, *University of California, Berkeley,* Chair

Daniel Freed, *University of Texas, Austin*

Edward Frenkel, *University of California, Berkeley*

Susan Friedlander, *University of Illinois, Chicago*

Robert Ghrist, *University of Illinois, Urbana-Champaign*

Martin Golubitsky, *University of Houston*

Andrew Granville, *University of Montreal*

Helmut Hoffer, *Courant Institute, New York University*

Carlos Kenig, *University of Chicago*

Izabella Laba, *University of British Columbia*

László Lovász, *Eotvos Lorand University*

Robert MacPherson, *Institute for Advanced Study*

Barry Mazur, *Harvard University*

Dusa McDuff, *State University of New York, Stony Brook*

Lauren Ancel Meyers, *University of Texas, Austin*

Kathleen O'Hara, *Mathematical Sciences Research Institute*

Michael Singer, *North Carolina State University*

Kannan Soundararajan, *Stanford University*

Madhu Sudan, *Massachusetts Institute of Technology*

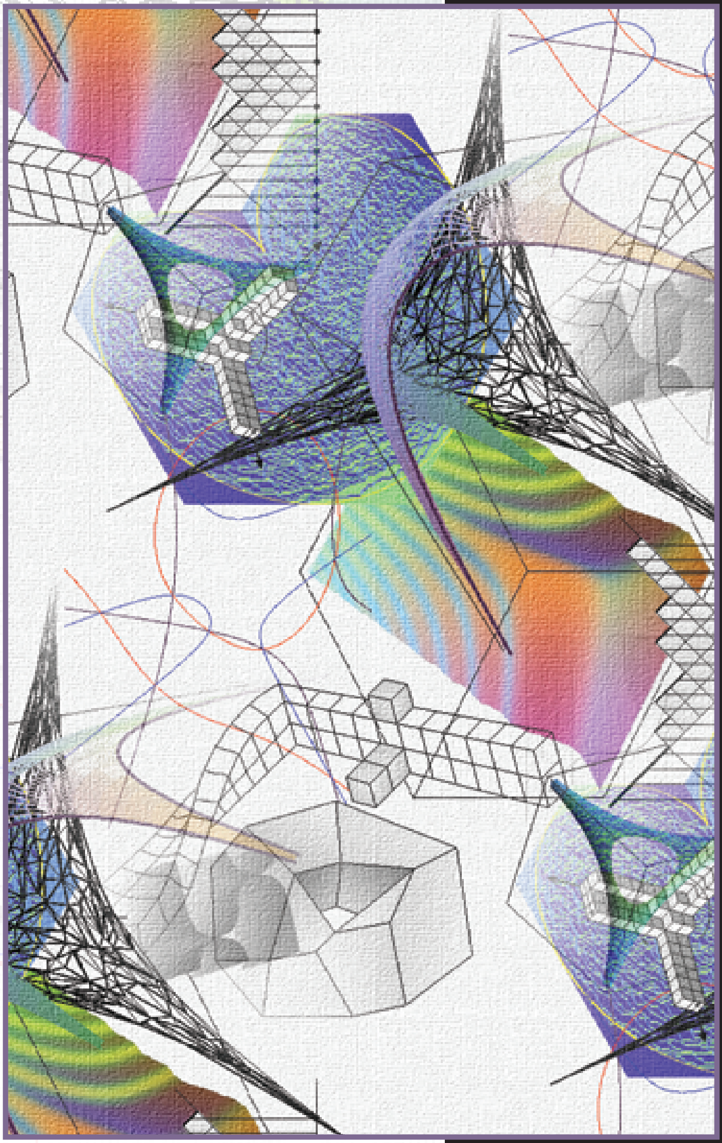
Ashkay Venkatesh, *Courant Institute, New York University*

Karen Vogtmann, *Cornell University*

Margaret Wright, *Courant Institute, New York University*

Andrei Zelevinsky, *Northeastern University*

Julius Zelmanowitz, *Mathematical Sciences Research Institute*



Cover photo associated with Uhlmann's talk courtesy of Mary Levin, UW Photography.

Cover graphic associated with Grassi's talk courtesy of Oliver Labs (www.AlgebraicSurface.net). The equation, "Barth's Sextic with 65 Nodes," is courtesy of Wolf Barth.

Cover graphic associated with Lawler's talk courtesy of Scott Sheffield, Courant Institute.

Cover graphic associated with Tao's talk courtesy of Development Centre for Ship Technology and Transport Systems, Duisburg, Germany.

The back cover graphic is reprinted courtesy of Andrei Okounkov.