AMERICAN MATHEMATICAL SOCIETY

CURRENT EVENTS BULLETIN

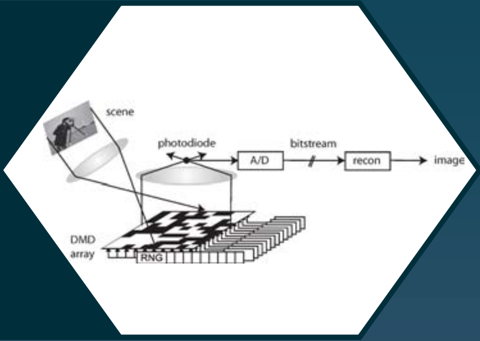
Wednesday, January 7, 2009, 1:00 PM to 5:00 PM
Joint Mathematics Meetings, Washington DC

Organized by David Eisenbud, University of California, Berkeley

1:00 PM

Matthew James Emerton

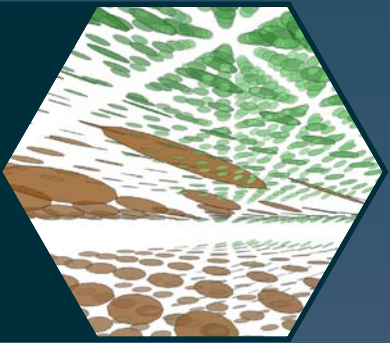
Topology, representation theory and arithmetic: Three-manifolds and the Langlands program



2:00 PM

Olga Holtz

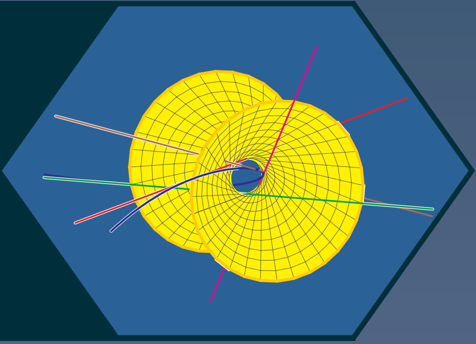
Compressive sensing: A paradigm shift in signal processing



3:00 PM

Michael Hutchings

From Seiberg-Witten theory to closed orbits of vector fields: Taubes's proof of the Weinstein conjecture



4:00 PM

Frank Sottile

Frontiers of reality in Schubert calculus

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.

TOPOLOGY, REPRESENTATION THEORY, AND ARITHMETIC: THREE-MANIFOLDS AND THE LANGLANDS PROGRAM

MATTHEW EMERTON

ABSTRACT. Using ideas from the Langlands program, F. Calegari and N. Dunfield have constructed a tower of finite covers whose members are closed hyperbolic rational homology 3-spheres, and whose injectivity radii grow without bound. The goal of this note is to sketch some of the ideas of the Langlands program, and to explain how they can be brought to bear on the study of hyperbolic 3-manifolds, and in particular, how they are applied in the construction of Calegari and Dunfield.

1. INTRODUCTION

Thurston has raised the following question regarding the topology of closed hyperbolic 3-manifolds (recall that *closed* means compact without boundary):

1.1. **Question.** If M is a closed connected hyperbolic 3-manifold, does M admit a finite cover whose first Betti number is positive?

To give this question some context, we recall some facts about the fundamental groups of closed hyperbolic 3-manifolds, as well as some terminology.

First, the facts: If M is a closed connected hyperbolic 3-manifold, then its fundamental group $\pi_1(M)$ is *infinite*, but *residually finite* (i.e. for any element $\gamma \in \pi_1(M) \setminus 1$, there exists a homomorphism $\pi_1(M) \to G$ with finite image such that the image of γ is non-trivial). Thus $\pi_1(M)$ admits many finite quotients, and correspondingly M admits many finite covers. Indeed, one can say something more precise: M admits finite covers of arbitrarily large *injectivity radius.*[1]

Next, the terminology: A closed connected orientable 3-manifold M is called a *rational homology sphere* if its first Betti number equals 0. By Poincaré duality, this implies that its second Betti number also equals 0, and thus that M has the same Betti numbers as the 3-sphere S^3. (Equivalently, the homology of M, computed with *rational* coefficients, coincides with that of S^3, whence the name.)

Returning now to Thurston's question we see that we can reformulate it in the following (negative) fashion:

1.2. **Question.** Can one find a closed hyperbolic 3-manifold that is a rational homology sphere, and all of whose finite covers are again rational homology spheres?

Received by the editors November 17, 2008.

2000 *Mathematics Subject Classification.* Primary .

The author was supported in part by NSF grant DMS-0701315.

[1] The injectivity radius of a closed Riemannian manifold N is one-half of the length of the shortest closed geodesic in N; thus it can be regarded as (one-half of) a minimal diameter of N.

In fact Thurston conjectured that the answer to his question in this form is *no*, or equivalently, that the answer to Question 1.1 is *yes.* This is the so-called *Virtual Positive Betti Number Conjecture* (the term virtual refers to the consideration of finite covers), which remains unproved at the moment. (We refer to [9] for a discussion of some of the literature and results related to this conjecture.)

Cooper raised the following question, related to Question 1.2 (see Problem 3.58 of [8]):

1.3. **Question.** Can one find a closed hyperbolic 3-manifold that is a rational homology sphere, and which admits finite covers of arbitrarily large injectivity radius that are again rational homology spheres?

Note that a negative answer to this question would yield a negative answer to Question 1.2 (since, as was recalled above, any closed hyperbolic manifold does admit covers of arbitrarily large injectivity radius), and hence a positive answer to Question 1.1, proving the Virtual Positive Betti Number Conjecture.

Unfortunately, the answer to Question 1.3 is in fact *yes*, and the goal of this note is to discuss a theorem of Frank Calegari and Nathan Dunfield to this effect [5, Thm. 1.4]:

1.4. **Theorem.** *There exists an infinite tower of finite covers*

$$\cdots \to M_n \to \cdots \to M_2 \to M_1 \to M_0,$$

each member of which is a closed hyperbolic rational homology sphere, and such that injectivity radius of M_n grows without bound as $n \to \infty$.

In one sense, this result was not a surprise: unlike Question 1.2, Cooper's Question 1.3 was actually expected to have a positive answer. What was surprising was the method of proof that Calegari and Dunfield gave of their Theorem 1.4: their proof relies on ideas from the Langlands program and the theory of Galois representations, topics that at first glance seem quite far removed from the topology of 3-manifolds.

We should note that Calegari and Dunfield's proof of Theorem 1.4 is contingent on certain other conjectures, not related to topology, but rather of an arithmetic nature. Namely, their proof relies on the Generalized Riemann Hypothesis, as well as on a particular case of Conjecture 1.7 below. Subsequently, Boston and Ellenberg [3] found an unconditional proof of Theorem 1.4. However, it is the original argument of Calegari and Dunfield that will be the focus of these notes, since it is this argument that exhibits a surprising link between the topology of 3-manifolds and questions of arithmetic.

In the remainder of this introduction, we give the briefest sketch of some of the ideas in the Langlands program, including a very rough statement of Langlands' reciprocity conjecture (Conjecture 1.7 below), which plays a key role in Calegari and Dunfield's proof of Theorem 1.4, before closing with an outline of the contents of the main body of this note.

1.5. **Langlands' reciprocity conjecture.** The Langlands program is an elaborate web of theorems and conjectures relating the representation theory of and harmonic analysis on certain Lie groups with arithmetic, and in particular, with representations of certain Galois groups. I will not try to make a precise statement of any of its tenets or conjectures here, but will content myself with briefest possible sketch of the ideas.

Recall that a complex number α is called an algebraic number if it is algebraic over $\mathbb{Q}$, i.e. if $f(\alpha) = 0$ for some polynomial $f(X) \in \mathbb{Q}[X]$. The set of all algebraic numbers forms a subfield $\overline{\mathbb{Q}} \subset \mathbb{C}$. Alternatively, one may define $\overline{\mathbb{Q}}$ to be the algebraic closure of $\mathbb{Q}$ in $\mathbb{C}$, or to the the union of all the finite subextensions F of $\mathbb{Q}$ in $\mathbb{C}$. Finite extensions F of $\mathbb{Q}$ are usually referred to as *number fields.* If F is any number field in $\mathbb{C}$, then $\overline{\mathbb{Q}}$ is also the algebraic closure of F in $\mathbb{C}$, and we may consider the so-called absolute Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$ of F; this is the group of all automorphisms of $\overline{\mathbb{Q}}$ that restrict to the identity on F. (It is in fact naturally a profinite group, rather than just a group, but we suppress this detail here.) The group $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$ is one of the characters in the Langlands story.

The other main characters in the story are automorphic eigenforms on reductive groups, the definition of which we now very briefly sketch. Suppose that $F \subset \overline{\mathbb{Q}}$ is a number field (i.e. of finite degree over $\mathbb{Q}$). Let G be a semi-simple or reductive linear algebraic group over F. (One can think of $\mathrm{GL}(n, F)$, although, as we will see below, there are other important examples too.) Let $G_{\mathbb{R}}$ denote the set of real points of G. (If $G = \mathrm{GL}(n, F)$ where $F = \mathbb{Q}(\alpha)$ is the number field obtained by adjoining the algebraic number α to $\mathbb{Q}$, and if the minimal polynomial of α has r_1 real roots and $2r_2$ complex roots, then $G_{\mathbb{R}} := \mathrm{GL}(n, \mathbb{R})^{r_1} \times \mathrm{GL}(n, \mathbb{C})^{r_2}$.) Let Γ be a *congruence subgroup* of the integer points of G. (See Subsections 4.2, 4.4, and 4.7 below for the definition of this notion in certain special cases; note in particular that Γ is then a discrete subgroup of $G_{\mathbb{R}}$.)

Consider the space $\mathcal{C}^\infty(\Gamma\backslash G_{\mathbb{R}})$ of smooth complex valued functions on the quotient $\Gamma\backslash G_{\mathbb{R}}$. This space is equipped with many commuting operators, namely the Casimir and higher Casimir operators (these are differential operators, analogous to a Laplacian, induced by the action of centre of the enveloping algebra of the Lie algebra of $G_{\mathbb{R}}$ on $\mathcal{C}^\infty(\Gamma\backslash G_{\mathbb{R}})$), and also the Hecke operators, which are indexed by (all but a finite number of) the prime ideals of the ring of integers of F.

1.6. **Definition.** An automorphic eigenform is a function $f \in \mathcal{C}^\infty(\Gamma\backslash G_{\mathbb{R}})$ which is a simultaneous eigenvector of all of the commuting operators discussed above, i.e. of all the Casimirs and Hecke operators, and which is slowly increasing at infinity.

We don't recall the precise definitions of the various operators alluded to above, or of the term "slowly increasing" as it is used in the above definition, referring the reader instead to the discussion of [2]. We will outline the definition of the Hecke operators in certain special cases in Subsection 4.9 below.

If f is an automorphic eigenform, then f determines a collection of eigenvalues, one eigenvalue for each of the Casimirs and each Hecke operator, which we will refer to as the *system of eigenvalues* attached to f.

Of importance for us is the following conjecture, which is a vaguely stated form of Langlands' Reciprocity Conjecture:

1.7. **Conjecture.** (a) *If $f \in \mathcal{C}^\infty(\Gamma\backslash G_{\mathbb{R}})$ is an automorphic eigenform, whose* (*appropriately normalized*) *eigenvalues under all the Casimir operators are integral, then the system of eigenvalues attached to f determines, and is in turn determined by, a certain representation of the Galois group* $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$.

(b) *Any representation of* $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$ *satisfying appropriate hypotheses is determined by an automorphic eigenform in the sense of part* (a).

This statement is overly simplified, and there are a myriad of details that we have omitted. Just to indicate some: (i) f should determine not just one representation, but in fact a whole family representations; we refer to subsection 5.1 below for some examples of such families. (ii) One can be precise about the group of matrices in which the values of the Galois representations associated to f should lie. In general, this requires a discussion of dual groups and L-groups (see e.g. [1]). In the case when $G = \mathrm{GL}(n, F)$, one can say that the associated Galois representations should be n-dimensional. (iii) There is an explicit description of the manner in which the system of eigenvalues attached to f, and the corresponding Galois representations, should determine each other. This description relies on, among other things, the Satake isomorphism (as explained e.g. in [6]). (iv) We have given no indication of what the "appropriate hypotheses" on a Galois representation in part (b) of the conjecture might be. For some examples of Galois representations satisfying these unspecified hypothesis, we again refer to Subsection 5.1.

To elaborate on all these points, and on the many others that we are omitting, would turn this brief note into a technical treatise, which is certainly not our goal. (In the case when $G = \mathrm{GL}(n, F)$ for a number field F, we refer to the article [12] of Taylor for a rather complete discussion.) Rather, we hope to give some indication as to how one can use Conjecture 1.7 to deduce concrete statements in mathematics, such as Theorem 1.4.

Let us close the present discussion by observing that (even in the vague form in which we have presented it), Conjecture 1.7 relates two kinds of objects, namely automorphic eigenforms and the Galois groups $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$, which *seem to have absolutely nothing to do with one another!* The first have to do with spectral theory on Lie groups, and the second to do with algebraic symmetries of algebraic numbers. It is the connection it provides between two totally different parts of mathematics that gives Conjecture 1.7 its force; it is also what makes the proof of even special cases of the conjecture so difficult. (See Subsection 5.4 below for a discussion of some of those special cases.)

1.8. **An outline of the paper.** In Section 2, we recall the basic facts about hyperbolic manifolds and hyperbolic spaces. In particular, we recall the connection between n-dimensional hyperbolic space and the Lie group $SO(n, 1)$. In the cases when $n = 2$ and 3, we recast this connection in terms of the more familiar groups $\mathrm{PSL}(2, \mathbb{R})$ and $\mathrm{PSL}(2, \mathbb{C})$.

In Section 3, we recall the basic facts concerning homology and cohomology of manifolds. After first considering the case of arbitrary dimension we then specialize the discussion to the cases of dimensions 2 and 3.

In Section 4, we introduce the notion of congruence quotients of $\mathbb{H}^2$ and $\mathbb{H}^3$. In particular, we define a particular tower of congruence quotients of $\mathbb{H}^3$ which satisfies the requirements of Theorem 1.4. (See Theorem 4.8.) Finally, we give a brief indication of how cohomology classes on congruence quotients give rise to automorphic forms, and outline the definition of the Hecke operators (in the context of cohomology).

In Section 5, we discuss the Langlands reciprocity conjecture for 2-dimensional Galois representations in more detail. We begin by giving some examples of 2-dimensional Galois representations, namely, those that arise from elliptic curves. We then describe the reciprocity conjecture for automorphic eigenforms and 2-dimensional Galois representations associated to $\mathbb{Q}$ and to quadratic imaginary number fields. Next, we very briefly sketch how a special case of the reciprocity conjecture was used by Andrew Wiles to prove Fermat's Last Theorem. We then explain how the reciprocity conjecture can be used to deduce Theorem 1.4. Finally, we explain how the reciprocity conjecture implies a positive answer to the virtual positive Betti number conjecture for arithmetic closed hyperbolic 3-manifolds.

These examples illustrate the different ways in which the reciprocity conjecture can be applied: in the proof of Fermat's Last Theorem, one uses the non-existence of certain kinds of automorphic forms to establish the non-existence of certain kinds of Galois representations, from which one in turn deduces Fermat's Last Theorem. By contrast, in the proof of Theorem 1.4, as we will see, one uses the non-existence of certain Galois representations to deduce the non-existence of certain automorphic forms, from which one in turn deduces that certain hyperbolic 3-manifolds are rational homology spheres. Finally, in the proof of the virtual positive Betti number conjecture in the arithmetic case, one uses the existence of certain Galois representations to force the existence of certain automorphic forms.

2. Hyperbolic Manifolds

In this section we discuss some of the basics of the theory of hyperbolic manifolds. We begin with a discussion in the general setting of n-dimensional hyperbolic manifolds, before specializing to the case of surfaces and 3-manifolds. We focus on explaining the connection with the Lie groups $\mathrm{SO}(n,1)$ (see in particular Subsection 2.4 below), since this gives the first indication that ideas from the Langlands program could be applied to the study of hyperbolic manifolds.

2.1. **Hyperbolic manifolds as quotients of hyperbolic space.** Let M be a connected complete n-dimensional smooth Riemannian manifold (with $n \geq 2$). We say that M is *hyperbolic* if M all the sectional curvatures of M are constant and negative. Rescaling the metric of M if necessary, we may and do assume that the sectional curvatures of M are then in fact all equal to -1.

The Riemannian metric on M pulls back to a Riemannian metric on the universal cover $\tilde{M}$ of M, which thus becomes a complete n-dimensional simply connected hyperbolic manifold. Such a manifold is unique, up to isometry, and we denote it by $\mathbb{H}^n$. It is referred to as *hyperbolic n-space.*

The original manifold M may be obtained as the quotient $\Gamma\backslash\mathbb{H}^n$ for a certain group of deck-transformations Γ acting on $\mathbb{H}^n$. Since the metric on $\mathbb{H}^n$ is simply pulled back from M, the group Γ is a group of isometries of $\mathbb{H}^n$, and it thus a discrete subgroup of the group $\mathrm{Isom}(\mathbb{H}^n)$ of all isometries of $\mathbb{H}^n$. If M is furthermore orientable, then Γ lies in the index 2 subgroup $\mathrm{Isom}^0(\mathbb{H}^n)$ of $\mathrm{Isom}(\mathbb{H}^n)$ consisting of orientation-preserving isometries.

2.2. **A quadric model of $\mathbb{H}^n$.** Hyperbolic n-space admits many models; in this subsection we describe one of them.

Let Q denote the quadratic form $x_1^2 + \cdots x_n^2 - x_{n+1}^2$ on $\mathbb{R}^{n+1}$. This quadratic forms induces a corresponding pseudo-Riemannian metric

$$g := dx_1^2 + \cdots dx_n^2 - dx_{n+1}^2$$

on $\mathbb{R}^{n+1}$. Let X denote the quadric

$$X := \{x \in \mathbb{R}^{n+1} \,|\, Q(x) = -1\} \subset \mathbb{R}^{n+1},$$

and let $X^+ := \{x \in X \,|\, x_{n+1} > 0\} \subset X$. (The quadric X is the union of two connected components, of which X^+ is one; the other is the subset X^- of X consisting of points for which $x_{n+1} < 0$.) If $g_{|X^+}$ denotes the restriction of g to

X^+, then $g_{|X^+}$ is a true Riemannian metric on X^+ (i.e. it is positive definite). Furthermore, the sectional are curvatures of X^+ are constant and negative. Since X^+ is simply connected (indeed, it is homeomorphic to $\mathbb{R}^n$), we find that X^+ provides a model for $\mathbb{H}^n$.

2.3. **A Lie theoretic description of $\mathbb{H}^n$.** Let $\mathrm{O}(n,1)$ denote the subgroup of $\mathrm{GL}(n+1,\mathbb{R})$ which preserves the quadratic form Q. The group $\mathrm{O}(n,1)$ then clearly preserves the quadric X, and one easily sees that it acts transitively on X.

If x is a point of X, and $\mathrm{O}(n,1)_x$ denotes the stabilizer of x in $\mathrm{O}(n,1)$, then $\mathrm{O}(n,1)_x$ acts faithfully by orthogonal transformations on the tangent hyperplane to X at x (orthogonal with respect to the positive definite quadratic form given by the metric $g_{|X}$), and in this manner one obtains an isomorphism $\mathrm{O}(n,1)_x \xrightarrow{\sim} \mathrm{O}(n)$, where $\mathrm{O}(n)$ denotes the usual orthogonal group of $n \times n$ matrices that preserve a positive definite n-dimensional quadratic form. Thus there is an identification

$$\mathrm{O}(n,1)/\,\mathrm{O}(n) \xrightarrow{\sim} X.$$

If we let $\mathrm{O}(n,1)^+$ denote the index 2 subgroup of $\mathrm{O}(n,1)$ consisting of transformations which take X^+ to itself, then we obtain a corresponding identification

$$\mathrm{O}(n,1)^+/\,\mathrm{O}(n) \xrightarrow{\sim} X^+ = \mathbb{H}^n, \tag{2.1}$$

and we also obtain an identification

$$\mathrm{O}(n,1)^+ \xrightarrow{\sim} \mathrm{Isom}(X^+) = \mathrm{Isom}(\mathbb{H}^n).$$

Any matrix in $\mathrm{O}(n,1)$ has determinant equal to ± 1. If we let $\mathrm{SO}(n,1)^+$ denote the subgroup of $\mathrm{O}(n,1)^+$ consisting of matrices of determinant 1, then $\mathrm{SO}(n,1)^+$ is identified with the index 2 subgroup $\mathrm{Isom}^0(\mathbb{H}^n)$ of orientation preserving isometries in $\mathrm{Isom}(\mathbb{H}^n)$. The identification (2.1) induces an identification

$$\mathrm{SO}(n,1)^+/\,\mathrm{SO}(n) \xrightarrow{\sim} X^+ = \mathbb{H}^n.$$

The group $\mathrm{SO}(n,1)^+$ is a connected semi-simple Lie group, and $\mathrm{SO}(n)$ is a maximal compact subgroup of $\mathrm{SO}(n,1)^+$. In general, the quotient of a connected semi-simple Lie group G by its maximal compact subgroup is referred to as the symmetric space associated to G. Thus $\mathbb{H}^n$ is the symmetric space associated to $\mathrm{SO}(n,1)^+$.

We can thus summarize the discussion of this section as follows: the group $\mathrm{Isom}^0(\mathbb{H}^n)$ of orientation-preserving isometries of $\mathbb{H}^n$ is isomorphic to the connected semi-simple Lie group $\mathrm{SO}(n,1)^+$, and $\mathbb{H}^n$ may be identified with the symmetric space associated to $\mathrm{SO}(n,1)^+$.

2.4. **Hyperbolic manifolds and discrete subgroups of $\mathrm{SO}(n,1)^+$.** If we combine the discussion of subsections 2.1 and 2.3, we find that any orientable complete hyperbolic n-manifold M may be written as a quotient

$$M \xrightarrow{\sim} \Gamma\backslash\,\mathrm{SO}(n,1)/SO(n),$$

where Γ is a discrete torsion-free subgroup of $\mathrm{SO}(n,1)$. (The torsion-free condition ensures that Γ acts properly discontinuously on $\mathrm{SO}(n,1)/SO(n) = \mathbb{H}^n$.)

2.5. **The upper half-space model for $\mathbb{H}^n$.** Another model for $\mathbb{H}^n$ is the n-dimensional upper half-space

$$\mathcal{H}^n := \{(x_1, \ldots, x_n) \in \mathbb{R}^n \,|\, x_n > 0\},$$

equipped with the metric $ds^2 := \dfrac{dx_1^2 + \ldots dx_n^2}{x_n^2}$. One computes that this is a metric of constant negative curvature, and so does indeed realize $\mathcal{H}^n$ as a model of $\mathbb{H}^n$. We will discuss this model in more detail in the two cases $n = 2$ and 3.

2.6. **The upper half-plane model of $\mathbb{H}^2$.** The group $\mathrm{SO}(2,1)^+$ is more familiar than it might seem. Indeed, there is an isomorphism[2]

$$\mathrm{PSL}(2,\mathbb{R}) := \mathrm{SL}(2,\mathbb{R})/\{\pm 1\} \xrightarrow{\sim} \mathrm{SO}(2,1)^+.$$

Thus there is also an isomorphism

$$\mathrm{PSL}(2,\mathbb{R}) \xrightarrow{\sim} \mathrm{Isom}^0(\mathbb{H}^2). \tag{2.2}$$

As we now explain, the upper half-plane model of $\mathbb{H}^2$ makes this isomorphism apparent.

Let us rewrite the upper half-plane $\mathcal{H}^2$ in the form

$$\mathcal{H}^2 := \{z = x + iy \in \mathbb{C} \,|\, y > 0\}$$

(i.e. we think of it as being the *complex upper half-plane*); the metric on $\mathcal{H}^2$ then becomes

$$ds^2 := \frac{dx^2 + dy^2}{y^2}.$$

The group $\mathrm{PSL}(2,\mathbb{R})$ is now seen to act on $\mathcal{H}^2$ as follows: an element $\gamma \in \mathrm{PSL}(2,\mathbb{R})$, represented by a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{R})$, acts on $\mathcal{H}^2$ via the formula

$$\gamma \cdot z := \frac{az + b}{cz + d}. \tag{2.3}$$

One easily computes that this action preserves the metric ds^2, and thus we obtain a concrete description of the isomorphism (2.2).

2.7. **The upper half-space model of $\mathbb{H}^3$.** The group $\mathrm{SO}(3,1)^+$ is also more familiar than it might seem. Indeed, there is an isomorphism[3]

$$\mathrm{PSL}(2,\mathbb{C}) := \mathrm{SL}(2,\mathbb{C})/\{\pm 1\} \xrightarrow{\sim} \mathrm{SO}(3,1)^+.$$

[2] This isomorphism can be understood conceptually as follows: $\mathrm{SL}(2,\mathbb{R})$ acts on its Lie algebra $\mathfrak{sl}_2$ (the space of 2×2-matrices of trace zero) via conjugation (the so-called adjoint action). This action factors through $\mathrm{PSL}(2,\mathbb{R})$, and preserves the quadratic form $\langle X, Y \rangle := \mathrm{Trace}(XY)$, which has signature $(2,1)$.

[3] Here is a theoretical description of this isomorphism: let V denote the two-dimensional complex vector space $\mathbb{C}^2$ equipped with the standard representation of $\mathrm{SL}(2,\mathbb{C})$, and let $\overline{V}$ denote $\mathbb{C}^2$ equipped with the complex conjugate action of $\mathrm{SL}(2,\mathbb{C})$. The tensor product $W := V \otimes_{\mathbb{C}} \overline{V}$ is then a 4-dimensional representation of $\mathrm{SL}(2,\mathbb{C})$. The subgroup $\{\pm 1\}$ acts trivially on W, and so W is in fact a representation of $\mathrm{PSL}(2,\mathbb{C})$. Furthermore, the character of W is real valued, and in fact W descends to a representation $W_{\mathbb{R}}$ of $\mathrm{PSL}(2,\mathbb{C})$ on a 4-dimensional real vector space. One computes that in the space of quadratic forms on $W_{\mathbb{R}}$, there is a unique line that is invariant under $\mathrm{PSL}(2,\mathbb{C})$, and that the non-zero quadratic forms in this line have signature $(3,1)$ and $(1,3)$. Thus $\mathrm{PSL}(2,\mathbb{C})$ preserves a quadratic form of signature $(3,1)$ on $W_{\mathbb{R}}$; this yields the stated isomorphism.

This isomorphism then induces an isomorphism

$$\mathrm{PSL}(2,\mathbb{R}) \xrightarrow{\sim} \mathrm{Isom}^0(\mathbb{H}^2). \tag{2.4}$$

As we now explain, the upper half-space model of $\mathbb{H}^3$ makes this isomorphism apparent.

Let us rewrite 3-dimensional upper half-space in the form

$$\mathcal{H}^3 := \{(z,t) \in \mathbb{C} \times \mathbb{R} \,|\, t > 0\}.$$

We can then identify the Riemann sphere $\mathbb{C}P^1 := \mathbb{C} \cup \{\infty\}$ with the "sphere at infinity" of $\mathcal{H}^3$, and the action of $\mathrm{Isom}^0(\mathbb{H}^3) = \mathrm{Isom}^0(\mathcal{H}^3)$ on $\mathcal{H}^3$ induces a corresponding action by conformal transformations on $\mathbb{C}P^1$. On the the other hand, from complex analysis we know that the group of conformal transformations of $\mathbb{C}P^1$ is identified with $\mathrm{PSL}(2,\mathbb{C})$, acting by linear fractional transformations (via the formula (2.3)). Thus we obtain a homomorphism $\mathrm{Isom}^0(\mathbb{H}^3) \to \mathrm{PSL}_2(\mathbb{C})$, which is in fact the inverse of the isomorphism (2.4).

3. The homology and cohomology of closed manifolds

As its title indicates, in this section, we describe some of the basic facts about the algebraic topology of closed manifolds.

3.1. **Homology.** If M is a closed connected n-manifold, then we can compute its homology groups $H_i(M,A)$ with coefficients in any abelian group A via singular chains. We will primarily be interested in the cases when A is one of the fields $\mathbb{Q}$ or $\mathbb{C}$. In this case the homology groups are actually $\mathbb{Q}$- or $\mathbb{C}$-vector spaces, and the universal coefficients theorem provides an isomorphism

$$\mathbb{C} \otimes_{\mathbb{Q}} H_i(M,\mathbb{Q}) \xrightarrow{\sim} H_i(M,\mathbb{C});$$

in particular, the dimension of H_i is independent of whether we use $\mathbb{Q}$ and $\mathbb{C}$; it is referred to as the ith Betti number of M, and denoted b_i. It is always finite.

If M is orientable, then Poincaré duality shows that $H_i(M,\mathbb{Q})$ and $H_{n-i}(M,\mathbb{Q})$ are naturally dual vector spaces (and similarly with $\mathbb{Q}$ replaced by $\mathbb{C}$). In particular $b_i = b_{n-i}$. Since there is no homology in negative degrees, we see that $b_i = 0$ if $i < 0$, and thus also $b_i = 0$ if $i > n$. Since M is connected, $b_0 = 1$. Thus $b_n = 1$ also.

3.2. **Definition.** The Euler characteristic of M is defined to be

$$\chi(M) := \sum_{i \geq 0} (-1)^i b_i.$$

Since b_i vanishes if $i > n$, this sum is well-defined.

3.3. *Remark.* If n is odd and M is orientable, then it follows from Poincaré duality (or more precisely, the formula $b_i = b_{n-i}$ for all i) that $\chi(M) = 0$.

If M is triangulable, i.e. is homeomorphic to a simplicial complex, then we may also compute homology simplicially, in terms of some given triangulation. The Euler-Poincaré formula then shows that we may also compute $\chi(M)$ as the alternating sum, for $i \geq 0$, of the number of simplices of dimension i appearing in some given triangulation of M.

We recall that closed surfaces and closed 3-manifolds are always triangulable.

The following lemma describes the behaviour of Euler characteristics under finite covering maps.

3.4. **Lemma.** *If $N \to M$ is a finite covering map of connected closed manifolds, of degree d say, then $\chi(N) = d\,\chi(M)$.*

Sketch of proof. We sketch a proof in the case when M is triangulable. (Since closed surfaces and closed 3-manifolds are always triangulable, this argument establishes the lemma in dimensions $n = 2$ or 3.) Fix a triangulation of M and use this triangulation to compute $\chi(M)$. Then, pull back this triangulation to N via the covering map and use this pulled-back triangulation to compute $\chi(N)$. Since the covering map has degree d, every simplex in the triangulation of M pulls back to d simplices in N. The formula follows. □

While the preceding lemma gives excellent control over the behaviour of the Euler characteristic in finite covers, it is significantly more difficult in general to control the behaviour of individual Betti numbers, as we will see.

3.5. **Cohomology.** If M is a closed connected n-manifold, then we can compute its cohomology group $H^i(M, A)$ with coefficients in any abelian group A via singular cochains. We will primarily be interested in the cases when A is one of the fields $\mathbb{Q}$ or $\mathbb{C}$. In this case the ith cohomology group is actually a $\mathbb{Q}$- or $\mathbb{C}$-vector space, and is naturally dual to the ith homology group. In particular, it has the same dimension, namely the ith Betti number b_i.

De Rham theory shows that we may also compute the complex cohomology spaces $H^i(M, \mathbb{C})$ using differential forms. If we fix a Riemannian metric g on M, then we may furthermore use Hodge theory to identify $H^i(M, \mathbb{C})$ with the space of harmonic i-forms on M. (The connection with analysis provided by de Rham theory and Hodge theory is one reason for considering cohomology as well as homology, even though from the point of view of the singular theory, they carry essentially the same information.)

If $N \to M$ is a finite cover, then pulling back cohomology classes induces a map $H^i(M, \mathbb{Q}) \to H^i(N, \mathbb{Q})$, which is in fact injective. (Indeed, if the cover is regular, with covering group G, then we can be more precise: by functoriality of cohomology, the group G acts on $H^i(N, \mathbb{Q})$, and $H^i(M, \mathbb{Q})$ is identified with the space of G-invariant elements in $H^i(N, \mathbb{Q})$.) This yields the following lemma:

3.6. **Lemma.** *If $N \to M$ is a finite covering map of closed manifolds, then the ith Betti number of N is greater than or equal to the ith Betti number of M.*

Note that this statement is much weaker than the very precise statement regarding Euler characteristics provided by Lemma 3.4. In the following subsections, we will discuss the extent to which we can improve it in the cases of surfaces and 3-manifolds.

3.7. **The topology of closed hyperbolic surfaces.** If M is a closed connected orientable surface, then $b_0 = b_2 = 1$, while $b_1 = 2g$, where g is the genus of M. Thus the Euler characteristic $\chi(M) = 2 - 2g$.

If M is equipped with a Riemannian metric with curvature K (a real-valued function on M, since M is a surface), then the *Gauss-Bonnet theorem* states that

$$\int_M K = 2\pi\,\chi(M).$$

In particular, if M is a hyperbolic manifold, so that $K \equiv -1$, then we find that

$$\chi(M) = -\frac{\operatorname{vol}(M)}{2\pi},$$

and so is negative; equivalently, the genus $g > 1$. (Conversely, if $\chi(M)$ is negative, then M admits a hyperbolic metric — indeed, a $(6g-6)$-dimensional moduli space of them.)

If N is a finite covering space of M, of degree d say, then Lemma 3.4 shows that $\chi(N) = d\,\chi(M)$. Consequently, if M has genus g, then N has genus $d(g-1)+1$. In particular, if M is hyperbolic, so that $g > 1$, then the genus of a degree d cover of M grows linearly with d (since $g > 1$), and hence so does the first Betti number. (Of course, b_0 and b_2 are both equal to 1 for any connected cover of M.)

3.8. **The topology of closed hyperbolic 3-manifolds.** If M is a closed connected orientable 3-manifold, then $b_0 = b_3 = 1$, while Poincaré duality shows that $b_1 = b_2$. In particular, as was noted in Remark 3.3, the Euler characteristic $\chi(M) = 0$. If N is any finite cover of M, then similarly $\chi(N) = 0$. Thus the formula of Lemma 3.4, while it is just as valid for for 3-manifolds as for surfaces, yields no information about the individual Betti numbers of N. Also, Lemma 3.6, while it shows that Betti numbers of N can't be less than those of M, does not imply that they must be greater than those of M.

Thus, unlike in the case of surfaces, it is not obvious that if one considers a tower of finite covers of M, then the first Betti number must increase as one moves up the tower. In fact, not only is is not obvious, it is not true! Indeed, Theorem 1.4 establishes the existence of an infinite tower of finite covers every member of which has $b_1 = 0$ (i.e. is a rational homology sphere).[4]

To prove Theorem 1.4, we need a tool that allows us to compute individual Betti numbers, at least for certain hyperbolic 3-manifolds. This tool will be provided by the Langlands reciprocity conjecture, applied in the context of arithmetic hyperbolic 3-manifolds.

4. Congruence quotients of $\mathbb{H}^2$ and $\mathbb{H}^3$

In this section we introduce the notion a congruence quotient of $\mathbb{H}^2$ or $\mathbb{H}^3$. The notion of a congruence quotient of $\mathbb{H}^n$ in fact make sense for any n, but we will focus on the cases $n = 2$ and 3, since then we can work with the more familiar groups $\mathrm{PSL}(2,\mathbb{R})$ and $\mathrm{PSL}(2,\mathbb{C})$, rather than the groups $\mathrm{SO}(n,1)$ (which are the groups we would have to deal with to study general values of n).

4.1. **First examples of discrete subgroups.** If we combine the discussions of Subsections 2.4, 2.6, and 2.7, we see that in order to find examples of hyperbolic surfaces or three folds, we have to find examples of discrete torsion-free subgroups of $\mathrm{PSL}(2,\mathbb{R})$ or $\mathrm{PSL}(3,\mathbb{C})$.

Some basic models for discrete objects inside continuous ones are provided by the inclusions $\mathbb{Z} \subset \mathbb{R}$ and $\mathbb{Z}[i] \subset \mathbb{C}$. (Here $\mathbb{Z}[i]$ denotes the ring of *Gaussian integers*, consisting of complex numbers whose real and imaginary parts are both

[4] We should point out that the existence of infinite towers of finite covers all of which are rational homology spheres was known prior to the work of Calegari and Dunfield in [5]; the new contribution of their work is to show that the tower can furthermore be chosen so that the injectivity radius grows without bound.

integers.) These immediately suggest examples of discrete subgroups in $\mathrm{PSL}(2,\mathbb{R})$ and $\mathrm{PSL}(2,\mathbb{C})$, namely $\mathrm{PSL}(2,\mathbb{Z})$ and $\mathrm{PSL}(2,\mathbb{Z}[i])$ respectively.

Unfortunately neither of these groups is torsion free, and so each of the quotients $\mathrm{PSL}(2,\mathbb{Z})\backslash\mathbb{H}^2$ and $\mathrm{PSL}(2,\mathbb{Z}[i])\backslash\mathbb{H}^3$ is an orbifold rather than a manifold. Also, neither of these quotients is compact.

4.2. **Congruence quotients of $\mathbb{H}^2$.** Although neither of the two groups $\mathrm{PSL}(2,\mathbb{Z})$ or $\mathrm{PSL}(2,\mathbb{Z}[i])$ is torsion-free, it is easy to construct closely related discrete subgroups which *are* torsion-free, the so-called *congruence subgroups.*

We begin by focusing on the $\mathrm{PSL}(2,\mathbb{Z})$ case. In fact it will be simpler to work with $\mathrm{SL}(2,\mathbb{Z})$, and we will do this from now on. Since $\mathrm{PSL}(2,\mathbb{Z})$ is a quotient of $\mathrm{SL}(2,\mathbb{Z})$, each of the subgroups we construct will have an associated image in $\mathrm{PSL}(2,\mathbb{Z})$, which will then be a discrete subgroup of $\mathrm{PSL}(2,\mathbb{R})$.

If n is any integer, then reduction modulo n induces a ring homomorphism $\mathbb{Z}\to\mathbb{Z}/n\mathbb{Z}$, and hence a homomorphism of groups

$$\mathrm{SL}(2,\mathbb{Z})\to\mathrm{SL}(2,\mathbb{Z}/n\mathbb{Z}). \tag{4.1}$$

The kernel of this map is denoted $\Gamma(n)$. It is referred to as the principal congruence subgroup of level n. It has finite index in $\mathrm{SL}(2,\mathbb{Z})$ (since $\mathrm{SL}(2,\mathbb{Z}/n\mathbb{Z})$ is finite), and is torsion-free if $n\geq 3$ (as its image in $\mathrm{PSL}(2,\mathbb{R})$). Thus we have produced an infinite family of discrete torsion-free subgroups of $\mathrm{PSL}(2,\mathbb{R})$, giving rise to an infinite family of hyperbolic surfaces. Note that if m divides n, then $\Gamma(m)\subset\Gamma(n)$ (with finite index), and so $\Gamma(m)\backslash\mathbb{H}^2$ is a finite cover of $\Gamma(n)\backslash\mathbb{H}^2$. Thus we also have lots of finite covering maps.

In fact, we will need to consider a slightly different family of congruence subgroups, usually denote $\Gamma_1(n)$, and defined as follows:

4.3. **Definition.** The group $\Gamma_1(n)\subset\mathrm{SL}(2,\mathbb{Z})$ is defined to be the preimage under (4.1) of the subgroup of upper triangular unipotent matrices in $\mathrm{SL}(2,\mathbb{Z}/n\mathbb{Z})$; i.e.

$$\Gamma_1(n):=\{\gamma\in\mathrm{SL}(2,\mathbb{Z}/n\mathbb{Z})\,|\,\gamma\equiv\begin{pmatrix}1&*\\0&1\end{pmatrix}\bmod N\}.$$

The groups $\Gamma_1(n)$ are torsion-free provided $n\geq 4$. Also, if m divides n, then $\Gamma_1(m)$ is a finite index subgroup of $\Gamma_1(n)$ Thus the quotients $\Gamma_1(n)\backslash\mathbb{H}^2$ give an infinite family of hyperbolic surfaces, with many finite covering maps between them.

In the number-theoretic literature, the group $\mathrm{SL}(2,\mathbb{Z})$ is referred to as the *modular group*, and the quotients $\Gamma(n)\backslash\mathbb{H}^2$ and $\Gamma_1(n)\backslash\mathbb{H}^2$ are called *modular curves.* In this note we will refer to them as *congruence quotients* (for the obvious reason: that they are quotients of $\mathbb{H}^2$ by congruence subgroups of $\mathrm{SL}(2,\mathbb{Z})$).

4.4. **Congruence quotients of $\mathbb{H}^3$.** One can define congruence subgroups of $\mathrm{PSL}(2,\mathbb{C})$ in an analogous manner to the case of $\mathrm{PSL}(2,\mathbb{R})$. Before doing so, we make some preliminary remarks.

The first remark is that the inclusion $\mathrm{SL}(2,\mathbb{C})\hookrightarrow\mathrm{GL}(2,\mathbb{C})$ induces an isomorphism

$$\mathrm{PSL}(2,\mathbb{C}):=\mathrm{SL}(2,\mathbb{C})/\{\pm1\}\xrightarrow{\sim}\mathrm{GL}(2,\mathbb{C})/\mathbb{C}^\times=:\mathrm{PGL}(2,\mathbb{C}).$$

(Here $\mathbb{C}^\times$, the multiplicative group of non-zero complex numbers, embeds into $\mathrm{GL}(2,\mathbb{C})$ as the subgroup of non-zero scalar matrices.) Thus, in order to construct discrete subgroups of $\mathrm{PSL}(2,\mathbb{C})$, it suffices to construct such subgroups of $\mathrm{GL}(2,\mathbb{C})$; we can then pass to their images in $\mathrm{PGL}(2,\mathbb{C})=\mathrm{PSL}(2,\mathbb{C})$. The reason for working with $\mathrm{GL}(2,\mathbb{C})$ rather than $\mathrm{SL}(2,\mathbb{C})$ at this point is dictated by the requirements of the paper [5] — it will allow us to apply the Langlands reciprocity conjecture in such a way as to construct a tower of rational homology spheres satisfying the requirements of Theorem 1.4.

The second remark is that $\mathbb{Z}[i]$ is not the only "discrete model" for the complex numbers that we have available for forming discrete subgroups of $\mathrm{GL}(2,\mathbb{C})$. If d is any positive square-free integer, then $F := \mathbb{Q}(\sqrt{-d})$ is a subfield of $\mathbb{C}$ with the property that $\mathbb{R}\otimes_{\mathbb{Q}} F \xrightarrow{\sim} \mathbb{C}$ (so F is to $\mathbb{C}$ as $\mathbb{Q}$ is to $\mathbb{R}$), and the ring of integers[5] $\mathcal{O}_F$ of F is a discrete subring of $\mathbb{C}$. (Such number fields F are called *quadratic imaginary.*)

For any choice of F, the group $\mathrm{GL}(2,\mathcal{O}_F)$ is a discrete subgroup of $\mathrm{GL}(2,\mathbb{C})$. If $\mathfrak{n}$ is a non-zero ideal in $\mathcal{O}_F$, then $\mathcal{O}_F/\mathfrak{n}$ is a finite ring, and we have the reduction map

$$\mathrm{GL}(2,\mathcal{O}_F)\to\mathrm{GL}(2,\mathcal{O}_F/\mathfrak{n}). \tag{4.2}$$

4.5. **Definition.** We define $\Gamma_1(\mathfrak{n})$ to be the subgroup of $\mathrm{GL}(2,\mathcal{O}_F)$ obtained as the preimage under (4.2) of the subgroup of upper triangular unipotent matrices in $\mathrm{GL}(2,\mathcal{O}_F/\mathfrak{n})$; i.e.

$$\Gamma_1(n) := \{\gamma\in\mathrm{GL}(2,\mathcal{O}_F/\mathfrak{n})\,|\,\gamma\equiv\begin{pmatrix}1 & *\\ 0 & 1\end{pmatrix} \bmod \mathfrak{n}\}.$$

Just as in the two-dimensional case, for a fixed choice of F, for all but finitely many ideal $\mathfrak{n}$, the group $\Gamma_1(\mathfrak{n})$ will have torsion-free image in $\mathrm{PSL}(2,\mathbb{C})$, and of course, if $\mathfrak{m}\subset\mathfrak{n}$ is an inclusion of non-zero ideals, then $\Gamma_1(\mathfrak{m})\subset\Gamma_1(\mathfrak{n})$. Thus we obtain an infinite collection of hyperbolic 3-manifolds, of the form $\Gamma_1(\mathfrak{n})\backslash\mathbb{H}^3$, with many finite covering maps between them. Just as in the surface case, we refer to these hyperbolic 3-manifolds as *congruence quotients* of $\mathbb{H}^3$.

4.6. **Noncompactness and unipotent elements.** The congruence quotients we have constructed so far, namely $\Gamma_1(n)\backslash\mathbb{H}^2$ and $\Gamma_1(\mathfrak{n})\backslash\mathbb{H}^3$, are important examples of hyperbolic manifolds, but they have the disadvantage, from the point of view of proving Theorem 1.4, that they are *not* compact (although they are of finite volume).

In the surface case, these quotients have ends that look like cylinders on a circle (i.e. are homeomorphic to $S^1\times\mathbb{R}$), while in the 3-manifold case they have ends that look like cylinders on a torus (i.e. are homeomorphic $S^1\times S^1\times R$).

The reason that the quotients $\Gamma_1(n)\backslash\mathbb{H}^2$ are not compact is that the groups $\Gamma_1(n)$ contain non-identity unipotent elements, such as the element $\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$. The presence of such elements forces the fundamental domains for these groups in $\mathbb{H}^2$ to contain vertical "strips" going off to infinity, which contribute cylindrical ends to the corresponding quotient surface. A similar phenomenon occurs with the groups $\Gamma_1(\mathfrak{n})$, which also contain unipotent elements.

4.7. **Compact congruence quotients of $\mathbb{H}^2$ and $\mathbb{H}^3$.** In this subsection we explain how to construct other kinds of congruence quotient of $\mathbb{H}^2$ and $\mathbb{H}^3$ which *are*

[5] The ring of integers $\mathcal{O}_F$ is equal to $\mathbb{Z}[\sqrt{-d}]$ if $d\equiv 1 \text{ or } 2 \bmod 4$, and $\mathbb{Z}[(1+\sqrt{-d})/2]$ if $d\equiv 3 \bmod 4$.

closed manifolds. To do this, we need to find discrete subgroups Γ which do *not* contain unipotent elements. We now explain how we can do this.

A matrix T in $\mathrm{SL}(2,\mathbb{R})$ or $\mathrm{SL}(2,\mathbb{C})$ is unipotent if and only if $T-1$ is a nilpotent element of the ring of matrices $\mathrm{M}(2,\mathbb{R})$ or $\mathrm{M}(2,\mathbb{C})$. Thus, in order to be able to define a notion of congruence subgroup that does not contain non-identity unipotent elements, we need to find an "integral model" for $\mathrm{M}(2,\mathbb{R})$ (say) which does not contain non-zero nilpotent elements. That is, we need to find a $\mathbb{Z}$-algebra A that contains no non-zero nilpotents, such that upon extending scalars to $\mathbb{R}$, we obtain an isomorphism $\mathbb{R}\otimes_{\mathbb{Z}} A \xrightarrow{\sim} \mathrm{M}(2,\mathbb{R})$. How can we do this?

The answer to this question is: via the theory of quaternion algebras! To see why, begin by recalling that Hamilton's ring of quaternions is the (associative, but non-commutative) 4-dimensional $\mathbb{R}$-algebra $\mathbb{H}$ (this is the traditional notation, but do not confuse it with a hyperbolic space!) generated by elements i and j with the commutation relations

$$i^2 = j^2 = -1, \quad ij = -ji.$$

(The elements $1, i, j$, and $k := ij$ form a basis for $\mathbb{H}$ as an $\mathbb{R}$-vector space.) The algebra $\mathbb{H}$ is a division algebra, and so in particular, contains no non-zero nilpotent elements, but there is an isomorphism $\mathbb{C}\otimes_{\mathbb{R}}\mathbb{H} \xrightarrow{\sim} \mathrm{M}(2,\mathbb{C})$.

Now, the theory of quaternion algebras over number fields allows us to greatly generalize Hamilton's construction. For example, the associative $\mathbb{Q}$-algebra D generated by elements i and j with the commutation relations

$$i^2 = 2, \quad j^2 = 3, \quad ij = -ji$$

is a 4-dimensional division algebra over $\mathbb{Q}$, and so in particular contains no non-zero nilpotents, but there is an isomorphism

$$\mathbb{R}\otimes_{\mathbb{Q}} D \xrightarrow{\sim} \mathrm{M}(2,\mathbb{R}), \tag{4.3}$$

given by

$$i \mapsto \begin{pmatrix} \sqrt{2} & 0 \\ 0 & -\sqrt{-2} \end{pmatrix}, \qquad j \mapsto \begin{pmatrix} 0 & 3 \\ 1 & 0 \end{pmatrix}.$$

Let B denote a maximal order in D; i.e. B is a maximal $\mathbb{Z}$-subalgebra of D that is finitely generated as a $\mathbb{Z}$-module. (So B is like a "ring of integers" of D, although, unlike in the case of rings of integers in number fields, B is not unique, but is only unique up to conjugation by a non-zero element of D.) We can use the isomorphism (4.3) to regard B as a subring of $\mathrm{M}(2,\mathbb{R})$, and then define a discrete subgroup Γ of $\mathrm{SL}(2,\mathbb{R})$ via $\Gamma := B\bigcap \mathrm{SL}(2,\mathbb{R})$. Since D, and hence B, contains no non-zero nilpotents, the group Γ contains no non-identity unipotent elements. Thus $\Gamma\backslash\mathbb{H}^2$ is *compact*.

Now Γ is not torsion-free, and so the quotient $\Gamma\backslash\mathbb{H}^2$ is an orbifold rather than a manifold. However, by considering the reduction modulo n maps

$$B \to B/nB$$

for natural numbers n, we may define congruence subgroups of Γ analogous to $\Gamma(n)$ or $\Gamma_1(n)$, and so obtain many torsion-free finite index subgroups of Γ. In this way, we can construct infinitely many closed hyperbolic surfaces, with many finite covering maps between them. (And this is just for one particular choice of D!) We refer to these surfaces, obtained by choosing an appropriate $\mathbb{Q}$-algebra D as above, and then constructing the associated group Γ and its congruence subgroups, as *compact congruence quotients* of $\mathbb{H}^2$.

We can similarly construct compact congruence quotients of $\mathbb{H}^3$. We first choose a quadratic imaginary number field F. We then construct a 4-dimensional division algebra D over F. Since $\mathbb{C}$ is algebraically closed, there will automatically be an isomorphism

$$\mathbb{C} \otimes_F D \xrightarrow{\sim} \mathrm{M}(2, \mathbb{C}). \tag{4.4}$$

We then choose a maximal $\mathcal{O}_F$-order B in D, and using (4.4) to regard B as a subring of $\mathrm{M}(2, \mathbb{C})$, we then set $\Gamma := B \bigcap \mathrm{GL}(2, \mathbb{C})$. For any non-zero ideal $\mathfrak{n} \subset \mathcal{O}_F$, we can use the reduction map $B \to B/\mathfrak{n}B$ to define congruence subgroups of Γ analogous to $\Gamma_1(\mathfrak{n})$. All but finitely many of these congruence subgroups will be torsion-free, and the corresponding quotients of $\mathbb{H}^3$ will then be *compact.* We refer to the closed hyperbolic 3-manifolds constructed in this manner as *compact congruence quotients* of $\mathbb{H}^3$.

We now give a concrete example, which is directly relevant to the proof of Theorem 1.4. Begin by taking $F = \mathbb{Q}(\sqrt{-2})$. Let $\pi := 1 - \sqrt{-2}$, $\overline{\pi} := 1 + \sqrt{-2}$. Note that π and $\overline{\pi}$ are prime elements of $\mathcal{O}_F := \mathbb{Z}[\sqrt{-2}]$, and that $\pi\overline{\pi} = 3$. Let D be the associate F-algebra generated by elements i and j satisfying commutation relations

$$i^2 = -1, j^2 = -3, ij = -ji.$$

Then D is a division algebra. We take B to be a maximal order in D, and define Γ as above. Finally, for any $n \geq 0$ we define a congruence subgroup Γ_n of Γ analogous to $\Gamma_1(\pi^n)$. (We refer to [5, §2] for the precise definition.) Following [5], we write

$$X[\overline{\pi}\pi^n] := \Gamma_n \backslash \mathbb{H}^3.$$

(The notation reflects the choice of congruence subgroup, together with the fact that the construction of D involves the number $3 = \overline{\pi}\pi$.) If n is sufficiently large than Γ_n is torsion-free, and thus $X[\overline{\pi}\pi^n]$ is a closed hyperbolic 3-manifold.

Calegari and Dunfield prove the following result [5, §2], which has Theorem 1.4 as an immediate consequence.

4.8. **Theorem.** (1) *The injectivity radius of* $X[\overline{\pi}\pi^n]$ *grows without bound as* $n \to \infty$.
(2) *Each* $X[\overline{\pi}\pi^n]$ *is a rational homology sphere.*

We will give the barest of sketches of the proof of this theorem in Subsection 5.7 below, which, as we have already emphasized, relies on the Langlands reciprocity conjecture.

4.9. **Automorphic forms, cohomology, and Hecke operators.** The reason that the reciprocity conjecture can be applied to the problem of computing the Betti numbers of congruence quotients is because there is a relationship between cohomology of congruence quotients and automorphic forms, which we now sketch.

Suppose that Γ is one of the (many) congruence subgroups of $\mathrm{SL}(2, \mathbb{R})$ (resp. $\mathrm{GL}(2, \mathbb{C})$) that we have defined. If the quotient $\Gamma\backslash\mathbb{H}^2$ (resp. $\Gamma\backslash\mathbb{H}^3$) is a closed manifold, then, as was remarked in Subsection 3.5, we may identify its ith cohomology group (with $\mathbb{C}$-coefficients) with the space of harmonic differential i-forms on $\Gamma\backslash\mathbb{H}^2$ (resp. $\Gamma\backslash\mathbb{H}^3$). (In fact, suitably interpreted, such a result holds true even when $\Gamma\backslash\mathbb{H}^2$ (resp. $\Gamma\backslash\mathbb{H}^3$) is not compact.) Now since $\Gamma\backslash\mathbb{H}^2 := \Gamma\backslash \mathrm{PSL}(2, \mathbb{R})/\mathrm{SO}(2)$ (resp. $\Gamma\backslash\mathbb{H}^3 := \Gamma\backslash \mathrm{PSL}(2, \mathbb{C})/\mathrm{SO}(3)$), we may pull-back a harmonic differential i-form on $\Gamma\backslash\mathbb{H}^2$ (resp. $\Gamma\backslash\mathbb{H}^3$) to obtain a certain kind of differential i-form on $\Gamma\backslash \mathrm{PSL}(2, \mathbb{R})$ (resp. $\Gamma\backslash \mathrm{PSL}(2, \mathbb{C})$), which may then be described explicitly in terms of a certain automorphic form on $\Gamma\backslash \mathrm{PSL}(2, \mathbb{R})$ (resp. $\Gamma\backslash \mathrm{PSL}(2, \mathbb{C})$). Since this automorphic forms arises from a harmonic i-form, it will automatically be an eigenvector for the Casimir and higher Casimir operators (that were discussed in the introduction).

It will not necessarily be an eigenvector for the Hecke operators. However, we can define an action of the Hecke operators directly on the cohomology of $\Gamma\backslash\mathbb{H}^2$ (resp. $\Gamma\backslash\mathbb{H}^3$). If we then begin with a cohomology class that *is* an eigenvector for the Hecke operators on cohomology, then the corresponding automorphic form will be an eigenvector for the Hecke operators as well, and so will be an automorphic eigenform, in the sense of the introduction.

The Hecke operators on cohomology are defined by certain Hecke correspondences.[6] We won't give the general description of these correspondence, but will content ourselves with describing them in the very simplest case, namely for the quotient $\mathrm{SL}(2,\mathbb{Z})\backslash\mathbb{H}^2$.

First, we need to define yet another species of congruence subgroup.

4.10. **Definition.** For any integer $n \geq 1$, the group $\Gamma_0(n) \subset \mathrm{SL}(2,\mathbb{Z})$ is defined to be the preimage under (4.1) of the subgroup of upper triangular matrices in $\mathrm{SL}(2,\mathbb{Z}/n\mathbb{Z})$; i.e.

$$\Gamma_0(n) := \{\gamma \in \mathrm{SL}(2,\mathbb{Z}/n\mathbb{Z}) \mid \gamma \equiv \begin{pmatrix} * & * \\ 0 & * \end{pmatrix} \bmod N\}.$$

Since $\Gamma_0(n) \subset \mathrm{SL}(2,\mathbb{Z}/n\mathbb{Z})$, there is a natural projection $\mathrm{pr} : \Gamma_0(n)\backslash\mathbb{H}^2 \to \mathrm{SL}(2,\mathbb{Z})\backslash H^2$.

An easy calculation shows that the matrix $\begin{pmatrix} 0 & 1 \\ n & 0 \end{pmatrix}$ normalizes $\Gamma_0(n)$. Thus the automorphism of $\mathbb{H}^2$ induced by this matrix induces a corresponding automorphism of $\Gamma_0(n)\backslash\mathbb{H}^2$, which we denote by w_n.

Now, for any prime p, the pth Hecke correspondence T_p is defined by the following diagram:

$$\mathrm{SL}(2,\mathbb{Z})\backslash\mathbb{H}^2 \xleftarrow{\mathrm{pr}} \Gamma_0(p)\backslash\mathbb{H}^2 \xrightarrow{w_p} \Gamma_0(p)\backslash\mathbb{H}^2 \xrightarrow{\mathrm{pr}} \mathrm{SL}(2,\mathbb{Z})\backslash\mathbb{H}^2,$$

or, in symbols, $T_p := \mathrm{pr} \circ w_p \circ \mathrm{pr}^{-1}$. (We obtain a correspondence, since pr is not one-one, and so pr^{-1} is multi-valued.) The operator induced by T_p on cohomology is called the pth Hecke operator. The T_p commute among themselves (essentially, because of the Chinese remainder theorem).

If we replace $\mathrm{SL}(2,\mathbb{Z})$ by $\Gamma_1(n)$ (for some $n \geq 1$), then we may form an analogous diagram for any p that does not divide n. Similarly, if $\Gamma\backslash\mathbb{H}^2$ is a compact congruence quotient, then we may construct an analogous diagram for all but finitely many p. In this way, we obtain a commuting family of operators on the cohomology of any such quotient, indexed by all but finitely many primes. One can make similar constructions after replacing $\mathbb{H}^2$ by $\mathbb{H}^3$.

As already noted, eigenvectors for the Hecke operators on cohomology of $\Gamma\backslash\mathbb{H}^2$ of $\Gamma\backslash\mathbb{H}^3$ give rise to automorphic eigenforms on $\Gamma\backslash \mathrm{PSL}(2,\mathbb{R})$ or $\Gamma\backslash \mathrm{PSL}(2,\mathbb{C})$.

[6] A correspondence is a multi-valued continuous map. Cohomology is not only contravariantly functorial under continuous maps, but under correspondences: heuristically, one pulls back a cohomology class under each choice of single valued "branch" of the correspondence, and then sums the results.

5. Two-dimensional Galois representations and the Langlands' reciprocity conjecture for GL_2

In this section we begin by explaining how an elliptic curve defined over a number field gives rise to a family of Galois representations, which encode Diophantine information related to the curve. These are the basic examples of families of Galois representations of the type that appear in Conjecture 1.7. We then state a slightly more precise form of this conjecture in the context of congruence subgroups of $\mathrm{PSL}(2, \mathbb{R})$ or $\mathrm{PSL}(2, \mathbb{C})$. Finally, we explain various implications of the conjecture, including Fermat's Last Theorem (which is now a theorem of Wiles, since, together with Taylor, he was able to prove the relevant case of the reciprocity conjecture in this context), Theorem 1.4 (proved conditionally on the reciprocity conjecture by Calegari and Dunfield [5], but then unconditionally by Boston and Ellenberg [3]), and the virtual positive Betti number conjecture in the case of congruence quotients.

5.1. **Elliptic curves.**

Let F be a number field. An *elliptic curve* E over F is an equation of the form $y^2 = x^3 + ax^2 + bx + c$, with $a, b, c \in F$, for which the cubic $x^3 + ax^2 + bx + c$ is separable (i.e. has distinct roots). Our goal in this subsection is to explain how E gives rise to representations of the Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$.

We begin by considering the set of complex solutions to E in $\mathbb{C}^2$. This set is a manifold[7] that is homeomorphic to a torus with 1 point removed. We can naturally adjoin this missing point as a "point at infinity" to the set of solutions of E (equivalently, consider the set of solutions not just in $\mathbb{C}^2$, but in the complex projective plane $\mathbb{C}P^2$). We denote this completed set of solutions by $E(\mathbb{C})$; it is topologically a torus.

Now a torus is homeomorphic to the quotient $\mathbb{R}^2/\mathbb{Z}^2$, and so admits not just the structure of a topological space, but the structure of an abelian topological group (thinking of it as a quotient of the additive topological group $\mathbb{R}^2$). We can then transport this group structure back to $E(\mathbb{C})$, and ask whether is has any intrinsic meaning.

The answer is *yes*: the set $E(\mathbb{C})$ does have an intrinsic abelian group structure. More precisely, the point at infinity is the identity for this group structure, while three points $P, Q, R \in E(\mathbb{C})$ sum to zero in this group structure precisely if they are collinear. One can check[8] that these two rules do indeed define an abelian group structure on $E(\mathbb{C})$. As we have more-or-less already stated, there is then an isomorphism of topological groups $E(\mathbb{C}) \xrightarrow{\sim} \mathbb{R}^2/\mathbb{Z}^2$.

If $n \geq 1$ is a positive integer, then the n-torsion subgroup of $\mathbb{R}^2/\mathbb{Z}^2$ (i.e. the kernel of multiplication by n) is isomorphic to $(\mathbb{Z}/n\mathbb{Z})^2$. Thus, if we write $E[n]$ to denote the n-torsion subgroup of $E(\mathbb{C})$, there is an isomorphism $E[n] \xrightarrow{\sim} (\mathbb{Z}/n\mathbb{Z})^2$.

Now, the points of $E[n]$ are obtained by solving an (increasingly elaborate, as n gets large) series of equations involving intersecting various lines with the elliptic curve E. Since the coefficients of the equation giving rise to E lie in F, we see

[7] It is here that the assumption that $x^3 + ax^2 + bx + c$ has distinct roots is used.

[8] This check is non-trivial. One relies very much on the fact that since the points in $E(\mathbb{C})$ are solutions to a *cubic* equation in x and y, any line meets $E(\mathbb{C})$ in exactly three points (counted with the correct multiplicities), by Bézout's Theorem. With this observation in hand one can check that the two stated rules give a well-defined commutative binary operation, with inverses. The verification of associativity remains a non-trivial application of Bézout's Theorem, together with some related projective geometry.

that this series of equations again has coefficients lying in F. Thus the coordinates of the points of $E[n]$ are algebraic over $\overline{\mathbb{Q}}$, i.e. $E[n] \subset E(\overline{\mathbb{Q}})$ (the subset of $E(\mathbb{C})$ consisting of points having algebraic number coordinates), and $E[n]$ is preserved under the natural action of $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$.

Now comes the punchline: the action of $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$ on $E[n]$ induces a homomorphism

$$\rho_{E,n} : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{Aut}(E[n]) \xrightarrow{\sim} \mathrm{Aut}\big((\mathbb{Z}/n\mathbb{Z})^2\big) = \mathrm{GL}(2, \mathbb{Z}/n\mathbb{Z}).$$

In short, the elliptic curve E gives rise to a family of two-dimensional representations $\rho_{E,n}$ of the Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$, with coefficients in the various rings $\mathbb{Z}/n\mathbb{Z}$.

5.2. **The Diophantine significance of $\rho_{E,n}$.** As in the preceding section, let E denote an elliptic curve over the number field F, given by the equation $y^2 = x^3 + ax^2 + bx + c$. Let $\mathcal{O}_F$ denote the ring of integers of F. Suppose that $\mathfrak{p}$ is a non-zero prime ideal of $\mathcal{O}_F$ satisfying the following two properties: (a) $\mathfrak{p}$ does not contain the denominators of any of a, b, or c; (b) $\mathfrak{p}$ does not divide the discriminant of $x^3 + ax^2 + bx + c$. Then, by virtue of (a), we may reduce a, b, and c modulo $\mathfrak{p}$, to obtain elements $\overline{a}$, $\overline{b}$, and $\overline{c}$ in the field $\mathbb{F} := \mathcal{O}_F/\mathfrak{p}$. By virtue of (b), the cubic $x^3 + \overline{a}x^2 + \overline{b}x + \overline{c}$ has distinct roots in $\mathbb{F}[x]$, and so $y^2 = x^3 + \overline{a}x^2 + \overline{b}x + \overline{c}$ defines an elliptic curve $\overline{E}$ over $\mathbb{F}$. Let $\overline{E}(\mathbb{F})$ denote the set of solutions in $\mathbb{F}$ to the equation defining $\overline{E}$ (including the one point at infinity), and write[9]

$$a_{\mathfrak{p}} := 1 + |\mathbb{F}| - |\overline{E}(\mathbb{F})|.$$

(Here we have written $|X|$ to denote the order of the finite set X.)

The quantities $a_{\mathfrak{p}}$ are of interest from a Diophantine point of view; they describe the number of the solutions to the various congruences $\overline{E}$ induced by the equation E. The following proposition shows that they may be recovered from the family of Galois representations $\rho_{E,n}$ attached to E. We first recall that algebraic number theory associates to each non-zero prime ideal a canonical element[10] $\mathrm{Frob}_{\mathfrak{p}} \in \mathrm{Gal}(\overline{\mathbb{Q}}/F)$.

5.3. **Proposition.** *For each prime $\mathfrak{p}$, the trace of $\rho_{E,n}(\mathrm{Frob}_{\mathfrak{p}})$ is congruence to $a_{\mathfrak{p}}$ mod n.*

The Cebotarev density theorem shows that the elements $\mathrm{Frob}_{\mathfrak{p}}$ are dense in the group $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$. Thus, knowing all the quantities $a_{\mathfrak{p}}$ is equivalent to knowing the characters of all the representations $\rho_{E,n}$. This in turn is essentially[11] equivalent

[9] The number of points on a projective line over $\mathbb{F}$ is equal to $|\mathbb{F}| + 1$ (there are the elements of $\mathbb{F}$ together with the point at infinity). The quantity $a_{\mathfrak{p}}$ may thus be regarded as measuring the extent to which number of points on the elliptic curve $\overline{E}$ over $\mathbb{F}$ deviates from the number of points on the projective line.

[10] In fact, $\mathrm{Frob}_{\mathfrak{p}}$ is not an element, but a conjugacy class of cosets. Precisely, if $D_{\mathfrak{p}} \subset \mathrm{Gal}(\overline{\mathbb{Q}}/F)$ is choice of decomposition group at $\mathfrak{p}$ — and such a choice is well-defined up to conjugation — then $\mathrm{Frob}_{\mathfrak{p}}$ is a canonically determined element of the quotient $D_{\mathfrak{p}}/I_{\mathfrak{p}}$, where $I_{\mathfrak{p}}$ denotes the inertia subgroup of $D_{\mathfrak{p}}$. In what follows, it will not cause any harm to speak as if $\mathrm{Frob}_{\mathfrak{p}}$ is in fact a well-determined element of $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$.

[11] Since $\rho_{E,n}$ is defined over the ring $\mathbb{Z}/n\mathbb{Z}$, which is not a field in general, it is not the case in general that $\rho_{E,n}$ is determined by its character. But this is a technical detail, which can safely be ignored for the sake of the present discussion.

to knowing all the Galois representations $\rho_{E,n}$. Thus the family of Galois representations $\rho_{E,n}$ is an algebraic package that encodes the collection of interesting Diophantine data $a_{\mathfrak{p}}$.

5.4. **Reciprocity.** We are now in a position to state a slightly more precise form of Conjecture 1.7. We let F denote either $\mathbb{Q}$ or $\mathbb{Q}(\sqrt{-d})$ for some square-free $d > 0$. In the first case, let Γ either be a congruence subgroup of $\mathrm{SL}(2, \mathbb{Z})$, or else a congruence subgroup associated to a 4-dimensional division algebra over $\mathbb{Q}$ as in Subsection 4.7. In the second case, let Γ either be a congruence subgroup of $\mathrm{GL}(2, \mathcal{O}_F)$, or else a congruence subgroup associated to a 4-dimensional division algebra over F as in Subsection 4.7.

In the first case, we consider automorphic eigenforms on $\Gamma\backslash \mathrm{PSL}(2, \mathbb{R})$. In the second case, we consider automorphic eigenforms on $\Gamma\backslash \mathrm{PSL}(2, \mathbb{C})$. In either case, there is a Hecke operator associated to all but finitely many non-zero prime ideals[12] $\mathfrak{p}$ in $\mathcal{O}_F$, and so if f is an automorphic Hecke eigenform, it has an associated Hecke eigenvalue $a_{\mathfrak{p}}$ for all but finitely many $\mathfrak{p}$.

5.5. **Conjecture.** (a) *If f is an automorphic Hecke eigenform, whose* (*suitably normalized*) *eigenvalue under the Casimir is integral, then the eigenvalues $a_{\mathfrak{p}}$ are algebraic integers, lying in the ring of integers $\mathcal{O}_L$ of some number field L, and for every non-zero ideal $\mathfrak{n} \subset \mathcal{O}_L$, there is a representation $\rho_{f,\mathfrak{n}} : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{GL}(2, \mathcal{O}_L/\mathfrak{n})$ with the property that $a_{\mathfrak{p}}$ is congruent mod $\mathfrak{n}$ to the trace of $\rho_{f,\mathfrak{n}}(\mathrm{Frob}_{\mathfrak{p}})$ for all but finitely many non-zero prime ideals $\mathfrak{p} \subset \mathcal{O}_F$.*

(b) *Suppose given a number field L, and a family of Galois representations $\rho_{\mathfrak{n}} : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{GL}(2, \mathcal{O}_L/\mathfrak{n})$, such that the traces of $\rho_{f,\mathfrak{n}}(\mathrm{Frob}_{\mathfrak{p}})$ are compatible* (*in an obvious sense*) *as $\mathfrak{n}$ varies, for all but finitely many non-zero prime ideals $\mathfrak{p} \subset \mathcal{O}_F$, and satisfying some other technical conditions which we suppress. Then there exists an automorphic Hecke eigenform f as in* (a) *such that $\rho_{\mathfrak{n}} = \rho_{f,\mathfrak{n}}$ for all $\mathfrak{n}$.*

If E is an elliptic curve over F, then we can take $L = \mathbb{Q}$, and the family of representations $\rho_{E,n}$ is one to which part (b) of the conjecture should apply. Thus every elliptic curve is conjectured to be associated to an automorphic eigenform, in the sense of the conjecture. Taking into account Proposition 5.3, we see that the Hecke eigenvalues of the eigenform encode the number of points on E modulo the various prime ideal $\mathfrak{p}$.

5.6. **Fermat's Last Theorem.** In the papers [13, 14], Taylor and Wiles proved part (b) of the reciprocity conjecture for the Galois representations arising from (most[13]) elliptic curves over the field $\mathbb{Q}$ of rational numbers. The conjecture in this case (which predates Langlands' more general conjecture, and was known as the Shimura-Taniyama, or Shimura-Taniyama-Weil, conjecture) famously implies Fermat's Last Theorem. Let us very briefly sketch the reason why (following the strategy due to Frey and Serre).

Fermat's Last Theorem for the exponent 3 was proved by Euler, and for the exponent 4 by Fermat himself. Suppose then that $u^p + v^p = w^p$ for some prime

[12] In fact, this is correct only if $\mathcal{O}_F$ has class number 1; otherwise, the situation is slightly more complicated. We suppress this technical detail.

[13] More precisely, they proved the conjecture for those Galois representations arising from so-called *semi-stable* elliptic curves over $\mathbb{Q}$. The conjecture for Galois representations arising from arbitrary elliptic curves was completely established in [4].

$p \geq 5$ and some integers u, v, and w. We can form the elliptic curve E with equation

$$y^2 = x(x - u^p)(x + v^p).$$

This elliptic curve gives rise to its Galois representations $\rho_{E,n}$, and hence, via the reciprocity conjecture, to an automorphic eigenform f. This automorphic eigenform will in turn contribute to the H^1 of some compact congruence quotient of $\mathbb{H}^2$. Moreover, a careful investigation of the properties of the Galois representations $\rho_{E,n}$, and in particular of the representation $\rho_{E,p}$, due to Ribet [10], allows one to compute this congruence quotient precisely.[14] A direct computation then shows that this particular congruence quotient has genus 0, and hence that $H^1 = 0$. Consequently, the eigenform can't exist, and thus neither can the solution to Fermat's Last Theorem.

5.7. **Rational homology spheres.** To prove Theorem 4.8, Calegari and Dunfield show that the reciprocity conjecture implies that each of the manifolds $X[\overline{\pi}\pi^n]$ has trivial first Betti number. The proof is by contradiction: Suppose that some $X[\overline{\pi}\pi^n]$ has positive first Betti number. Then there is a non-zero eigenvector for the Hecke operators in $H^1(X[\overline{\pi}\pi^n], \mathbb{C})$, whose harmonic representative will be an automorphic eigenform. Part (a) of the reciprocity conjecture implies that this eigenform gives rise to a family of Galois representations. Analyzing the properties of the resulting Galois representations, one finds that in fact they cannot exist![15] Thus b_1 must vanish after all.

We won't attempt to explain here the Galois-theoretic argument that rules out the existence of these representations. We mention only that it relies on the fact that $F := \mathbb{Q}(\sqrt{-2})$ has fairly small discriminant (-8), and that $\pi\overline{\pi} = 3$, which is a small prime.

Although great progress has been made on part (a) of the reciprocity conjecture in the case when F is quadratic imaginary (by Taylor, et. al. [7, 11]), it is not known in sufficient generality to make the argument of [5] unconditional. (Another, unconditional, proof of Theorem 1.4 has been found by Boston and Ellenberg.) Never the less, the reciprocity conjecture is certainly believed to be true, and the argument of Calegari and Dunfield gives a good indication of the range of its influence in diverse areas of mathematics.

5.8. **The virtual positive Betti number conjecture for congruence quotients of $\mathbb{H}^3$.** We close this note by observing that the reciprocity conjecture implies the virtual positive Betti number conjecture for compact congruence quotients

[14] One can reasonably ask where the property that (u, v, w) solves the Fermat equation is actually used. The answer is as follows: the discriminant of the cubic $x(x - u^p)(x + v^p)$ is equal to $-u^p v^p(u^p + w^p) = (-uvw)^p$, and in particular is a perfect pth power. Because of this, the representation $\rho_{E,p}$ is endowed with rather remarkable properties — too remarkable, as it turns out, for it to even exist.

This is not the place to explain carefully how the nature of the discriminant influences the properties of the representation $\rho_{E,p}$. But it may help to mention a simpler, but related, Galois-theoretic phenomenon: if $f(x) \in \mathbb{Q}[x]$ is an irreducible degree n polynomial, then the Galois group of the splitting field of $f(x)$ is a subgroup of the symmetric group S_n, and is typically equal to this group. However, if the discriminant of $f(x)$ is a square, then the Galois group in fact lies inside the alternating group A_n.

[15] In the non-existence proof as it is written in [5], the Generalized Riemann Hypothesis is also required. However, Calegari has informed me that in fact the argument can be made to work without it.

of $\mathbb{H}^3$. Fix $F = \mathbb{Q}(\sqrt{-d})$, and a 4-dimensional division algebra D over F, as in Subsection 4.7. If Γ_1 and Γ_2 are two congruence subgroups of $\mathrm{GL}(2, \mathbb{C})$ arising from these choices, then $\Gamma_1 \bigcap \Gamma_2$ has finite index in each of Γ_1 and Γ_2, and so $(\Gamma_1 \bigcap \Gamma_2)\backslash\mathbb{H}^3$ is a common finite cover of each of $\Gamma_1\backslash\mathbb{H}^3$ and $\Gamma_2\backslash\mathbb{H}^3$. Taking into account Lemma 3.6, it thus suffices to exhibit one Γ (arising from the particular choice of F and D) for which $\Gamma\backslash\mathbb{H}^3$ has positive first Betti number. For this, it suffices to exhibit a certain kind of non-zero automorphic eigenform. If one grants the reciprocity conjecture (in particular, part (b) of the conjecture), then to do this, it suffices in turn to write down certain kinds of Galois representations. Now one can write down lots of elliptic curves over the quadratic imaginary field F, so many in fact that one can find plenty of the required kinds of Galois representations. This proves the result.

References

1. Borel A., *Automorphic L-functions*, in Automorphic Forms, Representations, and L-functions (A. Borel, W. Casselman ed.), Proc. Symp. Pure Math. **33**, part 2, Amer. Math. Soc., Providence, RI (1979), 27–61.
2. Borel A., Jacquet H., *Automorphic forms and automorphic representations*, in Automorphic Forms, Representations, and L-functions (A. Borel, W. Casselman ed.), Proc. Symp. Pure Math. **33**, part 1, Amer. Math. Soc., Providence, RI (1979), 189–202.
3. Boston N., Ellenberg J. S., *Pro-p groups and towers of rational homology spheres*, Geom. Topol. **10** (2006), 331–334.
4. Breuil C., Conrad B., Diamond F., Taylor R., *On the modularity of elliptic curves over $\mathbb{Q}$: wild* 3*-adic exercises*, J. Amer. Math. Soc. **14** (2001), 843–939.
5. Calegari F., Dunfield N., *Automorphic forms and rational homology* 3*-spheres*, Geom. Topol. **10** (2006), 295–329.
6. Gross B. H., *On the Satake isomorphism*, in Galois representations in arithmetic algebraic geometry (Durham, 1996), London Math. Soc. Lecture Note Ser. **254**, Cambridge Univ. Press, Cambridge (1998), 223–237.
7. Harris M., Soudry D., Taylor R., *l-adic representations associated to modular forms over imaginary quadratic fields. II.*, Invent. Math. **114** (1993), 289–310.
8. Kirby R. C., *Problems in low-dimensional topology*, in Geometric topology (Athens, GA, 1993), AMS/IP Stud. Adv. Math. 2.2, Amer. Math. Soc., Providence, RI (1997), 35–473.
9. Reid A., *The geometry and topology of arithmetic hyperbolic* 3*-manifolds*, to appear in Proc. Symposium Topology, Complex Analysis and Arithmetic of Hyperbolic Spaces, Kyoto 2006, RIMS Kokyuroku Series.
10. Ribet K.A., *On modular representations of* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ *arising from modular forms*, Invent. Math. **100** (1990), 431–476.
11. Taylor R. L., *l-adic representations associated to modular forms over imaginary quadratic fields. II.*, Invent. Math. **116** (1994), 619–643.
12. Taylor R. L., *Galois representations*, Ann. Fac. Sci. Toulouse Math. (6) **13** (2004), 73–119.
13. Taylor R., Wiles A.J., *Ring-theoretic properties of certain Hecke algebras*, Ann. Math. **141** (1995), 553–572.
14. Wiles A.J., *Modular elliptic curves and Fermat's last theorem*, Ann. Math. **141** (1995), 443–551.

E-mail address: emerton@math.northwestern.edu

Mathematics Department, Northwestern University, 2033 Sheridan Rd., Evanston, IL 60208

E-mail address: emerton@math.northwestern.edu

Compressive sensing: a paradigm shift in signal processing

Olga V. Holtz
University of California-berkeley
& Technische Universität Berlin

Abstract

We survey a new paradigm in signal processing known as "compressive sensing". Contrary to old practices of data acquisition and reconstruction based on the Shannon-Nyquist sampling principle, the new theory shows that it is possible to reconstruct images or signals of scientific interest accurately and even exactly from a number of samples which is far smaller than the desired resolution of the image/signal, e.g., the number of pixels in the image. This new technique draws from results in several fields of mathematics, including algebra, optimization, probability theory, and harmonic analysis. We will discuss some of the key mathematical ideas behind compressive sensing, as well as its implications to other fields: numerical analysis, information theory, theoretical computer science, and engineering.

1 Introduction

Compressive sensing [45, 119] is a new concept in signal processing where one seeks to minimize the number of measurements to be taken from signals while still retaining the information necessary to approximate them well. The ideas have their origins in certain abstract results from functional analysis and approximation theory [79, 92] but were recently brought into the forefront by the work of Candés, Romberg and Tao [13, 15, 12] and Donoho [45] who constructed concrete algorithms and showed their promise in application.

Sparse approximation has been studied for nearly a century, and it has numerous applications. Temlyakov [111] locates the first example in a 1907 paper of Schmidt [104]. In the 1950s, statisticians launched an extensive investigation of another sparse approximation problem called subset selection in regression [87] and recently least angle regression [54, 113]. Later, approximation theorists began a systematic study of m-term approximation with respect to orthonormal bases and redundant systems [38, 111] and very recently in [25, 26].

Over the last decade, the signal processing community spurred by the work of Coifman et al. [28, 29] and Mallat et al. [84, 37, 36] has become interested in sparse representations for compression and analysis of audio [72], images [63] and video [90]. Sparsity criteria also arise in deconvolution [110], signal modeling [100], preconditioning [74], machine learning [70], de-noising [22], regularization [33, 35] and error correction [16, 19, 60, 58, 59, 61]. Most sparse approximation problems employ a linear model in which the collection of elementary signals is both linearly dependent and large. These models are often called redundant or overcomplete. Recent research suggests that overcomplete models offer a genuine increase in approximation power [95, 62]. Unfortunately, they also raise a serious challenge. How do we find a good representation of the input signal among the plethora of possibilities? One method is to select a parsimonious or sparse representation. The exact rationale for invoking sparsity may range from engineering to economics to philosophy. At least three justifications are commonly given:

1. It is sometimes known *a priori* that the input signal can be expressed as a short linear combination of elementary signals also contaminated with noise.
2. The approximation may have an associated cost that must be controlled. For example, the computational cost of evaluating the approximation depends on the number of elementary signals that participate. In compression, the goal is to minimize the number of bits required to store the approximation.
3. Some researchers cite Occam's Razor, "Pluralitas non est ponenda sine necessitate" (causes must not be multiplied beyond necessity).

Sparse approximation problems are computationally challenging because most reasonable sparsity measures are not convex. A formal hardness proof for one important class of problems independently appeared in [88] and [36]. A vast array of heuristic methods for producing sparse approximations have been proposed, but the literature contains few guarantees of their performance. The pertinent numerical techniques fall into at least three basic categories:

1. The convex relaxation approach replaces the nonconvex sparsity measure with a related convex function to obtain a convex programming problem. The convex program can be solved in polynomial time with standard software [8], and one expects that it will yield a good sparse approximation. More on that will be said in the sequel.
2. Greedy methods make a sequence of locally optimal choices in an effort to produce a good global solution to the approximation problem. This category includes forward selection procedures (such as matching pursuits), backward selection and others. Although these approaches sometimes succeed [31, 67, 69, 68, 115, 117, 118], they can also fail spectacularly [40, 22]. The monographs of Miller [87] and Temlyakov [111] taste the many flavors of greedy heuristic.
3. Specialized nonlinear programming software has been developed that attempts to solve sparse approximation problems directly using, for example, interior point methods [96]. These techniques are only guaranteed to discover a locally optimal solution though.

Several problems require solutions to be obtained from underdetermined systems of linear equations, i.e., systems with fewer equations than unknowns. Some example

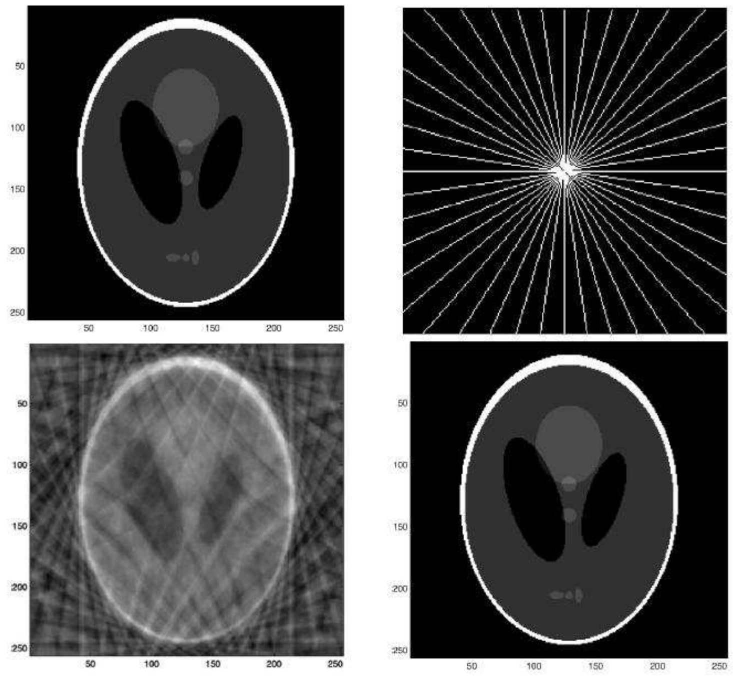


Figure 1: When Fourier coefficients of a testbed medical image known as the Logan–Shepp phantom (top left) are sampled along 22 radial lines in the frequency domain (top right), a naive, “minimal energy” reconstruction setting unobserved Fourier coefficients to 0 is marred by artifacts (bottom left). ℓ_1-reconstruction (bottom right) is exact.

of such problems arise in linear filtering signal processing, and inverse problems. For an underdetermined system of linear equations, if there is any solution, there are infinitely many solutions. In many applications, the “simplest“ solution is most acceptable. Such a solution is inspired by the minimalist principle of Occam's Razor. For example, if the parameters of a model are being estimated then among all models that explain the data equally well, the one with the minimum number of parameters is most desirable.

The notion that *sparse* signals–meaning signals with a small number of nonzero coefficients for a given basis (and no noise) one can (with high probability) be reconstructed exactly via ℓ_1-minimization is not exactly new. The idea was first expressed in 1986, by Fadil Santosa and William Symes [103]. But the full extent of the theory, including the robustness of the reconstruction procedure, is only now coming into full focus. One of the champions of this approach, David Donoho, coined the term *”compressed sensing“* to emphasize the fact that ℓ_1-minimization is not just a new way of massaging a ”complete“ set of measurements into a compact form, but rather a new way of thinking about how to measure things in the first place [45].

This new way of thinking has profoundly practical implications. Making measurements can be expensive, in terms of time, money, or (in the case of, say, x-rays) damage done to the object being imaged. Compressive sensing has the potential to provide substantial cost savings without sacrificing accuracy. In one impressive numerical experiment, Candès, Romberg, and Tao [12] showed that a 512×512-pixel test image, known as the Logan-Shepp phantom, can be reconstructed exactly from 512 Fourier coefficients sampled along 22 radial lines–with, in other words, more than 95% of the ostensibly relevant data missing (see Figure 1).

A host of practical applications are now being explored, including new sensing techniques, new analog-to-digital converters, and a new digital camera with a single photon detector, being developed by Kevin Kelly, Richard Baraniuk and the Digital Signal Processing group at Rice (dsp.rice.edu/cs/cscamera) [107, 121].

2 Mathematical foundations

2.1 Sparsity and undersampling

The celebrated Nyquist-Shannon-Whittaker sampling theorem shows that a signal with bandwidth 2Ω is completely determined by its uniform samples if and only if the samples are taken at least at the *Nyquist rate* Ω/π. This principle used to underlie all signal acquisition techniques used in practice, such as consumer electronics, medical imaging, analog-to-digital conversion and so on. Compressive sampling puts forward a novel sampling paradigm that replaces the notion of band-limited signals with that of sparse signals. This new notion allows for dramatically "undersampled" signals to be captured and manipulated using a very small amount of data. The point of this section is to explain the basic mathematics behind this new theory.

Suppose x is an unknown vector in $\mathbb{R}^N$ (a digital image or signal). We plan to sample x using n linear functionals of x and then reconstruct. We are interested in the case $n \ll N$, when we have many fewer measurements than the dimension of the signal space. Such situations arise in many applications. For example, in biomedical imaging, far fewer measurements are typically collected than the number of pixels in the image of interest. Further examples are provided by virtually any domain of science or technology where amounts of data are very large and costs of observation/acquisition/measurement are nontrivial.

The measurements y_k are obtained by sensing x against n vectors $\phi_k \in \mathbb{R}^N$. Thus $y_k = \langle x, \phi_k \rangle$ for $k = 1, \ldots, n$, or, equivalently

$$y = \Phi x \tag{1}$$

for some $n \times N$ *measurement/sensing matrix* Φ. Thus we arrive at an underdetermined system of linear equations, which, as is well known, in general has infinitely many solutions, so our problem is ill-posed. But suppose that our signal x is *sparse* or *compressible*, i.e., that is (essentially) depends only on a small number of degrees of freedom. To give a first impression of the theory, we in fact assume that the signal can be written exactly as a linear combination of only a few basis vectors.

Mathematically the problem can be formulated as follows. Given a matrix $\Phi \in \mathbb{R}^{n \times N}$ with many more columns than rows ($n \ll N$), and a vector $y \in \mathbb{R}^n$, find a

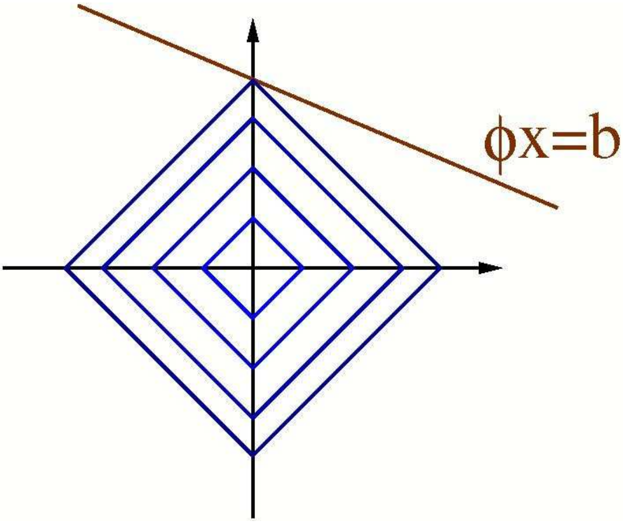


Figure 2: ℓ_1 minimization

vector $x \in \mathbb{R}^N$ with a minimum possible number of nonzero entries, i.e.,

$$\text{minimize } \|x\|_0 \text{ subject to } \Phi x = y \tag{2}$$

where $\|x\|_0$ is the number of nonzero entries of x [30]. By allowing noise ($\varepsilon \geq 0$), we obtain a variation of the problem (2):

$$\text{minimize } \|x\|_0 \text{ subject to } \|\Phi x - y\|_2 \leq \varepsilon, \tag{3}$$

These problems *per se* are NP-hard even for $\varepsilon = 0$, see [65, 88].

The classical, well studied, approach would be to minimize the 2-norm $\|x\|_2$ in the above problems, but this usually yields a solution vector x that is full, while for a sparse representation we would like to find a vector x with few nonzero entries.

The main approach taken in compressive sensing is to minimize the 1-norm $\|x\|_1$ instead.

$$\text{minimize } \|x\|_1 \text{ subject to } \Phi x = y, \tag{4}$$

and

$$\text{minimize } \|x\|_1 \text{ subject to } \|\Phi x - y\|_2 \leq \varepsilon, \tag{5}$$

respectively [22], where $\|x\|_1 := \sum_i |x_i|$ (See Figure 2). Very surprisingly, the ℓ_1 minimization yields the same result as the ℓ_0 minimization in many cases of practical interest. This phenomenon was initially observed by engineers and geophysicists, most notably Claerbout and Logan as early as 1970s (see [110]), and by Santosa and Symes in 1986 [103], as mentioned in the introduction. In the last five years or so, a series of papers [12, 41, 42, 46, 47, 56, 64, 73, 105] explained why ℓ_1 minimization can recover sparse signals in a variety of practical setups. In our next section, we give a few sample theorems about this remarkable phenomenon.

Finally, the ℓ_1 minimization problem can be efficiently solved by convex programming or by linear programming (LP) [8]. Most compressive sensing results due to Candès, Donoho, Romberg, Tao and others [41, 42, 12, 45, 25, 15] are all based on this method (see also [125, 126, 127]). Other approaches include greedy algorithms, for instance, the so-called *matching pursuit* introduced by Mallat and Zhang [84, 91, 114, 115]. Recently many variations on matching pursuit have been proposed, among which are orthogonal matching pursuit [91, 89], stagewise orthogonal matching pursuit [51], gradient pursuit [6], and others.

2.2 Incoherence and restricted isometry

Given an $n \times N$ matrix Φ, the first basic question is to determine whether Φ is good for compressive sensing, i.e., will lead to good recovery of sparse solutions to the equations $\Phi x = y$.

Candès and Tao [10]–[20] introduced a necessary condition that guarantees an estimate of its performance on classes of sparse vectors.

Definition ([19, 12, 13]). *A matrix Φ is said to satisfy the* Restricted Isometry Property (RIP) of order k with constant $\delta := \delta_k \in (0,1)$ *if*

$$(1-\delta_k)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\delta_k)\|x\|_2^2 \tag{6}$$

for any x such that $\|x\|_0 \leq k$.

It is straightforward to see that this condition can be reformulated as follows: Consider $n \times \#T$ matrices Φ_T formed by the columns of Φ with indices in the set T. Then the Gramian matrices

$$G_T := \Phi_T^t \Phi_T$$

are bounded and boundedly invertible on l_2 with bounds as in (6), uniform for all T of size $\#T = k$. Since each matrix G_T is symmetric and nonnegative definite, this is equivalent to each of these matrices having their eigenvalues in the interval $[1-\delta_k, 1+\delta_k]$. The role played by RIP becomes clear from the following result of Candès and Tao.

Theorem 1 ([18, 11]). *If the $n \times N$ matrix satisfies RIP of order $3k$ for some $\delta \in (0,1)$, then, for any vector $x \in \mathbb{R}^N$, the ℓ_1 minimization problem (4) has a solution x^* such that*

$$\|x - x^*\|_2 \leq C \cdot \frac{\|x - x_k\|_1}{\sqrt{k}}, \tag{7}$$

where x_k denotes the best k-sparse approximation to x and C denotes a constant.

The condition (7) means that the Φ's with higher values of k for which RIP is satisfied perform better in compressive sensing. For example, if an $n \times N$ matrix Φ has the restricted isometry property of order k, then its performance in l_2-norm on the unit ball of l_1^N is of order $C/\sqrt{k}$, and the optimal performance is achieved if Φ satisfies RIP of order $k = \Theta(n/\log(N/n))$ [39]. This is indeed achieved via various probabilistic constructions [41, 42, 43, 44, 48].

A primary matrix measure related to RIP is *mutual incoherence* [22, 118, 114, 115]:

$$\mathcal{M}(\Phi) := \max_{i \neq j} |(\Phi^t \Phi)_{i,j}|,$$

i.e., the maximum inner product of distinct columns of Φ. Since the columns of Φ are usually normalized to be of 2-norm 1, the mutual incoherence of a matrix is between 0 and 1.

This notion can be generalized [77] as follows: For a given normalized matrix $\Phi \in \mathbb{R}^{n \times N}$, its *k-mutual incoherence* $\mathcal{M}_k(\Phi)$ is defined by

$$\mathcal{M}_k(\Phi) := \max_{\#S \leq k} \max_{i \neq j} |(\Phi_S^T \Phi_S)_{i,j}|.$$

The mutual incoherences $\mathcal{M}_k$ are intimately related to the best constant δ_k with which the matrix Φ satisfies RIP of order k, but a full understanding of this connection has not been reached [77, 78].

A challenging aspect of RIP is its computational cost. Indeed, RIP is a property of the submatrices of a specific size. At present, no subexponential-time algorithm is known for testing RIP. Introducing other matrix measures may potentially help in effectively verifying the RIP or finding other, less demanding conditions for sparse recovery.

One such weaker condition has been introduced by Cohen, Dahmen and DeVore in [25]. To motivate their condition, we first recall that a pair (Φ, Δ) where Φ is a sensing matrix and Δ is a decoder, is called *instance-optimal of order k* for a normed space $(V, \|\cdot\|_V)$ if there exists an absolute constant C such that

$$\|x - \Delta(\Phi x)\|_V \leq C\|x - x_k\|_V.$$

The matrix Φ has the *null space property* in V if

$$\|x\|_V \leq c\|x - x_k\|_V \qquad \text{for all } x \text{ such that } \Phi x = 0.$$

The importance of the null space property can be seen from the following result:

Theorem 2 ([25, 26]). *Given an $n \times N$ matrix Φ, a norm $\|\cdot\|_V$ and a value k, the instance optimality in V with constant C_0 is equivalent to the null space property of Φ of order $2k$ with the constant $C_0/2$ in the sufficiency part and the same constant C_0 in the necessity part.*

Note that the null space property is preserved under row operations on the matrix Φ since, as its name suggests, it is simply a property of its null space. This property is therefore less rigid than the RIP and may allow for a more efficient verification.

2.3 Compressible signals

In practice, most signals may not be exactly sparse in a given basis but may concentrate near a sparse set. In fact, the most commonly used models in signal processing assume that the coefficients of the signal with respect to, say, a wavelet basis, decay rapidly away from their essential support. Smooth signals, images with bounded variation and those with bounded Besov norm are known to be of that type.

Given a nearly sparse signal x, denote by x_k its best k-sparse approximation, i.e., the vector obtained by keeping the k largest coefficients of x and discarding the rest. Candès, Romberg and Tao [12] showed that the initial signal can be recovered with error of order $\|x - x_k\|_1 / \sqrt{k}$ whenever the sensing matrix satisfies RIP of order $4k$ and the RIP constants δ_{3k} and δ_{4k} are not too close to 1.

Theorem 3 ([18]). *Let Φ satisfy RIP of order $4k$ with $\delta_{3k} + 3\delta_{4k} < 2$. Then, for any signal x, the solution x^* to (4) satisfies*

$$\|x^* - x\|_2 \leq C \cdot \frac{\|x - x_k\|_1}{\sqrt{k}},$$

with a well-behaved constant C.

A similar result holds [12] for stable recovery from imperfect measurements, i.e., in the setting of problem (5). All together, this indicates that ℓ_1 minimization stably recovers the largest k coeffients of a nearly k-sparse vector even in the presence of noise.

This result is in fact optimal for important classes of signals: Let x belong to the weak-ℓ_p ball or radius R, i.e., let the decreasing rearrangement of its coefficients $|x|_{(1)} \geq |x|_{(2)} \geq \cdots \geq |x|_{(N)}$ satisfy the condition

$$|x|_{(i)} \leq R \cdot i^{-1/p}, \qquad i = 1, \ldots, N.$$

This can be shown to imply

$$\|x - x_k\|_2 \leq C \cdot R \cdot k^{1/2-1/p} \quad \text{and} \quad \|x - x_k\|_1 \leq C \cdot R \cdot k^{1-1/p}$$

for some constant C. Moreover, for generic elements in weak-ℓ_p, no better estimates are obtainable. In other words, ℓ_1 recovery achieves an approximation error roughly as small as the error obtained by deliberately selecting the k largest coefficients of the signal.

2.4 Good sensing matrices

Most sampling algorithms developed so far in compressive sensing are based on randomization [17, 45]. Typically, the sensing matrices are produced by taking i.i.d. random variables with some given probability distribution and then normalizing their columns. Such matrices are guaranteed to perform well with very high probability, i.e., with the failure rate exponentially small in the size of the matrix [45]. Following [10], we mention three random constructions that are by now standard.

Random matrices with i.i.d. entries. Consider the matrix Φ with entries drawn independently at random from the Gaussian probability distribution with mean zero and variance $1/n$. Then [15, 42], with overwhelming probability, the ℓ_1 minimization (4) recovers k-sparse solutions whenever

$$k \leq \text{const} \cdot n / \log(N/n).$$

Fourier ensemble. Let Φ be obtained by randomly selecting n rows from the $N \times N$ discrete Fourier transform and renormalizing the columns so that they have 2-norm 1. If the rows are selected at random, then [15] as above, with overwhelming probability, the ℓ_1 minimization (4) recovers k-sparse vectors for

$$k \leq \text{const} \cdot n/(\log N)^6.$$

General orthogonal ensembles. Suppose Φ is obtained by selecting n rows from an $N \times N$ orthonormal matrix U and renormalizing the columns to be of unit length. In the rows are selected at random, then [15] k-sparse recovery by ℓ_1 minimization (4) is guaranteed with overwhelming probability provided that

$$k \leq \text{const} \cdot \frac{1}{\mathcal{M}^2(U)} \frac{n}{(\log N)^6}.$$

Note that the Fourier matrix U satisfies $\mathcal{M}(U) = 1$, so this is a generalization of the Fourier ensemble.

The natural problem already being addressed by several authors is how to achieve robust deterministic constructions of good CS matrices. Tao in [109] points out the importance of this problem, as well as its similarity to other derandomization problems from theoretical computer science and combinatorics. Several deterministic constructions are currently known (see, e.g., [39, 76]). However, the performance of matrices provided by these deterministic constructions is not yet on a par with that of matrices arising probabilistically.

To give several examples, DeVore in [39] proposes a construction of cyclic matrices using finite fields that satisfy RIP of order k for $k \leq C\sqrt{n}\log n/\log(N/n)$, which falls short of the above-mentioned range $k \leq Cn/\log(N/n)$ known for probabilistic constructions. Indyk in [76] and Xu and Hassibi in [123] propose another scheme for compressive sensing with deterministic performance guarantees based on bipartite expander graphs. Another flavor of randomness is introduced in [2] where random Toeplitz matrices are constructed with entries drawn independently from a given probability distribution.

2.5 Optimality and n-widths

The performance of the best sensing matrices Φ, which is presently achieved by random matrices with probabilistic guarantees, yields recovery of k-sparse vectors using n samples (so that the matrix Φ is $n \times N$) provided that

$$k \leq \text{const} \cdot n/\log(N/n).$$

In particular, a k-sparse vector can be recovered, say, by random projections, of dimension $O(k \cdot \log(N/k))$ [41].

For signals x in the weak-ℓ_p ball of radius R, ℓ_1 recovery gives the error [45]

$$\|x^* - x\|_2 \leq \text{const} \cdot R \cdot (n/\log(N/n))^{-1/p+1/2}.$$

It turns out that this performances cannot be improved even by using possibly adaptive sets of measurements and reconstruction algorithms.

The matter turns out to be closely related to the issue of the so-called Gelfand widths [92] known from approximation theory: For a class $\mathcal{F}$, let $E_n(\mathcal{F})$ be the best reconstruction error from n linear measurements

$$E_n(\mathcal{F}) := \inf \sup_{f \in \mathcal{F}} \|f - D(y)\|_2, \quad y = \Phi f,$$

where the infimum is over all sets of n linear functionals and all reconstruction algorithms D. The error $E_n(\mathcal{F})$ is essentially equal [92] to the Gelfand width of the class $\mathcal{F}$ defined as

$$d_n(\mathcal{F}) := \inf_V \{\sup_{f \in \mathcal{F}} \|P_V f\| : \operatorname{codim}(V) < n\},$$

where P_V is the orthogonal projector on the subspace V. Gelfand widths are known for many classes of interest. In particular, Kashin [79], Garnaev and Gluskin [66] showed that the Gelfand widths for the weak-ℓ_p ball of radius R satisfy

$$c \cdot R \cdot \left(\frac{\log(N/n)+1}{n}\right)^{-1/p+1/2} \leq d_n(\mathcal{F}) \leq C \cdot R \cdot \left(\frac{\log(N/n)+1}{n}\right)^{-1/p+1/2}.$$

for some universal constants c and C.

This shows that the recovery provided by compressive sensing techniques is in fact optimal for weak-ℓ_p norms in spite of being completely non-adaptive [25, 10]. This is one more indication of the great potential of compressive sensing in applications.

3 Connections with other fields

3.1 Statistical estimation

Candès [10] and Donoho [45] point out a number of connections of compressive sensing with ideas from statistics and coding theory. We briefly mention main ideas here.

In statistical estimation, the signal is assumed to be measured with stochastic errors

$$y = \Phi x + z$$

where z is a vector of i.i.d. (independent identically distributed) random variables with mean zero and variance σ^2. Very often, z is assumed to be Gaussian. The problem is again to recover x from y.

One seeks to design an estimator whose accuracy depends on the information content of the object x. The *Dantzig selector* [20] estimates x by solving the convex program

$$\text{minimize } \|\tilde{x}\|_1 \quad \text{subject to} \quad \sup_i |(\Phi^T r)_i| \leq \lambda\sigma$$

for some $\lambda > 0$, where r is the *residual* $r := y - \Phi\tilde{x}$. These ideas are very close to the so-called *lasso* approach [113, 54].

Analogously to ℓ_1 minimization in compressive sensing, the Danzig selector was shown [20] to recover sparse and compressible signals with the number of measurements much smaller than the dimension of x and within a logarithmic factor of the ideal mean squared error one would only achieve with an *oracle* supplying perfect information which coordinates are nonzero and which are above the noise level.

3.2 Error-correcting codes

In coding theory [60, 58, 59, 61], a vector x is transmitted to a remote receiver. The information x is encoded using an $n \times N$ matrix C with $n \ll N$. Gross errors may occur during transmission, so that a fraction of the entries of Cx is completely corrupted. The location and the damage done to those entries are unknown. It turns out that a constant fraction of errors with arbitrary magnitudes can still be corrected [19] by solving a suitable linear minimization problem. In fact, known methods recover the vector x exactly provided the fraction of the corrupted entries is not too big [20, 10].

3.3 Frame theory

The theory of compressive sensing matrices closely resembles the basic theory of frames [23, 32, 83, 124]. A countable collection of elements $\{f_i\}_{i\in I}$ is a *frame* for a Hilbert space H if there exist constants $0 < A \leq B < \infty$ (the *lower* and *upper frame bound*) such that, for all $g \in H$,

$$A\|g\|_H^2 \leq \sum_{i\in I} |\langle g, f_i\rangle|^2 \leq B\|g\|_H^2.$$

A frame is called *tight* if the upper and lower bounds are the same $A = B$. A frame is *bounded* if $\inf_{i\in I} \|f_i\|_H > 0$ (the condition $\sup_{i\in I} \|f_i\|_H < \infty$ follows automatically from the definition of a frame). A frame is *unit norm* if $\|f_i\|_H = 1$ for all $i \in I$. If $\{f_i\}_{i\in I}$ is a frame only for its closed linear span, it is called a *frame sequence.* A family $\{f_i\}_{i\in I}$ is a *Riesz basic sequence* for H if it is a *Riesz basis* for its closed linear span, i.e., if, for some constants $0 < A \leq B < \infty$ and for all sequences of scalars $\{c_i\}_{i\in I}$,

$$A\sum_{i\in I} |c_i|^2 \leq \|\sum_{i\in I} c_i f_i\|_H^2 \leq B\sum_{i\in I} |c_i|^2.$$

The analogy with the restricted isometry property is obvious, however, the latter is imposed only on submatrices formed from the original matrix.

This analogy must be worth pursuing in both directions, i.e., looking for applications of the theory and methodology of compressive sensing to frames and vice versa. Randomization techniques from compressive sensing could be of particular interest in attacking problems from frame theory (cf. [4, 116]).

4 Practical implications

Compressive sensing, and more generally the possibility of efficiently capturing sparse and compressible signals using a relatively small number of measurements, paves the way for a number of possible applications.

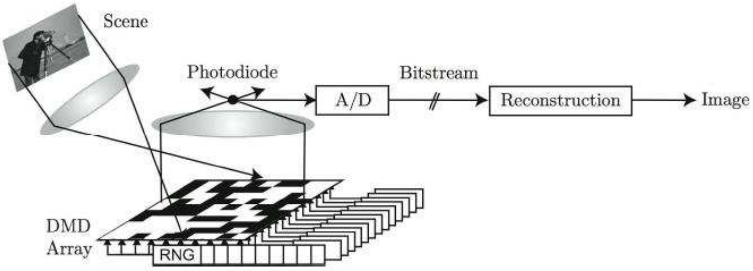


Figure 3: The scheme of the CS camera

Data acquisition. New physical sampling devices may be designed that directly record discrete low-rate incoherent measurements of the analog signal. This should be especially useful in situations where large collections of samples may be costly, difficult or impossible to obtain.

Data compression. The sparse basis in which the signal is to be represented may be unknown or unavailable. However, a randomly designed Φ is suitable for almost all signals. We stress that these protocols are nonadaptive to the signal and simply require to correlate it with a small number of other fixed vectors.

Inverse problems. The measurement system may have to satisfy rigid constraints such as in MR angiography and other MR setups, where Φ records a subset of the Fourier transform. However, if a sparse basis exists that is also incoherent with Φ, then efficient sensing is possible.

A particularly interesting example of successful implementation of compressive sensing methodology is provided by a digital camera newly developed by Richard Baraniuk and Kevin Kelly at Rice University (see dsp.rice.edu/cs/cscamera) [107, 121].

In the detector array of a conventional digital camera, each pixel performs an analog-to-digital conversion; for example, the detector on a 5-megapixel camera produces 5 million bits for each image. This large amount of data is then dramatically reduced through a compression algorithms (using wavelet or other techniques) so as not to overburden typical storage and transfer capacities.

Rather than collect 5 million pixels for an image, the new camera samples only a factor of about four times the 50,000 pixels that the jpg compression might typically output. These 200,000 single-pixel measurements provide an immediate 25-fold savings in data collected compared with 5 megapixels.

The camera developed at Rice replaces the CCD array with a digital-micromirror device (DMD). A sequence of random projections is performed on the micromirror array, so that the image "bounces off" of each random pattern in the sequence, and the reflected light from each pattern is collected sequentially with a photodiode sensor that acts as the single-pixel detector (see Figure 3). After taking a sequence of essentially time-multiplexed measurements, a specific ℓ_1 minimization algorithms decodes the picture out of the collected sequence of single-pixel measurements.

References

[1] G. Aubert, J. Aujol, Modeling very oscillating signals, application to image processing, Applied Mathematics and Optimization 51 (2) (2005) 163–182.

[2] W. Bajwa, J. Haupt, G. Raz, S. Wright, and R. Nowak, Toeplitz-structured compressed sensing matrices. IEEE Workshop on Statistical Signal Processing (SSP), Madison, Wisconsin, August 2007.

[3] R. Baraniuk. Optimal tree Approximation using Wavelets. In A. J. Aldroubi and M. Unser, editors, Wavelet Applications in Signal Processing, volume VII, pages 196–207, 1999. SPIE.

[4] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, A simple proof of the restricted isometry property for random matrices. To appear in Constr. Appr., 2007.

[5] L. A. Bassalygo, M. S. Pinsker, Complexity of an optimum nonblocking switching network without reconnections, Problems in Information Transmission, 9 (1) (1973) 84–87.

[6] T. Blumensath and M. E. Davies, Iterative thresholding for sparse approximations. Preprint, 2007.

[7] T. Blumensath and M. E. Davies, Gradient pursuits. Preprint, 2007.

[8] S. Boyd, and L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004. (available at http://www.stanford.edu/˜boyd/cvxbook/).

[9] A.M. Bruckstein, M. Elad, M. Zibolevsky, A non-negative and sparse enough solution of an underdetermined linear system of equations is unique, Submitted to IEEE Trans. Inform. Theory, 2007.

[10] E. J. Candès, Compressive sampling, Proceedings of the Int. Congress of Mathematics, 3, pp. 1433–1452, Madrid, Spain, 2006.

[11] E. J. Candès and P. Randall, Highly robust error correction by convex programming. Preprint, 2006.

[12] E. J. Candès, J. Romberg, and T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inform. Theory, 52(2) pp. 489–509, February 2006.

[13] E. J. Candès, and J. Romberg, Quantitative robust uncertainty principles and optimally sparse decompositions. Foundations of Comput. Math., 6(2), pp. 227–254, April 2006.

[14] E. J. Candès and J. Romberg, Sparsity and incoherence in compressive sampling. Inverse Problems, 23(3) pp. 969–985, 2007.

[15] E. J. Candès, and T. Tao, Near optimal signal recovery from random projections: Universal encoding strategies? IEEE Trans. Inform. Theory, 52(12), pp. 5406–5425, December 2006.

[16] E. J. Candès and T. Tao. Decoding by linear programming, IEEE Trans. Inform. Theory 51, 4203–4215, 2006.

[17] E. Candès and J. Romberg. Practical signal recovery from random projections, Proceedings of the SPIE, Volume 5674, pp. 76–86, 2005.

[18] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements, Communications on Pure and Applied Mathematics, Vol 59, No. 8, pp. 1207–1223, Aug 2006.

[19] E. J. Candès and T.Tao. Error Correction via linear programming, In Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS), pp. 295–308, 2005.

[20] E. Candès and T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n, To appear in Annals of Statistics.

[21] S. S. Chen, Basis Pursuit, Ph.D. Thesis, Department of Statistics, Stanford University, 1995.

[22] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit, SIAM J. Scientific Computing 20 (1999), 33–61.

[23] O. Christensen. An introduction to frames and Riesz bases. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, Inc., Boston, MA, 2003.

[24] A. Cohen, Numerical Analysis of Wavelet Methods, Studies in Mathematics and its Applications, vol. 32, Elsevier, Amsterdam, 2003.

[25] A. Cohen, W. Dahmen, and R. DeVore, Compressed sensing and best k-term approximation. Preprint, 2006.

[26] A. Cohen, W. Dahmen, and R. DeVore, Near optimal approximation of arbitrary vectors from highly incomplete measurements. Preprint, 2007.

[27] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore, Tree Approximation and Optimal Encoding. Appl. Comput. Harmon. Anal. 11 (2001), no. 2, 192–226.

[28] R. R. Coifman and Y. Meyer. Nouvelles bases orthonormées de $L^2(\mathbb{R})$ ayant la structure du système de Walsh. Manuscript, Mathematics Dept., Yale Univ., 1989.

[29] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best-basis selection. IEEE Trans. Inform. Theory, vol. 38, pp. 713–718, Mar. 1992.

[30] G. Cormode, M. Datar, P. Indyk and S. Muthukrishnan: Comparing Data Streams Using Hamming Norms (How to Zero In). IEEE Trans. Knowl. Data Eng. 15(3): 529–540 (2003).

[31] C. Couvreur and Y. Bresler. On the optimality of the Backward Greedy Algorithm for the subset selection problem. SIAM J. Matrix Anal. Appl., 21(3):797âĂŞ-808, 2000.

[32] I. Daubechies, Ten Lectures on Wavelets, CBMS-NSF Lecture Notes no. 61, SIAM, 1992.

[33] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm. Pure Appl. Math., vol. 57 (11), pages 1413–1457, 2004.

[34] I. Daubechies, M. Fornasier, and I. Loris, Accelerated projected gradient method for linear inverse problems with sparsity constraints. Preprint, 2007.

[35] I. Daubechies, G. Teschke and L. Vese. Iteratively solving linear inverse problems with general convex constraints, Inverse Problems and Imaging, 1(1), 29–46, 2007.

[36] G. Davis, S. Mallat, and M. Avellaneda. Greedy adaptive approximation. J. Constr. Approx., 13:57–98, 1997.

[37] G. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions. SPIE Journal of Optical Engineering 1994, 33(7):2183–2191.

[38] R. A. DeVore. Nonlinear approximation. Acta Numer., 7 (1998), 51–150.

[39] R. A. DeVore, Deterministic constructions of compressed sensing matrices. Preprint, 2007.

[40] R. DeVore and V. N. Temlyakov. Some remarks on greedy algorithms. Adv. Comput. Math., 5:173–187, 1996.

[41] D. L. Donoho, For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution, Comm. Pure Appl. Math. 59, no. 7 (2006), pp. 907âĂŞ-934.

[42] D. L. Donoho, For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution, Comm. Pure Appl. Math. 59, no. 6 (2006), pp. 797-âĂŞ829.

[43] D. L. Donoho, Neighborly polytopes and sparse solutions of undetermined linear equations. Preprint, 2005.

[44] D. L. Donoho, High-dimensional centrally-symmetric polytopes with neighborliness proportional to dimension. Disc. Comput. Geometry, 35(4) pp. 617–652, 2006.

[45] D. L. Donoho, Compressed Sensing, IEEE Trans. Inform. Theory. 52, n. 4, (2006), pp. 1289–1306.

[46] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. IEEE Trans. Inform. Theory, 47 (2001), 2845–2862.

[47] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization, Proc. Natl. Acad. Sci. USA 100 (2003), 2197–2202.

[48] D. L. Donoho, M. Elad, and V. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise, IEEE Trans. Inform. Theory 52, no. 1 (2006), pp. 6âĂŞ-18.

[49] D. L. Donoho and J. Tanner. Sparse nonnegative solutions of underdetermined linear equations by linear programming. Proc. National Academy of Sciences, 102(27), pp.9446–9451, 2005.

[50] D. Donoho and Y. Tsaig, Fast solution of $L1$-norm minimization problems when the solution may be sparse, Preprint, October 2006.

[51] D. L. Donoho, Y. Tsaig, I. Drori, and J. L. Starck, Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit, Preprint, March 2006.

[52] P. L. Dragotti, M. Vetterli, and T. Blu, Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang-Fix. IEEE Trans. on Signal Proc., 55(7), pp. 1741–1757, May 2007.

[53] M. F. Duarte, M. B. Wakin, and R. G. Baraniuk, Fast reconstruction of piecewise smooth signals from random projections, in Online Proc. SPARS05, Rennes, France, Nov. 2005.

[54] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, Least angle regression, Ann. Statist. 32, no. 2 (2004), 407–499.

[55] I. Ekeland, R. Temam, Convex analysis and variational problems, vol. 28, SIAM, Philadelphia, PA, 1999.

[56] M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. IEEE Trans. Inform. Theory, 48(9):2558–2567, 2002.

[57] H. W. Engl, M. Hanke, and A. Neubauer, Regularization of Inverse Problems, Kluwer, Dordrecht, 1996.

[58] J. Feldman. Decoding Error-Correcting Codes via Linear Programming. PhD thesis, Massachusetts Institute of Technology, 2003.

[59] J. Feldman, D. R. Karger, and M. J. Wainwright. LP decoding. In Proc. 41st Annual Allerton Conference on Communication, Control, and Computing, October 2003.

[60] J. Feldman and David R. Karger. Decoding turbo-like codes via linear programming. Proc. of the 43rd Symposium on Foundations of Computer Science, p.251–260, November 16-19, 2002.

[61] J. Feldman, M. J. Wainwright, and D. R. Karger. Using linear programming to decode linear codes. IEEE Trans. Inform. Theory 51(3): 954–972, 2005.

[62] P. Frossard and P. Vandergheynst. Redundant representations in image processing. In Proc. of the 2003 IEEE International Conference on Image Processing, 2003. Special session.

[63] P. Frossard, P. Vandergheynst, R. M. F. I Ventura, and M. Kunt. A posteriori quantization of progressive Matching Pursuit streams. IEEE Trans. Signal Proc., 52(2):525–535, Feb. 2004.

[64] J. J. Fuchs. On sparse representations in arbitrary redundant bases. IEEE Trans. Inform. Theory, 50(6):1341–1344, 2004.

[65] M. R. Garey and D. S. Johnson, Computers and Intractability. A Guide to the Theory of NP-Completeness, W. H. Freeman and Company, New York, 1979.

[66] A. Garnaev and E. Gluskin, The widths of a Euclidean ball, Dokl. Akad. Nauk USSR 277 (1984), 1048–1052; English transl. Soviet Math. Dokl. 30 (1984) 200–204.

[67] A. C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan and M. Strauss, Near-optimal sparse fourier representations via sampling, in Proc 34th ACM symposium on Theory of Computing, pp. 152–161, ACM Press, 2002.

[68] A.C. Gilbert, S. Muthukrishnan and M. Strauss, Improved Time Bounds for Near-Optimal Sparse Fourier Representation, to appear at the Wavelets XI conference in the SPIE Symposium on Optics & Photonics, 2005, San Diego, California, USA.

[69] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss, Approximation of Functions over Redundant Dictionaries Using Coherence, in Proceedings of 2003 SIAM Symposium on Discrete Algorithms SODA, pp. 243–252, 2003.

[70] F. Girosi. An equivalence between sparse approximation and Support Vector Machines. Neural Comput., 10(6):1455ï£¡-1480, 1998.

[71] I. F. Gorodnitsky and B. D. Rao, Sparse signal reconstruction from limited data using focuss : a reweighted norm minimization algorithm, IEEE Trans. Signal Proc., vol. 45, no. 3, pp. 600–616, March 1997.

[72] R. Gribonval and E. Bacry. Harmonic decomposition of audio signals with Matching Pursuit. IEEE Trans. Signal Proc., 51(1):101–111, 2003.

[73] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. IEEE Trans. Inform. Theory, 49(12):3320–3325, 2003.

[74] M. Grote and T. Huckle. Parallel preconditioning with sparse approximate inverses. SIAM J. Sci. Comput., 18(3):838–853, 1997.

[75] J. Haupt and R. Nowak, Information Theory, IEEE Trans. Inform. Theory, vol.52, no.9, pp.4036–4048, Sept. 2006.

[76] P. Indyk, Explicit constructions for compressed sensing of sparse signals. Symp. on Discrete Algorithms, 2008.

[77] S. Jokar, Mutual Incoherence, Restriceted Isometry Property and Kronecker Product of Matrices, Preprint, 2008.

[78] S. Jokar and M. Pfetsch, Exact and Approximate Sparse Solutions of Underdetermined Linear Equations, Matheon-Preprint 377, March 2007.

[79] B. Kashin, The widths of certain finite dimensional sets and classes of smooth functions, Izvestia 41(1977), 334–351.

[80] B. S. Kashin and V. N. Temlyakov, A remark on compressed sensing. Preprint, 2007.

[81] S. Kunis and H. Rauhut, Random Sampling of Sparse Trigonometric Polynomials II - Orthogonal Matching Pursuit versus Basis Pursuit. Foundations of Computational Mathematics, Springer New York, August 10, 2007.

[82] D. M. Malioutov, M. Çetin, and A.S. Willsky, Optimal Sparse Representations in General Overcomplete Bases. IEEE Int. Conf. Acoustics, Speech and Signal Processing, May 2004, Montreal, Canada.

[83] S. Mallat, A Wavelet Tour of Signal Processing. Boston, MA: Academic, 1998.

[84] S. Mallat and Z. Zhang. Matching Pursuits with time-frequency dictionaries. IEEE Trans. Signal Proc., 41(12):3397–3415, 1993.

[85] I. Maravic and M. Vetterli, Sampling and reconstruction of signals with finite rate of innovation in the presence of noise. IEEE Trans. Signal Proc., 53(8), pp. 2788–2805, August 2005.

[86] Y. Meyer, Oscillatory patterns in image processing and nonlinear evolution equations, University Lecture Series, Vol. 22, American Mathematical Society, Provindence, 2001.

[87] A. J. Miller. Subset Selection in Regression. Chapman and Hall, London, 2nd edition, 2002.

[88] B.K. Natarajan, Sparse Approximate Solutions to Linear Systems. SIAM J. Comput. 24: 227–234, 1995.

[89] D. Needell and R. Vershynin, Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. Preprint, 2007.

[90] T. Nguyen and A. Zakhor. Matching Pursuits based multiple description video coding for lossy environments. In Proceedings of the 2003 IEEE International Conference on Image Processing, Barcelona, 2003.

[91] Y. C. Pati, R. Rezaiifar, P. Krishnaprasad, Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, in: Proceedings of the 27th Annual Asilomar Conference in Signals, System and Computers, vol. 1, 1993, pp. 40-44.

[92] A. M. Pinkus, N-widths in Approximation Theory. Ergeb. Math. Grenzgeb. (3) 7, Springer-Verlag, Berlin 1985.

[93] A. M. Pinkus, On L^1-Approximation. Cambridge Tracts in Mathematics, Cambridge University Press, Vol. 93, 1989.

[94] R. Ramlau and G. Teschke. A Thresholding Iteration for Nonlinear Operator Equations with Sparsity Constraints. DFG-SPP-1114 Preprint, 2005.

[95] B. D. Rao and Y. Bresler. Signal processing with sparseness constraints. In Proc. of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 12-15 May 1998, vol. 3, 1861–1864.

[96] B. D. Rao and K. Kreutz-Delgado. An affine scaling methodology for best basis selection. IEEE Trans. Signal Proc., 47(1):187âĂŞ200, 1999.

[97] H. Rauhut, Random sampling of sparse trigonometric polynomials. Appl. Comput. Harmon. Anal. 22 (2007), no. 1, 16–42.

[98] H. Rauhut, Stability Results for Random Sampling of Sparse Trigonometric Polynomials. Preprint, 2006.

[99] H. Rauhut, K. Schass, and P. Vandergheynst, Compressed sensing and redundant dictionaries. Preprint, 2006.

[100] J. Rissanen. Modeling by shortest data description. Automatica, 14, 465-471, 1979.

[101] R. Rockafellar and R. Wets, Variational analysis, Springer-Verlag, Berlin, 1998.

[102] M. Rudelson and R. Vershynin. Geometric approach to error-correcting codes and reconstruction of signals. Technical report, Department of Mathematics, University of California, Davis, 2005.

[103] F. Santosa, and W.W. Symes, Linear inversion of band-limited reflection seismograms, SIAM J. Sci. Statist. Comput. 7 (1986), 1307–1330.

[104] E. Schmidt, Zur Theorie der linearen und nichtlinearen Integralgleichungen, Math. Ann. 63 (1907), 433–476.

[105] Y. Sharon, J. Wright, and Y. Ma, Computation and relaxation of conditions for equivalence between ℓ_1 and ℓ_0 minimization. Preprint, 2007.

[106] T. Strohmer, and R. Heath Jr. Grassmannian frames with applications to coding and communications. Appl. Comp. Harm. Anal., vol. 14(3): 257–275, 2003.

[107] D. Takhar, J. Laska, M. Wakin, M. Duarte, D. Baron, S. Sarvotham, K. Kelly, and R. Baraniuk, A new compressive imaging camera architecture using optical-domain compression. (Proc. of Computational Imaging IV at SPIE Electronic Imaging, San Jose, California, January 2006)

[108] T. Tao, An uncertainty principle for cyclic groups of prime order, Math. Res. Letters 12 (2005), 121–127.

[109] T. Tao, http://terrytao.wordpress.com/2007/07/02/open-question-deterministic-uup-matrices/.

[110] H. L. Taylor, S. C. Banks, and J. F. McCoy. Deconvolution with the ℓ_1 norm. Geophysics, 44(1):39–52, 1979.

[111] V.N. Temlyakov, Nonlinear methods of approximation, Found. Comput. Math. 3 (2003) 33–107.

[112] G. Teschke. Multi-Frame Representations in Linear Inverse Problems with Mixed Multi-Constraints. Appl. Comput. Harmon. Anal. 22 (2007), no. 1, 43–60.

[113] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Royal. Statist. Soc. B. 58, 267–288.

[114] J.A. Tropp. Greed is Good: Algorithmic Results for Sparse Approximation. IEEE Trans. Inform. Theory. 50 (11), Oct. 2004, pp. 2231–2242.

[115] J.A. Tropp. Just Relax: Convex programming methods for Subset Sleection and Sparse Approximation. IEEE Trans. Inform. Theory, vol. 51 (3), pp. 1030–1051, March, 2006.

[116] J. A. Tropp. The random paving property for uniformly bounded matrices. To appear in Studia Math, 2007.

[117] J. A. Tropp and A. C. Gilbert, Signal recovery from partial information via Orthogonal Matching Pursuit, IEEE Trans. Info. Theory. To appear, 2007.

[118] J. A. Tropp, A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss. Improved sparse approximation over quasi-incoherent dictionaries. In Proc. of the 2003 IEEE International Conference on Image Processing, Barcelona, 2003.

[119] Y. Tsaig and D. L. Donoho, Extensions of compressed sensing. Signal Processing, 86(3), pp. 549–571, March 2006.

[120] Y. Tsaig and D.L. Donoho, Breakdown of equivalence between the minimal ℓ_1-norm solution and the sparsest solution, Signal Processing, 86(3), pp. 533–548, March 2006.

[121] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, and R. Baraniuk, An Architecture for Compressive Imaging, Proc. International Conference on Image Processing – ICIP 2006, Atlanta, GA, Oct. 2006.

[122] M. Vetterli, P. Marziliano, and T. Blu, Sampling signals with finite rate of innovation, IEEE Trans. Signal Proc., vol. 50, no. 6, June 2002.

[123] W. Xu, B. Hassibi, Efficient compressive sensing with deterministic guarantees using expander graphs. IEEE Information Theory Workshop, Lake Tahoe, September 2007.

[124] R. Young, An introduction to nonharmonic Fourier series. Academic Press, New York, 1980.

[125] Y. Zhang. Solution-recovery in ℓ_1-norm for non-square linear systems: deterministic conditions and open questions. Technical report TR05-06, Department of computational and Applied Mathematics, Rice University, Houston, TX, 2005.

[126] Y. Zhang. A Simple Proof for Recoverability of ℓ_1-Minimization: Go Over or Under? Technical report TR05-09, Department of computational and Applied Mathematics, Rice University, Houston, TX, 2005.

[127] Y. Zhang. A Simple Proof for Recoverability of ℓ_1-Minimization (II): the Nonnegativity Case. Technical report TR05-10, Department of computational and Applied Mathematics, Rice University, Houston, TX, 2005.

TAUBES'S PROOF OF THE WEINSTEIN CONJECTURE IN DIMENSION THREE

MICHAEL HUTCHINGS

Abstract. Does every smooth vector field on a closed three-manifold, for example the three-sphere, have a closed orbit? No, according to counterexamples by K. Kuperberg and others. On the other hand there is a special class of vector fields, called Reeb vector fields, which are associated to contact forms. The three-dimensional case of the Weinstein conjecture asserts that every Reeb vector field on a closed oriented three-manifold has a closed orbit. This conjecture was recently proved by Taubes using Seiberg-Witten theory. We will give an introduction to the Weinstein conjecture, the main ideas in Taubes's proof, and the bigger picture into which it fits.

Taubes's proof of the Weinstein conjecture involves a lot of material: not only the 86-page paper proving the conjecture itself, but also, for example, substantial parts of the book by Kronheimer and Mrowka constructing Seiberg-Witten Floer homology ($\sim$800 pages), as well as ideas from Taubes's ealier work on Seiberg-Witten invariants of symplectic 4-manifolds ($\sim$400 pages). As a consequence, the present article, despite its length, is not so much an explanation of the proof as rather an introduction to it and some of the relevant background. Hopefully this will at least provide a good warmup for reading Taubes's paper to learn the full story. Another excellent source of information on this subject is the online video archive from the June 2008 MSRI hot topics workshop [31].

The contents of the present article are as follows. In §1–§3 we introduce the statement of the Weinstein conjecture and discuss some examples. In §4–§6 we discuss a natural strategy for approaching the Weinstein conjecture, which proves it in many but not all cases, and provides background for Taubes's work. In §7 we give an overview of the big picture surrounding Taubes's proof of the Weinstein conjecture. In §8–§9 we bring in necessary material from Seiberg-Witten theory. In §10 we finally give an outline of Taubes's proof.

1. Statement of the Weinstein conjecture

The Weinstein conjecture asserts that certain vector fields must have closed orbits. Before stating the conjecture at the end of this section, we first give a semi-historical discussion of its origins. This is only "semi-historical", because only a sample of the relevant works will be cited, and not always in chronological order.

1.1. Closed orbits of vector fields.

Let Y be a closed manifold (all manifolds in this article are smooth unless otherwise stated), and let V be a vector field on Y. A *closed orbit* of V is a map $\gamma : \mathbb{R}/T\mathbb{Z} \to Y$, for some $T > 0$, satisfying the

2000 *Mathematics Subject Classification.* 57R17,57R57,53D40.

Partially supported by NSF grant DMS-0806037.

ordinary differential equation

$$\frac{d\gamma(t)}{dt} = V(\gamma(t)).$$

Given a closed manifold Y, we can ask: Does every vector field on Y have a closed orbit? If the Euler characteristic $\chi(Y) \neq 0$, then by the Poincaré-Hopf index theorem every vector field on Y has zeroes, which give rise to constant closed orbits. In this article we will mainly focus on the case where Y is three-dimensional. Then $\chi(Y) = 0$, and as a result it is relatively easy to construct vector fields on Y with no zeroes. But understanding nonconstant closed orbits is much harder.

Already for $Y = S^3$, the question of whether all vector fields have closed orbits is very difficult. It turns out that the answer is no. Examples of vector fields on S^3 with no closed orbit, with increasing degrees of regularity, were constructed by P. Schweizer (C^1), J. Harrison (C^2), K. Kuperberg (C^∞), G. Kuperberg and K. Kuperberg (real analytic), and G. Kuperberg (C^1 and volume preserving), see [40, 17, 29, 28, 27]. Thus to guarantee the existence of a closed orbit one needs to make some additional assumption on the vector field.

1.2. **Hamiltonian vector fields.** A very important class of vector fields, of interest in mechanics and dynamical systems, are Hamiltonian vector fields. We briefly recall the definition of these. Let (M^{2n}, ω) be a symplectic manifold. This means that ω is a closed 2-form on M such that $\omega^n \neq 0$ everywhere, or equivalently ω defines a nondegenerate bilinear form on the tangent space T_xM at each point $x \in M$. If $H : M \to \mathbb{R}$ is a smooth function, the associated *Hamiltonian vector field* is the vector field X_H on M characterized by $\omega(X_H, \cdot) = dH$. This exists and is unique by the nondegeneracy condition. The flow of X_H preserves H, because

$$X_H(H) = dH(X_H) = \omega(X_H, X_H) = 0.$$

So if $E \in \mathbb{R}$ is a regular value of H, then the level set $H^{-1}(E)$ is a $(2n-1)$-dimensional submanifold of M, which we will call a "regular level set", and X_H is a smooth vector field on it.

Must the Hamiltonian vector field X_H have a closed orbit on every regular level set? Results of Hofer-Zehnder and Struwe using symplectic capacities, see eg [19, 32], show that under favorable circumstances the answer is "almost every". For example, suppose $M = \mathbb{R}^{2n}$ with coordinates $x_1, \ldots, x_n, y_1, \ldots, y_n$ and the standard symplectic form

$$\omega = \sum_{i=1}^{n} dx_i \wedge dy_i.$$

Theorem 1.1 (Hofer-Zehnder, Struwe, 1990). *Let $H : \mathbb{R}^{2n} \to \mathbb{R}$ be a proper smooth function. Then the vector field X_H has a closed orbit on $H^{-1}(E)$ for almost every E in the interval $[-1, 1]$ (in the measure-theoretic sense).*

But not necessarily for every E. For example, Ginzburg-Gürel [15] proved that there is a proper C^2 Hamiltonian on $\mathbb{R}^4$ with a regular level set with no closed characteristic. And C^∞ examples are known in higher dimensions. So even for Hamiltonian vector fields on $\mathbb{R}^{2n}$ one needs a further assumption to guarantee the existence of closed orbits.

1.3. **Weinstein's conjecture.** To see what kind of assumption might be needed, let Y be a hypersurface in a symplectic manifold (M,ω) which is a regular level set of a Hamiltonian $H : M \to \mathbb{R}$. Observe that the existence of a closed orbit of X_H on Y depends only on the hypersurface Y and not on the Hamiltonian H. For suppose $K : M \to \mathbb{R}$ is another Hamiltonian which also has Y as a regular level set. Then $dK|_Y = f dH|_Y$ for some nonvanishing function $f : Y \to \mathbb{R}$, so $X_K = fX_H$ on Y. Thus the periodic orbits of X_K and X_H on Y differ only in their parametrizations. In fact one can describe the existence problem for periodic orbits on Y without referring to a Hamiltonian at all by noting that the Hamiltonian vector field on Y always lives in the rank 1 subbundle

$$L_Y := \operatorname{Ker}(\omega|_Y) \subset TY,$$

called the *charateristic foliation*. Thus the existence of a closed orbit on Y for any Hamiltonian having it as a regular level set is equivalent to the existence of a closed curve tangent to L_Y, sometimes called a *closed characteristic*.

Under what circumstances must a hypersurface in a symplectic manifold have a closed characteristic? In the late 1970's, Weinstein [48] showed that in $\mathbb{R}^{2n}$ with the standard symplectic form, if Y is a convex compact hypersurface, then Y has a closed characteristic. Rabinowitz [37] showed that one can replace "convex" in the statement by "star-shaped", meaning transverse to the radial vector field

$$\rho := \frac{1}{2}\sum_{i=1}^{n}(x_i\partial_{x_i} + y_i\partial_{y_i}).$$

Now the existence of a closed characteristic is invariant under symplectomorphisms of $\mathbb{R}^{2n}$, while the star-shaped condition is not. This suggests that there should be a more general condition on the hypersurface Y that guarantees the existence of a closed characteristic. Weinstein [49] proposed such a condition as follows.

Definition 1.2. A hypersurface Y in a symplectic manifold (M,ω) is of *contact type* if there is a 1-form λ on Y such that $d\lambda = \omega|_Y$, and $\lambda(v) \neq 0$ for all nonzero $v \in L_Y$.

This condition is clearly invariant under symplectomorphisms of (M,ω).

If Y is a star-shaped hypersurface in $\mathbb{R}^{2n}$, then Y is of contact type, because one can take

$$\lambda = \frac{1}{2}\sum_{i=1}^{n}(x_i dy_i - y_i dx_i)|_Y.$$

This works because if $v \in L_Y$ is nonzero, then the star-shaped condition implies that $\omega(\rho, v) \neq 0$, while it follows from the above formulas that $\omega(\rho, v) = \lambda(v)$, so $\lambda(v) \neq 0$.

More generally, a *Liouville vector field* on a symplectic manifold (M,ω) is a vector field ρ such that $\mathcal{L}_\rho\omega = \omega$. It turns out that a hypersurface Y in (M,ω) is of contact type if and only if there exists a Liouville vector field defined in a neighborhood of Y which is transverse to Y. In particular, given such a vector field ρ, the contact type condition is fulfilled by $\lambda = \omega(\rho,\cdot)|_Y$.

One could now conjecture that if Y is a compact hypersurface of contact type in a symplectic manifold (M,ω), then Y has a closed characteristic. This is essentially what Weinstein conjectured in [49]. (The original conjecture in [49] also assumed that $H^1(Y;\mathbb{R}) = 0$.)

1.4. **Contact forms.** In fact one can remove the symplectic manifold (M,ω) from the statement as follows. Let Y be an oriented manifold of dimension $2n-1$. A *contact form* on Y is a 1-form λ on Y such that

$$\lambda \wedge (d\lambda)^{n-1} > 0$$

everywhere. A contact form λ determines a vector field R, called the *Reeb vector field*, characterized by

$$d\lambda(R,\cdot) = 0, \qquad \lambda(R) = 1.$$

If Y is a hypersurface of contact type in a symplectic manifold, then the 1-form λ appearing in the definition of contact type is a contact form, and the associated Reeb vector field is tangent to the characteristic foliation L_Y. Conversely, any manifold with a contact form (Y,λ) arises as a hypersurface of contact type in an associated symplectic manifold. Namely, the *symplectization* of (Y,λ) is the manifold $\mathbb{R}\times Y$ with the symplectic form $\omega = d(e^s\lambda)$, where s denotes the $\mathbb{R}$ coordinate. The slice $\{0\}\times Y$ is a hypersurface of contact type in $\mathbb{R}\times Y$, with λ fulfilling the definition of contact type. In conclusion, the conjecture in the previous paragraph is equivalent to the following:

Weinstein Conjecture. *Let Y be a closed oriented odd-dimensional manifold with a contact form λ. Then the associated Reeb vector field has a closed orbit.*

1.5. **Some terminology.** Below, to save space, we usually say "Reeb orbit" instead of "closed orbit of the Reeb flow". Two Reeb orbits $\gamma,\gamma' : \mathbb{R}/T\mathbb{Z} \to Y$ are considered equivalent if they differ by reparametrization, ie precomposition with a translation of $\mathbb{R}/T\mathbb{Z}$. If $\gamma : \mathbb{R}/T\mathbb{Z} \to Y$ is a Reeb orbit and k is a positive integer, then the k-fold *iterate* of γ is the pullback of γ to $\mathbb{R}/kT\mathbb{Z}$. A Reeb orbit γ is embedded if and only if it is not the k-fold iterate of another Reeb orbit where $k>1$.

2. Basic examples in $2n-1$ dimensions

2.1. **Hypersurfaces in $\mathbb{R}^{2n}$.** The Weinstein conjecture for compact hypersurfaces of contact type in $\mathbb{R}^{2n}$ was proved in 1987 by Viterbo [47]. In fact the almost-existence result in Theorem 1.1 is a generalization of this. To see why, let Y be a compact hypersurface in $\mathbb{R}^{2n}$ of contact type. As mentioned previously, there exists a Liouville vector field ρ defined on a neighborhood of Y which is transverse to Y. Flowing the hypersurface Y along the vector field ρ for a small time ε gives another hypersurface Y_ε and a diffeomorphism $Y \simeq Y_\varepsilon$. Since ρ is transverse to Y, there exists $\delta > 0$ such that the different hypersurfaces Y_ε for $|\epsilon| < \delta$ are disjoint and sweep out a neighborhood that can be identified in the obvious way with $(-\delta,\delta)\times Y$. Also, the Liouville condition $\mathcal{L}_\rho\omega = \omega$ implies that the diffeomorphism $Y \simeq Y_\varepsilon$ respects the characteristic foliations. Thus Y_ε has a closed characteristic for either all ε or none. But Theorem 1.1 implies that Y_ε has a closed characteristic for almost every ε. Hence there is a closed characteristic for every ε, and in particular for $\varepsilon = 0$.

2.2. **Cotangent bundles.** Another important example of a manifold with a contact form is the unit cotangent bundle of a Riemannian manifold.

To start, let Q be a smooth manifold. There is a canonical 1-form λ on the cotangent bundle T^*Q, defined as follows. Let $\pi : T^*Q \to Q$ denote the projection.

If $q \in Q$, and if $p \in T_q^*Q$, then $\lambda : T_{(q,p)}T^*Q \to \mathbb{R}$ is the composition

$$T_{(q,p)}T^*Q \xrightarrow{\pi_*} T_qQ \xrightarrow{p} \mathbb{R}.$$

More explicitly, if $q_1, \ldots, q_n$ are local coordinates on Q, and if we write a cotangent vector in this coordinate patch as $p = \sum_{i=1}^n p_i dq_i$, then this gives local coordinates $q_1, \ldots, q_n, p_1, \ldots, p_n$ on T^*Q, and in these coordinates

$$\lambda = \sum_{i=1}^n p_i dq_i.$$

Now suppose that Q has a Riemannian metric. This induces a metric on T^*Q, and we consider the unit cotangent bundle

$$ST^*Q = \{p \in T^*Q \mid |p| = 1\}.$$

The restriction of λ to ST^*Q is a contact form. Indeed ST^*Q is a hypersurface of contact type in the symplectic manifold $(T^*Q, d\lambda)$, with transverse Liouville vector field $\rho = \sum_{i=1}^n p_i \partial_{p_i}$. One can further show that the associated Reeb vector field agrees with the geodesic flow, see eg [14, §1.5]. Thus Reeb orbits in ST^*Q are equivalent to closed geodesics in Q. If Q is compact, then so is ST^*Q, and the Weinstein conjecture in this case is equivalent to the classical Lyusternik-Fet theorem asserting that every compact Riemannian metric has at least one closed geodesic.

2.3. **Prequantization spaces.** Another important example of manifolds with contact forms is given by circle bundles, or "prequantization spaces". Let (Σ, ω) be a symplectic manifold of dimension $2n-2$, and suppose that the cohomology class $[\omega] \in H^2(\Sigma; \mathbb{R})$ is the image of an integral class $e \in H^2(\Sigma; \mathbb{Z})$. Let $\pi : V \to \Sigma$ be the principal S^1-bundle with first Chern class e. This means that there is a free S^1 action on V whose quotient is Σ, and $e \in H^2(\Sigma; \mathbb{Z})$ is the primary obstruction to finding a section $\Sigma \to V$. Let R denote the derivative of the S^1 action; this is a vector field on V which is tangent to the fibers. There is a unique connection 1-form λ on V whose curvature equals ω. These conditions mean that λ is invariant under the S^1 action, $\lambda(R) = 1$, and $d\lambda = \pi^*\omega$. It follows that λ is a contact form on V whose Reeb vector field is R. In particular, the Reeb orbits are the fibers and their iterates.

The fact that Reeb orbits appear here in $(2n-2)$-dimensional smooth families is a special feature arising from the symmetry of the picture. For a "generic" contact form on a manifold Y the Reeb orbits are isolated, in the sense that if γ is a Reeb orbit of length T which goes through a point $x \in Y$, then there is no other Reeb orbit through a point close to x with length close to T. For example, on a circle bundle as above, one can get rid of most of the Reeb orbits by a perturbation in which one replaces the contact form λ by

$$\lambda' = (1 + \pi^*H)\lambda,$$

where $H : \Sigma \to \mathbb{R}$ is a smooth function whose first derivative is small. A short calculation then shows that the new Reeb vector field is given by

$$R' = (1 + \pi^*H)^{-1}R + (1 + \pi^*H)^{-2}\widetilde{X}_H,$$

where X_H is the Hamiltonian vector field on Σ determined by H, and $\widetilde{X}_H$ denotes its horizontal lift, ie the unique vector field on V with $\pi_*\widetilde{X}_H = X_H$ and $\lambda(\widetilde{X}_H) = 0$. So for this new contact form, the only fibers that are Reeb orbits are the fibers over the critical points of H. (There may also be some much longer Reeb orbits that cover closed orbits of X_H.) But the Weinstein conjecture here asserts that there is no way to eliminate all of the remaining Reeb orbits without introducing new ones. The Weinstein conjecture in this case can be proved using contact homology [11, §2.9], about which we will have more to say later.

3. The Weinstein conjecture in three dimensions

We now restrict attention to the three-dimensional case. The main goal of this article is to explain Taubes's proof of the Weinstein conjecture in dimension three:

Theorem 3.1 (Taubes)**.** *If Y is a closed oriented three-manifold with a contact form, then the associated Reeb vector field has a closed orbit.*

3.1. Contact structures in three dimensions.

To gain an appreciation for this result and its proof, it will help to recall a bit about contact structures on three-manifolds. For much more about this subject we refer the reader to [12] and [14].

Recall that a contact form on a closed oriented three-manifold Y is a 1-form λ such that $\lambda \wedge d\lambda > 0$ everywhere. The associated *contact structure* is the 2-plane field $\xi = \operatorname{Ker}(\lambda)$. This has an orientation induced from the orientation of Y and the direction of the Reeb vector field. In general one defines a contact structure[1] to be an oriented 2-plane field which is the kernel of some contact form. A contact structure is a "totally nonintegrable" 2-plane field, which means that in a sense it as far as possible from being a foliation: the kernel of λ is a foliation if and only if $\lambda \wedge d\lambda \equiv 0$.

Different contact forms can give rise to the same contact structure. To be precise, if λ is a contact form, then λ' is another contact form giving rise to the same contact structure if and only if $\lambda' = f\lambda$ where $f : Y \to \mathbb{R}$ is a positive smooth function. For a given contact structure, the Reeb vector field depends on the choice of contact form, but it is always transverse to the contact structure.

A (three-dimensional closed) *contact manifold* is a pair (Y, ξ) where Y is a closed oriented three-manifold and ξ is a contact structure on Y. Two contact manifolds (Y, ξ) and (Y, ξ') are *isomorphic*, or *contactomorphic*, if there is an orientation-preserving diffeomorphism $\phi : Y \to Y'$ such that ϕ_* sends ξ to ξ' preserving the orientiations. Two contact structures ξ and ξ' on Y are *isotopic* if there is a one-parameter family of contact structures $\{\xi_t \mid t \in [0,1]\}$ on Y such that $\xi_0 = \xi$ and $\xi_1 = \xi'$. Gray's stability theorem asserts that ξ and ξ' are isotopic if and only if there is a contactomorphism between them which is isotopic to the identity.

It turns out that any contact structure on a 3-manifold is locally isomorphic to the "standard contact structure" on $\mathbb{R}^3$, which is the kernel of the contact form

$$\lambda = dz - y\,dx. \tag{3.1}$$

In fact any contact form is locally isomorphic to this one. The contact structure defined by (3.1) is invariant under translation in the z direction. The contact planes are horizontal along the x axis, but rotate as one moves in the y direction; the total rotation angle as y goes from $-\infty$ to $+\infty$ is π. The Reeb vector field associated to λ is simply

$$R = \partial_z.$$

[1] Strictly speaking this is a "co-oriented contact structure". There is also a notion of unoriented contact structure which is only locally the kernel of a contact form.

In particular there are no Reeb orbits. Of course this does not contradict the Weinstein conjecture since $\mathbb{R}^3$ is not compact, but it does indicate that any proof of the Weinstein conjecture will need to use global considerations.

3.2. **Some classification results.** It turns out that every closed oriented three-manifold admits a contact structure. In fact, two much stronger results were proved by Eliashberg and Giroux, which we describe here and in §3.4 respectively.

To state Eliashberg's result, define an *overtwisted disk* in a contact 3-manifold (Y, ξ) to be a smoothly embedded closed disk $D \subset Y$ such that for each $y \in \partial D$ we have $T_y\partial D \subset \xi_y$ but $T_yD \neq \xi_y$. A contact 3-manifold is called *overtwisted* if it contains an overtwisted disk; otherwise it is called *tight*.

An example of an overtwisted contact structure on $\mathbb{R}^3$ is the kernel of the contact form given in cylindrical coordinates by

$$\lambda = \cos r\, dz + r \sin r\, d\theta.$$

This contact structure is invariant under translation in the z direction. Here the contact planes are vertical on the z axis, but rotate infinitely many times as one moves out from the z axis along a horizontal ray. An overtwisted disk can be obtained by taking a horizontal disk of radius r, where r is a positive number such that $\tan r = -r$, and then bending it down slightly so that it is not horizontal at the boundary.

On the other hand, the standard contact structure defined by (3.1) is tight, although this is less trivial to prove.

Theorem 3.2 (Eliashberg [10]). *For any closed oriented 3-manifold Y, the inclusion of the set of overtwisted contact structures on Y into the set of oriented 2-plane fields on Y is a homotopy equivalence.*

A detailed exposition of the proof may be found in [14]. In particular, this theorem implies that overtwisted contact structures modulo isotopy are equivalent to homotopy classes of oriented 2-plane fields. The latter always exist, because an oriented 3-manifold has trivial tangent bundle. (As far as I know, in higher dimensions it is not completely known which manifolds admit contact structures. For example it was only in 2002 that odd-dimensional tori were shown to admit contact structures, by Bourgeois [4].)

The classification of tight contact structures is more complicated, and a subject of ongoing research. In particular, the map from tight contact structures to homotopy classes of oriented 2-plane fields is in general neither injective nor surjective. We will see in §6.6 below that a given homotopy class of oriented 2-plane fields may contain more than one isotopy class of tight contact structures. On the other hand, Colin-Giroux-Honda have shown that on any given closed oriented 3-manifold there are only finitely many homotopy classes of oriented 2-plane fields that contain tight contact structures [8] (even though there are always infinitely many homotopy classes of oriented 2-plane fields, see Remark 8.3 below).

3.3. **Simple examples.** To get some examples of contact forms on 3-manifolds, recall from §1.3 that any star-shaped hypersurface Y in $\mathbb{R}^4$ has a contact form obtained by restricting the form

$$\lambda = \frac{1}{2}\sum_{i=1}^{2}(x_i dy_i - y_i dx_i)$$

to Y. The resulting contact structure on S^3 is tight, and a theorem of Eliashberg asserts that all tight contact structures on S^3 are isotopic to this one. If Y is the unit sphere $|z_1|^2 + |z_2|^2 = 1$ (here we are writing $z_k = x_k + iy_k$), then the Reeb vector field is tangent to the fibers of the Hopf fibration $S^3 \to S^2$. In particular, there is a family of Reeb orbits parametrized by S^2. This is in fact a special case of the circle bundle example that we considered previously, and in particular it is not "generic". If one replaces the sphere with the ellipsoid

$$\frac{|z_1|^2}{a_1^2} + \frac{|z_2|^2}{a_2^2} = 1$$

where a_1, a_2 are real numbers with a_1/a_2 irrational, then there are just two embedded Reeb orbits, namely the circles $z_1 = 0$ and $z_2 = 0$. The Weinstein conjecture says that we can not further modify the contact form to eliminate these two remaining orbits without introducing new ones.

Remark 3.3. Examples of contact forms on closed three-manifolds with only two embedded Reeb orbits (at least in which all Reeb orbits are nondegenerate, see §6.2) are rare: a generalization of the Weinstein conjecture proved in [25] asserts that these only exist on spheres and lens spaces.

We next consider some examples of contact forms on the 3-torus T^3. Write $T^3 = (\mathbb{R}/2\pi\mathbb{Z})^3$ with coordinates x, y, z. For each positive integer n, define a contact form λ_n on T^3 by

$$\lambda_n := \cos(nz)dx + \sin(nz)dy. \tag{3.2}$$

The associated Reeb vector field is given by

$$R_n = \cos(nz)\partial_x + \sin(nz)\partial_y. \tag{3.3}$$

We can regard T^3 as a T^2-bundle over S^1, where z is the coordinate on S^1 and x, y are the fiber coordinates. The Reeb vector field is then a linear vector field tangent to each fiber, whose slope rotates as z increases. Whenever the slope is rational, the fiber is foliated by Reeb orbits. That is, there is a circle of embedded Reeb orbits for each z such that $\tan(nz) \in \mathbb{Q} \cup \{\infty\}$. Again this is a non-generic situation, and it turns out for any such z, one can perturb the contact form so that the corresponding circle of Reeb orbits disintegrates into just two Reeb orbits. We will see in §6 below that the contact structures $\xi_n := \operatorname{Ker}(\lambda_n)$ are pairwise non-contactomorphic, although they all represent the same homotopy class of oriented 2-plane fields. Also, they are all tight. Note that ξ_1 is isomorphic to the canonical contact structure on the unit contangent bundle of T^2 with a flat metric.

Remark 3.4. One can use the above example to illustrate that Reeb vector fields are somewhat special. To see how, note first that if R is the Reeb vector field associated to a contact form λ on a 3-manifold, then R is volume preserving with respect to the volume form $\lambda \wedge d\lambda$, because the definition of Reeb vector field implies that the Lie derivative $\mathcal{L}_R\lambda = 0$. But not every volume-preserving vector field is a Reeb vector field. In fact the Reeb vector field (3.3) can easily be perturbed to a volume-preserving vector field with no closed orbits (which of course cannot be a Reeb vector field). Namely, consider the vector field

$$V = \cos(nz)\partial_x + (\sin(nz) + \varepsilon_1)\partial_y + \varepsilon_2\partial_z$$

where $\varepsilon_2 \neq 0$ and $\varepsilon_1/\varepsilon_2$ is irrational. Suppose that $(x(t), y(t), z(t))$ is a trajectory of V. Then $z(t) = z(0) + \varepsilon_2 t$, so if this is a closed orbit then the period must be $2\pi k/\varepsilon_2$ for some positive integer k. But the path $(x(t), y(t))$ moves in the sum of a circular motion with period $2\pi/(n\varepsilon_2)$ and a vertical motion of speed ε_1, so we have $(x(2\pi k/\varepsilon_2), y(2\pi k/\varepsilon_2)) = (0, 2\pi k\varepsilon_1/\varepsilon_2)$. Thus our assumption that $\varepsilon_1/\varepsilon_2$ is irrational implies that there is no closed orbit.

3.4. **Open book decompositions.** There is a useful classification of all contact three-manifolds, not just the overtwisted ones, in terms of open book decompositions.

Let Σ be a compact oriented connected surface with nonempty boundary. Let $\phi : \Sigma \to \Sigma$ be an orientation-preserving diffeomorphism which is the identity near the boundary. One can then define a closed oriented three-manifold

$$\begin{aligned} Y_\phi &:= [0,1] \times \Sigma / \sim, \\ (1,x) &\sim (0,\phi(x)) \quad \forall x \in \Sigma, \\ (t,x) &\sim (t',x) \qquad \forall x \in \partial\Sigma,\ t,t' \in [0,1], \end{aligned}$$

called an *open book*. The image of a set $\{t\} \times \Sigma$ in Y_ϕ is called a *page*. The boundaries of the different pages are all identified with each other, to an oriented link in Y_ϕ called the *binding*. The map ϕ is called the *monodromy* of the open book. An *open book decomposition* of a closed oriented three-manifold Y is a diffeomorphism of Y with an open book Y_ϕ as above.

Definition 3.5. A contact structure ξ on a closed three-manifold Y is *compatible* with, or *supported* by, an open book decomposition $Y \simeq Y_\phi$ if ξ is isotopic to a contact structure given by a contact form λ such that:

- The Reeb vector field is tangent to the binding (oriented positively).
- The Reeb vector field is transverse to the interior of each page, intersecting positively.

Example 3.6. The standard contact structure on the unit sphere described in §3.3 is compatible with an open book decomposition of S^3 in which Σ is a disk and ϕ is the identity map.

A classical theorem of Alexander asserts that every closed oriented connected three-manifold has an open book decomposition. A short argument by Thurston-Winkelnkemper shows that every open book decomposition has a compatible contact structure. This is a nice way to show that all closed oriented three-manifolds admit contact structures. In fact the compatible contact structure is determined up to isotopy by the open book. Moreover, a theorem of Giroux asserts that two open books determine isotopic contact structures if and only if they are related to each other by "positive stabilizations". For more about this see [16, 13].

3.5. **Previous results on the 3d Weinstein conjecture.** There is a long history of work proving the Weinstein conjecture for various classes of contact three-manifolds. For example: Hofer proved that for all contact three-manifolds (Y, ξ) in which ξ is overtwisted or $\pi_2(Y) \neq 0$, for all contact forms with kernel ξ, there exists a contractible Reeb orbit [18]. Abbas-Cieliebak-Hofer proved the Weinstein conjecture for all contact three-manifolds supported by an open book in which the pages have genus zero [1]. Colin-Honda used linearized contact homology (see §6 below) to prove the Weinstein conjecture for contact three-manifolds supported by open books in which the monodromy is periodic, and in many cases where the monodromy is pseudo-Anosov [9]. In fact, for many contact structures supported by open books with pseudo-Anosov monodromy, they proved a much stronger statement: that for any contact form, there are infinitely many free homotopy classes of loops that contain an embedded Reeb orbit.

4. Some strategies for proving the Weinstein conjecture

One naive strategy for proving the Weinstein conjecture might be to try the following:

- Define some kind of count of Reeb orbits with appropriate signs.
- Show that this count is a topological invariant.
- Calculate this invariant and show that it is nonzero.

This strategy is too simple for at least two reasons. First of all, often there are actually infinitely many embedded Reeb orbits, so it is not clear how to obtain a well-defined count of them. Second, even if the above difficulty can be overcome, the signed count might be zero, despite the existence of some Reeb orbits with opposite signs that cannot be eliminated. In general, as one deforms the contact form, pairs of Reeb orbits can be created or destroyed (more complicated bifurcations such as period-doubling are also possible), and one needs some way of keeping track of when this can happen.

A more refined strategy, which avoids the above two problems, is as follows:

- Define some kind of chain complex which is generated by Reeb orbits, such that there are differentials between pairs of Reeb orbits that can potentially be destroyed in a bifurcation.
- Show that the homology of this chain complex is a topological invariant.
- Compute this homology and show that it is nontrivial.

It turns out that there does exist a chain complex along these lines which is sufficient to prove the Weinstein conjecture. The origin of this chain complex is a bit of a long story, which we will now attempt to explain.

5. Prototype for a chain complex: Morse homology

The prototype for the type of chain complex we want to consider is Morse homology, which we now review. We will not give any proofs, as these require a fair bit of analysis; a good reference for this is [38]. There is also an interesting history of the development of the Morse complex, for which we refer the reader to [3].

5.1. **Morse functions.** Let X be an n-dimensional closed smooth manifold and let $f : X \to \mathbb{R}$ be a smooth function. A *critical point* of f is a point $p \in X$ such that $0 = df_p : T_pX \to \mathbb{R}$. The basic goal of Morse theory is to relate the critical points of f to the topology of X. A first question is, on a given X, what is the minimum number of critical points that a smooth function f can have? Of course f must have at least two critical points, namely a minimum and a maximum, since we are assuming that X is compact and has no boundary. But usually there must be more.

To make this question easier, we can require that the critical points of f be “generic”, in a sense which we now specify. If $p \in X$ is a critical point of $f : X \to \mathbb{R}$, define the *Hessian*

$$H(f,p) : T_pX \otimes T_pX \longrightarrow \mathbb{R} \tag{5.1}$$

as follows. Let $\psi : X \to T^*X$ denote the section corresponding to *df*. Then $H(f,p)$ is the composition

$$T_pX \xrightarrow{d\psi_p} T_{(p,0)}T^*X = T_pX \oplus T_p^*X \xrightarrow{\pi} T_p^*X,$$

where π denotes the projection onto the second factor. To be more explicit, if $(x_1,\ldots,x_n)$ are local coordinates on X centered at p, then

$$H(f,p)\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right) = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

In particular, $H(f,p)$ is a symmetric bilinear form. If one chooses a metric on X, then the Hessian can be identified with a self-adjoint operator

$$H(f,p) : T_pX \longrightarrow T_pX. \tag{5.2}$$

The critical point p is said to be *nondegenerate* if the Hessian pairing (5.1) is nondegenerate, or equivalently the Hessian operator (5.2) does not have zero as an eigenvalue, or equivalently the graph of *df* in T^*X is transverse to the zero section at $(p,0)$. In particular, nondegenerate critical points are isolated in X.

We say that f is a *Morse function* if all of its critical points are nondegenerate. One can show that "generic" smooth functions are Morse. More precisely, the set of Morse functions is open and dense in the set of all smooth functions $f : X \to \mathbb{R}$, with the C^∞ topology. We can now ask, what is the minimum possible number of critical points of a Morse function on X?

One could start by counting the critical points with signs. If p is a nondegenerate critical point of f, define the *index* of p, denoted by $\operatorname{ind}(p)$, to be the maximal dimension of a subspace on which the Hessian pairing (5.1) is negative definite, or equivalently the number of negative eigenvalues, counted with multiplicity, of the Hessian operator (5.2). For example, a local minimum has index 0, and a local maximum has index n. More generally, the Morse lemma asserts that if p is a nondegenerate critical point of index i, then there are local coordinates $(x_1,\ldots,x_n)$ centered at p in which

$$f = f(p) - x_1^2 - \cdots - x_i^2 + x_{i+1}^2 + \cdots + x_n^2. \tag{5.3}$$

It turns out that the appropriate sign with which to count an index i critical point is $(-1)^i$. In fact, if f is a Morse function on X, and if c_i denotes the number of critical points of index i, then

$$\sum_{i=0}^{n} (-1)^i c_i = \chi(X).$$

One can prove this by choosing a metric on X, applying the Poincaré-Hopf index theorem to the resulting gradient vector field ∇f, and checking that the sign of the zero of ∇f at an index i critical point is $(-1)^i$. In particular, the number of critical points is at least $|\chi(X)|$. But if $\chi(X) = 0$ then this tells us nothing.

5.2. **The classical approach.** One can obtain better lower bounds on the number of critical points of a Morse function $f : X \to \mathbb{R}$ by bringing in more dynamical information. The classical approach to Morse theory is to define, for a real number a, the subset

$$X_a := \{x \in X \mid f(x) \le a\}$$

and study how the topology of X_a changes as a increases. If a is a regular value of f then X_a is a manifold with boundary. One can further show that if $a < b$ are regular values of f, then X_b is obtained from X_a by attaching one i-handle for each index i critical point in $f^{-1}[a, b]$. One can relate the resulting handle decomposition of X to the homology of X. If b_i denotes the rank of $H_i(X)$, then one finds that $c_i \ge b_i$, and more generally one obtains the *Morse inequalities*:

$$c_i - c_{i-1} + c_{i-2} - \cdots + (-1)^i c_0 \ge b_i - b_{i-1} + b_{i-2} - \cdots + (-1)^i b_0. \tag{5.4}$$

This approach can be extended to the energy functional on the (infinite dimensional) loop space of a Riemannian manifold, in order to relate closed geodesics to the topology of the loop space. For more about classical Morse theory see eg [34].

5.3. **The Morse complex.** A more modern approach, which is a better model for the particular infinite-dimensional generalizations that we will need later, is to consider "gradient flow lines" between critical points and package these into a chain complex. Here is how this works.

Choose a metric g on X and use it to define the gradient vector field ∇f. If p and q are critical points, a downward *gradient flow line* from p to q is a map $\gamma : \mathbb{R} \to X$ such that

$$\frac{d\gamma(s)}{ds} = \nabla f(\gamma(s))$$

and $\lim_{s\to+\infty} \gamma(s) = p$ and $\lim_{s\to-\infty} \gamma(s) = q$. Let $\mathcal{M}(p, q)$ denote the set of downward gradient flow lines from p to q.

If the metric g is generic, then $\mathcal{M}(p, q)$ is naturally a manifold of dimension

$$\dim \mathcal{M}(p, q) = \operatorname{ind}(p) - \operatorname{ind}(q). \tag{5.5}$$

To see why, for $s \in \mathbb{R}$ let $\psi_s : X \to X$ denote the time s flow of the vector field ∇f. The *descending manifold* of a critical point p is the set

$$\mathcal{D}(p) := \{x \in X \mid \lim_{s\to+\infty} \psi_s(x) = p\}.$$

Informally, this is the set of points in X that "can be reached by downward gradient flow starting at p". Similarly, the *ascending manifold* of a critical point q is defined to be

$$\mathcal{A}(q) := \{x \in X \mid \lim_{s\to-\infty} \psi_s(x) = q\}.$$

One can show that the descending manifold $\mathcal{D}(p)$ is a smoothly embedded open ball in X of dimension $\operatorname{ind}(p)$, and also $T_p\mathcal{D}(p)$ is the negative eigenspace of the Hessian (5.2). For example, in the local coordinates given by the Morse Lemma in which (5.3) holds, if one chooses the metric to be Euclidean near the origin, then near the origin $\mathcal{D}(p)$ agrees with the linear subspace $(x_{i+1} = \cdots = x_n = 0)$. Likewise, $\mathcal{A}(q)$ is a smoothly embedded open ball of dimension $n - \operatorname{ind}(q)$.

It follows from the definitions that there is a bijection

$$\mathcal{M}(p, q) \xrightarrow{\simeq} \mathcal{D}(p) \cap \mathcal{A}(q)$$

sending a flow line γ to the point $\gamma(0) \in X$. Consequently, if $\mathcal{D}(p)$ is transverse to $\mathcal{A}(q)$, then equation (5.5) follows by dimension counting. The pair (f,g) is said to be *Morse-Smale* if $\mathcal{D}(p)$ is transverse to $\mathcal{A}(q)$ for every pair of critical points p,q. One can show that for a given Morse function f, for generic metrics g the pair (f,g) is Morse-Smale. Henceforth we always assume that this condition holds. Observe also that $\mathbb{R}$ acts on $\mathcal{M}(p,q)$ by precomposition with translations, and if $p \neq q$ then this action is free, so that

$$\dim(\mathcal{M}(p,q)/\mathbb{R}) = \operatorname{ind}(p) - \operatorname{ind}(q) - 1.$$

In particular, if $\operatorname{ind}(q) \geq \operatorname{ind}(p)$ then $\mathcal{M}(p,q)$ is empty (except when $p = q$), and if $\operatorname{ind}(q) = \operatorname{ind}(p) - 1$ then $\mathcal{M}(p,q)/\mathbb{R}$ is discrete.

We now define the *Morse complex* $C_*^{\text{Morse}}(X,f,g)$ as follows. The chain module in degree i is the free $\mathbb{Z}$-module generated by the index i critical points:

$$C_i^{\text{Morse}}(X,f,g) := \mathbb{Z}\{p \in X \mid df_p = 0,\ \operatorname{ind}(p) = i\}.$$

The differential $\partial : C_i^{\text{Morse}}(X,f,g) \to C_{i-1}^{\text{Morse}}(X,f,g)$ is defined by counting gradient flow lines as follows: if p is an index i critical point, then

$$\partial p := \sum_{\operatorname{ind}(p)-\operatorname{ind}(q)=1} \# \frac{\mathcal{M}(p,q)}{\mathbb{R}} \cdot q.$$

Here '#' denotes a signed count. We will not say more about what these signs are, except to note that the signs depend on choices of orientations of the descending manifolds of the critical points, but the chain complexes resulting from different sign choices are canonically isomorphic to each other.

To show that this chain complex is well-defined, one must prove that ∂ is well-defined, ie $\mathcal{M}(p,q)/\mathbb{R}$ is finite whenever $\operatorname{ind}(p) - \operatorname{ind}(q) = 1$, and one must also prove that $\partial^2 = 0$. The first step is to prove a compactness theorem which asserts that given critical points $p \neq q$, any sequence $\{\gamma_n\}_{n=1}^{\infty}$ in $\mathcal{M}(p,q)/\mathbb{R}$ has a subsequence which converges in an appropriate sense to a "k-times broken flow line" from p to q. This is a tuple $(\widehat{\gamma}_0, \ldots, \widehat{\gamma}_k)$ for some $k \geq 0$, such that there are critical points $p = r_0, r_1, \ldots, r_{k+1} = q$ for which $\widehat{\gamma}_i \in \mathcal{M}(r_i, r_{i+1})/\mathbb{R}$ is a nonconstant flow line. If $\operatorname{ind}(p) - \operatorname{ind}(q) = 1$, then the Morse-Smale condition implies that there are no k-times broken flow lines with $k > 0$, so $\mathcal{M}(p,q)$ is compact, and hence finite, so ∂ is well-defined.

To prove that $\partial^2 = 0$, one shows that if $\operatorname{ind}(p) - \operatorname{ind}(q) = 2$, then $\mathcal{M}(p,q)/\mathbb{R}$ has a compactification to a compact oriented 1-manifold with boundary $\overline{\mathcal{M}(p,q)/\mathbb{R}}$, whose boundary as an oriented manifold is

$$\partial\overline{\mathcal{M}(p,q)/\mathbb{R}} = \bigcup_{\operatorname{ind}(p)-\operatorname{ind}(r)=1} \frac{\mathcal{M}(p,r)}{\mathbb{R}} \times \frac{\mathcal{M}(r,q)}{\mathbb{R}}. \tag{5.6}$$

Granted (5.6), counting the points on both sides with signs gives

$$\#\partial\overline{\mathcal{M}(p,q)/\mathbb{R}} = \sum_{\operatorname{ind}(p)-\operatorname{ind}(r)=1} \langle \partial p, r\rangle \langle \partial r, q\rangle. \tag{5.7}$$

Here $\langle \partial p, r\rangle \in \mathbb{Z}$ denotes the coefficient of r in ∂p. Thus the right hand side of (5.7) is, by definition, the coefficient $\langle \partial^2 p, q\rangle$. On the other hand a fundamental theorem of differential topology, namely that a compact oriented 1-manifold has zero boundary points counted with signs, implies that the left hand side of (5.7) is zero.

To justify (5.6), two additional ingredients are needed. First, there is a gluing theorem asserting that each broken flow line $(\widehat{\gamma}_0, \widehat{\gamma}_1)$ with $\widehat{\gamma}_0 \in \mathcal{M}(p,r)/\mathbb{R}$ and $\widehat{\gamma}_1 \in \mathcal{M}(r,q)/\mathbb{R}$ can be patched to an unbroken flow line in a unique end of $\mathcal{M}(p,q)/\mathbb{R}$. Second, one has to show that $\mathcal{M}(p,q)/\mathbb{R}$ can be oriented so that the orientations on both sides of (5.6) agree.

5.4. **Morse homology.** The *Morse homology* $H_*^{\text{Morse}}(X,f,g)$ is the homology of the above chain complex.

Example 5.1. Consider a Morse function $f : S^2 \to \mathbb{R}$ with two index 2 critical points x_1, x_2, one index 1 critical point y, and one index 0 critical point z. One can visualize f as the height function on a "heart-shaped" sphere embedded in $\mathbb{R}^3$. Pick any metric g on S^2; it turns out that (f,g) will automatically be Morse-Smale in this example. There is (up to reparametrization) a unique downward gradient flow line from each x_i to y. There are two gradient flow lines from y to z. The latter turn out to have opposite signs, and so for suitable orientation choices the Morse complex is given by

$$C_2^{\text{Morse}} = \mathbb{Z}\{x_1, x_2\}, \quad C_1^{\text{Morse}} = \mathbb{Z}\{y\}, \quad C_0^{\text{Morse}} = \mathbb{Z}\{z\},$$
$$\partial x_1 = y, \quad \partial x_2 = -y, \quad \partial y = 0.$$

Thus $H_2^{\text{Morse}} \simeq \mathbb{Z}$, generated by $x_1 + x_2$; $H_1^{\text{Morse}} = 0$; and $H_0^{\text{Morse}} = \mathbb{Z}$, generated by z.

The above example illustrates a fundamental theorem in the subject:

Theorem 5.2. *There is a canonical isomorphism between Morse homology and singular homology,*

$$H_*^{\text{Morse}}(X,f,g) \simeq H_*(X).$$

An immediate corollary is that there must be enough critical points to generate a chain complex whose homology is $H_*(X)$. For example, c_i must be greater than or equal to the rank of $H_i(X)$, and more generally it is not hard to deduce the Morse inequalities (5.4).

A brief outline of one proof of Theorem 5.2 (which requires considerable work to justify) is as follows. If p is a critical point, then the descending manifold $\mathcal{D}(p)$ has a compactification $\overline{\mathcal{D}(p)}$ using broken flow lines, with a natural map $e : \overline{\mathcal{D}(p)} \to X$ extending the inclusion $\mathcal{D}(p) \to X$, such that $e(\overline{\mathcal{D}(p)} \setminus \mathcal{D}(p))$ is contained in the union of the descending manifolds of the critical points with index less than that of p. Furthermore, $\overline{\mathcal{D}(p)}$ is homeomorphic to a closed ball of dimension $\operatorname{ind}(p)$, so that the balls $\overline{\mathcal{D}(p)}$ together with the maps e give X the structure of a CW complex with one cell for each critical point. Finally, the resulting cellular chain complex agrees with the Morse complex.

5.5. **Continuation maps.** We will later construct analogues of Morse homology which generally do not have interpretations in terms of previously known invariants such as singular homology. What is most important here as a model for these later constructions is that the Morse homology $H_*^{\text{Morse}}(X,f,g)$ is a topological invariant of X which does not depend on f or g. One can prove this directly, without making the comparison with singular homology, as follows.

Let (f_0, g_0) and (f_1, g_1) be two Morse-Smale pairs. Let $\{(f_s, g_s) \mid s \in \mathbb{R}\}$ be a smooth family of pairs of functions and metrics on X such that $(f_s, g_s) = (f_0, g_0)$

for $s \le 0$ and $(f_s, g_s) = (f_1, g_1)$ for $s \ge 1$. We do not (and in general cannot) assume that the pair (f_s, g_s) is Morse-Smale for all s. One now defines a map

$$\Phi : C_*^{\mathrm{Morse}}(X, f_1, g_1) \longrightarrow C_*^{\mathrm{Morse}}(X, f_0, g_0),$$

called the *continuation map*, as follows. If p_0 is an index i critical point of f_0 and p_1 is an index i critical point of f_1, then $\langle \Phi(p_1), p_0 \rangle$ is a signed count of maps $\gamma : \mathbb{R} \to X$ satisfying

$$\frac{d\gamma(s)}{ds} = \nabla f_s(\gamma(s))$$

and $\lim_{s \to -\infty} \gamma(s) = p_0$ and $\lim_{s \to +\infty} \gamma(s) = p_1$. Here the gradient of f_s is computed using the metric g_s. Similarly to the proof that ∂ is well defined and $\partial^2 = 0$, one can show that if the family of metrics $\{g_s\}$ is generic then Φ is a well-defined chain map. For example, if the family $\{(f_s, g_s)\}$ is constant then Φ is the identity map. Finally, related considerations show that if

$$\Phi' : C_*^{\mathrm{Morse}}(X, f_0, g_0) \longrightarrow C_*^{\mathrm{Morse}}(X, f_1, g_1)$$

is the continuation map induced by a generic homotopy in the reverse direction, then the compositions $\Phi\Phi'$ and $\Phi'\Phi$ are chain homotopic to the respective identity maps. Thus Φ induces an isomorphism on homology.

5.6. **Floer homology.** Floer homology is a kind of Morse homology for certain functionals on infinite dimensional manifolds. Roughly speaking, the Floer homology in each case is the homology of a chain complex which is generated by critical points of the functional, which are usually geometric objects of interest, and whose differential counts "gradient flow lines", which are solutions of an elliptic partial differential equation. There are now a number of different flavors of Floer theory. The versions originally introduced by Floer are Floer theory for symplectomorphisms (which is generated by fixed points of a symplectomorphism and whose differential counts certain holomorphic cylinders), more generally Floer theory for Lagrangian intersections (which is generated by intersection points of two Lagrangian submanifolds of a symplectic manifold and whose differential counts certain holomorphic strips), and instanton Floer homology of a three-manifold Y (which is generated by flat $SU(2)$-connections on Y and whose differential counts anti-self-dual connections on $\mathbb{R} \times Y$). These provided inspiration for the definitions of the versions of Floer theory that are relevant to the Weinstein conjecture. But to keep a long story from getting even longer, we now jump ahead to the latter.

6. First attempt at a chain complex: cylindrical contact homology

Let Y be a closed oriented 3-manifold and let λ be a contact form on Y. We would like to define an analogue of the Morse complex which is generated by Reeb orbits, and whose differential counts an appropriate notion of "flow lines" between them. Although the analogy with Morse homology breaks down slightly, this idea leads naturally to the definition of cylindrical homology. This theory can be used to prove the Weinstein conjecture in many cases. Although cylindrical contact homology can be defined for contact manifolds of any odd dimension, for definiteness we stick to the three-dimensional case.

6.1. **The symplectic action functional.** We would like to find a functional on the (smooth) free loop space whose critical points are Reeb orbits, and try to define an analogue of Morse homology for this functional.

There is indeed a natural functional $\mathcal{A}$ along these lines, called the *symplectic action*. If $\gamma : S^1 \to Y$ is a smooth loop, the symplectic action of γ is defined simply by integrating the contact form along γ, namely

$$\mathcal{A}(\gamma) := \int_{S^1} \gamma^*\lambda = \int_{S^1} \lambda(\gamma'(t))dt$$

where $\gamma'(t) := d\gamma(t)/dt$.

The symplectic action is a smooth function on the Frechet manifold $\mathcal{X}$ of smooth maps $S^1 \to Y$. Its differential can be computed as follows. The tangent space to $\mathcal{X}$ at a smooth loop $\gamma : S^1 \to Y$ is the space of smooth sections $\dot{\gamma}$ of the bundle γ^*TY over γ. We then have

$$d\mathcal{A}_\gamma(\dot{\gamma}) = \int_{S^1} d\lambda(\dot{\gamma}(t), \gamma'(t))dt. \tag{6.1}$$

It follows that γ is a critical point of $\mathcal{A}$ if and only if $\gamma'(t)$ is in the kernel of $d\lambda$, ie proportional to the Reeb vector field, for all t.

6.2. **Nondegenerate Reeb orbits.** We now explain the appropriate analogue of nondegenerate critical point in this context.

For $t \in \mathbb{R}$, let $\psi_t : Y \to Y$ denote the diffeomorphism obtained by flowing along the Reeb vector field R for time t. The definition of Reeb vector field implies that the Reeb flow preserves the contact form, ie $\mathcal{L}_R\lambda = 0$. It follows that the derivative of ψ_t is a contactomorphism. That is, for each $y \in Y$ the linear map $d\psi_t : T_yY \to T_{\psi(y)}Y$ restricts to a map on contact planes $\xi_y \to \xi_{\psi(y)}$. Moreover, the latter map preserves the symplectic forms given by $d\lambda$.

Now let $\gamma : \mathbb{R}/T\mathbb{Z} \to Y$ be a Reeb orbit. Then for each $t \in \mathbb{R}/T\mathbb{Z}$, we have a linear symplectic automorphism

$$d\psi_T : \xi_{\gamma(t)} \xrightarrow{\simeq} \xi_{\gamma(t)},$$

called the *linearized return map*. Another way to describe this map is as follows. Let D be a small embedded disk in Y centered at $\gamma(t)$ and transverse to γ, such that $T_{\gamma(t)}D = \xi_{\gamma(t)}$. For $x \in D$ close to the center, there is a unique point in D which is reached by following the Reeb flow for a time close to T. This gives a partially defined "return map" $\phi : D \to D$ which is defined near the origin. The derivative of this map at the origin is the linearized return map.

We say that the Reeb orbit γ is *nondegenerate* if the linearized return map does not have 1 as an eigenvalue. This does not depend on the choice of $t \in \mathbb{R}/T\mathbb{Z}$, because the linearized return maps for different t are conjugate to each other. If the Reeb orbit γ is nondegenerate then it is isolated, because Reeb orbits close to γ give rise to fixed points of the map ϕ, and the condition that $1 - d\phi$ is invertible at the origin implies that ϕ has no fixed points near the origin.

One can show that for a given contact structure ξ, for generic contact forms λ, all Reeb orbits are nondegenerate. We will always assume unless otherwise stated that all Reeb orbits are nondegenerate.

One can classify (nondegenerate) Reeb orbits into three types, according to the eigenvalues λ, λ^{-1} of the linearized return map:

- *elliptic*: $\lambda, \lambda^{-1} = e^{\pm 2\pi i\theta}$.

- *positive hyperbolic*: $\lambda, \lambda^{-1} > 0$.
- *negative hyperbolic*: $\lambda, \lambda^{-1} < 0$.

6.3. **Holomorphic cylinders.** The appropriate analogue of "gradient flow line" in this context is a certain kind of holomorphic cylinder. We now explain what these are.

In general, recall that a *complex structure* on an even-dimensional real vector bundle $E \to X$ is a bundle map $J : E \to E$ satisfying $J^2 = -1$. An *almost complex structure* on an even-dimensional manifold X is a complex structure J on the tangent bundle TX. A *holomorphic curve* in (X, J) is a map $u : \Sigma \to X$ where Σ is a surface with an almost complex structure j (ie a Riemann surface), and $J \circ du = du \circ j$. Two holomorphic curves $u : (\Sigma, j) \to X$ and $u' : (\Sigma', j') \to X$ are considered equivalent if there is a biholomorphic map $\phi : (\Sigma, j) \to (\Sigma', j')$ with $u = u' \circ \phi$. If u is an embedding then the equivalence class of u is determined by its image. That is, an embedded holomorphic curve in (X, J) is just a 2-dimensional submanifold $C \subset X$ such that $J(TC) = TC$.

Returning now to the situation of interest:

Definition 6.1. Let Y be a three-manifold with a contact form λ. An almost complex structure J on the 4-manifold $\mathbb{R} \times Y$ is *admissible* if:

(1) J sends ξ to itself, rotating positively with respect to the orientation given by $d\lambda$.
(2) If s denotes the $\mathbb{R}$ coordinate on $\mathbb{R} \times Y$, then $J(\partial_s) = R$.
(3) J is invariant under the $\mathbb{R}$ action on $\mathbb{R} \times Y$ that translates s.

Note that the space of such J is nonempty and contractible. Indeed, the choice of such a J is equivalent to the choice of a complex structure on the 2-plane bundle ξ over Y which rotates positively with respect to $d\lambda$. Fix an admissible almost complex structure on Y below.

Observe that if γ is an embedded Reeb orbit, then $\mathbb{R} \times \gamma$ is a holomorphic cylinder in $\mathbb{R} \times Y$. This follows from condition (2) above. More generally, we can study holomorphic curves in $\mathbb{R} \times Y$ that are asymptotic to such $\mathbb{R}$-invariant cylinders as the $\mathbb{R}$ coordinate goes to plus or minus infinity. To define what we mean by this, consider $[0, \infty) \times S^1$ or $(-\infty, 0] \times S^1$ with coordinates s, t, with the almost complex structure j sending ∂_s to ∂_t. Let $\pi_{\mathbb{R}} : \mathbb{R} \times Y \to \mathbb{R}$ and $\pi_Y : \mathbb{R} \times Y \to Y$ denote the two projections. If $u : \Sigma \to \mathbb{R} \times Y$ is a holomorphic curve and if γ is a Reeb orbit (not necessarily embedded), we define a *positive end* of u at γ to be an end of Σ which can be parametrized as $[0, \infty) \times S^1$ with the almost complex structure j as above, such that $\lim_{s\to\infty} \pi_{\mathbb{R}}(s, \cdot) = \infty$, and $\lim_{s\to\infty} \pi_Y(s, \cdot)$ is a reparametrization of γ. Likewise, a *negative end* of u at γ is an end of Σ which can be parametrized as $(-\infty, 0] \times S^1$, with the almost complex structure j as above, such that $\lim_{s\to-\infty} \pi_{\mathbb{R}}(s, \cdot) = -\infty$, and $\lim_{s\to-\infty} \pi_Y(s, \cdot)$ is a reparametrization of γ.

If γ_+ and γ_- are two Reeb orbits, define $\mathcal{M}_0(\gamma_+, \gamma_-)$ to be the set of holomorphic cylinders in $\mathbb{R} \times Y$ that have a positive end at γ_+ and a negative end at γ_-. It turns out that these holomorphic cylinders are the appropriate "gradient flow lines" from γ_+ to γ_-. Note that there is an $\mathbb{R}$ action on $\mathcal{M}_0(\gamma_+, \gamma_-)$ given by translating the $\mathbb{R}$ coordinate on $\mathbb{R} \times Y$. This action is free except on the $\mathbb{R}$-invariant cylinders $\mathbb{R} \times \gamma$ in $\mathcal{M}_0(\gamma, \gamma)$. (This is not to be confused with the $\mathbb{R} \times S^1$ action on the set of holomorphic maps $\mathbb{R} \times S^1 \to \mathbb{R} \times Y$ given by compositions with translations of the domain, which we have already modded out by in our definition of holomorphic curve.)

We will not explain the analogy between holomorphic cylinders and gradient flow lines, except to note the following fundamental lemma:

Lemma 6.2. *If there exists a holomorphic cylinder $u \in \mathcal{M}_0(\gamma_+, \gamma_-)$, then*

$$\mathcal{A}(\gamma_+) \geq \mathcal{A}(\gamma_-).$$

Equality holds if and only if $\gamma_+ = \gamma_-$ and the image of u is an $\mathbb{R}$-invariant cylinder.

Proof. Let $u : \mathbb{R} \times S^1 \to \mathbb{R} \times Y$ be a holomorphic cylinder in $\mathcal{M}_0(\gamma_+, \gamma_-)$. By Stokes' theorem,

$$\mathcal{A}(\gamma_+) - \mathcal{A}(\gamma_-) = \int_{\mathbb{R}\times S^1} u^* d\lambda.$$

By the definition of admissible almost complex structure, $u^* d\lambda \geq 0$ at each point in $\mathbb{R} \times S^1$, with equality only where u is tangent to $\mathbb{R}$ cross the Reeb direction. □

6.4. **The chain complex.** We can now define an analogue of the Morse complex in this setting. We will give a "quick and dirty" definition to save space; for the more general context into which this definition fits, see [11].

To start, for reasons we will explain below, one must discard certain "bad" Reeb orbits for the construction to work:

Definition 6.3. A Reeb orbit γ is said to be *bad* if it is the k-fold iterate of a negative hyperbolic orbit with k even. Otherwise γ is said to be *good*.

Now fix $\Gamma \in H_1(Y)$. Define $CC(Y, \lambda, \Gamma)$ to be the free $\mathbb{Q}$-module generated by the good Reeb orbits γ representing the homology class Γ. One then defines a differential

$$\partial : CC(Y, \lambda, \Gamma) \longrightarrow CC(Y, \lambda, \Gamma)$$

as follows. Fix a generic admissible almost complex structure J on $\mathbb{R} \times Y$. If γ_+ is a good Reeb orbit, then

$$\partial\gamma_+ := \sum_{\gamma_-} k_{\gamma_-} n(\gamma_+, \gamma_-)\gamma_-. \tag{6.2}$$

Here the sum is over good Reeb orbits γ_-, and k_γ denotes the unique positive integer such that γ is the k_γ-fold iterate of an embedded Reeb orbit. Meanwhile, $n(\gamma_+, \gamma_-) \in \mathbb{Q}$ is a signed count of holomorphic cylinders in $\mathcal{M}(\gamma_+, \gamma_-)/\mathbb{R}$ that live in zero-dimensional moduli spaces. Multiply covered cylinders are counted with weight ±1 divided by the covering multiplicity. The homology of this chain complex, when defined (see below), is called *cylindrical contact homology*, and we denote it by $CH(Y, \xi, \Gamma)$.

The following is a special case of a result to be proved in [7], see [9, §3.2] for the statement, asserting that a more general theory called "linearized contact homology" is well-defined.

Theorem 6.4. *Suppose there are no contractible Reeb orbits. Then ∂ is well-defined*[2]*, $\partial^2 = 0$, and the homology $CH(Y, \xi, \Gamma)$ depends only on Y, the contact*

[2] The expert reader may worry that even for generic J, multiply covered holomorphic cylinders might have smaller index than the cylinders that they cover, making ∂ undefined. It turns out that this does not happen for holomorphic cylinders in the symplectization of a contact 3-manifold. However this is an issue in defining the continuation maps and chain homotopies needed to prove the invariance statement in Theorem 6.4.

structure ξ, and the homology class Γ, and not on the contact form λ or almost complex structure J.

A few comments are in order. First, the factors of k_γ in (6.2) are needed to make $\partial^2 = 0$ work, because when one glues two holomorphic cylinders along a Reeb orbit γ which is the k-fold iterate of an embedded Reeb orbit, there are k different ways to glue. This is also why bad Reeb orbits need to be discarded: these k different gluings all have the same sign when γ is good, but have cancelling signs when γ is bad. Finally, the assumption that there are no contractible Reeb orbits ensures that the necessary compactness arguments go through, by ruling out bubbling off of holomorphic discs.

6.5. **The index.** Unlike Morse homology, cylindrical contact homology is not $\mathbb{Z}$-graded. Rather, it is relatively $\mathbb{Z}/d(2c_1(\xi))$-graded, where $d(2c_1(\xi))$ denotes the divisibility of $2c_1(\xi)$ in $H^2(Y;\mathbb{Z})$ mod torsion. This means that any two generators γ_+ and γ_- have a well-defined relative grading, which can be regarded as the grading difference between γ_+ and γ_-, and which is an element of $\mathbb{Z}/d(2c_1(\xi))$. In this sense the differential ∂ decreases the grading by 1. The reason why there is no absolute grading analogous to the Morse index is that the analogue of the Hessian in this setting has infinitely many negative and infinitely many positive eigenvalues. Nonetheless it still makes sense to define the relative grading of γ_+ and γ_- to be the expected dimension of the moduli space of holomorphic cylinders $\mathcal{M}_0(\gamma_+, \gamma_-)$ that represent some relative homology class Z. This is given by a topological formula which we will not state here. It is only defined modulo $d(2c_1(\xi))$, because if Z' is a different relative homology class of cylinder then the corresponding expected dimensions differ by $\langle 2c_1(\xi), Z - Z'\rangle$.

There is also a canonical absolute $\mathbb{Z}/2$-grading: a Reeb orbit has odd grading if it is positive hyperbolic, and even grading if it is elliptic or negative hyperbolic. The differential ∂ also has degree -1 with respect to this $\mathbb{Z}/2$-grading.

6.6. **Examples.** (1) Consider the contact form λ_n on T^3 defined in (3.2). Recall that all Reeb orbits represent homology classes of the form $(a, b, 0) \in H_1(T^3)$ with $(a, b) \neq (0, 0)$. As a consequence, the cylindrical contact homology $CH_*(T^3, \xi_n, \Gamma)$ is nonzero only for Γ of this form. Fix such a class $\Gamma = (a, b, 0)$. All Reeb orbits γ in the homology class Γ have symplectic action

$$\mathcal{A}(\gamma) = 2\pi\sqrt{a^2 + b^2}.$$

So by Lemma 6.2, there are no non-$\mathbb{R}$-invariant holomorphic cylinders between them. Now the cylindrical contact homology is not yet defined because λ_n is a "Morse-Bott" contact form whose Reeb orbits are not isolated but rather appear in one-parameter families. But one can show, see [5], that one can perturb λ_n to a contact form λ_n' such that each of the n circles of Reeb orbits in the homology class Γ splits into two Reeb orbits (one elliptic and one positive hyperbolic), there are no other Reeb orbits in the class Γ, and the differential on $CC_*(Y, \lambda_n', \Gamma)$ vanishes for any choice of admissible almost complex structure J. We conclude that the $\mathbb{Z}/2$-graded cylindrical contact homology is given by

$$CH_{\text{even}}(Y, \xi_n, (a, b, 0)) \simeq CH_{\text{odd}}(Y, \xi_n, (a, b, 0)) \simeq \mathbb{Q}^n.$$

It now follows from Theorem 6.4 that the different contact structures ξ_n are pairwise non-contactomorphic. Also, they all satisfy the Weinstein conjecture. Because for any contact form λ with $\xi_n = \mathrm{Ker}(\lambda)$, either there is no contractible Reeb orbit, in which case the cylindrical contact homology is well-defined and nonzero, or else there is a contractible Reeb orbit. In fact there is a generalization of cylindrical contact homology, called *linearized contact homology*, which is sometimes defined even when there are contractible Reeb orbits, by adding certain correction terms to the cylindrical contact homology differential. Linearized contact homology can be used in this example to prove that there is always a Reeb orbit in the homology class $(a, b, 0)$.

(2) Let us compute the cylindrical contact homology of the irrational ellipsoid (S^3, λ) from §3.3. Here of course we must take $\Gamma = 0$. Strictly speaking Theorem 6.4 is not applicable here because all Reeb orbits are contractible, but in fact the cylindrical contact homology is still defined in this example because all Reeb orbits satisfy a certain Conley-Zehnder index condition.

Denote the two embedded Reeb orbits by γ_1 and γ_2. These are elliptic. Let γ_i^k denote the k-fold iterate of γ_i. The chain complex $CC_*(S^3, \lambda, 0)$ has a relative $\mathbb{Z}$-grading, and because $\Gamma = 0$ there is in fact a canonical way to normalize it to an absolute $\mathbb{Z}$-grading. This grading is given as follows: there are positive irrational numbers ϕ_1 and ϕ_2 with $\phi_1\phi_2 = 1$ (if the ellipsoid is close to a sphere then ϕ_1 and ϕ_2 are close to 1) such that the grading of γ_i^k is

$$|\gamma_i^k| = 2\lfloor k(1 + \phi_i)\rfloor. \tag{6.3}$$

It is an exercise to deduce from (6.3) that there is one generator of each positive even grading. Hence the differential vanishes for degree reasons, and

$$CH_*(S^3, \xi, 0) \simeq \begin{cases} \mathbb{Q}, & * = 2, 4, \ldots, \\ 0, & \text{otherwise.} \end{cases}$$

Here ξ denotes the contact structure determined by λ, which as mentioned previously is the unique tight contact structure on S^3. See [6] for some applications of contact homology to the Reeb dynamics of other contact forms determining this contact structure.

(3) As mentioned previously, Colin-Honda used linearized contact homology to prove the Weinstein conjecture in many cases. However as far as I know it is not currently known whether linearized contact homology can be used to prove the Weinstein conjecture for all tight contact three-manifolds. (In the overtwisted case linearized contact homology is never defined, but the failure of linearized contact homology to be defined implies the existence of a contractible Reeb orbit, reproducing Hofer's result.) Taubes's proof of the Weinstein conjecture for all contact three-manifolds needs Seiberg-Witten theory.

7. The big picture surrounding Taubes's proof of the Weinstein conjecture

7.1. Seiberg-Witten invariants of four-manifolds.

The Seiberg-Witten invariants (and the conjecturally equivalent Ozsvath-Szabo invariants) are the most powerful tool currently available for distinguishing smooth four-manifolds. To briefly outline what these are, let X be a closed oriented connected smooth four-manifold. If $b_2^+(X) > 1$, then the Seiberg-Witten invariant of X is, after certain orientation choices have been made, a function

$$SW : \mathrm{Spin}^c(X) \longrightarrow \mathbb{Z}.$$

Here $b_2^+(X)$ denotes the maximal dimension of a subspace of $H_2(X;\mathbb{R})$ on which the intersection pairing is positive definite. Also $\operatorname{Spin}^c(X)$ denotes the set of spin-c structures on X. This is an affine space over $H^2(X;\mathbb{Z})$, which we will say more about in §8.1. Given a spin-c structure $\mathfrak{s}$, the Seiberg-Witten invariant $SW(X,\mathfrak{s})$ is defined by appropriately counting solutions to the Seiberg-Witten equations on X; for a detailed definition see eg [35]. We will not write down the four-dimensional Seiberg-Witten equations here, but we will write down the three-dimensional Seiberg-Witten equations in §8.3. The Seiberg-Witten invariants depend only on the diffeomorphism type of X, and can distinguish many pairs of smooth four-manifolds that are homeomorphic but not diffeomorphic. The Seiberg-Witten invariants are conjecturally equivalent to the earlier Donaldson invariants that count solutions ot the anti-self-dual Yang-Mills equations, but the Seiberg-Witten equations are analytically better behaved and as a result the Seiberg-Witten invariants are generally easier to work with.

7.2. **Taubes's "SW=Gr" theorem.** Suppose now that our four-manifold X is symplectic. A major result of Taubes from the 1990's asserts that the Seiberg-Witten invariants of X are equivalent to a certain count of holomorphic curves.

To describe this result, let ω denote the symplectic form on X. It turns out that ω defines a bijection

$$\imath_\omega : \operatorname{Spin}^c(X) \xrightarrow{\simeq} H_2(X).$$

Now let J be a generic, ω-tame almost complex structure on X; the tameness condition means that $\omega(v, Jv) > 0$ for all nonzero tangent vectors v. Taubes then defines a "Gromov invariant"

$$Gr : H_2(X) \longrightarrow \mathbb{Z}$$

roughly as follows. For each $Z \in H_2(X)$, the integer $Gr(X,Z)$ is a count of certain J-holomorphic curves C in X representing the homology class Z. The curves C that are counted are required to be embedded, except that they may include multiple covers of tori of square zero. Such a curve C is not required to be connected, but each component of C must live in a zero-dimensional moduli space. Each such holomorphic curve C is counted with a certain integer weight. (The weight is ±1, except when C includes multiply covered tori, in which case the weight is given by a somewhat complicated recipe.) Taubes's theorem is now:

Theorem 7.1 (Taubes). *Let X be a closed connected symplectic 4-manifold with $b_2^+(X) > 1$. Then for each $\mathfrak{s} \in \operatorname{Spin}^c(X)$ we have*

$$SW(X,\mathfrak{s}) = Gr(X, \imath_\omega(\mathfrak{s})).$$

The proof of this theorem is given in [41]; for an introduction to it see [23].

7.3. **Three-dimensional Seiberg-Witten theory.** Now let Y be a closed oriented 3-manifold. The set of spin-c structures on Y, denoted by $\operatorname{Spin}^c(Y)$, is an affine space over $H^2(Y;\mathbb{Z})$. A spin-c structure on Y determines a product spin-c structure on $\mathbb{R}\times Y$. There are then various ways to define topological invariants of the 3-manifold Y by studying solutions to the Seiberg-Witten equations on the noncompact 4-manifold $\mathbb{R}\times Y$.

To start, one can consider $\mathbb{R}$-invariant solutions to the Seiberg-Witten equations on $\mathbb{R}\times Y$. These are equivalent to solutions to the three-dimensional Seiberg-Witten equations on Y, which we will discuss in §8.3. When $b_1(Y) > 0$, one can count these solutions with signs to obtain the three-dimensional *Seiberg-Witten invariant* $SW : \mathrm{Spin}^c(Y) \to \mathbb{Z}$. (When $b_1(Y) = 1$, for "torsion" spin-c structures, ie spin-c structures whose first Chern class is torsion, see §8.1, this invariant depends on the choice of one of two possible "chambers".) As shown in [33, 46], this invariant agrees with the Turaev torsion of Y, which generalizes the Alexander polynomial of a knot, and which is explicitly computable in terms of the determinants of certain matrices associated to a triangulation of Y.

To get more interesting invariants of Y, one observes that solutions to the Seiberg-Witten equations on Y are critical points of a certain functional $\mathcal{F}$ on a configuration space associated to Y. Moreover, solutions to the Seiberg-Witten equations on $\mathbb{R} \times Y$ (not necessarily $\mathbb{R}$-invariant) are equivalent to gradient flow lines of this functional. It turns out that one can then define a version of Morse homology for the functional $\mathcal{F}$, called *Seiberg-Witten Floer homology*. This construction is highly nontrivial and has been carried out by Kronheimer-Mrowka [26]. Roughly speaking, the Seiberg-Witten Floer homology is the homology of a chain complex which is generated by solutions to the Seiberg-Witten equations on Y, and whose differential counts solutions to the Seiberg-Witten equations on $\mathbb{R}\times Y$. When $b_!(Y) > 0$, for non-torsion spin-c structures, the Euler characteristic of the Seiberg-Witten Floer homology agrees with the Seiberg-Witten invariant discussed above. However Seiberg-Witten Floer homology can be defined without any assumption on $b_1(Y)$.

In fact there are two basic versions of Seiberg-Witten Floer theory that one can define, depending on how one treats "reducibles", see §8.4. The first is denoted by $\check{HM}_*(Y)$, and pronounced "HM-to"; this assigns a relatively graded homology group $\check{HM}_*(Y, \mathfrak{s})$ to each spin-c structure $\mathfrak{s}$ on Y, which is conjecturally isomorphic to the Heegaard Floer homology $HF_*^+(Y, \mathfrak{s})$ defined in [36]. The second is pronounced "HM-from", denoted by $\widehat{HM}_*(Y, \mathfrak{s})$, and conjecturally isomorphic to the Heegaard Floer homology $HF_*^-(Y, \mathfrak{s})$. For non-torsion spin-c structures, $\check{HM}_*$ and $\widehat{HM}_*$ are the same. For any spin-c structure $\mathfrak{s}$, there is a canonical isomorphism

$$\check{HM}_*(-Y, \mathfrak{s}) = \widehat{HM}^{-*}(Y, \mathfrak{s}),$$

where $-Y$ denotes Y with its orientation reversed, and $\widehat{HM}^*$ denotes the cohomological version of $\widehat{HM}_*$ obtained by dualizing the chain complex.

7.4. **Embedded contact homology.** Suppose now that our three-manifold Y is equipped with a contact form λ. Recall that the four-manifold $\mathbb{R} \times Y$ then has a symplectic form $d(e^s\lambda)$, where s denotes the $\mathbb{R}$ coordinate. It is natural to seek an analogue of Taubes's "SW=Gr" theorem for the noncompact symplectic four-manifold $\mathbb{R} \times Y$. That is, one would like to understand the Seiberg-Witten Floer homology of Y in terms of holomorphic curves in $\mathbb{R} \times Y$. For this purpose it is appropriate to use an admissible almost complex structure as in Definition 6.1.

The analogy suggests that the Seiberg-Witten Floer homology of Y should be isomorphic to the homology of a chain complex whose differential counts certain (mostly) embedded holomorphic curves in $\mathbb{R}\times Y$, and which is generated by certain $\mathbb{R}$-invariant holomorphic curves in $\mathbb{R} \times Y$, that is to say unions of Reeb orbits. The resulting theory is called *embedded contact homology*, or ECH for short. There is some resemblance between ECH and cylindrical contact homology; but among other differences, ECH does not require the holomorphic curves that are counted to be cylinders, while cylindrical contact homology does not require them to be embedded.

To say a bit more about what ECH is, assume as usual that all Reeb orbits are nondegenerate.

Definition 7.2. An *orbit set* is a finite set of pairs $\alpha = \{(\alpha_i, m_i)\}$ where:

- The α_i's are distinct embedded Reeb orbits.
- The m_i's are positive integers.

The homology class of α is defined by

$$[\alpha] := \sum_i m_i[\alpha_i] \in H_1(Y).$$

The orbit set α is *admissible* if:

- $m_i = 1$ whenever α_i is hyperbolic (see §6.2).

Given $\Gamma \in H_1(Y)$, the embedded contact homology $ECH_*(Y, \lambda, \Gamma)$ is the homology of a chain complex which is freely generated over $\mathbb{Z}$ by admissible orbit sets α with $[\alpha] = \Gamma$. The differential counts certain (mostly) embedded holomorphic curves in $\mathbb{R} \times Y$. It has a relative $\mathbb{Z}/d(c_1(\xi) + 2\,\mathrm{PD}(\Gamma))$ grading. The precise definitions of the grading and the differential are somewhat subtle; we refer the reader to [22] for an overview and to [20, 24] for more details.

Example 7.3. Consider again the irrational ellipsoid (S^3, λ) as discussed in §6.6. Of course we must take $\Gamma = 0$. The generators of the ECH chain complex have the form $\gamma_1^{m_1}\gamma_2^{m_2}$ where γ_1 and γ_2 are the two embedded Reeb orbits, m_1 and m_2 are nonnegative integers, and $\gamma_1^{m_1}\gamma_2^{m_2}$ is shorthand for the orbit set consisting of the pair (γ_1, m_1) (when $m_1 \neq 0$) together with the pair (γ_2, m_2) (when $m_2 \neq 0$). In this case the chain complex has a relative $\mathbb{Z}$-grading, and since $\Gamma = 0$ it has a canonical refinement to an absolute $\mathbb{Z}$-grading such that the grading of the empty set is zero. It is shown in [25] that the grading is given by

$$|\gamma_1^{m_1}\gamma_2^{m_2}| = 2\left(m_1 + m_2 + m_1m_2 + \sum_{i=1}^{2}\sum_{k=1}^{m_i}\lfloor k\phi_i\rfloor\right)$$

where ϕ_1 and ϕ_2 are as in §6.6. One can deduce from this formula that there is one generator of each nonnegative even grading, so that

$$ECH_*(S^3, \lambda, 0) \simeq \begin{cases} \mathbb{Z}, & * = 0, 2, \ldots, \\ 0, & \text{otherwise} \end{cases}$$

In general it turns out that ECH depends only on Y, the contact structure ξ, and the homology class Γ, and not on the choice of contact form λ or admissible almost complex structure J. This follows from a much stronger result recently proved by Taubes [45], which is the analogue of SW=Gr in this setting:

Theorem 7.4 (Taubes). *Let Y be a closed oriented 3-manifold with a contact form λ such that all Reeb orbits are nondegenerate. Then for each $\Gamma \in H_1(Y)$ there is an isomorphism*

$$ECH_*(Y, \lambda, \Gamma) \simeq \widehat{HM}^{-*}(Y, \mathfrak{s}_\xi + \mathrm{PD}(\Gamma)), \tag{7.1}$$

up to a grading shift.

Here ξ denotes the contact structure determined by λ, and $\mathfrak{s}_\xi$ is a spin-c structure determined by ξ, see Example 8.2 below. In fact both sides of (7.1) have canonical absolute gradings by homotopy classes of oriented 2-plane fields, see [26, 21], and it is natural to conjecture that the isomorphism (7.1) respects these gradings.

7.5. **Significance for the Weinstein conjecture.** To prove the Weinstein conjecture, it is enough to show that the embedded contact homology is nontrivial. More precisely, if (Y, λ) were a counterexample to the Weinstein conjecture, then we would have

$$ECH(Y, \lambda, \Gamma) \simeq \left\{ \begin{array}{ll} \mathbb{Z}, & \Gamma = 0, \\ 0, & \Gamma \neq 0. \end{array} \right. \tag{7.2}$$

Here the $\mathbb{Z}$ corresponds to the empty set of Reeb orbits. (Incidentally the empty set is an important ECH generator; it is always a cycle in the ECH chain complex by the argument in Lemma 6.2, and its homology class in ECH conjecturally agrees with the invariants of contact structures in Seiberg-Witten and Heegaard Floer homology.)

However, by Theorem 7.4, the ECH cannot be trivial as in (7.2), because the Seiberg-Witten Floer homology is always infinitely generated:

Theorem 7.5 (Kronheimer-Mrowka). *Let Y be a closed oriented 3-manifold and let $\mathfrak{s}$ be a spin-c structure with $c_1(\mathfrak{s})$ torsion. Then $\widehat{HM}^*(Y, \mathfrak{s})$ is nonzero for infinitely many values of the grading $*$, which are bounded from above.*

Here $c_1(\mathfrak{s})$ denotes the first Chern class of the spin-c structure, which is defined in §8.1 below. In terms of the correspondence between $\operatorname{Spin}^c(Y)$ and $H_1(Y)$ in (7.1), we have

$$c_1(\mathfrak{s}_\xi + \mathrm{PD}(\Gamma)) = c_1(\xi) + 2\,\mathrm{PD}(\Gamma).$$

It turns out that $c_1(\xi)$ is always divisible by 2 in $H^2(Y; \mathbb{Z})$. Thus Theorems 7.4 and 7.5 imply the following more precise version of the Weinstein conjecture:

- *Let Y be a closed oriented 3-manifold with a contact form λ such that all Reeb orbits are nondegenerate. Let $\Gamma \in H_1(Y)$ such that $c_1(\xi) + 2\,\mathrm{PD}(\Gamma) \in H^2(Y; \mathbb{Z})$ is torsion. (Such Γ always exist.) Then there is a nonempty admissible orbit set α with $[\alpha] = \Gamma$.*

Remark 7.6. The fact that ECH is infinitely generated does not imply that there are infinitely many embedded Reeb orbits, as shown by the irrational ellipsoid example. On the other hand we are not aware of any example of a closed oriented three-manifold, other than a sphere or a lens space, with a contact form with only finitely many embedded Reeb orbits.

In fact one does not need the full force of the isomorphism in Theorem 7.4 to prove the Weinstein conjecture; rather one just needs a way of passing from generators of Seiberg-Witten Floer homology to ECH generators. This is what Taubes's original proof of the Weinstein conjecture in [42] establishes, yielding a proof of the following theorem, which is slightly different than the statement above:

Theorem 7.7 (Taubes). *Let Y be a closed oriented 3-manifold with a contact form λ. Let $\Gamma \in H_1(Y)$ such that $c_1(\xi) + 2\,\mathrm{PD}(\Gamma) \in H^2(Y; \mathbb{Z})$ is torsion. (Such Γ always exist.) Then there is a nonempty orbit set α with $[\alpha] = \Gamma$.*

In a sequel [43], Taubes uses more nontriviality results for Seiberg-Witten Floer homology to find nonempty orbit sets in some other homology classes. In [25], the full force of the isomorphism in Theorem 7.4 is used to obtain some slight improvements on the Weinstein conjecture.

The rest of this article will attempt to describe Taubes's proof of Theorem 7.7.

8. The three-dimensional Seiberg-Witten equations

It is time to introduce the three-dimensional Seiberg-Witten equations. These equtions originated in physics, and if one is not conversant in the relevant physics then they may seem unmotivated. However, they turn out to be exquisitely tuned to produce well-behaved moduli spaces out of which one can extract topological invariants.

8.1. Spin-c structures.

Let Y be a closed oriented connected 3-manifold, and choose a Riemannian metric on Y.

Definition 8.1. A *spin-c structure* on Y is a pair $\mathfrak{s} = (\mathbb{S}, \rho)$ where $\mathbb{S}$ is a rank 2 Hermitian vector bundle on Y, called the *spinor bundle*, and

$$\rho : TY \longrightarrow \operatorname{End}(\mathbb{S})$$

is a bundle map, called *Clifford multiplication*, such that:

(1) If $a, b \in T_yY$, then

$$\rho(a)\rho(b) + \rho(b)\rho(a) = -2\langle a, b\rangle.$$

(2) If e_1, e_2, e_3 is an oriented orthonormal frame for T_yY, then

$$\rho(e_1)\rho(e_2)\rho(e_3) = 1.$$

Properties (1) and (2) of Clifford multiplication are equivalent to the following: For each $y \in Y$, there is an oriented orthonormal frame e_1, e_2, e_3 for T_yY, and a basis for $\mathbb{S}_y$, in which Clifford multiplication is given by the Pauli matrices

$$\rho(e_1) := \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \qquad \rho(e_2) := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad \rho(e_3) := \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}. \tag{8.1}$$

A section ψ of the spinor bundle $\mathbb{S}$ is often called a *spinor*.

Example 8.2. An oriented 2-plane field $\xi \subset TY$ determines a spin-c structure $\mathfrak{s}_\xi$ as follows. The spinor bundle is defined by

$$\mathbb{S} := \underline{\mathbb{C}} \oplus \xi. \tag{8.2}$$

where ξ is regarded as a Hermitian line bundle using its orientation and the Riemannian metric on Y, while $\underline{\mathbb{C}}$ denotes the trivial complex line bundle. The Clifford multiplication is defined at a point $y \in Y$ by the equations (8.1), where e_1, e_2, e_3 are an oriented orthonormal frame for T_yY such that e_2, e_3 is an oriented orthonormal frame for ξ_y, and the basis for $\mathbb{S}_y$ is given in terms of the decomposition (8.2) by $(1, e_2)$.

Two spin-c structures $(\mathbb{S}, \rho)$ and $(\mathbb{S}', \rho')$ are isomorphic if there is a Hermitian vector bundle isomorphism $\phi : \mathbb{S} \overset{\simeq}{\to} \mathbb{S}'$ such that $\rho'(v)\circ\phi = \phi\circ\rho(v)$ for every tangent vector v. Let $\operatorname{Spin}^c(Y)$ denote the set of isomorphism classes of spin-c structures on Y. This does not depend on the choice of Riemannian metric on Y. There is an action of $H^2(Y;\mathbb{Z})$ on $\operatorname{Spin}^c(Y)$ defined as follows: Given $\alpha \in H^2(Y;\mathbb{Z})$, let

L_α denote the complex line bundle on Y with $c_1(L_\alpha) = \alpha$, assign it a Hermitian metric, and define

$$\alpha \cdot (\mathbb{S}, \rho) := (\mathbb{S} \otimes L_\alpha, \rho \otimes 1).$$

It turns out that this action is free and transitive, so that $\operatorname{Spin}^c(Y)$ is an affine space over $H^2(Y;\mathbb{Z})$.

Remark 8.3. If $\mathcal{P}(Y)$ denotes the set of homotopy classes of oriented 2-plane fields on Y, then the map $\mathcal{P}(Y) \to \operatorname{Spin}^c(Y)$ defined in Example 8.2 is surjective. Two oriented 2-plane fields give rise to isomorphic spin-c structures if and only if they are homotopic over the 2-skeleton of Y (for some triangulation). Given a spin-c structure $\mathfrak{s}$, its inverse image in $\mathcal{P}(Y)$ is an affine space over $\mathbb{Z}/d(c_1(\mathfrak{s}))$. Here $c_1(\mathfrak{s})$ denotes the first Chern class of the spinor bundle, and d denotes its divisibility in $H^2(Y;\mathbb{Z})$ mod torsion, which is always an even integer.

Here is an alternate definition of a spin-c structure, on an n-dimensional oriented manifold Y with a Riemannian metric. Let $F \to Y$ denote the *frame bundle*, whose fiber over $y \in Y$ is the set of orientation-preserving linear isometries $\mathbb{R}^n \overset{\simeq}{\to} Y_y$. Note that $\mathrm{SO}(n)$ acts on the right on F by precomposition with automorphisms of $\mathbb{R}^n$. This action makes F into a principal $\mathrm{SO}(n)$-bundle over Y. Now the Lie group $\operatorname{Spin}^c(n)$ is defined by

$$\operatorname{Spin}^c(n) := \operatorname{Spin}(n) \times_{\mathbb{Z}/2} \mathrm{U}(1).$$

Here $\operatorname{Spin}(n)$ is the connected double cover of $\mathrm{SO}(n)$; and $\mathbb{Z}/2$ acts on $\operatorname{Spin}(n)$ as the nontrivial covering transformation, and on $\mathrm{U}(1)$ as multiplication by -1. A spin-c structure on Y is then defined to be a lift of F to a principal $\operatorname{Spin}^c(n)$-bundle, ie a principal $\operatorname{Spin}^c(n)$-bundle $\widetilde{F} \to Y$ together with a map $\widetilde{F} \to F$ which commutes with the group actions and the projections to Y.

When $n = 3$, this definition is equivalent to the previous one. In particular, given a lift $\widetilde{F}$ of the frame bundle, the spinor bundle and Clifford multiplication are recovered as follows. We can identify $\operatorname{Spin}(3) = \mathrm{SU}(2)$ and $\operatorname{Spin}^c(3) = \mathrm{U}(2)$. The spinor bundle is then associated to $\widetilde{F}$ via the fundamental representation of $\mathrm{U}(2)$. On the other hand the tangent bundle of Y is associated to $\widetilde{F}$ by the representation of $\mathrm{U}(2)$ on $\mathbb{R}^3$ given by the projection $\operatorname{Spin}^c(3) \to \mathrm{SO}(3)$. Clifford multiplication is then defined on these associated bundles using the model linear map $\mathbb{R}^3 \to \operatorname{End}(\mathbb{C}^2)$ that sends the standard basis vectors of $\mathbb{R}^3$ to the three matrices in (8.1).

8.2. The Dirac operator.

Let $\mathfrak{s} = (\mathbb{S}, \rho)$ be a spin-c structure.

Definition 8.4. A *spin-c connection* on $\mathfrak{s}$ is a Hermitian connection on $\mathbb{S}$ such that the associated covariant deriviative ∇_A is compatible with Clifford multiplication in the following sense: If v is a section of TY and if ψ is a section of $\mathbb{S}$, then

$$\nabla_A(\rho(v)\psi) = \rho(\nabla v)\psi + \rho(v)\nabla_A\psi.$$

Here ∇v denotes the covariant derivative of v with respect to the Levi-Civita connection on TY.

It follows from the above definition that any two spin-c connections differ by an imaginary-valued 1-form on Y. It is not hard to show that spin-c connections exist. In fact a spin-c connection is equivalent to a Hermitian connection on the $\mathrm{U}(1)$-bundle $\det(\mathbb{S})$. One can see this by using the alternate definition of spin-c structure and noting that the Lie algebra of $\mathrm{U}(2)$ is the sum of the Lie algebras of

SO(3) and of U(1). A connection on $\mathbb{S}$ is then determined by a connection on TY (which we take to be the Levi-Civita connection) and a connection on the complex line bundle associated to the determinant map U(2) → U(1), namely $\det(\mathbb{S})$. So in the notation ∇_A, we regard A as a connection on $\det(\mathbb{S})$. From this perspective, adding an imaginary-valued 1-form a to the connection A on $\det(\mathbb{S})$ adds $a/2$ to the associated spin-c connection ∇_A on $\mathbb{S}$.

Definition 8.5. Given a connection A on $\det(\mathbb{S})$, define the *Dirac operator* D_A to be the composition

$$C^\infty(Y,\mathbb{S}) \xrightarrow{\nabla_A} C^\infty(Y, T^*X\otimes\mathbb{S}) \xrightarrow{\rho} C^\infty(Y,\mathbb{S}).$$

Here the Clifford action is extended to cotangent vectors using the metric.

A key property of the Dirac operator is that its square is the "connection Laplacian", plus some zeroth order terms involving curvature. More precisely, it satisfies the Bochner-Lichnerowitz-Weitzenbock formula

$$D_A^2\psi = \nabla_A^*\nabla_A\psi + \frac{s}{4}\psi + \frac{1}{2}\rho(F_A)\psi. \tag{8.3}$$

Here s denotes the scalar curvature of Y, which is a real-valued function; F_A denotes the curvature of the connection A on $\det(\mathbb{S})$, which is an imaginary-valued closed 2-form on Y; and Clifford multiplication is extended to differential forms by the rule

$$\rho(a\wedge b) := \frac{1}{2}\left(\rho(a)\rho(b) + (-1)^{\deg(a)\deg(b)}\rho(b)\rho(a)\right).$$

Indeed, the motivation for introducing the original version of the Dirac operator (on Minkowski spacetime) was to find a square root of the Laplacian, in order to study the quantum theory of the electron.

Another important property is that the Dirac operator is formally self-adjoint: if ψ_1 and ψ_2 are two spinors, then

$$\int_Y \langle D_A\psi_1,\psi_2\rangle d\,\mathrm{vol} = \int_Y \langle\psi_1, D_A\psi_2\rangle d\,\mathrm{vol}.$$

For much more about Dirac operators, see eg [2, 30].

8.3. **The Seiberg-Witten equations.** Fix a spin-c structure $(\mathbb{S},\rho)$. The Seiberg-Witten equations concern a pair (A,ψ) where A is a connection on $\det(\mathbb{S})$ and ψ is a section of $\mathbb{S}$.

Definition 8.6. The (unperturbed) *Seiberg-Witten equations* for the pair (A,ψ) are

$$\begin{aligned} D_A\psi &= 0,\\ *F_A &= \langle\rho(\cdot)\psi,\psi\rangle. \end{aligned}$$

Here '$*$' denotes the Hodge star. Note that it follows from the properties of Clifford multiplication that if ψ is any spinor then $\langle\rho(\cdot)\psi,\psi\rangle$ is an imaginary-valued 1-form.

Remark 8.7. Conventions for the Seiberg-Witten equations vary in the literature. In particular, one could multiply one side of the second equation by a positive constant, and the solutions to the equations would be equivalent via rescaling the spinor. However the sign is crucial: switching the sign in the second equation would ruin the key compactness properties of the equations.

One also needs to consider certain perturbations of the equations. In particular, one often needs to make small perturbations in order to obtain transversality, while Taubes's proof of the Weinstein conjecture will involve a large perturbation.

Definition 8.8. Let μ be a real closed 2-form on Y. The *Seiberg-Witten equations with perturbation* μ for the pair (A,ψ) are

$$\begin{aligned} D_A\psi &= 0, \\ *F_A &= \langle\rho(\cdot)\psi,\psi\rangle + i*\mu. \end{aligned} \tag{8.4}$$

8.4. **Gauge transformations.** The equations (8.4) have a large amount of symmetry. Namely, the Seiberg-Witten equations are defined on the *configuration space*

$$\mathcal{C} := \operatorname{Conn}(\det(\mathbb{S})) \times C^\infty(Y,\mathbb{S}),$$

where $\operatorname{Conn}(\det(\mathbb{S}))$ denotes the set of Hermitian connections on $\det(\mathbb{S})$. Define the *gauge group*

$$\mathcal{G} := C^\infty(Y,S^1).$$

This can be regarded as the automorphism group of the spin-c structure $(\mathbb{S},\rho)$. As such it has a natural action on the configuration space $\mathcal{C}$ defined as follows: If $g : Y \to S^1$ is in $\mathcal{G}$, then regarding S^1 as the unit circle in $\mathbb{C}$, one defines

$$g\cdot(A,\psi) := (A - 2g^{-1}dg, g\psi).$$

Lemma 8.9. *The set of solutions to the Seiberg-Witten equations* (8.4) *is invariant under the action of the gauge group* $\mathcal{G}$.

Proof. The curvature equation is invariant because the curvature of a connection is invariant under gauge transformations. The Dirac equation is invariant because

$$\begin{aligned} D_{A-2g^{-1}dg}(g\psi) &= \rho((\nabla_A - g^{-1}dg)g\psi) \\ &= \rho(dg\otimes\psi + g\nabla_A\psi - g^{-1}dg\otimes g\psi) \\ &= \rho(g\nabla_A\psi) \\ &= gD_A\psi. \end{aligned} \tag{8.5}$$

□

Two solutions to the Seiberg-Witten equations are called *gauge equivalent* if they differ by the action of an element of $\mathcal{G}$. In general, one counts solutions only modulo gauge equivalence.

Observe that the action of $\mathcal{G}$ on $\mathcal{C}$ is free, except that configurations $(A,0)$ have S^1 stabilizer given by the constant maps $Y\to S^1$. To keep track of this, a configuration (A,ψ) is called *reducible* if $\psi\equiv 0$, and *irreducible* otherwise. Reducible solutions exist if and only if $[\mu] = -2\pi c_1(S)$ in $H^2(Y;\mathbb{R})$. In this case the set of reducible solutions modulo gauge equivalence can be identified with the torus $H^1(Y;\mathbb{R})/2\pi H^1(Y;\mathbb{Z})$.

9. Seiberg-Witten Floer homology

We now briefly review some of what we will need to know about Seiberg-Witten Floer homology, following [26].

9.1. **The Chern-Simons-Dirac functional.** We begin by realizing the solutions to the Seiberg-Witten equations as the critical points of a functional. The Seiberg-Witten Floer theory will then be some kind of Morse homology for this functional.

Let Y be a closed connected oriented 3-manifold with a Riemannian metric and a spin-c structure $\mathfrak{s} = (\mathbb{S}, \rho)$. Fix a real closed 2-form μ for use in defining the perturbed Seiberg-Witten equations (8.4). Also fix a reference connection A_0 on $\det(\mathbb{S})$, so that if A is any other connection on $\det(\mathbb{S})$ then $A - A_0$ is an imaginary-valued 1-form on Y.

Definition 9.1. Define the (perturbed) *Chern-Simons-Dirac functional*

$$\mathcal{F} : \mathcal{C} \to \mathbb{R}$$

by

$$\mathcal{F}(A,\psi) := -\frac{1}{8}\int_Y (A - A_0)\wedge(F_A + F_{A_0} - 2i\mu) + \frac{1}{2}\int_Y \langle D_A\psi, \psi\rangle d\,\mathrm{vol}. \tag{9.1}$$

Lemma 9.2. *(A,ψ) is a critical point of $\mathcal{F}$ if and only if (A,ψ) satisfies the perturbed Seiberg-Witten equations* (8.4).

Proof. Let $(\dot{A},\dot{\psi})$ be a tangent vector to the configuration space $\mathcal{C}$ at (A,ψ). This means that $\dot{A}$ is an imaginary-valued 1-form and $\dot{\psi}$ is a spinor. We compute

$$\begin{aligned} d\mathcal{F}_{(A,\psi)}(\dot{A},\dot{\psi}) = & -\frac{1}{8}\int_Y \dot{A}\wedge(F_A + F_{A_0} - 2i\mu) - \frac{1}{8}\int_Y (A - A_0)\wedge d\dot{A} \\ & + \frac{1}{4}\int_Y \langle\rho(\dot{A})\psi,\psi\rangle d\,\mathrm{vol} + \frac{1}{2}\int_Y \langle D_A\dot{\psi},\psi\rangle d\,\mathrm{vol} + \frac{1}{2}\int_Y \langle D_A\psi,\dot{\psi}\rangle d\,\mathrm{vol}. \end{aligned}$$

Applying Stokes' theorem to the second term, using the definition of Clifford multiplication to manipulate the third term, and applying self-adjointness of the Dirac operator to the fourth term, we obtain

$$d\mathcal{F}_{(A,\psi)}(\dot{A},\dot{\psi}) = -\frac{1}{4}\int_Y \dot{A}\wedge(F_A - i\mu - *\langle\rho(\cdot)\psi,\psi\rangle) + \int_Y \mathrm{Re}\langle D_A\psi,\dot{\psi}\rangle d\,\mathrm{vol}.$$

This vanishes for all $\dot{A}$ and $\dot{\psi}$ if and only if (A,ψ) satisfy the perturbed Seiberg-Witten equations (8.4). □

We now consider the behavior of $\mathcal{F}$ under gauge transformations. Recall that the set of homotopy classes of maps $Y \to S^1$ can be identified with $H^1(Y;\mathbb{Z})$. Denote the homotopy class of a map g by $[g]$. If $g : Y \to S^1$ is smooth, and if we regard S^1 as the unit circle in $\mathbb{C}$, then the image of $[g]$ in $H^1(Y;\mathbb{R})$ is the cohomology class of the 1-form $(2\pi i)^{-1}g^{-1}dg$.

Lemma 9.3. *If $g : Y \to S^1$ is a gauge transformation, then*

$$\mathcal{F}(g\cdot(A,\psi)) - \mathcal{F}(A,\psi) = \pi\int_{[Y]} [g] \smile (2\pi c_1(\mathfrak{s}) + [\mu]).$$

Proof. We compute, using (8.5), that

$$\begin{aligned} \mathcal{F}(g\cdot(A,\psi)) - \mathcal{F}(A,\psi) = & -\frac{1}{8}\int_Y (-2g^{-1}dg)\wedge(F_A + F_{A_0} - 2i\mu) \\ & + \frac{1}{2}\int_Y \langle gD_A\psi, g\psi\rangle d\,\mathrm{vol} - \frac{1}{2}\int_Y \langle D_A\psi,\psi\rangle d\,\mathrm{vol}. \end{aligned}$$

The second line vanishes. To process the first line, recall that F_A and F_{A_0} both represent the class $-2\pi i c_1(\mathfrak{s})$. The lemma follows. □

In particular, $\mathcal{F}$ is gauge invariant if and only if $[\mu] = -2\pi c_1(\mathfrak{s})$ in $H^2(Y;\mathbb{R})$.

9.2. **Seiberg-Witten Floer homology.** Roughly speaking, Seiberg-Witten Floer homology is the Morse homology of the functional $\mathcal{F}$ on $\mathcal{C}/\mathcal{G}$, where the perturbation 2-form μ is taken to be exact. (One can also define versions of Seiberg-Witten Floer homology when μ is closed but not exact, but these have different properties.) The detailed construction is carried out in [26]. Some basic issues to keep in mind are the following:

(1) When $c_1(\mathfrak{s})$ is not torsion, the functional $\mathcal{F}$ is not gauge invariant, so it is not actually a real-valued functional on $\mathcal{C}/\mathcal{G}$. However, it follows from Lemma 9.3 that the differential $d\mathcal{F}$ is a well-defined closed 1-form on $\mathcal{C}/\mathcal{G}$, whose periods are integers (times $2\pi^2$), so $\mathcal{F}$ can be regarded as an S^1-valued functional on $\mathcal{C}/\mathcal{G}$. As such, its Morse theory can still be defined; and in the present case no Novikov ring is needed, because there is a monotonicity formula relating the ambiguity in $\mathcal{F}$ and the ambiguity in the index.

(2) The quotiented configuration space $\mathcal{C}/\mathcal{G}$ on which $\mathcal{F}$ is defined is not a manifold in any natural sense, because $\mathcal{G}$ does not act freely on the reducibles. On the other hand, $\mathcal{C}/\mathcal{G}$ is the quotient of a manifold by an S^1 action. Namely, if one fixes a point $y_0 \in Y$ and defines $\mathcal{G}_0 := \{g \in \mathcal{G} \mid g(y_0) = 1\}$, then $\mathcal{C}/\mathcal{G}_0$ is a manifold, and $\mathcal{C}/\mathcal{G} = (\mathcal{C}/\mathcal{G}_0)/S^1$. So one wants to define some kind of S^1-equivariant homology of $\mathcal{F}$ on $\mathcal{C}/\mathcal{G}_0$.

The approach taken by Kronheimer-Mrowka is, roughly speaking, to blow up the singularities of $\mathcal{C}/\mathcal{G}$, so as to obtain a manifold-with-boundary $\widetilde{\mathcal{C}/\mathcal{G}}$, where the boundary arises from the reducibles. The gradient flow of $\mathcal{F}$ induces a (partially defined) flow on $\widetilde{\mathcal{C}/\mathcal{G}}$, which is tangent to the boundary.

There is now a finite-dimensional model for how to proceed. Let X be a finite dimensional compact manifold with boundary, and let (f,g) be a Morse-Smale pair on X such that the gradient flow is tangent to the boundary. In this context, as explained in [26, §2], there are three versions of Morse homology one can define, which fit into a long exact sequence:

$$\overline{H}_*^{\text{Morse}}(X,f,g) \longrightarrow \check{H}_*^{\text{Morse}}(X,f,g) \longrightarrow \widehat{H}_*^{\text{Morse}}(X,f,g) \longrightarrow \overline{H}_{*-1}^{\text{Morse}}(X,f,g) \longrightarrow$$

Here $\overline{H}_*$ is just the Morse homology of the boundary. The version $\check{H}_*$ is the homology of a chain complex which is freely generated over $\mathbb{Z}$ by interior critical points and "boundary stable" critical points on the boundary, and whose differential counts certain configurations of flow lines between them. The version $\widehat{H}_*$ is similar but its generators include "boundary unstable" critical points on the boundary instead. The above exact sequences turns out to agree with the relative homology exact sequence

$$H_*(\partial X) \longrightarrow H_*(X) \longrightarrow H_*(X,\partial X) \longrightarrow H_{*-1}(\partial X) \longrightarrow \cdots$$

Carrying out an analogue of this construction on the blown-up configuration space now gives three versions of Seiberg-Witten Floer homology, which fit into a long exact sequence:

$$\overline{HM}_*(Y,\mathfrak{s}) \longrightarrow \check{HM}_*(Y,\mathfrak{s}) \longrightarrow \widehat{HM}_*(Y,\mathfrak{s}) \longrightarrow \overline{HM}_{*-1}(Y,\mathfrak{s}) \longrightarrow \cdots$$

The version $\overline{HM}$ comes entirely from the reducibles and can be computed (at least over $\mathbb{Q}$) in terms of classical algebraic topology, specifically the triple cup product on $H^*(Y)$. As such, $\overline{HM}_*$ may seem less interesting than the other two versions

$\check{HM}_*$ and $\widehat{HM}_*$. However the computation of $\overline{HM}_*$ is needed (in conjunction with the above exact sequence) to prove the nontriviality result of Theorem 7.5, which plays a key role in the proof of the Weinstein conjecture.

(3) The three versions of Seiberg-Witten Floer homology above all have a relative $\mathbb{Z}/d(c_1(\mathfrak{s}))$-grading. So if $c_1(\mathfrak{s})$ is torsion then the Seiberg-Witten Floer homologies are relatively $\mathbb{Z}$-graded. As mentioned previously, there is in fact an absolute grading by homotopy classes of oriented 2-plane fields, which is compatible with the map $\mathcal{P}(Y) \to \operatorname{Spin}^c(Y)$ that was discussed in Remark 8.3. However for our purposes it is enough to just regard the grading as taking values in some affine space over $\mathbb{Z}/d(c_1(\mathfrak{s}))$.

(4) Heuristically the differentials in the Seiberg-Witten Floer chain complexes count gradient flow lines of $\mathcal{F}$, but in fact some abstract perturbations of the equations are needed to obtain the transversality needed to count solutions. The perturbations are explained in detail in [26] and will be suppressed in the exposition here.

10. Outline of Taubes's proof

We now have the background in place to describe Taubes's proof of the Weinstein conjecture. Here we follow Taubes's paper [42] and MSRI lectures [31]. This will be just an outline, with the proofs of many key results omitted.

10.1. **Geometric setup.** Let Y be a closed oriented connected 3-manifold with a contact form λ. Fix a Riemannian metric on Y such that $|\lambda| = 1$ and $d\lambda = 2 * \lambda$. We can do this because $\lambda \wedge d\lambda > 0$.

Denote[3] the oriented 2-plane field $\xi = \operatorname{Ker}(\lambda)$, regarded as a Hermitian line bundle, by K^{-1}. Recall that ξ determines a distinguished spin-c structure $\mathfrak{s}_\xi$, in which $\mathbb{S} = \underline{\mathbb{C}} \oplus K^{-1}$. Any spin-c structure $\mathfrak{s}$ is obtained from $\mathfrak{s}_\xi$ by tensoring with a Hermitian line bundle E, so that

$$\mathbb{S} = E \oplus K^{-1}E. \tag{10.1}$$

In this decomposition, E is the $+i$ eigenspace of Clifford multiplication by λ, while $K^{-1}E$ is the $-i$ eigenspace. The significance of E is that Taubes's Theorem 7.4 ultimately shows that the Seiberg-Witten Floer cohomology $\widehat{HM}^{-*}(Y, \mathfrak{s})$ is isomorphic to $ECH_*(Y, \lambda, \Gamma)$ where Γ is Poincare dual to $c_1(E)$.

For any spin-c structure as in (10.1), connections on $\det(\mathbb{S})$ have the form $A_0 + 2A$ where A_0 is a reference connection on K^{-1} while A is a connection on E. Also, Taubes picks out a canonical connection A_0 on K^{-1}, such that with respect to the above decomposition, $D_{A_0}(1, 0) = 0$. Henceforth, think of spin connections as being determined this way by the canonical connection A_0 on K^{-1} together with a connection A on E.

10.2. **Taubes's perturbation.** The idea of Taubes's proof of the Weinstein conjecture is to deform the Seiberg-Witten equations by a sequence of increasingly large perturbations, and to use a sequence of solutions to the perturbed equations (provided by the known nontriviality of Seiberg-Witten Floer homology), to yield a nonempty collection of Reeb orbits.

[3] This notation is carried over from the SW=Gr story, where K denotes the canonical bundle of a symplectic 4-manifold. For a contact 3-manifold, K is the canonical bundle of the symplectization.

The basic perturbation of the Seiberg-Witten equations considered by Taubes is

$$\begin{aligned} *F_A &= r(\langle\rho(\cdot)\psi,\psi\rangle - i\lambda) + i\overline{\omega}, \\ D_A\psi &= 0. \end{aligned} \tag{10.2}$$

Here $r \geq 1$ (the deformation involves taking $r \to \infty$), and $\overline{\omega}$ denotes the harmonic 1-form whose Hodge star represents the image of $\pi c_1(K^{-1})$ in $H^2(Y;\mathbb{R})$. Also remember that now A denotes a connection on E, and ψ is a section of $E \oplus K^{-1}E$. Taubes's equations (10.2) are equivalent to a case of the perturbed Seiberg-Witten equations (8.4), via rescaling the spinor by a factor of $\sqrt{r}$ and taking μ to be the exact 2-form

$$\mu = -rd\lambda - iF_{A_0} + 2*\overline{\omega}.$$

Some parts of the argument involve further perturbations of the equations (10.2) in order to obtain necessary transversality. For simplicity, we will suppress these in the exposition below.

10.3. **Convergence to Reeb orbits.**

There is now the following theorem which finds Reeb orbits from a sequence of solutions to (10.2).

Theorem 10.1. *(special case of* [42, Thm. 2.1]*) Let (r_n, ψ_n, A_n) be a sequence of solutions to the equations* (10.2) *with $r_n \to \infty$. Suppose that:*

1. *There is a constant $c > 0$ with $\sup_Y(1 - |\psi_n|) > c$.*
2. *There is a constant $C < \infty$ with $i\int_Y \lambda \wedge F_{A_n} < C$.*

Then there exists a nonempty orbit set a with $[a] = \mathrm{PD}(c_1(E))$.

This is a three-dimensional analogue of an earlier theorem of Taubes for symplectic four-manifolds, part of the SW=Gr story, that obtains holomorphic curves from sequences of solutions to the Seiberg-Witten equations perturbed using the symplectic form. The idea of the proof of Theorem 10.1 is to write $\psi_n = (\alpha_n, \beta_n)$, where α_n is a section of E and β_n is a section of $K^{-1}E$, and show that one can pass to a subsequence such that $\alpha_n^{-1}(0)$ converges as a current to a nonempty orbit set. This will then, of course, represent the Poincare dual of $c_1(E)$. In fact, when n is large, $|\beta_n|$ will be close to zero everywhere, while $|\alpha_n|$ will be close to 1 except near its zero set, which is where the curvature F_{A_n} will be concentrated.

Assumption (1) is needed to avoid solutions with α_n nonvanishing, which can exist when $c_1(E) = 0$. In fact, if $c_1(E) = 0$, then for any $c > 0$, if r is sufficiently large then there exists a unique (up to gauge equivalence) solution (A,ψ) to the perturbed Seiberg-Witten equations (10.2) with $1 - |\psi| < c$ everywhere, see Lemma 10.2 below. This solution has $|\alpha|$ close to 1 and $|\beta|$ close to zero everywhere; it corresponds to the ECH generator given by the empty set of Reeb orbits.

The idea of assumption (2) is that when n is large, $i\int_Y \lambda \wedge F_{A_n}$ is approximately 2π times the symplectic action of the orbit set to which $\alpha_n^{-1}(0)$ is converging. A uniform upper bound on this integral is needed in order to obtain an orbit set of finite length. Without condition (2) one can still obtain an invariant set for the Reeb flow, but this might not contain any Reeb orbit.

Note also that assumption (2) guarantees that (A_n, ψ_n) is irreducible when n is sufficiently large, because it follows from the equations (10.2) that if $(A, 0)$ is a reducible solution to (10.2) then $i\int_Y \lambda \wedge F_A$ is a linear, increasing function of r.

10.4. **Avoiding the empty set.** Now fix E such that $c_1(K^{-1}) + 2c_1(E)$ is torsion in $H^2(Y;\mathbb{Z})$. Let $\mathfrak{s}$ denote the corresponding spin-c structure. Kronheimer-Mrowka's Theorem 7.5 guarantees the existence of solutions to the perturbed equations (10.2) for all $r \geq 1$. To complete the proof of the Weinstein conjecture, we need to find a sequence of such solutions with $r \to \infty$ such that conditions (1) and (2) in Theorem 10.1 are satisfied.

One can achieve condition (1) using the following lemma:

Lemma 10.2.

- *If $c_1(E) \neq 0$, then there is a constant $c > 0$ such that if r is sufficiently large, and if (A,ψ) is a solution to the equations* (10.2)*, then there exist points in Y where $1 - |\psi| \geq 1 - c/\sqrt{r}$.*
- *If $c_1(E) = 0$, then there is a constant $c > 0$ such that if r is sufficiently large then there exists a unique (up to gauge equivalence) solution (A_0,ψ_0) to the equations* (10.2) *such that $1 - |\psi_0| \leq c$. In fact $1 - |\psi_0| \leq c/\sqrt{r}$. The grading of (A_0,ψ_0) in the Seiberg-Witten Floer chain complex is independent of r.*

This means that if $c_1(E) \neq 0$ then any sequence of solutions will automatically satisfy condition (1); and if $c_1(E) = 0$ then a sequence of solutions will satisfy condition (1) as long as we avoid the one bad gauge equivalence class (A_0,ψ_0),which we can easily do since we know from Theorem 7.5 that the Seiberg-Witten Floer homology is nonzero in infinitely many gradings.

10.5. **Three functionals.** The hardest part of the proof is to achieve condition (2) in Theorem 10.1. This is a new problem which does not arise in the four-dimensional SW=Gr story. (On a symplectic four-manifold (X,ω), one needs an analogue of condition (2) in which λ replaced by the symplectic form ω; but there the quantity that needs to be bounded is constant because the symplectic form is closed.)

The first step is to write the Chern-Simons-Dirac functional $\mathcal{F}$ in (9.1), of which Seiberg-Witten Floer homology is the Morse homology, as the sum of two other functionals. To do so, fix a reference connection A_1 on E.

Definition 10.3. If A is a connection on E, define the *Chern-Simons functional* $cs(A)$ by

$$cs(A) := -\int_Y (A - A_1) \wedge (F_A + F_{A_1} - 2i * \varpi).$$

Note that this is gauge invariant thanks to our assumption that $2c_1(E) + c_1(K^{-1})$ is torsion. Also, define the *energy*

$$\mathcal{E}(A) := i\int_Y \lambda \wedge F_A.$$

This is the quantity that we want to control.

Observe now that for a given r, if in the definition of $\mathcal{F}$ we take our reference connection on $\det(\mathbb{S})$ to be $A_0 + 2A_1$, then we have

$$\mathcal{F}(A,\psi) = \frac{1}{2}\left(cs(A) - r\mathcal{E}(A)\right) + \frac{r}{2}\int_Y \langle D_A\psi,\psi\rangle d\,\mathrm{vol},$$

up to the addition of an r-dependent constant. Since adding a constant to $\mathcal{F}$ does not affect its Morse homology, we will ignore this constant and take the above equation to be the new definition of $\mathcal{F}$. In particular, if (A,ψ) is a solution to the perturbed Seiberg-Witten equations (10.2), then the three functionals in play are related by

$$\mathcal{F}(A,\psi) = \frac{1}{2}\left(cs(A) - r\mathcal{E}(A)\right). \tag{10.3}$$

10.6. **A piecewise smooth family of solutions.** The next step is:

Lemma 10.4. *One can choose for each r sufficiently large a solution $(A(r),\psi(r))$ to the equations* (10.2) *such that:*

- *$(A(r),\psi(r))$ is a piecewise smooth function of r.*
- *$\mathcal{F}(A(r),\psi(r))$ is a continuous function of r.*
- *For all r at which $(A(r),\psi(r))$ is smooth as a function of r, the grading of $(A(r),\psi(r))$ in the Seiberg-Witten Floer chain complex is well-defined and independent of r.*
- *If $c_1(E) = 0$, then $(A(r),\psi(r))$ is not gauge equivalent to the solution (A_0,ψ_0) described in Lemma 10.2.*

The idea of the proof of Lemma 10.4 is as follows. First, one shows that for any given grading, if r is sufficiently large then all generators of the chain complex defining $\widehat{HM}_*(Y,\mathfrak{s})$ with that grading are irreducible. This is proved using a spectral flow estimate related to Proposition 10.7 below. Thus, for any given range of gradings, if r is sufficiently large then the differential in the Seiberg-Witten Floer chain complex just counts (perturbed) gradient flow lines of $\mathcal{F}$, without the subtleties arising from reducibles.

Now by Theorem 7.5, there is a nonzero class σ in the Seiberg-Witten Floer homology $\widehat{HM}_*(Y,\mathfrak{s})$ whose grading is not the same as that of (A_0,ψ_0). Fix such a class σ.

For any given r, a chain representing the class σ can be expressed as $\sum_i n_i c_i$, where n_i is a nonzero integer and c_i is a critical point of $\mathcal{F}$ for each i in some finite set. Define $f(r)$ to be the minimum, over all chains $\sum_i n_i c_i$ representing σ, of $\max_i \mathcal{F}(\sigma_i)$. The idea is then to define $(A(r),\psi(r))$ to be the maximal $\mathcal{F}$ critical point in a representative of σ realizing the minimum $f(r)$. One can choose this $(A(r),\psi(r))$ to vary piecewise smoothly with r, jumping when the "Morse complex" undergoes a bifurcation involving the critical point $(A(r),\psi(r))$.

To complete the proof of Lemma 10.4, one needs to show that the function $f(r)$ defined above is continuous. Taubes does so by explicitly studying the bifurcations that can happen in a generic one-parameter family of "Morse complexes". It might also be possible to prove this by estimating that the continuation maps that relate the Seiberg-Witten Floer homologies for nearby values of r, compare §5.5, do not increase the functional $\mathcal{F}$ too much. This method has been used to prove analogous continuity results in symplectic Floer homology, see eg [39, §2.4].

Lemma 10.5. *Let $\{(A(r),\psi(r))\}$ be a piecewise smooth family from Lemma 10.4. Then*

$$\frac{d}{dr}\mathcal{F}(A(r),\psi(r)) = -\frac{1}{2}\mathcal{E}(A(r)).$$

Proof. This follows from a general principle: If X is a smooth manifold (finite or infinite dimensional), if $f : \mathbb{R}\times X \to \mathbb{R}$ is a smooth function, and if $\{x(t)\}$ is a smooth family of critical points of $f_t := f(t,\cdot)$ on X defined for t in some interval, then

$$\frac{d}{dt} f_t(x(t)) = \frac{\partial f}{\partial t}(t, x(t)).$$

To prove this one uses the chain rule to compute

$$\frac{d}{dt} f(t, x(t)) = \frac{\partial f}{\partial t}(t, x(t)) + df_t\left(\frac{dx(t)}{dt}\right),$$

and notes that the second term on the right vanishes because $x(t)$ is a critical point of f_t. □

10.7. **The energy dichotomy.** Let $\{(A(r), \psi(r)\}$ be a piecewise smooth family given by Lemma 10.4. To prove the Weinstein conjecture, by Theorem 10.1 we just need to show that there is a sequence $r_n \to \infty$ such that the energy $\mathcal{E}(A(r_n))$ is bounded. The next step is to show that if this is not the case, then there is a sequence $r_n \to \infty$ such that the energy $\mathcal{E}(A(r_n))$ grows at least linearly, and the Chern-Simons functional grows quadratically. The last step will be to show that this quadratic growth of the Chern-Simons functional leads to a contradiction.

Lemma 10.6. *Let $\{(A(r), \psi(r)\}$ be a piecewise smooth family given by Lemma 10.4. Then one of the following two alternatives holds:*

1. *There is a sequence $r_n \to \infty$ and a constant $C < \infty$ such that $\mathcal{E}(A(r_n)) < C$ for all n.*
2. *There is a sequence $r_n \to \infty$ and a constant $c > 0$ such that $\mathcal{E}(A(r_n)) \geq cr_n$ and $cs(A(r_n)) \geq cr_n^2$ for all n.*

The proof of Lemma 10.6 uses an inequality

$$|cs(A(r))| \leq c_0 r^{2/3} |\mathcal{E}(A(r))|^{4/3}. \tag{10.4}$$

The proof of (10.4) is too long to be included here, but it involves a priori estimates on solutions to the Seiberg-Witten equations, starting with (8.3) and the maximum principle. Granted this, we can now give:

Proof of Lemma 10.6. Introduce the shorthand $cs(r) := cs(A(r))$, $\mathcal{E}(r) := \mathcal{E}(A(r))$ and $\mathcal{F}(r) := \mathcal{F}(A(r), \psi(r))$. We can assume without loss of generality that $\mathcal{E}(r) > 0$ for all r sufficiently large (since otherwise case (1) holds). Now fix $\varepsilon_0 \in (0, 1/5)$. We consider two cases.

Case A: There is a sequence $r_n \to \infty$ with

$$cs(r_n) \geq \varepsilon_0 r_n \mathcal{E}(r_n)$$

for all n. It follows in this case from the inequality (10.4) that alternative (2) holds.

Case B: For all r sufficiently large,

$$cs(r) < \varepsilon_0 r \mathcal{E}(r). \tag{10.5}$$

In this case we will show that alternative (1) holds.

To do so, define

$$v(r) := \mathcal{E}(r) - \frac{cs(r)}{r} = -\frac{2\mathcal{F}}{r}.$$

It then follows from Lemma 10.5 that

$$\frac{dv}{dr} = \frac{cs}{r^2}. \tag{10.6}$$

On the other hand, the hypothesis (10.5) is equivalent to

$$\mathcal{E} < (1 - \varepsilon_0)^{-1} v. \tag{10.7}$$

It now follows from (10.6), (10.5) and (10.7) that

$$\frac{dv}{dr} < \frac{\varepsilon v}{r},$$

where $\varepsilon := (1-\varepsilon_0)^{-1}\varepsilon_0 < 1/4$. Therefore

$$v < c_1 r^{\varepsilon} \tag{10.8}$$

for some constant c_1. On the other hand, by (10.4), (10.7) and (10.8), we have

$$cs < c_2 r^{2/3+(4/3)\varepsilon}$$

for some constant c_2. Putting this back into (10.6), we get

$$\frac{dv}{dr} < c_2 r^{(4/3)(\varepsilon-1)}.$$

Since $\varepsilon < 1/4$, the exponent in the above inequality is less than -1. Consequently the above inequality can be integrated to show that v is bounded from above. Then $\mathcal{E}$ is also bounded from above by (10.7). □

10.8. **Controlling the Chern-Simons functional.** The last step in Taubes's proof of the Weinstein conjecture is the following proposition relating the Seiberg-Witten Floer grading to the Chern-Simons functional. To state it, if (A,ψ) is a solution to the perturbed Seiberg-Witten equations (10.2), let $\deg(A,\psi)$ denote its grading in the Seiberg-Witten Floer chain complex, and for r large let (A_0,ψ_0) denote the distinguished Seiberg-Witten Floer generator given by Lemma 10.2.

Proposition 10.7. *(special case of* [42, Prop. 5.1]*) There exists $\kappa > 0$ such that for all r sufficiently large, if (A,ψ) is a solution to* (10.2)*, then*

$$\left|\deg(A,\psi) - \deg(A_0,\psi_0) + \frac{1}{4\pi^2}cs(A)\right| < \kappa r^{31/16}.$$

This is proved using a new estimate on the spectral flow of Dirac operators (the latter determines the relative grading in Seiberg-Witten Floer homology). The proof in the present case is given in [42, §5], and a higher-dimensional generalization is proved in [44].

10.9. **Conclusion.** To prove the Weinstein Conjecture, more specifically Theorem 7.7, let $\Gamma \in H_1(Y)$ such that $c_1(\xi) + 2\,\mathrm{PD}(\Gamma)$ is torsion, and let E be the line bundle with $c_1(E) = \mathrm{PD}(\Gamma)$. Let $\{(A(r),\psi(r)\}$ be a piecewise smooth family given by Lemma 10.4. Alternative (2) in Lemma 10.6 is impossible by Proposition 10.7, because $\deg(A,\psi)$ and $\deg(A_0,\psi_0)$ are independent of r. So alternative (1) in Lemma 10.6 holds. Then we have a sequence $r_n \to \infty$ such that condition (2) in Theorem 10.1 holds. Condition (1) in Theorem 10.1 also holds by Lemma 10.2. Thus Theorem 10.1 applies to produce the desired nonempty orbit set.

References

1. C. Abbas, K. Cieliebak, and H. Hofer, *The Weinstein conjecture for planar contact structures in dimension three*, Comment. Math. Helv. **80** (2005), 771-793.
2. N. Berline, E. Getzler, and M. Vergne, *Heat kernels and Dirac operators*, Springer Verlag, 1990.
3. R. Bott, *Morse theory indomitable*, IHES Publ. Math. (1988) 99–114.
4. F. Bourgeois, *Odd dimensional tori are contact manifolds*, Int. Math. Res. Not. **2002**, no. 30, 1571–1574.

5. F. Bourgeois, *A Morse-Bott approach to contact homology*, Symplectic and contact topology: interactions and perspectives, 55–77, Fields Inst. Commum. 35, AMS, 2003.
6. F. Bourgeois, K. Cieliebak, and T. Ekholm, *A note on Reeb dynamics on the tight 3-sphere*, J. Modern Dynamics **1** (2007), 597–613.
7. F. Bourgeois, T. Ekholm, and Y. Eliashberg, in preparation.
8. V. Colin, E. Giroux, and K Honda, *Finitude homotopique et isotopique des structures de contact tendues*, arXiv:0805.3051.
9. V. Colin and K. Honda, *Reeb vector fields and open book decompositions*, arXiv:0809.5088.
10. Y. Eliashberg, *Classification of overtwisted contact structures on 3-manifolds*, Invent. Math. **98** (1989), 623–637.
11. Y. Eliashberg, A. Givental, and H. Hofer, *Introduction to symplectic field theory*, GAFA 2000, Special Volume, Part II, 560–673.
12. J. Etnyre, *Introductory lectures on contact geometry*, Proc. Sympos. Pure Math. 71, pp. 81–107, Amer. Math. Soc. 2003.
13. J. Etnyre, *Lectures on open book decompositions and contact structures*, Floer homology, gauge theory, and low-dimensional topology, 103–141, Clay Math. Proc. 5, AMS, 2006.
14. H. Geiges, *An introduction to contact topology*, Cambridge University Press, 2008.
15. V. Ginzburg and B. Gürel, *A C^2-smooth counterexample to the Hamiltonian Seifert conjecture in $\mathbb{R}^4$*, Ann. of Math. **158** (2003), 953–976.
16. E. Giroux, *Géometrie de contact: de la dimension trois vers les dimensions supérieures*, Proceedings of the ICM, Beijing, 2002, Vol. II, 405–414.
17. J. Harrison, *C^2 counterexamples to the Seifert conjecture*, Topology **27** (1998), 249–278.
18. H. Hofer, *Pseudoholomorphic curves in symplectizations with applications to the Weinstein conjecture in dimension three*, Invent. Math. **114** (1993), 515–563.
19. H. Hofer and E. Zehnder, *Symplectic invariants and Hamiltonian dynamics*, Birkhäuser, 1994.
20. M. Hutchings, *An index inequality for embedded pseudoholomorphic curves in symplectizations*, J. Eur. Math. Soc. **4** (2002), 313–361.
21. M. Hutchings, *The embedded contact homology index revisited*, arXiv:0805.1240, to appear in the Yashafest proceedings.
22. M. Hutchings and M. Sullivan, *Rounding corners of polygons and the embedded contact homology of T^3*, Geom. Topol. **10** (2006), 269–266.
23. M. Hutchings and C.H. Taubes, *An introduction to the Seiberg-Witten equations on symplectic four-manifolds*, Symplectic geometry and topology (Park City, UT, 1997), 103–142, AMS, 1999.
24. M. Hutchings and C.H. Taubes, *Gluing pseudoholomorphic curves along branched covered cylinders I*, J. Symplectic Geom. **5** (2007), 43–137.
25. M. Hutchings and C.H. Taubes, *The Weinstein conjecture for stable Hamiltonian structures*, arXiv:0809.0140.
26. P.B. Kronheimer and T.S. Mrowka, *Monopoles and 3-manifolds*, Cambridge University Press, 2008.
27. G. Kuperberg, *A volume-preserving counterexample to the Seifert conjecture*, Comment. Math. Helv. **71** (1996), 70–97.
28. G. Kuperberg and K. Kuperberg, *Generalized counterexamples to the Seifert conjecture*, Ann. of Math. **143** (1996), 547–576.
29. K. Kuperberg, *A smooth counterexample to the Seifert conjecture*, Ann. of Math. **140** (1994), 723–732.
30. H. Lawson and M. Michelson, *Spin geometry*, Princeton University Press, 1989.
31. MSRI Hot Topics Workshop, *Contact structures, dynamics and the Seiberg-Witten equations in dimension 3*, June 2008, videos at www.msri.org.
32. D. McDuff and D. Salamon, *Introduction to symplectic topology*, 2nd edition, Oxford University Press, 1998.
33. G. Meng and C.H. Taubes, *$\underline{SW}$ =Milnor torsion*, Math. Res. Lett. **3** (1996), 661–674.
34. J. Milnor, *Morse theory*, Princeton University Press, 1969.
35. J. Morgan, *The Seiberg-Witten equations and applications to the topology of smooth four-manifolds*, Mathematical Notes 44, Princeton Uinv. Press, 1996.
36. P. Ozsváth and Z. Szabó, *Holomorphic disks and topological invariants for closd three-manifolds*, Ann. of Math. **159** (2004), 1027–1158.

37. P. Rabinowitz, *Periodic solutions of Hamiltonian systems*, Comm. Pure Appl. Math **31** (1978), 157–184.
38. M. Schwarz, *Morse homology*, Progress in Mathematics, Birkhäuser, 1993.
39. M. Schwarz, *On the action spectrum for closed symplectially aspherical manifolds*, Pac. J. Math. **193** (2000), 419–461.
40. P. Schweizer, *Counterexamples to the Seifert conjecture and opening closed leaves of foliations*, Ann. of Math. **100** (1974), 386–400.
41. C.H. Taubes, *Seiberg-Witten and Gromov invariants for symplectic 4-manifolds*, First International Press Lecture Series 2, International Press, 2000.
42. C.H. Taubes, *The Seiberg-Witten equations and the Weinstein conjecture*, Geom. Topol. **11** (2007), 2117–2202.
43. C.H. Taubes, *The Seiberg-Witten equations and the Weinstein conjecture II: More closed integral curves for the Reeb vector field*, arXiv:math/0702366.
44. C.H. Taubes, *Asymptotic spectral flow for Dirac operators*, Comm. Anal. Geom. **15** (2007), 569–587.
45. C.H. Taubes, *Embedded contact homology and Seiberg-Witten Floer homology I-IV*, preprints, 2008.
46. V. Turaev, *A combinatorial formulation for the Seiberg-Witten invariants of* 3*-manifolds*, Math. Res. Lett. **5** (1998), 583–598.
47. C. Viterbo, *A proof of Weinstein's conjecture in* $\mathbb{R}^{2n}$, Annales de l'Institut Henri Poincaré-Anal. Non Linéaire, **4** (1987), 337–356.
48. A. Weinstein, *Periodic orbits for convex Hamiltonian systems*, Ann. of Math. **108** (1978), 507–518.
49. A. Weinstein, *On the hypothesis of Rabinowitz' periodic orbit theorems*, J. Differential Equations **33** (1979), 353–358.

Mathematics Department, 970 Evans Hall, University of California, Berkeley CA 94720

E-mail address: `hutching@math.berkeley.edu`

FRONTIERS OF REALITY IN SCHUBERT CALCULUS

FRANK SOTTILE

Abstract. The theorem of Mukhin, Tarasov, and Varchenko (formerly the Shapiro conjecture for Grassmannians) asserts that all (*a priori* complex) solutions to certain geometric problems from the Schubert calculus are actually real. Their proof is quite remarkable, using ideas from integrable systems, Fuchsian differential equations, and representation theory. Despite this advance, the original Shapiro conjecture is not yet settled. While it is false as stated, it has several interesting and not quite understood modifications and generalizations that are likely true.

These notes will introduce the Shapiro conjecture for Grassmannians, give some idea of its proofs and consequences, its links to other subjects, and sketch its extensions.

Introduction

While it is not unusual for a univariate polynomial f with real coefficients to have *some* real roots—under reasonable assumptions we expect $\sqrt{\deg f}$ real roots [20]—it is rare for a polynomial to have all of its roots be real. In fact, the primary example from nature that comes to mind is when f is the characteristic polynomial of a symmetric matrix, as all eigenvalues of a symmetric matrix are real.

Similarly, when a system of real polynomial equations has finitely many (*a priori* complex) solutions, we expect some, but likely not all, solutions to be real. In fact, upper bounds on the number of real solutions [1, 18] sometimes ensure that not all solutions can be real. As before, the primary example that comes to mind of a system with only real solutions is the system of equations for the eigenvectors and eigenvalues of a real symmetric matrix.

Here is another system of polynomial equations which also turns out to have only real solutions. The Wronskian of univariate polynomials $f_0, \ldots, f_n \in \mathbb{C}[t]$ is the determinant

$$\det \begin{pmatrix} f_0(t) & f_1(t) & \cdots & f_n(t) \\ f_0'(t) & f_1'(t) & \cdots & f_n'(t) \\ \vdots & \vdots & \ddots & \vdots \\ f_0^{(n)}(t) & f_1^{(n)}(t) & \cdots & f_n^{(n)}(t) \end{pmatrix} .$$

Up to a scalar multiple, the Wronskian depends only upon the linear span P of the polynomials $f_0, \ldots, f_n$ in the vector space $\mathbb{C}[t]$ of all polynomials. This scaling retains only the information of the roots of the Wronskian and their multiplicities. Recently, Mukhin, Tarasov, and Varchenko [22] proved the remarkable (but seemingly innocuous) result.

2000 *Mathematics Subject Classification.* 14M15, 14N15.

Key words and phrases. Schubert calculus, Bethe ansatz.

Work of Sottile supported by NSF grant DMS-0701050.

Theorem 1. *If the Wronskian of a space P of polynomials has only real roots, then P has a basis of real polynomials.*

While not immediately apparent, those $(n+1)$-dimensional subspaces P of $\mathbb{C}[t]$ with a given Wronskian W are the solutions to a system of polynomial equations which depend on the roots of W. In Section 1, we explain how the Shapiro conjecture for Grassmannians is equivalent to Theorem 1.

The proof of Theorem 1 uses the Bethe ansatz for the Gaudin model on certain modules (representations) of the Lie algebra $\mathfrak{sl}_{n+1}\mathbb{C}$. This is a method to decompose a module of $\mathfrak{sl}_{n+1}\mathbb{C}$ into irreducible submodules that is compatible with a family of commuting operators called the Gaudin Hamiltonians. It includes a set-theoretic map from the spaces of polynomials with a given Wronskian to certain Bethe eigenvectors for the Gaudin Hamiltonians. A coincidence of numbers, from the Schubert calculus on one hand and from representation theory on the other, implies that this map is a bijection. As the Gaudin Hamiltonians are symmetric with respect to the positive definite Shapovalov form, their eigenvectors and eigenvalues are real. Theorem 1 follows as eigenvectors with real eigenvalues must come from real spaces of polynomials. We describe this in Sections 2, 3, and 4.

The geometry behind the statement of Theorem 1 appears in many other guises, some of which we describe in Section 5. These include linear series on the projective line [5] and rational curves with prescribed flexes [17]. A special case of the Shapiro conjecture concerns a remarkable statement about rational functions with prescribed critical points, and was proved in this form by Eremenko and Gabrielov [8]. They showed that a rational function whose critical points lie on a circle in the Riemann sphere maps that circle to another circle.

A generalization of Theorem 1 by Mukhin, Tarasov, and Varchenko [24] implies the following particularly attractive statement from matrix theory. Let $\beta_1,\dots,\beta_n$ be distinct real numbers, $\alpha_1,\dots,\alpha_n$ be complex numbers, and consider the matrix

$$Z := \begin{pmatrix} \alpha_1 & (\beta_1-\beta_2)^{-1} & \cdots & (\beta_1-\beta_n)^{-1} \\ (\beta_2-\beta_1)^{-1} & \alpha_2 & \cdots & (\beta_2-\beta_n)^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ (\beta_n-\beta_1)^{-1} & (\beta_n-\beta_2)^{-1} & \cdots & \alpha_n \end{pmatrix}.$$

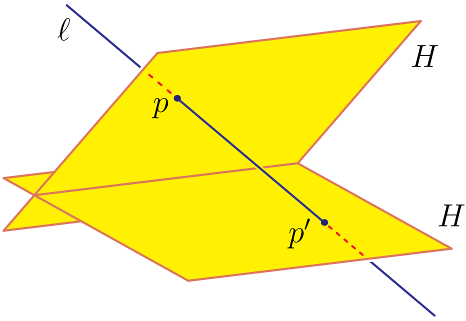
Theorem 2. *If Z has only real eigenvalues, then $\alpha_1,\dots,\alpha_n$ are real.*

Unlike its proof, the statement of Theorem 2 has nothing to do with Schubert calculus or representations of $\mathfrak{sl}_{n+1}\mathbb{C}$ or integrable systems, and it remains a challenge to prove it directly.

The statement and proof of Theorem 1 is only part of this story. Theorem 1 settles (for Grassmannians) a conjecture about Schubert calculus originally made by Boris Shapiro and Michael Shapiro in 1993/4. While this Shapiro conjecture is false for most other flag manifolds, there are appealing corrections and generalizations supported by theoretical evidence and also by overwhelming computational evidence. We sketch this in Section 6.

There are now three proofs [14, 22, 25] of the Shapiro conjecture for Grassmannians, all passing through integrable systems and representation theory. In the second proof [14], the solutions are identified with certain representations of a real rational Cherednik algebra [11], and reality follows as these representations are necessarily real. The third proof [25] provides a surprising and deep connection between the Schubert calculus and the representation theory of $\mathfrak{sl}_{n+1}\mathbb{C}$. We will only treat the first proof in these notes. During 2009, these notes will be expanded into a more complete treatment of this remarkable conjecture and its proofs.

First steps: the problem of four lines. We close this Introduction by illustrating the Schubert calculus and the Shapiro conjecture with some beautiful geometry. Consider the set of all lines in three-dimensional space. This set (a Grassmannian) is four-dimensional, which we may see by counting degrees of freedom for a line ℓ as follows. Fix planes H and H' that meet ℓ in points p and p'.



Since each point p and p' has two degrees of freedom to move within its plane, we see that the line ℓ enjoys four degrees of freedom.

Similarly, the set of lines that meet a fixed line is three-dimensional. More parameter counting tells us that if we fix four lines, then the set of lines that meet each of our fixed lines will be zero-dimensional. That is, it consists of finitely many lines. The Schubert calculus gives algorithms to determine this number of lines. We instead use elementary geometry to show that this number is 2.

The Shapiro conjecture asserts that if the four fixed lines are chosen in a particular way, then both solution lines will be real. This special choice begins by specifying a twisted cubic curve, γ. While any twisted cubic will do, we'll take the one with parameterization

$$\gamma\ :\ t\ \longmapsto (6t^2-1,\ \tfrac{7}{2}t^3+\tfrac{3}{2}t,\ \tfrac{3}{2}t-\tfrac{1}{2}t^3)\,. \tag{1}$$

Our fixed lines will be four lines tangent to γ.

We understand the lines that meet our four tangent lines by first considering lines that meet three tangent lines. We are free to fix the first three tangent points to be any of our choosing, for instance, $\gamma(-1)$, $\gamma(0)$, and $\gamma(1)$. Then the three lines $\ell(-1)$, $\ell(0)$, and $\ell(1)$ tangent at these points have parameterizations

$$(-5+s, 5-s, -1)\,,\quad (-1, s, s)\,,\quad \text{and}\quad (5+s, 5+s, 1)\quad \text{for } s\in\mathbb{R}.$$

These lines all lie on the hyperboloid H of one sheet defined by

$$x^2-y^2+z^2\ =\ 1\,, \tag{2}$$

which has two rulings by families of lines. The lines $\ell(-1)$, $\ell(0)$, and $\ell(1)$ lie in one family, and the other family consists of the lines meeting $\ell(-1)$, $\ell(0)$, and $\ell(1)$. This family is drawn on the hyperboloid H in Figure 1.

The lines that meet $\ell(-1)$, $\ell(0)$, $\ell(1)$, and a fourth line $\ell(s)$ will be those in this second family that also meet $\ell(s)$. In general, there will be two such lines, one for each point of intersection of line $\ell(s)$ with H, as H is defined by the quadratic polynomial (2). The remarkable geometric fact is that every such tangent line, $\ell(s)$ for $s \not\in \{-1, 0, 1\}$, will meet the hyperboloid in two real points. We illustrate this when $s = 0.31$ in Figure 1, highlighting the two solution lines.

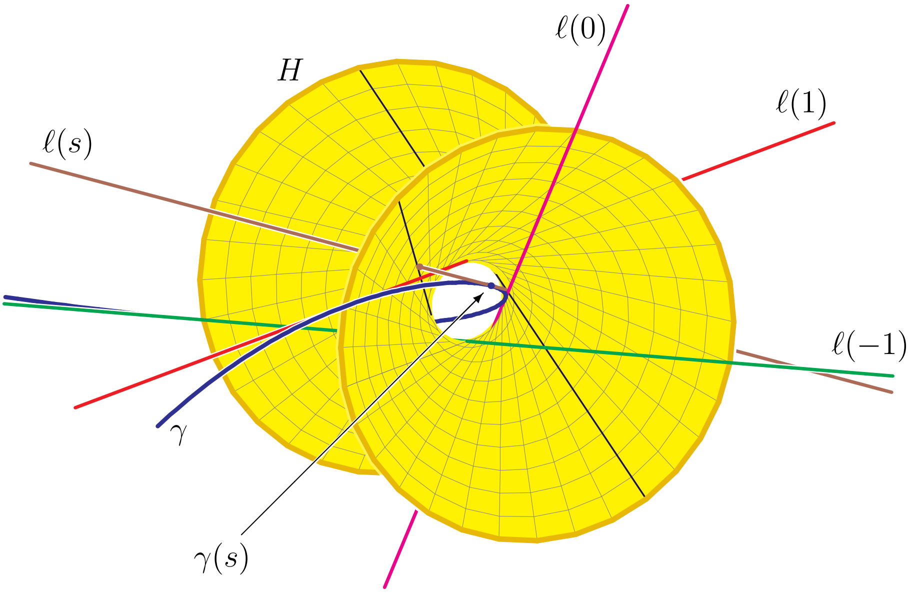


FIGURE 1. The problem of four lines.

The Shapiro conjecture and its extensions claim that this reality always happens: If the conditions for a Schubert problem are chosen relative to a rational normal curve (here, the twisted cubic curve γ of (1)), then all solutions will be real. When the Schubert problem comes from a Grassmannian (like this problem of four lines), the Shapiro conjecture is true—this is the theorem of Mukhin, Tarasov, and Varchenko. For most other flag manifolds, it is known to fail, but in very interesting ways.

Our plan is to explain this conjecture more precisely for Grassmannians, outline the first proof of Mukhin, Tarasov, and Varchenko, and then give some of its consequences. Along the way we will discuss some special cases of the conjecture which at first glance do not seem to have any relation to Schubert calculus or representation theory. We conclude with a sketch of the emerging landscape of conjectures for other flag manifolds which generalize and correct the Shapiro conjecture.

Acknowledgments. We thank the many people who have helped us to better learn this story and to improve this exposition. In particular, we thank Eugene Mukhin, Alexander Varchenko, Aaron Lauve, Zach Teitler, and Nickolas Hein.

1. The Shapiro conjecture for Grassmannians

Let $\mathbb{C}_d[t]$ be the set of complex polynomials of degree at most d in the indeterminate t, a vector space of dimension $d+1$. Fix a positive integer $n < d$, and let $\text{Grass}_{n,d}$ be the set of all $(n+1)$-dimensional linear subspaces P of $\mathbb{C}_d[t]$. This *Grassmannian* is a complex manifold of dimension $(n+1)(d-n)$ [15, Ch. 1.5].

The main character in our story is the Wronski map, which associates a point $P \in \mathrm{Grass}_{n,d}$ to the Wronskian of a basis for P. If $\{f_0(t), \dots, f_n(t)\}$ is a basis for P, its Wronskian is the determinant of the derivatives of the basis,

$$\mathrm{Wr}(f_0, \dots, f_n) \;:=\; \det \begin{pmatrix} f_0 & f_1 & \cdots & f_n \\ f_0' & f_1' & \cdots & f_n' \\ \vdots & \vdots & \ddots & \vdots \\ f_0^{(n)} & f_1^{(n)} & \cdots & f_n^{(n)} \end{pmatrix}, \tag{1.1}$$

which is a nonzero polynomial of degree of at most $(n+1)(d-n)$. This does not quite define a map $\mathrm{Grass}_{n,d} \to \mathbb{C}_{(n+1)(d-n)}[t]$, as choosing a different basis for P multiplies the Wronskian by a nonzero constant. If we consider the Wronskian up to a nonzero constant, we obtain the *Wronski map*

$$\mathrm{Wr} \;:\; \mathrm{Grass}_{n,d} \;\longrightarrow\; \mathbb{P}(\mathbb{C}_{(n+1)(d-n)}[t]) \;\simeq\; \mathbb{P}^{(n+1)(d-n)}, \tag{1.2}$$

where $\mathbb{P}(V)$ denotes the projective space consisting of all 1-dimensional linear subspaces of a vector space V.

We restate Theorem 1, the simplest version of the Theorem of Mukhin, Tarasov, and Varchenko [22].

Theorem 1. *If the Wronskian of a space P of polynomials has only real roots, then P has a basis of real polynomials.*

The problem of four lines is a special case of Theorem 1 when $d = 3$ and $n = 1$. To see this, note that if we apply an affine function $a + bx + cy + dz$ to the curve $\gamma(t)$ of (1), we obtain a cubic polynomial in $\mathbb{C}_3[t]$, and every cubic polynomial comes from a unique affine function. A line ℓ in $\mathbb{C}^3$ (actually in $\mathbb{P}^3$) is cut out by a two-dimensional space of affine functions, which gives a 2-dimensional space P_ℓ of polynomials in $\mathbb{C}_3[t]$, and hence a point $P_\ell \in \mathrm{Grass}_{n,d}$.

It turns out that the Wronskian of a point $P_\ell \in \mathrm{Grass}_{n,d}$ is a quartic polynomial with a root at $s \in \mathbb{C}$ if and only if the line ℓ corresponding to P_ℓ meets the line $\ell(s)$ tangent to the curve γ at $\gamma(s)$. Thus a line ℓ meets four lines tangent to γ at real points if and only if the corresponding point $P_\ell \in \mathrm{Grass}_{n,d}$ has a Wronskian whose roots are these four points. Since these points are real, Theorem 1 implies that P_ℓ has a basis of real polynomials. Thus ℓ is cut out by real affine functions, and hence is real.

1.1. Geometric form of the Shapiro Conjecture.

Let $P \in \mathrm{Grass}_{n,d}$ be a subspace. We consider the order of vanishing at a point $s \in \mathbb{C}$ of polynomials in a basis for P. There will be a minimal order a_0 of vanishing for these polynomials. Suppose that f_0 vanishes to this order. Subtracting an appropriate multiple of f_0 from each of the other polynomials, we may assume that they vanish to order greater than a_0 at s. Let a_1 be the minimal order of vanishing at s of these remaining polynomials. Continuing in this fashion, we obtain a basis $f_0, \dots, f_n$ of P and a sequence

$$0 \le a_0 < a_1 < \cdots < a_n \le d\,,$$

where f_i vanishes to order a_i at s. We call this sequence $\mathbf{a}_P(s)$ the *ramification* of P at s. For a sequence $\mathbf{a} : 0 \le a_0 < a_1 < \cdots < a_n \le d$, write $\Omega^\circ_{\mathbf{a}}(s)$ for the set of points $P \in \mathrm{Grass}_{n,d}$ with ramification $\mathbf{a}$ at s, which is a Schubert cell of $\mathrm{Grass}_{n,d}$. It has codimension

$$|\mathbf{a}| \;:=\; a_0 \;+\; a_1{-}1 \;+\; \cdots \;+\; a_n{-}n\,,$$

as may be seen by expanding the basis $f_0,\dots,f_n$ of P in the basis $\{(t-s)^i \mid i=0,\dots,d\}$ of $\mathbb{C}_d[t]$. Since $f_j^{(i)}$ vanishes to order at least a_j-i at s and $f_i^{(i)}$ vanishes to order exactly a_i-i at s, we see that the Wronskian of a subspace $P\in\Omega^\circ_{\mathbf{a}}(s)$ vanishes to order exactly $|\mathbf{a}|$ at s.

Let $\mathrm{Grass}^\circ_{n,d}$ consist of subspaces $P\in\mathrm{Grass}_{n,d}$ that have a basis $f_0,\dots,f_n$ where f_i has degree $d{-}n{+}1$. This is an open subset of $\mathrm{Grass}_{n,d}$. When $P\in\mathrm{Grass}^\circ_{n,d}$, we obtain the Plücker formula for the total ramification of a general subspace P of $\mathbb{C}_d[t]$,

$$\dim\mathrm{Grass}_{n,d} \;=\; \sum_{s\in\mathbb{C}} |\mathbf{a}_P(s)|\,. \tag{1.3}$$

In general, the total ramification of P is bounded by the dimension of $\mathrm{Grass}_{n,d}$. (One may also define ramification at infinity for subspaces $P\not\in\mathrm{Grass}^\circ_{n,d}$, to obtain the Plücker formula in full generality.) If $\mathbf{a}_P(s)=0<1<\cdots<n$, so that $|\mathbf{a}_P(s)|=0$, then we say that P is *unramified* at s. In this language, Theorem 1 states that if a subspace $P\in\mathrm{Grass}_{n,d}$ is ramified only at real points, then P is real in that it has a basis of real polynomials.

Let us introduce some more geometry. Let W be the Wronskian of P. Then

$$P \;\in\; \bigcap_{s\colon W(s)=0} \Omega^\circ_{\mathbf{a}_P(s)}(s)\,,$$

and this intersection consists of all subspaces with Wronskian W. In particular, P lies in the intersection of the closures of these Schubert cells, which we now describe. For each $s\in\mathbb{C}$, $\mathbb{C}_d[t]$ has a complete flag of subspaces

$$F_\bullet(s) \;:\; \mathbb{C}\cdot(t{-}s)^d \;\subset\; \mathbb{C}_1[t]\cdot(t{-}s)^{d-1} \;\subset\; \cdots \;\subset\; \mathbb{C}_{d-1}[t]\cdot(t{-}s) \;\subset\; \mathbb{C}_d[t]\,.$$

More generally, a flag $F_\bullet$ is a sequence of subspaces

$$F_\bullet \;:\; F_1 \;\subset\; F_2 \;\subset\; \cdots \;\subset\; F_d \;\subset \mathbb{C}_d[t]\,,$$

where F_i has dimension i. For a sequence $\mathbf{a}$ and a flag $F_\bullet$, the *Schubert variety*

$$\{P\in\mathrm{Grass}_{n,d} \mid \dim\big(P\cap F_{d+1-a_j}\big)\geq n{+}1-j,\ \text{for } j=0,1,\dots,n\}\,, \tag{1.4}$$

is a subvariety of $\mathrm{Grass}_{n,d}$, written $\Omega_{\mathbf{a}}F_\bullet$. It consists of linear subspaces P having special position (encoded by $\mathbf{a}$) with respect to the flag $F_\bullet$. Since $\dim(P\cap F_{d+1-i}(s))$ counts the number of linearly independent polynomials in P that vanish to order at least i at s, we see that $\Omega^\circ_{\mathbf{a}}(s)\subset\Omega_{\mathbf{a}}F_\bullet(s)$. More precisely, $\Omega_{\mathbf{a}}F_\bullet(s)$ is the closure of the Schubert cell $\Omega^\circ_{\mathbf{a}}(s)$ and it is the disjoint union of $\Omega^\circ_{\mathbf{b}}(s)$ for $\mathbf{b}\geq\mathbf{a}$, where $\geq$ is componentwise comparison.

Given ramification sequences $\mathbf{a}^{(1)},\dots,\mathbf{a}^{(m)}$ and flags $F_\bullet^{(1)},\dots,F_\bullet^{(m)}$, the intersection

$$\Omega_{\mathbf{a}^{(1)}}F_\bullet^{(1)} \bigcap \Omega_{\mathbf{a}^{(2)}}F_\bullet^{(2)} \bigcap \cdots \bigcap \Omega_{\mathbf{a}^{(m)}}F_\bullet^{(m)} \tag{1.5}$$

consists of those linear subspaces $P \in G$ having specified position $\mathbf{a}^{(i)}$ with respect to the flag $F_\bullet^{(i)}$, for each $i = 1, \ldots, m$. Kleiman [19] showed that if the flags $F_\bullet^{(i)}$ are general, then the intersection (1.5) is (generically) transverse.

A *Schubert problem* is a list $\mathbf{A} := (\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)})$ of sequences satisfying

$$(n{+}1)(d{-}n)\ (=\ \dim \mathrm{Grass}_{n,d}\,)\ =\ |\mathbf{a}^{(1)}| + \cdots + |\mathbf{a}^{(m)}|\,.$$

Given a Schubert problem, Kleiman's Theorem implies that a general intersection (1.5) will be zero-dimensional and thus consist of finitely many points. By transversality, the number $\delta(\mathbf{A})$ of these points is independent of choice of general flags. The Schubert calculus, through the Littlewood-Richardson rule [12], gives algorithms to determine $\delta(\mathbf{A})$.

We mention an important special case. Let $\boldsymbol{\iota} := 0 < 1 < \cdots < n{-}1 < n{+}1$ be the unique ramification sequence with $|\boldsymbol{\iota}| = 1$, and suppose that $(\boldsymbol{\iota}, \ldots, \boldsymbol{\iota})$ is a Schubert problem, so that $(n{+}1)(d{-}n)$ is the number of occurrences of $\boldsymbol{\iota}$. Write $\boldsymbol{\iota}_{n,d}$ for this sequence. Schubert [30] gave the following formula

$$\delta(\boldsymbol{\iota}_{n,d})\ =\ [(n{+}1)(d{-}n)]!\,\frac{1!2!\cdots n!}{(d{-}n)!(d{-}n{+}1)!\cdots d!}\ . \tag{1.6}$$

By (1.3), the total ramification $(\mathbf{a}_P(s) \mid |\mathbf{a}_P(s)| > 0)$ of a subspace $P \in \mathrm{Grass}^\circ_{n,d}$ is a Schubert problem. Let W be the Wronskian of P. We would like the intersection containing P

$$\bigcap_{s\colon W(s)=0} \Omega_{\mathbf{a}_P(s)} F_\bullet(s) \tag{1.7}$$

to be transverse and zero-dimensional. However, Kleiman's Theorem does not apply, as the flags $F_\bullet(s)$ for s a root of W are not generic. For example, in the problem of four lines, if the Wronskian is $t^4 - t$, then the corresponding intersection (1.7) of Schubert varieties is not transverse. (This worked out in detail in [5, §9].)

We can see that this intersection (1.7) is however always zero-dimensional. Note that any positive-dimensional subvariety meets $\Omega_{\boldsymbol{\iota}} F_\bullet$, for any flag $F_\bullet$. (This is because, for example, $\Omega_{\boldsymbol{\iota}} F_\bullet$ is a hyperplane section of $\mathrm{Grass}_{n,d}$ in its Plücker embedding into projective space.) In particular, if the intersection (1.7) is not zero-dimensional, then given a point $s \in \mathbb{P}^1$ with $W(s) \neq 0$, there will be a point P' in (1.7) which also lies in $\Omega_\iota F_\bullet(s)$. But then the total ramification of P' does not satisfy the Plücker formula (1.3), as its ramification strictly contains the total ramification of P.

A consequence of this argument is that the Wronski map (1.2) is a finite map. In particular, all of its fibers are finite. The intersection number $\delta(\boldsymbol{\iota}_{n,d})$ in (1.6) is an upper bound for the cardinality of a fiber. By Sard's Theorem, this upper bound is obtained for generic Wronskians.

Theorem 1.8. *There are finitely many spaces of polynomials $P \in \mathrm{Grass}_{n,d}$ with a given Wronskian. For a general polynomial $W(t)$ of degree $(n{+}1)(d{-}n)$, there are exactly $\delta(\boldsymbol{\iota}_{n,d})$ spaces of polynomials with Wronskian $W(t)$.*

When W has distinct roots, these spaces of polynomials are exactly the points in the intersection (1.7), where $\mathbf{a}_P(s) = \boldsymbol{\iota}$ at each root s of W. A limiting argument, in which the roots of the Wronskian are allowed to collide one-by-one, proves a local form of Theorem 1.

Theorem 1.9 ([32]). *If the roots of a polynomial $W(t)$ of degree $(n+1)(d-n)$ are real, distinct, and sufficiently clustered together, then there are $\delta(\boldsymbol{\iota}_{n,d})$ spaces of polynomials with Wronskian $W(t)$, so that the intersection* (1.7) *is transverse, and each such space of polynomials is real.*

We noted that the intersection (1.7) is not transverse when $d = 3$, $n = 1$, and $W(t) = t^4 - t$. It turns out that it is always transverse when the roots of the Wronskian are distinct and real. This is the stronger form of the Theorem of Mukhin, Tarasov, and Varchenko.

Theorem 1.10 ([25]). *For any Schubert problem $(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)})$ and any distinct real numbers $s_1, \ldots, s_m$, the intersection*

$$\Omega_{\mathbf{a}^{(1)}}F_\bullet(s_1) \bigcap \Omega_{\mathbf{a}^{(2)}}F_\bullet(s_2) \bigcap \cdots \bigcap \Omega_{\mathbf{a}^{(m)}}F_\bullet(s_m) \tag{1.11}$$

is transverse and consists solely of real points.

This theorem (without the transversality) is the original statement of the conjecture of Shapiro and Shapiro for Grassmannians, which was posed in exactly this form to the author in May 1995. The Shapiro conjecture was first discussed and studied in detail in [33], where significant computational evidence was presented (see also [35] and [28]). The results and computations in [33], as well as the result of Theorem 1.9, highlighted the key role that transversality seemed to play in the conjecture. This conjecture also appeared in [31].

We will not discuss the proof of Theorem 1.10, except to remark that its main ingredient is an isomorphism between algebraic objects associated to the intersection (1.11) and to certain representation-theoretic data. This isomorphism provides a very deep link between Schubert calculus for the Grassmannian and the representation theory of $\mathfrak{sl}_{n+1}\mathbb{C}$.

We will however sketch the proof of Theorem 1 in the next three sections.

2. Spaces of polynomials with given Wronskian

Theorem 1.8 enables the reduction of Theorem 1 to a special case. Since the Wronski map is finite, a standard limiting argument (given for example in Section 1.3 of [22] or Remark 3.4 of [33]) shows that it suffices to prove Theorem 1 when the Wronskian has distinct real roots that are sufficiently general. Since $\delta(\boldsymbol{\iota}_{n,d})$ is the upper bound for the number of spaces of polynomials with given Wronskian, it suffices to construct this number of distinct spaces of real polynomials with a given Wronskian, when the Wronskian has distinct real roots that are sufficiently general. In fact, this is exactly what Mukhin, Tarasov, and Varchenko do.

Theorem 1′. *If $s_1, \ldots, s_{(n+1)(d-n)}$ are generic real numbers, there exist exactly $\delta(\boldsymbol{\iota}_{n,d})$ distinct real vector spaces of polynomials P with Wronskian $\prod_i (t - s_i)$.*

The proof proceeds by first constructing $\delta(\boldsymbol{\iota}_{n,d})$ distinct spaces of polynomials with a given Wronskian having generic complex roots, which we describe in Section 2.1. This uses a Fuchsian differential equation given by the critical points of a remarkable symmetric function, called the master function. Critical points of the master function are also used in the Bethe ansatz for the Gaudin model, which is a method for decomposing a representation V of $\mathfrak{sl}_{n+1}\mathbb{C}$ into irreducibles that is compatible with the action of certain commuting operators, called the Gaudin Hamiltonians. In particular, a critical point of the master function gives a Bethe eigenvector of the Gaudin Hamiltonians which is also a highest weight vector for an irreducible submodule of V. This is described in Section 3, where the eigenvalues of the Gaudin Hamiltonians on a Bethe vector are shown to be the coefficients of the Fuchsian differential equation giving the corresponding spaces of polynomials. Finally, the reality of the space of polynomials follows as the Gaudin Hamiltonians are real symmetric operators when the Wronskian has only real roots. This implies that the eigenvalues are real, and thus the Fuchsian differential equation and the corresponding space of polynomials is also real. In all, this is an extraordinary proof.

2.1. **Critical points of master functions.** The construction of $\delta(\boldsymbol{\iota}_{n,d})$ spaces of polynomials with a given Wronskian begins with the critical points of a symmetric rational function that arose in the study of hypergeometric solutions to the Knizhnik-Zamolodchikov equations [29], as well as the Bethe ansatz method for the Gaudin model. (See §3.)

The master function $\Phi(\mathbf{x};\mathbf{s})$ depends upon a point $\mathbf{s} := (s_1,\dots,s_{(n+1)(d-n)}) \in \mathbb{C}^{(n+1)(d-n)}$, whose coordinates will be the roots of our Wronskian W, and an additional $\binom{n+1}{2}(d-n)$ complex variables

$$\mathbf{x} \;:=\; (x^{(0)}_1,\dots,x^{(0)}_{d-n},\, x^{(1)}_1,\dots,x^{(1)}_{2(d-n)},\,\dots\,,\, x^{(n-1)}_1,\dots,x^{(n-1)}_{n(d-n)})\,.$$

Each set of variables $\mathbf{x}^{(i)} := (x^{(i)}_1,\dots,x^{(i)}_{(i+1)(d-n)})$ will be the roots of certain intermediate Wronskians.

Define the *master function* $\Phi(\mathbf{x};\mathbf{s})$ by the (rather formidable) formula

$$\frac{\displaystyle\prod_{i=0}^{n-1}\ \prod_{1\le j<k\le (i+1)(d-n)} (x^{(i)}_j - x^{(i)}_k)^2}{\displaystyle\prod_{j=1}^{n(d-n)}\ \prod_{k=1}^{(n+1)(d-n)} (x^{(n-1)}_j - s_k)\ \prod_{i=0}^{n-2}\ \prod_{j=1}^{(i+1)(d-n)}\ \prod_{k=1}^{(i+2)(d-n)} (x^{(i)}_j - x^{(i+1)}_k)}\,. \tag{2.1}$$

This is separately symmetric in each set of variables $\mathbf{x}^{(i)}$. This master function has a much simpler formulation which we give below (2.4).

The critical points of the master function are solutions to the system of equations

$$\frac{1}{\Phi}\frac{\partial}{\partial x^{(i)}_j}\Phi(\mathbf{x};\mathbf{s}) \;=\; 0 \qquad\text{for}\quad i=0,\dots,n{-}1,\quad j=1,\dots,(i{+}1)(d{-}n)\,. \tag{2.2}$$

When the parameters $\mathbf{s}$ are generic, these *Bethe ansatz equations* turn out to have finitely many solutions. The master function is invariant under the group

$$\mathcal{S} \;:=\; \mathcal{S}_{d-n}\times\mathcal{S}_{2(d-n)}\times\;\cdots\;\times\mathcal{S}_{n(d-n)}\,,$$

where $\mathcal{S}_m$ for the group of permutations of $1,\dots,n$, and the factor $\mathcal{S}_{(i+1)(d-n)}$ permutes the variables in $\mathbf{x}^{(i)}$. Thus $\mathcal{S}$ acts on the critical points. The invariants of this action are polynomials whose roots are the coordinates of the critical points. Given a critical point $\mathbf{x}$, define monic polynomials $\mathbf{p}_{\mathbf{x}} := (p_0,\dots,p_{n-1})$ where the components $\mathbf{x}^{(i)}$ of $\mathbf{x}$ are the roots of p_i,

$$p_i \;:=\; \prod_{j=1}^{(i+1)(d-n)} (t - x_j^{(i)}) \qquad \text{for} \quad i = 0, \dots, n{-}1\,. \tag{2.3}$$

Also write p_n for the Wronskian, the monic polynomial with roots $\mathbf{s}$. The master function is greatly simplified by this notation. Recall that the discriminant $\mathrm{Discr}(f)$ of a polynomial f is the square of the product of differences of its roots and the resultant $\mathrm{Res}(f,g)$ is the product of all differences of the roots of f and g [4]. Then the formula for the master function (2.1) is

$$\Phi(\mathbf{x};\mathbf{s}) \;=\; \prod_{i=0}^{n} \mathrm{Discr}(p_i) \Big/ \prod_{i=0}^{n-1} \mathrm{Res}(p_i, p_{i+1})\ . \tag{2.4}$$

The connection between the critical points of the master function and spaces of polynomials is through a Fuchsian differential equation of order $n{+}1$ that has only polynomial solutions. Given (an orbit of) a critical point $\mathbf{x}$ represented by the list of polynomials $\mathbf{p}_{\mathbf{x}}$ and write p_n for the Wronskian W, define the *fundamental differential operator $D_{\mathbf{x}}$ of the critical point* $\mathbf{x}$ by

$$\Bigl(\frac{d}{dt} - \ln'\Bigl(\frac{p_n}{p_{n-1}}\Bigr)\Bigr) \cdots \Bigl(\frac{d}{dt} - \ln'\Bigl(\frac{p_1}{p_0}\Bigr)\Bigr)\Bigl(\frac{d}{dt} - \ln'(p_0)\Bigr)\,, \tag{2.5}$$

where $\ln'(f) := \frac{d}{dt}\ln f$. Write $V_{\mathbf{x}}$ for the kernel of $D_{\mathbf{x}}$, which we call the *fundamental space of the critical point* $\mathbf{x}$.

Example 2.6. Since

$$\Bigl(\frac{d}{dt} \;-\; \ln'(p)\Bigr)\, p \;=\; \Bigl(\frac{d}{dt} \;-\; \frac{p'}{p}\Bigr)\, p \;=\; p' - \frac{p'}{p}p \;=\; 0\,,$$

we see that p_0 is a solution of $D_{\mathbf{x}}$. It is instructive to look at $D_{\mathbf{x}}$ and $V_{\mathbf{x}}$ when $n=1$. Suppose that f a solution to $D_{\mathbf{x}}$ that is linearly independent from p_0. Then

$$0 \;=\; \Bigl(\frac{d}{dt} - \ln'\Bigl(\frac{W}{p_0}\Bigr)\Bigr)\Bigl(\frac{d}{dt} - \ln'(p_0)\Bigr)\, f \;=\; \Bigl(\frac{d}{dt} - \ln'\Bigl(\frac{W}{p_0}\Bigr)\Bigr)\Bigl(f' - \frac{p_0'}{p_0}f\Bigr)\,.$$

This implies that

$$\frac{W}{p_0} \;=\; f' - \frac{p_0'}{p_0}f\,,$$

or rather that $W = f'p_0 - p_0'f = \mathrm{Wr}(f,p_0)$, so that the kernel of $D_{\mathbf{x}}$ is a space of functions with Wronskian W.

Mukhin and Varchenko showed that what we just saw is always the case, and much more.

Theorem 2.7 ([26], Section 5)**.** *Suppose that $V_{\mathbf{x}}$ is the fundamental space of the critical point $\mathbf{x}$ of the master function Φ whose parameters $\mathbf{s}$ are roots of a polynomial W.*

(1) *The critical point $\mathbf{x}$ is recovered from $V_{\mathbf{x}}$ as follows. Suppose that $f_0,\dots,f_n$ are monic polynomials in $V_{\mathbf{x}}$ with $\deg f_i = d{-}n+i$. Then, up to scalar multiples, the polynomials $p_0,\dots,p_{n-1}$ in the sequence $\mathbf{p}_{\mathbf{x}}$ are*

$$f_0\,,\ \mathrm{Wr}(f_0,f_1)\,,\ \mathrm{Wr}(f_0,f_1,f_2)\,,\ \dots\,,\ \mathrm{Wr}(f_0,\dots,f_{n-1})\,.$$

(2) *$V_{\mathbf{x}}$ is a space of polynomials of degree d and dimension $n{+}1$ lying in G° with Wronskian W.*

Statement (1) is quite general; it generalizes Example 2.6 and gives a recipe for writing the differential operator with kernel generated by sufficiently differentiable functions $f_0(t), f_2(t), \dots, f_n(t)$. It follows from some interesting identities among Wronskians shown in the Appendix of [26]. Statement 2 is the deeper of the two. Together these imply that the kernel V of an operator of the form (2.5) is a space of polynomials with Wronskian W if and only if the the polynomials $p_0, \dots, p_{n-1}$ come from the critical points of the master function (2.1) corresponding to W.

Thus there is an injection from $\mathcal{S}$-orbits of critical points of the master function Φ with parameters $\mathbf{s}$ to spaces of polynomials in $\mathrm{Grass}^\circ_{n,d}$ whose Wronskian has roots $\mathbf{s}$. Mukhin and Varchenko also showed that when $\mathbf{s}$ is generic, this is in fact a bijection.

Theorem 2.8 (Theorem 6.1 in [27])**.** *For generic complex numbers $\mathbf{s}$, the master function Φ has $\delta(\boldsymbol{\iota}_{n,d})$ distinct orbits of critical points and all critical points are nondegenerate.*

The structure (but not of course the details) of their proof is remarkably similar to the structure of the proof of Theorem 1.9 (given in [32]); they allow the parameters to collide one-by-one, and show how the orbits of critical points behave. Ultimately, they obtain the same recursion as in [32], which mimics the Pieri formula for the branching rule for tensor products of representations of $\mathfrak{sl}_{n+1}$ with its fundamental representation V_{ω_n}. This same structure is also found in the main argument in [7]. In fact, this is the same recursion in $\mathbf{a}$ that Schubert established for intersection numbers $\delta(\mathbf{a}, \boldsymbol{\iota}, \dots, \boldsymbol{\iota})$, and then solved to obtain the formula (1.6).

3. The Bethe ansatz for the Gaudin model

The Bethe ansatz is intended to give an explicit decomposition of a representation V of $\mathfrak{sl}_{n+1}\mathbb{C}$ into irreducible submodules that is also compatible with the action of a family of commuting operators on V, called the Gaudin Hamiltonians. These commuting operators constitute an integrable system. Its development, justification, and refinements are the subject of a large body of work, a small part of which we mention. One unintended consequence (besides the proof of the Shapiro conjecture) is a deeper link between Schubert calculus on the Grassmannian $\mathrm{Grass}_{n,d}$ and representation theory of $\mathfrak{sl}_{n+1}\mathbb{C}$ than had been known previously.

3.1. **Representations of $\mathfrak{sl}_{n+1}\mathbb{C}$.** The Lie algebra $\mathfrak{sl}_{n+1}\mathbb{C}$ (or simply $\mathfrak{sl}_{n+1}$) is the space of $(n{+}1) \times (n{+}1)$-matrices with zero trace. It has a decomposition

$$\mathfrak{sl}_{n+1} \;=\; \mathfrak{n}_- \oplus \mathfrak{h} \oplus \mathfrak{n}_+\,,$$

where $\mathfrak{n}_+$ ($\mathfrak{n}_-$) are the strictly upper (lower) triangular matrices, and $\mathfrak{h}$ consists of the diagonal matrices with zero trace.

As $\mathfrak{h}$ is commutative, any representation V of $\mathfrak{sl}_{n+1}$ decomposes into joint eigenspaces of $\mathfrak{h}$, called *weight spaces*,

$$V \;=\; \bigoplus_{\mu\in\mathfrak{h}^*} V[\mu]\,,$$

where, for $v \in V[\mu]$ and $h \in \mathfrak{h}$, we have $h.v = \mu(h)v$. The possible weights μ of representations lie in the integral *weight lattice*. Positive weights are those that are integral linear combinations of weights of the representation $\mathfrak{n}_+$. The weight lattice has a distinguished basis of *fundamental weights* $\omega_1, \ldots, \omega_n$ that generate the cone of *dominant weights* (a subcone of the positive weights).

The irreducible representations of $\mathfrak{sl}_{n+1}$ enjoy the following classification. An irreducible representation V has a unique highest weight μ. That is, if π is another weight of V, then $\mu - \pi$ is positive. Furthermore, μ is dominant. This highest weight space $V[\mu]$ is 1-dimensional and it generates V, and any two irreducible modules with the same highest weight are isomorphic. Write V_μ for the *highest weight module* with highest weight μ. Lastly, there is one highest weight module for each dominant weight.

The highest weight space $V_\mu[\mu]$ of V_μ is also distinguished as the set of vectors in V_μ that are annihilated by the nilpotent subalgebra $\mathfrak{n}_+$ of $\mathfrak{sl}_{n+1}$. More generally, if V is any representation of $\mathfrak{sl}_{n+1}$ and μ is a weight, then the *singular vectors* in V of weight μ, written $\mathrm{sing}(V[\mu])$, are the vectors in $V[\mu]$ annihilated by $\mathfrak{n}_+$. If $v \in \mathrm{sing}(V[\mu])$ is nonzero, then the submodule $\mathfrak{sl}_{n+1}.v$ it generates is isomorphic to the highest weight module V_μ. Thus V decomposes as a direct sum of submodules generated by the singular vectors,

$$V \;=\; \bigoplus_{\mu} \mathfrak{sl}_{n+1}.\mathrm{sing}(V[\mu])\,, \tag{3.1}$$

so that the multiplicity of the highest weight module V_μ in V is simply the dimension of this space of singular vectors of weight μ.

When V is a tensor product of highest weight modules, the Littlewood-Richardson rule [12] gives formulas for the dimensions of the spaces of singular vectors. Since this is formally the same rule as used to determine the number of points in an intersection (1.5) of Schubert varieties coming from a Schubert problem, these geometric intersection numbers are equal to the dimensions of spaces of singular vectors. In particular, if $V_{\omega_1} \simeq \mathbb{C}^{n+1}$ is the defining representation of $\mathfrak{sl}_{n+1}$ and $V_{\omega_n} = \wedge^n V_{\omega_1} = V^*_{\omega_1}$ (these are the first and last fundamental representations of $\mathfrak{sl}_{n+1}$), then

$$\dim \mathrm{sing}(V_{\omega_n}^{\otimes(n+1)(d-n)}[0]) \;=\; \delta(\boldsymbol{\iota}_{n,d})\,. \tag{3.2}$$

3.2. **The Gaudin model.** The Bethe ansatz is a conjectural method to obtain this decomposition (3.1) by giving an explicit basis for $\mathrm{sing}(V[\mu])$, which is also an eigenbasis for a family of commuting operators on V. For us, V is the tensor product $V_{\omega_n}^{\otimes m}$, and the family of commuting operators are the Gaudin Hamiltonians. These depend upon m distinct complex numbers $s_1, \ldots, s_m$ and a complex variable t.

For each $i, j = 1, \ldots, n{+}1$, let $E_{i,j} \in \mathfrak{sl}_{n+1}$ be the matrix that has all entries 0, except a 1 in row i and column j. For each such pair (i,j) consider the differential operator $X_{i,j}(t)$ acting on $V_{\omega_n}^{\otimes m}$-valued functions of t,

$$X_{i,j}(t) \;:=\; \delta_{i,j}\frac{d}{dt} \;-\; \sum_{k=1}^{m} \frac{E^{(k)}_{j,i}}{t - s_k}\;,$$

where $E^{(k)}_{j,i}$ acts on tensors in $V^{\otimes m}_{\omega_n}$ by $E_{j,i}$ in the kth factor and by the identity in other factors. Define a differential operator acting on $V^{\otimes m}_{\omega_n}$-valued functions of t,

$$\mathbf{M} \;:=\; \sum_{\sigma\in\mathcal{S}} |\sigma|\; X_{1,\sigma(1)}(t)\; X_{2,\sigma(2)}(t)\;\cdots\; X_{n+1,\sigma(n+1)}(t)\;,$$

where $\mathcal{S}$ is the group of permutations of $\{1,\dots,n{+}1\}$ and $|\sigma| = \pm$ is the sign of a permutation $\sigma\in\mathcal{S}$. Write $\mathbf{M}$ in standard form

$$\mathbf{M} \;=\; \frac{d^{n+1}}{dt^{n+1}} \;+\; M_1(t)\frac{d^n}{dt^n} \;+\;\cdots\;+\; M_{n+1}(t)\,.$$

These coefficients $M_1(t),\dots,M_{n+1}(t)$ are called the *Gaudin Hamiltonians*. They are linear operators that depend rationally on t and act on $V^{\otimes m}_{\omega_n}$. We collect together some of their properties.

Theorem 3.3. *Suppose that $s_1,\dots,s_m$ are distinct complex numbers. Then*

(1) *The Gaudin Hamiltonians commute, that is, $[M_i(u), M_j(v)] = 0$ for all $i,j = 1,\dots,n{+}1$ and $u,v\in\mathbb{C}$.*

(2) *The Gaudin Hamiltonians centralize the action of $\mathfrak{sl}_{n+1}$ on $V^{\otimes m}_{\omega_n}$.*

Proofs of these statements may be found in [21], as well as Propositions 7.2 and 8.3 in [23]. A consequence of the second assertion is that the Gaudin Hamiltonians preserve the weight space decomposition of the singular vectors of $V^{\otimes m}_{\omega_n}$. Since they commute with each other, the singular vectors of $V^{\otimes m}_{\omega_n}$ have a basis of common eigenvalues. The Bethe ansatz is a method to write down these joint eigenvectors and their eigenvalues.

3.3. The Bethe ansatz for the Gaudin model.

The Bethe ansatz for the Gaudin model begins with a rational function, called a universal weight function, that takes values in a weight space $V^{\otimes m}_{\omega_n}[\mu]$,

$$v\;:\;\mathbb{C}^l\times\mathbb{C}^m\;\longmapsto\;V^{\otimes m}_{\omega_n}[\mu]\;.$$

This *universal weight function* was introduced in [29] to solve the Knizhnik-Zamolodchikov equations with values in $V^{\otimes m}_{\omega_n}[\mu]$. When the arguments $(\mathbf{x},\mathbf{s})$ are the critical points of a master function, the vector $v(\mathbf{x},\mathbf{s})$ is both singular and an eigenvector of the Gaudin Hamiltonians. (This master function is a generalization of the one defined by (2.1).) The Bethe ansatz conjecture asserts that the vectors $v(\mathbf{x},\mathbf{s})$ form a basis for the space of singular vectors.

For us, $m=(n{+}1)(d{-}n)$, $l=\binom{n+1}{2}(d{-}n)$, and $\mu=0$. Then the universal weight function is a map

$$v\;:\;\mathbb{C}^{\binom{n+1}{2}(d-n)}\;\longmapsto\;V^{\otimes(n+1)(d-n)}_{\omega_n}[0]\,.$$

For these notes, we omit the definition of $v(\mathbf{x},\mathbf{s})$.

While $v(\mathbf{x},\mathbf{s})$ is a rational function of $\mathbf{x}$ and hence not globally defined, it turns out (Lemma 2.1 of [27]) that if the coordinates of $\mathbf{s}$ are distinct and $\mathbf{x}$ is a critical point of the master function (2.1), then the vector $v(\mathbf{x},\mathbf{s})\in V^{\otimes(n+1)(d-n)}_{\omega_n}[0]$ is well-defined and it is in fact a singular vector. Such a vector $v(\mathbf{x},\mathbf{s})$ when $\mathbf{x}$ is a critical point of the master function a *Bethe vector*. Mukhin and Varchenko also prove the following, which is the second part of Theorem 6.1 in [27].

Theorem 3.4. *When $\mathbf{s} \in \mathbb{C}^{(n+1)(d-n)}$ is general, the Bethe vectors form a basis of the space* $\mathrm{sing}(V_{\omega_n}^{\otimes(n+1)(d-n)}[0])$.

The reason to introduce these Bethe vectors is that they are the joint eigenvectors of the Gaudin Hamiltonians.

Theorem 3.5 (Theorem 9.2 in [23]). *For any critical point* $\mathbf{x}$ *of the master function* (2.1)*, the Bethe vector* $v(\mathbf{x}, \mathbf{s})$ *is a joint eigenvector of the Gaudin Hamiltonians* $M_1(t), \dots, M_{n+1}(t)$. *The corresponding eigenvalues* $\mu_1(t), \dots, \mu_{n+1}(t)$ *are given by the formula*

$$\begin{aligned}(3.6)\qquad & \frac{d^{n+1}}{dt^{n+1}} \;+\; \mu_1(t)\frac{d^n}{dt^n} \;+\; \cdots \;+\; \mu_n(t)\frac{d}{dt} \;+\; \mu_{n+1}(t) \;=\; \\ & \Big(\frac{d}{dt} + \ln'(p_0)\Big)\Big(\frac{d}{dt} + \ln'\Big(\frac{p_1}{p_0}\Big)\Big) \;\cdots\; \Big(\frac{d}{dt} + \ln'\Big(\frac{p_n}{p_{n-1}}\Big)\Big)\Big(\frac{d}{dt} + \ln'\Big(\frac{W}{p_n}\Big)\Big)\,,\end{aligned}$$

where $(p_0(t), \dots, p_n(t))$ *are the polynomials* (2.3) *associated to the critical point* $\mathbf{x}$ *and* $W(t)$ *is the polynomial with roots* $\mathbf{s}$.

Observe that (3.6) is similar to the formula (2.5) for the differential operator $D_{\mathbf{x}}$ of the critical point $\mathbf{x}$. This similarity is made more precise if we replace the Gaudin Hamiltonians by a different set of operators. Consider the differential operator formally conjugate to $(-1)^{n+1}M$,

$$\begin{aligned} K \;&=\; \frac{d^{n+1}}{dt^{n+1}} \;-\; \frac{d^n}{dt^n}M_1(t) \;+\; \cdots \;+\; (-1)^n\frac{d}{dt}M_n(t) \;+\; (-1)^{n+1}M_{n+1}(t) \\ &=\; \frac{d^{n+1}}{dt^{n+1}} \;+\; K_1(t)\frac{d^n}{dt^n} \;+\; \cdots \;+\; K_n(t)\frac{d}{dt} \;+\; K_{n+1}(t)\;.\end{aligned}$$

These coefficients $K_i(t)$ are operators on $V_{\omega_n}^{\otimes(n+1)(d-n)}$ that depend rationally on t, and are also called the Gaudin Hamiltonians. Here are the first three,

$$\begin{aligned} K_1(t) \;&=\; -M_1(t)\,, \qquad\qquad K_2(t) \;=\; M_2(t) \;-\; nM_1'(t)\,, \\ K_3(t) \;&=\; -M_3(t) \;+\; (n-1)M_2''(t) \;-\; \binom{n}{2}M_1'''(t)\,,\end{aligned}$$

and in general $K_i(t)$ is a differential polynomial in $M_1(t), \dots, M_{n+1}(t)$.

These operators also commute, $[K_i(u), K_j(v)] = 0$ for all i, j, u, v, and they also commute with the $\mathfrak{sl}_{n+1}$-action on $V_{\omega_n}^{\otimes(n+1)(d-n)}$, and the Bethe vector $v(\mathbf{x}, \mathbf{s})$ is also a joint eigenvector of these new Gaudin Hamiltonians $K_i(t)$. The corresponding eigenvalues $\lambda_1(t), \dots, \lambda_{n+1}(t)$ are given by the formula

$$\begin{aligned}(3.7)\qquad & \frac{d^{n+1}}{dt^{n+1}} \;+\; \lambda_1(t)\frac{d^n}{dt^n} \;+\; \cdots \;+\; \lambda_n(t)\frac{d}{dt} \;+\; \lambda_{n+1}(t) \;=\; \\ & \Big(\frac{d}{dt} - \ln'\Big(\frac{W}{p_{n-1}}\Big)\Big)\Big(\frac{d}{dt} - \ln'\Big(\frac{p_{n-1}}{p_n}\Big)\Big) \cdots \Big(\frac{d}{dt} - \ln'\Big(\frac{p_1}{p_0}\Big)\Big)\Big(\frac{d}{dt} - \ln'(p_0)\Big)\,,\end{aligned}$$

which is just the fundamental differential operator $D_{\mathbf{x}}$ of the critical point $\mathbf{x}$.

Corollary 3.8. *Suppose that* $\mathbf{s} \in \mathbb{C}^{(n+1)(d-n)}$ *is generic.*

(1) *The set of Bethe vectors form an eigenbasis of* $\mathrm{sing}(V_{\omega_n}^{\otimes(n+1)(d-n)}[0])$ *for the Gaudin Hamiltonians* $K_1(t), \dots, K_{n+1}(t)$.

(2) *The Gaudin Hamiltonians $K_1(t),\dots,K_{n+1}(t)$ have simple spectrum in that eigenvalues of the Gaudin Hamiltonians separate the basis of eigenvectors.*

Statement (1) follows from Theorems 3.4 and 3.5. For Statement (2), suppose that two Bethe vectors $v(\mathbf{x},\mathbf{s})$ and $v(\mathbf{x}',\mathbf{s})$ have the same eigenvalues. By (3.7), the corresponding fundamental differential operators would be equal, $D_{\mathbf{x}} = D_{\mathbf{x}'}$. But this implies that the fundamental spaces coincide, $V_{\mathbf{x}} = V_{\mathbf{x}'}$. By Theorem 2.7 the fundamental space determines the orbit of critical points, so the critical points $\mathbf{x}$ and $\mathbf{x}'$ lie in the same orbit, which implies that $v(\mathbf{x},\mathbf{s}) = v(\mathbf{x}',\mathbf{s})$.

All that remains is to show that the space $V_{\mathbf{x}}$ is real.

4. Shapovalov form and the proof of the Shapiro conjecture

The last step in the proof of Theorem 1 is to show that if $\mathbf{s}\in\mathbb{R}^{(n+1)(d-n)}$ is generic and $\mathbf{x}$ a critical point of the master function (2.1), then the fundamental space $V_{\mathbf{x}}$ of the critical point $\mathbf{x}$ has a basis of real polynomials. As promised in the introduction, the reason for this reality is that the eigenvectors and eigenvalues of a symmetric matrix are real.

We begin with the Shapovalov form. The map $\tau\colon E_{ij}\mapsto E_{ji}$ induces an antiautomorphism on $\mathfrak{sl}_{n+1}$. Given a highest weight module V_μ and a nonzero vector v in $V_\mu[\mu]$, the *Shapovalov form* $\langle\cdot,\cdot\rangle$ on V_μ is defined recursively by

$$\langle v,v\rangle \;=\; 1 \qquad\text{and}\qquad \langle g.u,v\rangle \;=\; \langle u,\tau(g).v\rangle\,,$$

for $g\in\mathfrak{sl}_{n+1}$ and $u,v\in V$.

For example, if $V_{\omega_1} = \mathbb{C}^{n+1}$ is the defining representation of $\mathfrak{sl}_{n+1}$ with basis $e_0,\dots,e_n$, and we set $v := e_n$, then $\langle e_i,e_j\rangle = \delta_{ij}$. Thus the Shapovalov form is the standard Euclidean inner product on V_{ω_1}. As V_{ω_n} is the linear dual of V_{ω_1}, the Shapovalov form on V_{ω_n} is also the standard Euclidean inner product. In general, this Shapovalov form is nondegenerate on V_μ and positive definite on the real part of V_μ.

The Shapovalov form on V_{ω_n} induces a symmetric bilinear form, also called the Shapovalov form, on the tensor product $V_{\omega_n}^{\otimes(n+1)(d-n)}$. This tensor Shapovalov form is also positive definite on the real part of $V_{\omega_n}^{\otimes(n+1)(d-n)}$.

Theorem 4.1 (Proposition 9.1 in [23])**.** *The Gaudin Hamiltonians are symmetric with respect to the tensor Shapovalov form,*

$$\langle K_i(t).u,\, v\rangle \;=\; \langle u,\, K_i(t).v\rangle\,,$$

for all $i=1,\dots,n{+}1$, $t\in\mathbb{C}$, and $u,v\in V_{\omega_n}^{\otimes(n+1)(d-n)}$.

We give the most important consequence of this result for our story.

Corollary 4.2. *When the parameters $\mathbf{s}$ and variable t are real, the Gaudin Hamiltonians $K_1(t),\dots,K_{n+1}(t)$ are real linear operators which are simultaneously diagonalizable with real spectrum.*

Proof. From the definition of the Gaudin Hamiltonians $M_1(t),\dots,M_{n+1}(t)$, we see that they are real linear operators which act on the real part of $V_{\omega_n}^{\otimes(n+1)(d-n)}$. The same is then also true of the Gaudin Hamiltonians $K_1(t),\dots,K_{n+1}(t)$. But these are symmetric with respect to the positive definite Shapovalov form. Consequently, the are simultaneously diagonalizable with real spectrum.

Proof of Theorem 1. Suppose that $\mathbf{s} \in \mathbb{R}^{(n+1)(d-n)}$ is general. By Corollary 4.2, the Gaudin Hamiltonians for $t \in \mathbb{R}$ acting on $\operatorname{sing}(V_{\omega_n}^{(n+1)(d-n)}[0])$ are symmetric operators on a Euclidean space, and so have real eigenvectors and eigenvalues. The Bethe vectors $v(\mathbf{x}, \mathbf{s})$ for critical points $\mathbf{x}$ of the master function with parameters $\mathbf{s}$ form an eigenbasis for the Gaudin Hamiltonians. As $\mathbf{s}$ is general, the eigenvalues are distinct by Corollary 3.8 (2), and so the Bethe vectors must be real.

Given a critical point $\mathbf{x}$, the eigenvalues $\lambda_1(t), \ldots, \lambda_{n+1}(t)$ of the Bethe vectors are then real rational functions, and so the fundamental differential operator $D_{\mathbf{x}}$ has real coefficients. But then the fundamental space $V_{\mathbf{x}}$ of polynomials is real.

In this way, we see that each of the $\delta(\boldsymbol{\iota}_{n,d})$ spaces of polynomials $V_{\mathbf{x}}$ whose Wronskian has roots $\mathbf{s}$ that were constructed in Section 2 is in fact real. This proves Theorem 1.

5. Applications of the Shapiro conjecture

Theorem 1 and its stronger version, Theorem 1.10, have a number of other applications in mathematics. Some are straightforward, such as linear series on $\mathbb{P}^1$ with real ramification. Others are much less so, such as Schützenberger evacuation in algebraic combinatorics. Here, we discuss two applications which are in the first class, namely maximally inflected curves and rational functions with real critical points.

5.1. **Maximally inflected curves.**

One of the earliest occurrences of the central mathematical object of these notes, spaces of polynomials with prescribed ramification, was in algebraic geometry, as these are linear series $P \subset H^0(\mathbb{P}^1, \mathcal{O}(d))$ on $\mathbb{P}^1$ with prescribed ramification. Their connection to Schubert calculus originated in work of Castelnuovo in 1889 [3] on g-nodal rational curves, and this was important in Brill-Noether theory (see Ch. 5 of [16]) and the Eisenbud-Harris theory of limit linear series [5, 6].

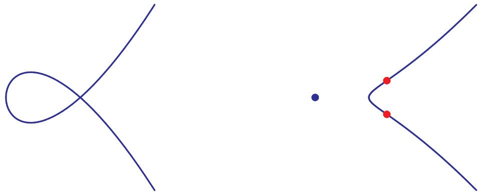
A linear series P on $\mathbb{P}^1$ of degree d and dimension $n+1$ (subspace in $\mathrm{Grass}_{n,d}$) gives rise to a degree d map

$$\varphi : \mathbb{P}^1 \longrightarrow \mathbb{P}^n = \mathbb{P}(P^*) \tag{5.1}$$

of $\mathbb{P}^1$ to projective space. We will call this map a curve. The linear series if ramified at points $s \in \mathbb{P}^1$ where the curve φ is not convex (the jets $\varphi(s), \varphi'(s), \ldots, \varphi^{(n)}(s)$ do not span $\mathbb{P}^n$). Call such a point s a *flex* of the curve (5.1).

A curve is real when P is real. It is *maximally inflected* if all of its flexes are real. The study of these curves was initiated in [17], where restrictions on the topology of plane maximally inflected curves were established. Specifically, there is a lower bound on the number of isolated singularities (and hence an upper bound on the number of nodes) of a maximally inflected plane curve.

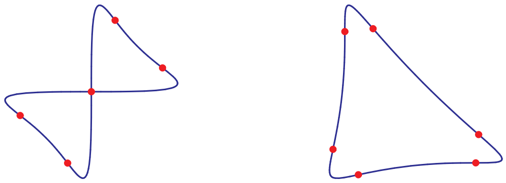
For example, there are two types of cubic curves, which are distinguished by their singular points. The singular point of the curve on the left is a node and connected to the rest of the curve, while the singular point on the other curve is isolated from the rest of the curve.



$y^2 = x^3 + x^2$ $\qquad$ $y^2 = x^3 - x^2$

While both curves have one of their three flexes at infinity, only the curve on the right has its other two flexes real (the dots) and is therefore maximally inflected. A nodal cubic cannot be maximally inflected.

Similarly, a maximally inflected quartic has either 1 or 0 of its (necessarily 3) singular points a node, and necessarily 2 or 3 solitary points. We draw the two types of maximally inflected quartics having six flexes, without their solitary points.



While many constructions of maximally inflected curves were known, Theorem 1, and in particular Theorem 1.10, show that there are many maximally inflected curves: Any curve $\varphi\colon \mathbb{P}^1 \to \mathbb{P}^n$ whose ramification lies in $\mathbb{RP}^1$ must be real and is therefore maximally inflected.

5.2. **Rational functions with real critical points.** A special case of Theorem 1, proved earlier by Eremenko and Gabrielov, serves to illustrate the breadth of mathematical areas touched by this Shapiro conjecture. When $n = 1$, we may associate a rational function $\varphi_P := f_1(t)/f_2(t)$ to a basis $\{f_1(t), f_2(t)\}$ of a vector space $P \in \mathrm{Grass}_{n,d}$ of polynomials. Different bases give different rational functions, but they all differ from φ_P by a fractional linear transformation of the image $\mathbb{P}^1$. We say that such rational functions are *equivalent*.

The critical points of any such rational function are the points of the domain $\mathbb{P}^1$ where the derivative of φ_P,

$$d\varphi_P \ := \ \frac{f_1' f_2 - f_1 f_2'}{f_2^2} \ = \ \frac{1}{f_2^2} \cdot \det \begin{pmatrix} f_1 & f_2 \\ f_1' & f_2' \end{pmatrix} \ ,$$

vanishes. That is, at the roots of the Wronskian. Eremenko and Gabrielov [8] prove the following result about the critical points of rational functions.

Theorem 5.2. *A rational function φ whose critical points lie on a circle in $\mathbb{P}^1$ maps that circle to a circle.*

To see that this is equivalent to Theorem 1 when $n = 1$, note that we may apply a change of variables to φ so that its critical points lie on the circle $\mathbb{RP}^1 \subset \mathbb{P}^1$. Similarly, the image circle may be assumed to be $\mathbb{RP}^1$. Reversing these coordinate changes establishes the equivalence.

The proof used methods specific to rational functions. Goldberg showed [13] that there are at most $c_d := \frac{1}{d}\binom{2d-2}{d-1}$ rational functions of degree d with a given collection of $2d-2$ simple critical points. If the critical points of a rational function φ of degree d lie on a circle $C \subset \mathbb{CP}^1$ and if φ maps C to C, then $\varphi^{-1}(C)$ forms a graph on the Riemann sphere with nodes the $2d{-}2$ critical points, each of degree 4, and each having two edges along C and one edge on each side of C. It turns out that there are also c_d such abstract graphs. (In fact, c_d is Catalan number, which counts many objects in combinatorics.) Eremenko and Gabrielov essentially constructed such a rational function φ for each such graph and choice of critical points on C. Since c_d is the upper bound for the number of such rational functions, this construction gives all rational functions with given set of critical points and thus proves Theorem 5.2. More recently, Eremenko and Gabrielov have found an elementary proof of this result, which uses an induction similar to that described after Theorem 2.8, but that has unfortunately never been published [9].

6. Extensions of the Shapiro conjecture

The proofs of different Bethe ansätze for other models (other integrable systems) and other Lie algebras, which is ongoing work of Mukhin, Tarasov, and Varchenko, and others, leads to generalizations of Theorem 1. One such is given in an appendix of [22], where it is conjectured that orbits of critical points of generalized master functions are real. This is the analog of the consequence of Theorem 1 and Theorem 2.7 (1) that the polynomials p_i are real, which is that new conjecture for the Lie algebra $\mathfrak{sl}_{n+1}$. In that appendix, it is noted that this generalization of the Shapiro conjecture is true for $\mathfrak{sp}_{2n}$ and $\mathfrak{so}_{2n+1}$, by the results in Section 7 of [26].

In [24], the Bethe ansatz for the XXX model is used to prove an analog of Theorem 1 for spaces of quasipolynomials (functions of the form $e^{\lambda_i x} f_i(x)$ with $\lambda_i \in \mathbb{R}$) whose discrete Wronskian has only simple real roots separated by at least the step size used in the discrete Wronskian. There surely is more to come.

Likewise the Shapiro conjecture, that an intersection of Schubert varieties in the Grassmannian given by the special flags $F_\bullet(s)$ consists only of real points, makes sense for other flag manifolds. In this more general setting, it is known to fail, but in a very interesting way. When it fails, we can modify it to give a conjecture that holds under scrutiny, and the Shapiro conjecture also admits some appealing generalizations. We briefly describe some of this story.

6.1. Lagrangian and Orthogonal Grassmannians.

Lagrangian and orthogonal Grassmannians are two varieties closely related to the classical Grassmannian. For each of these, the Shapiro conjecture is particularly easy to state.

The (odd) orthogonal Grassmannian, begins with a non-degenerate symmetric bilinear form $\langle\cdot,\cdot\rangle$ on $\mathbb{C}^{2n+1}$. This vector space has a basis $e_1,\ldots,e_{2n+1}$ such that

$$\langle e_i,\, e_{2n+2-j}\rangle \;=\; \delta_{i,j}\,.$$

The *(odd) orthogonal Grassmannian $OG(n)$* is the set of all n-dimensional subspaces V of $\mathbb{C}^{2n+1}$ that are *isotropic* in that $\langle V,V\rangle = 0$. These subspaces have maximal dimension among all isotropic vector spaces. This variety has dimension $\binom{n+1}{2}$.

The Shapiro conjecture for $OG(n)$ begins with a particular rational normal curve γ having parametrization

$$t \longmapsto e_1 + te_2 + \frac{t^2}{2}e_3 + \cdots + \frac{t^n}{n!}e_{n+1} - \frac{t^{n+1}}{(n+1)!}e_{n+2} \\ + \frac{t^{n+2}}{(n+2)!}e_{n+3} - \cdots + (-1)^n\frac{t^{2n}}{(2n)!}e_{2n+1}\,.$$

This has special properties with respect to the form $\langle\cdot,\cdot\rangle$. For $t\in\mathbb{C}$, define the flag $F_\bullet(t)$ in $\mathbb{C}^{2n+1}$ by

$$F_i(t) := \operatorname{Span}\{\gamma(t),\,\gamma'(t)\,,\ldots,\,\gamma^{(i-1)}(t)\}\,.$$

Then $F_\bullet(t)$ is *isotropic* in that

$$\langle F_i(t),\, F_{2n+1-i}(t)\rangle = 0\,.$$

More generally, an isotropic flag $F_\bullet$ of $\mathbb{C}^{2n+1}$ is a flag such that $\langle F_i, F_{2n+1-i}\rangle = 0$. The *Schubert variety $X_{\boldsymbol\lambda}F_\bullet$* of $OG(n)$ is defined by a Schubert index $\boldsymbol\lambda$ and an isotropic flag $F_\bullet$. Write $|\boldsymbol\lambda|$ for its codimension. A Schubert problem is a list $(\boldsymbol\lambda_1,\ldots,\boldsymbol\lambda_m)$ of Schubert indices such that

$$|\boldsymbol\lambda_1| + |\boldsymbol\lambda_2| + \cdots + |\boldsymbol\lambda_m| = \dim OG(n) = \binom{n+1}{2}.$$

We state the Shapiro conjecture for $OG(n)$.

Conjecture 6.1. *If $(\boldsymbol\lambda_1,\ldots,\boldsymbol\lambda_m)$ is a Schubert problem for $OG(n)$ and $s_1,\ldots,s_m$ are distinct real numbers, then the intersection*

$$X_{\boldsymbol\lambda_1}F_\bullet(s_1) \bigcap X_{\boldsymbol\lambda_2}F_\bullet(s_2) \bigcap \cdots \bigcap X_{\boldsymbol\lambda_m}F_\bullet(s_m)$$

is transverse with all points real.

Besides optimism based upon the validity of the Shapiro conjecture for Grassmannians, the evidence for Conjecture 6.1 comes in two forms. A local version, analogous to Theorem 1.9, is true [34], and several tens of thousands of instances have been checked with a computer.

There is a similar story but with a different outcome for the Lagrangian Grassmannian. Let $\langle\cdot,\cdot\rangle$ be a nondegenerate skew symmetric bilinear form on $\mathbb{C}^{2n}$. This vector space has a basis $e_1,\ldots,e_{2n}$ such that

$$\langle e_i,\, e_{2n+1-j}\rangle = \left\{\begin{array}{cl} \delta_{i,j} & \text{if } i\leq 2n \\ -\delta_{i,j} & \text{if } i>2n \end{array}\right. .$$

The *Lagrangian Grassmannian $LG(n)$* is the set of all *isotropic* n-dimensional subspaces V of $\mathbb{C}^{2n}$. These subspaces have maximal dimension among all isotropic vector spaces, and are typically called Lagrangian subspaces. This variety has dimension $\binom{n+1}{2}$.

For the Shapiro conjecture for $LG(n)$, we have the rational normal curve γ with parametrization

$$\begin{aligned} t \longmapsto e_1 + te_2 + \frac{t^2}{2}e_3 + \cdots + \frac{t^n}{n!}e_{n+1} - \frac{t^{n+1}}{(n+1)!}e_{n+2} \\ + \frac{t^{n+2}}{(n+2)!}e_{n+3} - \cdots + (-1)^{n-1}\frac{t^{2n-1}}{(2n-1)!}e_{2n}\,. \end{aligned}$$

For $t \in \mathbb{C}$, define the flag $F_\bullet(t)$ in $\mathbb{C}^{2n+1}$ by

$$F_i(t) \ := \ \mathrm{Span}\{\gamma(t),\, \gamma'(t)\,, \ldots,\, \gamma^{(i-1)}(t)\}\,.$$

Then $F_\bullet(t)$ is *isotropic* in that

$$\langle F_i(t),\, F_{2n-i}(t)\rangle \ = \ 0\,.$$

More generally, an isotropic flag $F_\bullet$ of $\mathbb{C}^{2n}$ is a flag such that $\langle F_i, F_{2n-i}\rangle = 0$. The Schubert variety $X_{\boldsymbol{\lambda}}F_\bullet$ of $LG(n)$ is defined by a Schubert index $\boldsymbol{\lambda}$ and an isotropic flag $F_\bullet$. It has codimension $|\boldsymbol{\lambda}|$. A Schubert problem is a list $(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m)$ such that

$$|\boldsymbol{\lambda}_1| + |\boldsymbol{\lambda}_2| + \cdots + |\boldsymbol{\lambda}_m| \ = \ \dim LG(n) \ = \ \binom{n+1}{2}\,.$$

Belkale and Kumar [2] define a notion of Levi movability for Schubert conditions, which has the following geometric interpretation. Each Schubert variety $X_{\boldsymbol{\lambda}}F_\bullet$ of $LG(n)$ is the intersection of $LG(n)$ with a Schubert variety $\Omega_{\mathbf{a}(\boldsymbol{\lambda})}F_\bullet$ of the Grassmannian of n planes in $\mathbb{C}^{2n}$. The index $\boldsymbol{\lambda}$ is *Levi movable* when these two Schubert varieties have the same codimension in their respective Grassmannians. A *Levi movable* Schubert problem is one made up of Levi movable Schubert indices.

The obvious generalization of Theorem 1 and Conjecture 6.1 to $LG(n)$ turns out to be false. We offer a modification that we believe is true.

Conjecture 6.2. *If $(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m)$ is a Schubert problem for $LG(n)$ and $s_1, \ldots, s_m$ are distinct real numbers, then the intersection*

$$X_{\boldsymbol{\lambda}_1}F_\bullet(s_1) \bigcap X_{\boldsymbol{\lambda}_2}F_\bullet(s_2) \bigcap \cdots \bigcap X_{\boldsymbol{\lambda}_m}F_\bullet(s_m)$$

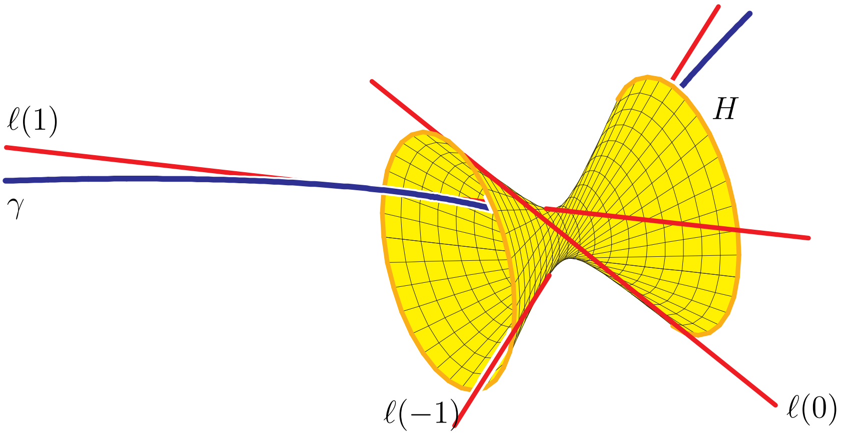
is transverse. If $(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m)$ is Levi movable, then all points of intersection are real, but if it is not Levi movable, then no point in the intersection is real.

The strongest evidence in favor of Conjecture 6.2 is that it is true when the Schubert problem $(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m)$ is Levi movable. This follows from the definition of Levi movable and the Shapiro conjecture for Grassmannians. Further evidence is that if each $\boldsymbol{\lambda}_i$ is simple in that $|\boldsymbol{\lambda}| = 1$, then a local version, similar to Theorem 1.9 but without transversality, is true. That is, if the s_i are sufficiently clustered, then no point in the intersection is real [34]. Lastly, several tens of thousands of instances have been checked with a computer.

6.2. **Monotone conjecture for flag manifolds.** The original Shapiro conjecture was for Schubert varieties in the classical (type-A) flag manifold. This conjecture fails for the first non-trivial Schubert problem on a flag variety that is not a Grassmannian. Consider the geometric problem of partial flags $\ell \subset \Lambda$ in 3-dimensional space where ℓ is required to meet three fixed lines and Λ is required to contain two fixed points.

This is just the problem of four lines in disguise. Suppose that p and q are the two fixed points that Λ is required to contain. Then Λ contains the secant line $\overline{p,q}$ spanned by these two points. Since $\ell \subset \Lambda$, we see that ℓ must meet $\overline{p,q}$. As ℓ must also meet three lines, this problem reduces to the problem of four lines. In this way, there are two solutions to this Schubert problem.

Now let us investigate the original Shapiro conjecture for this Schubert problem, which posits that both flags $\ell \subset \Lambda$ will be real, if we require that ℓ meets three fixed tangent lines to a rational curve and Λ contains two fixed points of the rational curve. Let γ be the rational normal curve (1) from the Introduction and suppose that the three fixed lines of our problem are its tangent lines $\ell(-1)$, $\ell(0)$, and $\ell(1)$. These line lie on the hyperboloid H with equation (2). Here is another view of these lines, the curve γ, and the hyperboloid.



If we require ℓ to meet the three tangent lines $\ell(-1)$, $\ell(0)$, and $\ell(1)$ and Λ to contain the two points $\gamma(v)$ and $\gamma(w)$ of γ, then ℓ also meets the line $\lambda(v,w)$ spanned by these two points. As in the Introduction, the lines ℓ that we seek will come from points where the secant line $\lambda(v,w)$ meets H.

Figure 2 shows an expanded view down the throat of the hyperboloid, with a secant line $\lambda(v,w)$ that meets the hyperboloid in two points. For these points $\gamma(v)$

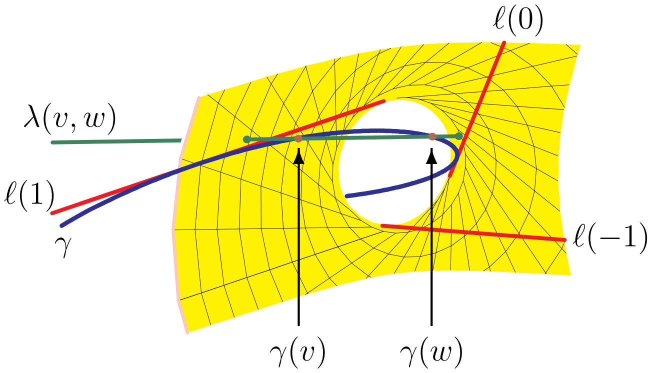


FIGURE 2. A secant line meeting H

and $\gamma(w)$ there will be two real flags $\ell \subset \Lambda$ satisfying our conditions. This is consistent with the Shapiro conjecture.

In contrast, Figure 3 shows a secant lines $\lambda(v, w)$ that does not meet the hyperboloid in any real points. For these points $\gamma(v)$ and $\gamma(w)$, neither flag $\ell \subset \Lambda$ satisfying our conditions is real. This is not consistent with the Shapiro conjecture, so we see that Shapiro conjecture does not hold for this Schubert problem, and so it is false.

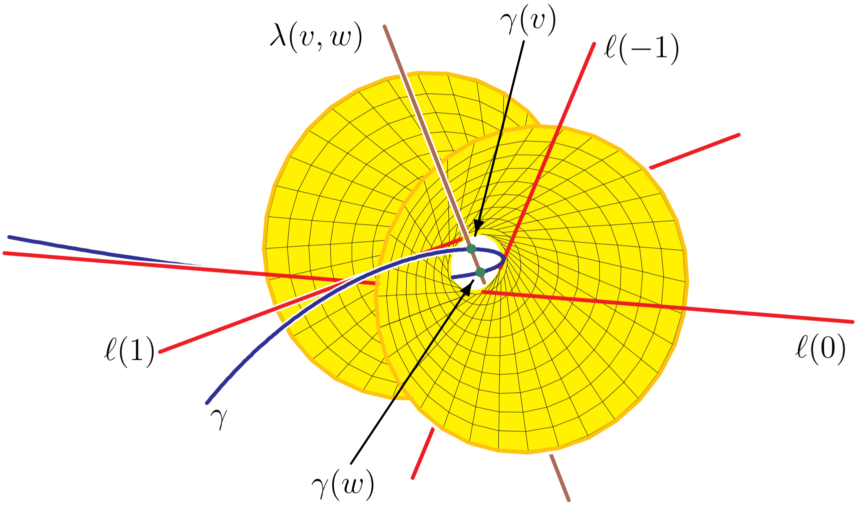


FIGURE 3. A secant line not meeting H

This failure is however quite interesting. If we label the points $-1, 0, 1$ with 1 (conditions on the line) and v, w by 2 (conditions on the plane), then along γ they occur in order

$$11122 \text{ in Figure 2} \quad \text{and} \quad 11212 \text{ in Figure 3.}$$

The sequence for Figure 2 is *monotone* and in this case both solutions are always real. This example suggests a way to correct the Shapiro conjecture, which we call the *monotone conjecture*.

Specifically, let $\mathbf{n}\colon 0 < n_1 < \cdots < n_m < d$ be a sequence of integers. The manifold $\mathbb{F}\ell_{\mathbf{n},d}$ of flags of type $\mathbf{n}$ is the set of all sequences of subspaces

$$E_\bullet \ : \ E_{n_1} \ \subset \ E_{n_2} \ \subset \ \cdots \ \subset \ E_{n_m} \ \subset \ \mathbb{C}_d[t]$$

with $\dim E_{n_i} = n_i$. The forgetful map $E_\bullet \mapsto E_{n_i}$ induces a projection

$$\pi_i \ : \ \mathbb{F}\ell_{\mathbf{n},d} \ \longmapsto \ \mathrm{Grass}_{n_i,d}$$

to a Grassmannian. A *Grassmannian Schubert variety* is a subvariety of $\mathbb{F}\ell_{\mathbf{n},d}$ of the form $\pi_i^{-1}\Omega_{\mathbf{a}}F_\bullet$. Write $X_{(\mathbf{a},n_i)}F_\bullet$ for this Grassmannian Schubert variety and call $(\mathbf{a}, n_i)$ a Grassmannian Schubert condition.

A *Grassmannian Schubert problem* is a list

$$(\mathbf{a}^{(1)}, n^{(1)}),\ (\mathbf{a}^{(2)}, n^{(2)}),\ \ldots,\ (\mathbf{a}^{(m)}, n^{(m)}), \tag{6.3}$$

of Grassmannian Schubert conditions satisfying $|\mathbf{a}^{(1)}| + \cdots + |\mathbf{a}^{(m)}| = \dim \mathbb{F}\ell_{\mathbf{n},d}$. We assume that the conditions (6.3) of a Grassmannian Schubert problem are sorted in that

$$n^{(1)} \leq n^{(2)} \leq \cdots \leq n^{(m)}.$$

We state the monotone conjecture.

Conjecture 6.4. *Let* $\big((\mathbf{a}^{(1)}, n^{(1)}), \ldots, (\mathbf{a}^{(m)}, n^{(m)})\big)$ *be a Grassmannian Schubert problem for the flag variety* $\mathbb{F}\ell_{\mathbf{n},d}$*. Whenever* $s_1 < s_2 < \cdots < s_m$ *are real numbers, the intersection*

$$X_{(\mathbf{a}^{(1)},n^{(1)})}F_\bullet(s_1) \bigcap X_{(\mathbf{a}^{(2)},n^{(2)})}F_\bullet(s_2) \bigcap \cdots \bigcap X_{(\mathbf{a}^{(m)},n^{(m)})}F_\bullet(s_m),$$

is transverse with all points of intersection real (when it is nonempty).

There is a lot of evidence in support of this monotone conjecture. First, the Shapiro conjecture for Grassmannians is the special case of the monotone conjecture when $m = 1$, for then $\mathbb{F}\ell_{\mathbf{n},d} = \mathrm{Grass}_{n_1,d}$, and the monotonicity condition $s_1 < \cdots < s_m$ is empty as any reordering of the list of Schubert conditions remains sorted.

But there is more. This conjecture was formulated in [28], where the failure of reality in our example was noted. That project utilized some serious computer investigation of the monotone conjecture. This computer experimentation used over 15 gigaHertz-years of computing, solving over 500 million polynomial systems representing intersections of Schubert varieties in over 1100 different enumerative problems on 27 different flag manifolds. Some of this computation studied intersections of Schubert varieties that were not necessarily monotone and that did not always involve Grassmannian Schubert conditions. This experimentation discovered that such an intersection is not necessarily transverse if the monotone condition is violated. More interesting, the intersection may not be zero-dimensional (for any $s_1, \ldots, s_m \in \mathbb{C}$) if the Schubert problem does not involve Grassmannian Schubert conditions.

A third piece of evidence for the monotone conjecture was provided by Eremenko, et. al [10], who showed that it is true for two-step flag manifolds, when $\mathbf{n} = d{-}1 < d$. This result is a special case of their main theorem, which asserts the reality of a rational function φ with prescribed critical points on $\mathbb{RP}^1$ and prescribed coincidences $\varphi(v) = \varphi(w)$, when v, w are real. Their proof was based on the results of [8].

Phrasing their result in terms of $\mathrm{Grass}_{d-1,d}$ shows that it is a generalization of the Shapiro conjecture, where we replace the flags $F_\bullet(s)$ by more general secant flags. Geometrically, the flag $F_\bullet(s)$ is the flag of subspaces osculating the rational normal curve γ. A *secant flag* $F_\bullet$ is one where every subspace F_i of $F_\bullet$ is spanned by its points of intersection with γ. Secant flags $F^1_\bullet, \ldots, F^m_\bullet$ are *disjoint* if there exist disjoint intervals $I_1, \ldots, I_m$ of γ such that the subspaces in flag $F^i_\bullet$ meet γ at points of I_i. The main result of [8] is that an intersection of Schubert varieties in $\mathrm{Grass}_{d-1,d}$ given by disjoint secant flags is transverse with all points real.

This result motivates the following *secant conjecture*.

Conjecture 6.5. *If* $(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ *is a Schubert problem for* $\mathrm{Grass}_{n,d}$ *and* $F^1_\bullet, \ldots, F^m_\bullet$ *are disjoint secant flags, then the intersection*

$$\Omega_{\mathbf{a}_1}F^1_\bullet \bigcap \Omega_{\mathbf{a}_2}F^2_\bullet \bigcap \cdots \bigcap \Omega_{\mathbf{a}_m}F^m_\bullet$$

is transverse with all points real.

References

[1] Daniel J. Bates, Frédéric Bihan, and Frank Sottile, *Bounds on the number of real solutions to polynomial equations*, Int. Math. Res. Not. IMRN (2007), no. 23, Art. ID rnm114, 7.

[2] Prakash Belkale and Shrawan Kumar, *Eigenvalue problem and a new product in cohomology of flag varieties*, Invent. Math. **166** (2006), no. 1, 185–228.

[3] G. Castelnuovo, *Numero delle involuzioni razionali gaicenti sopra una curva di dato genere*, Rendi. R. Accad. Lincei **4** (1889), no. 5, 130–133.

[4] David Cox, John Little, and Donal O'Shea, *Ideals, varieties, and algorithms*, third ed., Undergraduate Texts in Mathematics, Springer, New York, 2007, An introduction to computational algebraic geometry and commutative algebra.

[5] D. Eisenbud and J. Harris, *Divisors on general curves and cuspidal rational curves*, Invent. Math. **74** (1983), no. 3, 371–418.

[6] D. Eisenbud and J. Harris, *When ramification points meet*, Invent. Math. **87** (1987), 485–493.

[7] A. Eremenko and A. Gabrielov, *Degrees of real Wronski maps*, Discrete Comput. Geom. **28** (2002), no. 3, 331–347.

[8] ———, *Rational functions with real critical points and the B. and M. Shapiro conjecture in real enumerative geometry*, Ann. of Math. (2) **155** (2002), no. 1, 105–129.

[9] ———, *Elementary proof of the B. and M. Shapiro conjecture for rational functions*, 2005, `arXiv:math/0512370`.

[10] A. Eremenko, A. Gabrielov, M. Shapiro, and A. Vainshtein, *Rational functions and real Schubert calculus*, Proc. Amer. Math. Soc. **134** (2006), no. 4, 949–957 (electronic).

[11] Pavel Etingof and Victor Ginzburg, *Symplectic reflection algebras, Calogero-Moser space, and deformed Harish-Chandra homomorphism*, Invent. Math. **147** (2002), no. 2, 243–348.

[12] William Fulton, *Young tableaux*, London Mathematical Society Student Texts, vol. 35, Cambridge University Press, Cambridge, 1997, With applications to representation theory and geometry.

[13] Lisa R. Goldberg, *Catalan numbers and branched coverings by the Riemann sphere*, Adv. Math. **85** (1991), no. 2, 129–144.

[14] I. Gordon, E. Horozov, and M. Yakimov, *The real loci of Calogero-Moser spaces, representations of rational Cherednik algebras and the Shapiro conjecture*, 2007, `arXiv:math/0711.4336`.

[15] P. Griffiths and J. Harris, *Principles of algebraic geometry*, J. Wiley and Sons, 1978.

[16] Joe Harris and Ian Morrison, *Moduli of curves*, Graduate Texts in Mathematics 187, Springer-Verlag, 1998.

[17] Viatcheslav Kharlamov and Frank Sottile, *Maximally inflected real rational curves*, Mosc. Math. J. **3** (2003), no. 3, 947–987, 1199–1200.

[18] A. G. Khovanskiĭ, *Fewnomials*, Translations of Mathematical Monographs, vol. 88, American Mathematical Society, Providence, RI, 1991.

[19] Steven L. Kleiman, *The transversality of a general translate*, Compositio Math. **28** (1974), 287–297.

[20] E. Kostlan, *On the distribution of roots of random polynomials*, From Topology to Computation: Proceedings of the Smalefest (Berkeley, CA, 1990), Springer, New York, 1993, pp. 419–431.

[21] P. P. Kulish and E. K. Sklyanin, *Quantum spectral transform method. Recent developments*, Lecture Notes in Phys., vol. 151, Springer, Berlin, 1982, pp. 61–119.

[22] E. Mukhin, V. Tarasov, and A. Varchenko, *The B. and M. Shapiro conjecture in real algebraic geometry and the Bethe ansatz*, 2005, Annals of Mathematics, to appear.

[23] ———, *Bethe eigenvectors of higher transfer matrices*, J. Stat. Mech. Theory Exp. (2006), no. 8, P08002, 44 pp. (electronic).

[24] ———, *On reality property of Wronski maps*, 2007, `arXiv:math/0710.5856`.

[25] ———, *Schubert calculus and representations of general linear group*, 2007, `arXiv:math/0711.4079`.

[26] E. Mukhin and A. Varchenko, *Critical points of master functions and flag varieties*, Commun. Contemp. Math. **6** (2004), no. 1, 111–163.

[27] Evgeny Mukhin and Alexander Varchenko, *Norm of a Bethe vector and the Hessian of the master function*, Compos. Math. **141** (2005), no. 4, 1012–1028.

[28] Jim Ruffo, Yuval Sivan, Evgenia Soprunova, and Frank Sottile, *Experimentation and conjectures in the real Schubert calculus for flag manifolds*, Experiment. Math. **15** (2006), no. 2, 199–221.

[29] Vadim V. Schechtman and Alexander N. Varchenko, *Arrangements of hyperplanes and Lie algebra homology*, Invent. Math. **106** (1991), no. 1, 139–194.

[30] H. Schubert, *Anzahl-Bestimmungen für lineare Räume beliebiger Dimension*, Acta. Math. **8** (1886), 97–118.

[31] V. Sedykh and B. Shapiro, *On two conjectures concerning convex curves*, Internat. J. Math. **16** (2005), no. 10, 1157–1173.

[32] Frank Sottile, *The special Schubert calculus is real*, Electron. Res. Announc. Amer. Math. Soc. **5** (1999), 35–39 (electronic).

[33] ______, *Real Schubert calculus: polynomial systems and a conjecture of Shapiro and Shapiro*, Experiment. Math. **9** (2000), no. 2, 161–182.

[34] ______, *Some real and unreal enumerative geometry for flag manifolds*, Michigan Math. J. **48** (2000), 573–592, Dedicated to William Fulton on the occasion of his 60th birthday.

[35] Jan Verschelde, *Numerical evidence for a conjecture in real algebraic geometry*, Experiment. Math. **9** (2000), no. 2, 183–196.

Department of Mathematics, Texas A&M University, College Station, TX 77843, USA

E-mail address: sottile@math.tamu.edu

URL: www.math.tamu.edu/~sottile

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 8, 2008 (San Diego, California)

Günther Uhlmann, University of Washington
Invisibility

Antonella Grassi, University of Pennsylvania
Birational Geometry: Old and New

Gregory F. Lawler, University of Chicago
Conformal Invariance and 2-d Statistical Physics

Terence C. Tao, University of California, Los Angeles
Why are Solitons Stable?

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

2009 CURRENT EVENTS BULLETIN
Committee

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

James Demmel, *University of California, Berkeley*

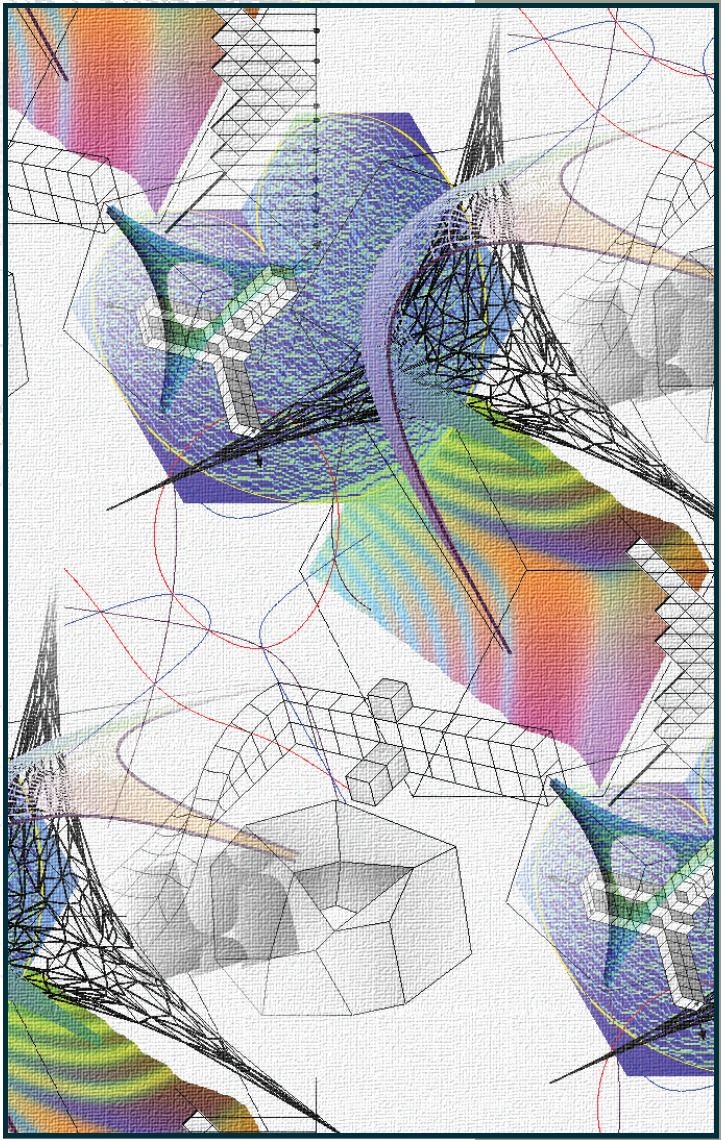
Eric Friedlander, *University of Southern California*

Susan Friedlander, *University of Southern California*

Andrew Granville, *University of Montreal*

Helmut Hofer, *Courant Institute, New York University*

Izabella Laba, *University of British Columbia*



Cover graphic associated with Emerton's talk courtesy of Jarke J. vanWijk using SeifertView, developed in cooperation with Arjeh Cohen. The image shows a Seifert surface of a Whitehead link.

Cover graphic associated with Hutchings' talk courtesy of Kenneth Baker, University of Miami

Cover graphic associated with Holtz's talk courtesy of Richard Baraniuk and Kevin Kelly, Rice University.

Cover graphic associated with Sottile's talk courtesy of Frank Sottile.

The back cover graphic is reprinted courtesy of Andrei Okounkov.