The Thom Conjecture

with an introduction to Seiberg-Witten gauge theory

An essay submitted to the
Department of Pure Mathematics and Mathematical Statistics
at the University of Cambridge
for Part III of the Mathematical Tripos 2014-15.

Contents

Acknowledgments. The material below is hardly original. I mainly drew from John Morgan's excellent and concise introduction to Seiberg-Witten theory [9], and from Liviu Nicolaescu's comprehensive study of the equations on 4-manifolds [11]. For introductory overview and emphases I appreciated the survey [5] by Michael Hutchings and Clifford Taubes, and the very nice overview in [13] by Zoltán Szabó, as well as some discussion of these equations in Donaldson's notes on geometric analysis [3]. In the introduction, some of my historical remarks are drawn from Varadarajan's thought provoking piece [18] on the history of "connections." For my understanding of the big picture on gauge theory I am indebted to Terence Tao's exposition [16] titled "What is a gauge?" and for the details to José Figueroa-O'Farrill's lecture notes on the subject [12]. This last heavily influenced the presentation in §1.1 below. I also used in a very limited way the books by Nakahara [10] and Kronheimer and Mrowka [7]; the former for information on gauge theory and the definition of the principle symbol; the latter for the monopole equations on cylinders. A few thoughts below come directly from Lawson and Michelsohn's book [8] on spin geometry, and Dietmar Salamon's manuscript [14] on the Seiberg-Witten invariants.

I would like to express my thanks firstly to Jacob Rasmussen for setting this essay, and for his generosity in spending time discussing geometry and answering my questions. I would also thank my Part III friend David Reutter for endless conversations about geometry and physics, and especially for my introduction to gauge theory.

I am very grateful to my undergraduate advisor Prof. Stephen Lovett for initiating my interest in differential geometry, and I thank my parents and family for support and encouragement during my adventure in England.

1 Introduction

Gauge theory in mathematics is the study of connections on principle bundles. In conceptual terms, a gauge is a sort of coordinate system for a family of similar objects. Given some family $\{P_x\}$ indexed by $x \in X$, a gauge is a collection of isomorphisms $\Phi_x : P_x \to H$, where the "standard object" H is well understood, and the isomorphisms are compatible with the structure of X; then (H, Φ_x) gives coordinates for P_x. Often in physics one describes a state of affairs by specifying, for each "place" in the world ($x \in X$), the space (P_x) of possible configurations of physical parameters at that place. Physically meaningful quantities should not depend on a choice of coordinates. A different gauge is given by another collection Φ'_x, from which it follows that the set of coordinate choices is just the symmetry group (or *structure group* or *gauge group*), $\mathrm{Aut}(H) =: G$. Assuming that all these objects are smooth gives a Lie group for G and a fiber bundle over a manifold $P \to X$. Actions of G on other spaces give rise to associated bundles (they "transform" in the same way), and in particular the self-action of G gives the *principle G-bundle.* All other associated bundles are naturally associated to the principle bundle, justifying its name.

While in general one begins with *a priori* relations between elements of X, e.g., topological proximity, the construction above does not specify any relation (other than smooth compatibility) between elements in fibers $\tilde{x} \in P_x$ and $\tilde{y} \in P_y$ for $x \neq y$. This is the role of a *connection*: a procedure to "lift" the relations between x and y to relations between $\tilde{x}$ and $\tilde{y}$. In the case of a fiber bundle $P \to X$, a path γ in the base manifold X from x to y is lifted by a connection to a unique path in P from a starting point $\tilde{\gamma}(0) = \tilde{x} \in P_x$ running through P_y. The data of $\tilde{\gamma}(0)$ is transported along γ. For the tangent bundle $TX \to X$, this is the meaning of parallel transport.

In historical terms, the need to understand connections of tangent spaces was realized fairly early, starting with Christoffel in the mid 19th century, and genuine parallel transport was introduced a bit later with Levi-Civita and Ricci in a 1900 paper. While the general definition in terms of lifting curves to the total space of a principle bundle was only fully prescribed by Charles Ehresmann in 1950, the real insight to connect something other than the tangent bundle came from Hermann Weyl's 1910's work on the geometry of spacetime and electromagnetism. Weyl recognized the need to "connect" the conformal *scale* of the metric in Einstein's theory of gravity (structure group $\mathbb{R}_{>0}$): length invariance in parallel transport (the Levi-Civita assumption) seemed mathematically and physically artificial. Abandoning that assumption opened the door to the more general notion of connection that would be crucial in finding the subtle geometric invariants that gauge theory would eventually produce.

Gauge theory continued to increase in importance as particle physics developed after the mid-1950's, eventually underlying Yang-Mills theory and thereby the Standard Model; but another conceptual leap came about when Simon Donaldson, building on Michael Atiyah's 1970's work on the Yang-Mills equations, realized that their solutions (connections on a principle bundle with structure group $SU(n)$) contain extraordinarily deep information about the global differential-topology of the underlying manifold. In 1980 he used these to define invariants able to distinguish differentiable structures of the same topological manifold.

Fourteen years later, Nathan Seiberg and Ed Witten introduced another set of gauge-theoretic equations in [15] and [19] (hereafter the SW equations) also arising from work in the geometry of quantum field theory. The abelian $U(1)$ structure group of these equations makes them much easier to manage than the Yang-Mills equations of Donaldson theory, and so it is somewhat surprising that they yield similarly profound results. One such is a confirmation of the Thom Conjecture:

Theorem 1.1 (Kronheimer and Mrowka [6], 1994). *Let Σ be an oriented 2-manifold smoothly embedded in $\mathbb{CP}^2$ so as to represent the same homology class as an algebraic curve of degree d. Then the genus g of Σ satisfies $g \geq (d-1)(d-2)/2$.*

An algebriac curve of degree d smoothly embedded in $\mathbb{CP}^2$ has genus given by strict equality in the above expression, so in this sense algebraic curves *minimize the genus* in any homology class.

The goal of this essay is to introduce the Seiberg-Witten invariants, including some brief snapshots of the rather sophisticated underlying machinery, and then to show how these invariants are used to prove Theorem 1.1. In the remaining sections of this introduction we define the main objects involved, and give some of their properties necessary to understanding the SW invariants. The first section is a more precise presentation of the gauge machinery discussed above; since this is the foundation of the subject we do spend some time there. The SW invariants are in terms of the moduli space of solutions to an elliptic first order PDE, very closely related to the *Dirac* equation (which Dirac wrote down to describe electrons in QED, conjugate solutions of which led to the discovery of positrons). Accordingly in §1.2 we describe how to put the Dirac equation on a manifold. A Dirac operator can be defined as an operator which *squares to a Laplacian*, and according to the "Bochner method" via the Weitzenböck presentation, two Laplacians differ by a zeroth order term involving the background curvature. This will be used in §2.3.1 with the preexisting *connection Laplacian* to bound solutions and thereby prove compactness of the moduli space. Then in §1.3 we develop the notion of *Spinor bundles* associated by representation to certain geometric *Spin* and *Spin^c* principle bundles called Spin and Spin^c structures. These can be seen as naturally induced by the requirements of a Dirac operator: the principle symbol defines a quadratic form begetting a Clifford algebra. Alternatively one could define Clifford bundles first and put the Dirac operator in their terms: the theory of spin bundles is contained in the theory of Clifford bundles, and Clifford bundles arise naturally in Riemannian geometry from the quadratic form on the tangent bundle. But we choose to emphasize the Dirac operator and take the former route (cf. Lawson and Michelson [8] for the latter). The section also includes a description of the Spin^c structures that are induced on the cross sections of cylinders equipped with Spin^c structures, and those induced by almost complex structures; both are used in the proof of Theorem 1.1.

Section 2 presents the Seiberg-Witten equations and invariant, the latter of which is in essence the homology class of the moduli space of solutions. To show that the moduli space is smooth, one applies a genericity argument to a perturburbation parameter (which we note lives in $H^2_+(X;\mathbb{R})$). For a smooth Riemannian 4-manifold (X, g), perturbation η, and Spin^c structure σ, the moduli space of solutions $\mathcal{M}(\sigma, g, \eta)$ has several nice properties including:

(1) $\mathcal{M}(\sigma, g, \eta)$ is compact.

(2) For generic η and $b_2^+(X) > 0$, $\mathcal{M}(\sigma, g, \eta)$ is a smooth, orientable manifold of finite dimension.

(3) If additionally $b_2^+(X) > 1$, then we can find a generic path (g_t, η_t) between generic points (g_0, η_0) and (g_1, η_1) in the space of metrics and perturbation parameters such that the *parametrized* moduli space $W := \{\mathcal{M}(\sigma, g_t, \eta_t)\}$ is also a smooth compact manifold with boundary $\partial W = \mathcal{M}(\sigma, g_1, \eta_1) - \mathcal{M}(\sigma, g_0, \eta_0)$.

Here b_2^+ is the positive second Betti number; the maximal dimension of a subspace $H^2_+(X;\mathbb{R})$ of $H^2(X;\mathbb{R})$ on which the pairing $Q(\alpha,\beta) = \int_X \alpha \wedge \beta$ is positive definite.

The cobordism in the third point leads to the remarkable fact that the homology class of $\mathcal{M}(\sigma, g, \eta)$ is *independent of the metric and perturbation.* The Seiberg-Witten invariant is an integral function of this homology class, given by integrating a natural volume form over the moduli space.

The procedure used to establish (2) and (3) is standard for such problems. The linearized equations are elliptic, and a version of Sard's theorem for Banach manifolds gives the genericity argument (also for the parametrized space). The dimension is given by Fredholm theory and straightforward application of the Atiyah-Singer index formula. The truly special feature of these equations is that $\mathcal{M}$ is always *compact.* This results from a uniform bound on the solutions stemming ultimately from the Bochner-type formula converting the Dirac Laplacian to a connection Laplacian (plus curvature terms). Accordingly we give a reasonably complete proof of compactness in §2.3.1, and sketches for most of the other facts.

In §2.4 we discuss the SW invariant, including a simplified version when $\dim(\mathcal{M}) = 0$ that we will use to prove the Thom Conjecture. Unfortunately the manifold for that conjecture has $b_2^+(\mathbb{CP}^2) = 1$ so the third point above doesn't strictly apply. This mild complication is discussed in §2.4.2.

Section 3 covers each piece of the proof of Theorem 1.1, describing in some detail SW solutions on cylinders and the process of "stretching the neck." A summary of the overall argument of the proof is as follows: we first take a smoothly embedded surface $\Sigma \hookrightarrow \mathbb{CP}^2$ in the same homology class as a degree d algebraic curve. After taking the d^2th blowup $\mathbb{CP}^2 \# d^2\overline{\mathbb{CP}}^2$ we can glue Σ to d^2 spheres to get a surface $\tilde{\Sigma}$ with the same genus and trivial self-intersection, and from that trivial normal bundle and tubular neighborhood. We can put a sequence of metrics on this object that "stretch" the tubular neighborhood radially, and show that for sufficiently stretched neighborhood the SW invariant is non-trivial; that is, solutions exist. In the meantime we can show that solutions on the very stretched space lead to (translation invariant) solutions on an infinite cylinder, now with $S^1 \times \tilde{\Sigma}$ as cross section. But these in turn imply a bound on the genus of $\tilde{\Sigma}$ in terms of its class and thereby the original degree, leading to the result.

1.1 Gauge machinery

Definition 1.2. A principle G-bundle $\pi : P \to X$ on a smooth manifold X is a fiber bundle with Lie group G as both the typical fiber *and* the structure group (acting on the left).

Then the fibers form G-torsors for the *right* action of G, which commutes with the action of the structure group.

Definition 1.3. Given a representation $\rho : G \to \mathrm{Aut}(F)$ for some space F, the associated bundle has fibers F and structure group G via the action ρ. It can be written $P \times_\rho F = P \times F / \sim$ where $(pg, f) \sim (p, \rho(g)f)$.

In what follows it will be useful to think of these in terms of local *gluing* data. The basic object is a *gluing G cocycle*, which is a (good) open cover $\{U_\alpha\}$ of X and collection of maps

$$g_{\alpha\beta} : U_{\alpha\beta} := U_\alpha \cap U_\beta \to G$$

for non-trivial double overlaps such that $g_{\alpha\alpha} \equiv 1$ and which satisfy the *cocycle condition* on triple overlaps:

$$g_{\gamma\alpha} = g_{\gamma\beta} \cdot g_{\beta\alpha} \qquad \forall x \in U_{\alpha\beta\gamma}.$$

Then the principle bundle is formed by gluing $G \times U_\alpha$ to $G \times U_\beta$ along their overlap via $(g, x) \sim (g_{\alpha\beta} \cdot g, x)$. With a representation $\rho : G \to \mathrm{Aut}(F)$ one forms the associated bundle by gluing $F \times U_\alpha$ to $F \times U_\beta$ along their overlap via $(f, x) \sim (\rho(g_{\alpha\beta})f, x)$.

Definition 1.4. A *connection* on a smooth principle G bundle $\pi : P \to X$ over a smooth manifold X is a choice of splitting

$$T_p P = V_p \oplus H_p$$

of the tangent bundle of P where V_p is the tangent space of the fiber $\pi^{-1}(\pi(p))$ called the *vertical* space, and the choice of complement H_p will be called the *horizontal* space. The splitting should be invariant under the G action, and vary smoothly. The projections are labeled v_p and h_p.

The point is that a connection lifts the tangent space of the underlying manifold into the total space by the identification $H_p \cong T_{\pi(p)} X$. Clearly vector fields and curves on X lift also.

The action of G on P gives a vector field $\sigma(X)$ on P for each $X \in \mathfrak{g}$, the Lie algebra of G. This action is fiberwise so $\sigma_p(X) \in V_p$, and it in fact yields an isomorphism $\sigma_p : \mathfrak{g} \cong V_p$. We then have:

Definition 1.5. The $\mathfrak{g}$-valued *connection 1-form* is given by the map

$$\omega_p : T_pP \xrightarrow{v_p} V_p \xrightarrow{\sigma_p^{-1}} \mathfrak{g}.$$

One easily checks that invariance of the splitting under the G action results in adjoint transformation of the 1-form:

$$\omega_p(\nu \cdot g) = g^{-1}\omega_p(\nu)g.$$

There is a unique *Maurer-Cartan* $\mathfrak{g}$-valued 1-form θ_g on G which is $\mathbf{id} : T_eG \to \mathfrak{g}$ at the identity e and is also left-invariant. One then finds that the restriction of ω_p to fibers gives θ.

In terms of gluing data $g_{\alpha\beta} : U_{\alpha\beta} \to G$, a connection is specified by $\mathfrak{g}$-valued 1-forms A_α which satisfy the gluing-compatibility condition

$$A_\beta = g_{\alpha\beta}^*\theta + g_{\alpha\beta}^{-1}A_\alpha g_{\alpha\beta}. \tag{1.1}$$

Given an associated vector bundle $P \times_\rho F$, we recover the gluing data of the associated covariant derivative as the map defined by

$$\nabla^\alpha = d + \rho_*(A_\alpha),$$

where d is the differential of a section considered as a collection $s_\alpha : U_\alpha \to F$, with $ds \in \Omega^1(U_\alpha)\otimes F$, and ρ_* is the differential at $e \in G$ so that $\rho_*(A_\alpha) \in \Omega^1(U_\alpha) \otimes \mathrm{Aut}(F)$. This operator measures the deviation of a section from being a "horizontal section," or the lift of a curve on X.

The *space of connections* $\mathcal{A}$ is just given by the possibilities for $\{A_\alpha\}$, so by taking the difference of two connections $\{A_\alpha\}$ and $\{\tilde{A}_\alpha\}$ in 1.1 we see that $\mathcal{A}$ is an affine space modeled on $\Omega^1(X; \mathrm{Ad}\, P)$.

Definition 1.6. The *curvature* of a connection is the pullback to horizontal vectors of the derivative of the connection:

$$\Omega = h^*d\omega.$$

Plugging in arbitrary vectors gives

$$\Omega(u, v) = -\omega([hu, hv])$$

since $\omega = 0$ on H_p, so $\Omega \equiv 0$ if and only if $[hu, hv] \equiv 0$; this tells us that Ω measures the *failure of integrability* of the distribution $\{H_p\}$. One can also check the *structure equation*

$$\Omega = d\omega + \frac{1}{2}[\omega, \omega]$$

(where $\frac{1}{2}[\cdot, \cdot]$ is at once half the Lie bracket on $\mathfrak{g}$ and the whole wedge on 1-forms), and the *Bianchi* *identity:*

$$h^* d\Omega = 0.$$

Locally the curvature is given by a collection of $\mathfrak{g}$-valued 2-forms F_{A_α} defined by:

$$F_{A_\alpha} = dA_\alpha + \frac{1}{2}[A_\alpha, A_\alpha]$$

which glue according to:

$$F_{A_\beta} = g_{\alpha\beta} F_{A_\alpha} g_{\alpha\beta}^{-1}.$$

These form a global 2-form $F_A \in \Omega^2(X; \mathrm{Ad}\, P)$.

Example 1.7. Seiberg-Witten gauge theory involves connections on complex line bundles $L \to X$, which will induce connections on the other relevant bundles. Assume an $L \to X$ has the *abelian* $U(1)$ structure group (from a Hermitian metric) and is given by gluing cocycles

$$z_{\alpha\beta} : U_{\alpha\beta} \to U(1).$$

A connection is then given locally by a set of $\mathbb{C}$-valued 1-forms A_α obeying

$$A_\beta = z_{\alpha\beta}^{-1} dz_{\alpha\beta} + A_\alpha,$$

and since the Lie algebra is $\mathbf{i}\mathbb{R}$ we can write $A_\alpha = \mathbf{i}\theta_\alpha$ for a real 1-form θ_α. The curvature is just $\{F_{A_\alpha}\} = \{dA_\alpha\}$.

One uses *invariant polynomials*—homogeneous polynomials on the Lie algebra which are invariant under the adjoint action—to construct the invariant *characteristic classes* of a fiber bundle. We will require the Chern class, which is the characteristic polynomial of the curvature 2-form, in local coordinates given by

$$\det\left(t\mathbf{1} + \frac{i}{2\pi}F_A\right) = t^n + c_1(F_A)t^{n-1} + c_2(F_A)t^{n-2} + \dots.$$

For line bundles only $c_1(F_A)$ isn't necessarily trivial. Also notice that in general $c_1(F_A) = \frac{i}{2\pi}\,\mathrm{tr}(F_A)$; considering again line bundles, all of this is captured in the important expression

$$c_1(F_A) = \frac{i}{2\pi}F_A.$$

One also knows by the Chern-Weil theorem that the c_j are closed and their cohomology classes don't depend on the connection, so we have a class $[c_1(L)]$; when the structure group is $U(n)$ this class is actually *integral,* so we may consider it in $H^2(X;\mathbb{R})$ or $H^2(X;\mathbb{Z})$.

The last major piece of general gauge theory that we will need is the notion of *change of gauge.* We define a gauge transformation to be an automorphism $\Phi : P \to P$, that is, a G-equivariant diffeomorphism which respects the fibers. A gauge transformation is given by a collection of local maps

$$\phi_\alpha : U_\alpha \to G$$

that again must glue consistently. A short calculation shows that the appropriate condition is

$$\phi_\beta = g_{\alpha\beta}\phi_\alpha g_{\alpha\beta}^{-1},$$

whence $\{\phi_\alpha\}$ defines a section $\phi \in C^\infty(X; \mathrm{Ad}\, P)$. We therefore have for the group of gauge transformations:

$$\mathcal{G} \cong \Omega^0(X; \mathrm{Ad}\, P);$$

when G is abelian this is just $\mathcal{G} \cong \Omega^0(X; G)$, and for $U(1)$ it is $\mathcal{G} \cong \Omega^0(X; S^1)$.

These act naturally on $\mathcal{A}$ as follows: Given a G-invariant distribution $\{H_p\}$ and transformation $\Phi \in \mathcal{G}$, the distribution $\Phi_* H_p$ is also G-invariant and defines another connection. Again one can work out the effect in terms of local data:

$$A_\alpha^\Phi = \phi_\alpha^* \theta + \phi_\alpha A_\alpha \phi_\alpha^{-1}, \qquad F_{A_\alpha}^\Phi = \phi_\alpha F_{A_\alpha} \phi_\alpha^{-1}. \tag{1.2}$$

Notice by comparing this to 1.1 that any object defined in terms of the $\{A_\alpha\}$ which is invariant under change of gauge will be consistently defined across overlaps, and thereby exist globally. Note also that the action doesn't change the norm of the curvature; indeed for abelian G the curvature doesn't change at all.

1.2 The Dirac operator

We turn now to some facts about partial differential operators on manifolds, beginning with a definition. Let (X, g) be a smooth, oriented Riemannian manifold, and $E, F \to X$ a pair of Hermitian vector bundles over X.

Definition 1.8. A *partial differential operator (PDO) of order* $\leq k$ is a $\mathbb{C}$-linear map on sections $T : C^\infty(E) \to C^\infty(F)$ such that for any $f \in C^\infty(X)$ the commutator $[T, f]$ is a PDO of order $\leq k-1$. A PDO of order 0 is an element of $\mathrm{Hom}(E, F)$.

These operators can be written in local coordinates in the usual way, with coefficients matrix-valued functions of position $x \in X$. Intuition for the above inductive definition is that $[T, f]$ removes the highest order term (modifying slightly the lower terms).

Definition 1.9. Let $x \in X$, $\xi \in T_x^* X$, and $s \in \pi_E^{-1}(x)$. Let $\tilde{s} \in C^\infty(X, E)$ be any section with

$\tilde{s}(x) = s$ and let $f \in C^\infty(X)$ be arbitrary such that $f(x) = 0$ and $df(x) = \xi$. The *principle symbol* of a PDO $T : C^\infty(E) \to C^\infty(F)$ of order n is given by the map

$$\sigma(T, \xi)s = \frac{1}{n!} T(f^n \tilde{s})|_p.$$

The point is that all but the leading term is removed by f^n since $f(x) = 0$. We will use three facts about the principle symbol. First, one can check that it composes with composition of operators. Second, if the linear bundle map defined by $\sigma(T, \xi)$ is invertible whenever $\xi \neq 0$, we say the operator T is *elliptic,* and we can avail ourselves of the strong existence and regularity results for elliptic PDE. The third is our next essential definition:

Definition 1.10. A *generalized Laplacian* is an operator $L : C^\infty(X, E) \to C^\infty(X, E)$ with the symbol:

$$\sigma(L, \xi) = -|\xi|^2 \mathbf{1}_E. \tag{1.3}$$

Pairing sections (with the Hermitian forms) and integrating over X lets us define formal adjoints of operators in the usual way. A Hermitian connection ∇ on $E \to X$ is a first-order PDO, with formal adjoint ∇^*. Together they define the *connection Laplacian*

$$\nabla^*\nabla : C^\infty(X, E) \to C^\infty(X, E).$$

A key feature of generalized Laplacians is that they are the same to nonzero order; this means they can all be expressed in terms of a connection Laplacian and a remainder:

Proposition 1.11 (Weitzenböck presentation)**.** *Let $L : C^\infty(X, E) \to C^\infty(X, E)$ be a formally self-adjoint generalized Laplacian. Then we can find a Hermitian connection and symmetric $\mathcal{R} \in \mathrm{End}(E)$ such that*

$$L = \nabla^*\nabla + \mathcal{R}.$$

Proof. See [11, Prop. 1.2.8] or [2] for two proofs. □

We now define the Dirac operator:

Definition 1.12. A *Dirac operator* $D : C^\infty(X, E) \to C^\infty(X, F)$ is a PDO of first order such that $D^2 = D^*D$ is a generalized Laplacian.

Expressing $\mathcal{R}$ in terms of geometric quantities like the curvature is known generally as *Böchner's method,* and there are nearly as many formulas of "Böchner-type" as there are different kinds of Laplacians. We shall in particular make heavy use of the one by Lichnerowicz (given below after introducing spinors) that relates the Dirac Laplacian to a connection Laplacian.

We conclude the section by stating a Kato inequality that will be necessary later for the compactness proof:

Proposition 1.13. *Let us be given a Hermitian vector bundle $E \to X$ over a Riemannian manifold (X, g). Let Δ_X be the Hodge–de Rham Laplacian. Suppose ∇_E is a Hermitian connection, and $\Delta_E := \nabla_E^* \nabla_E$ its connection Laplacian. Then the (almost everywhere) bound obtains:*

$$\Delta_X |u|^2 \leq 2 \operatorname{Re}\langle \Delta_E u, u \rangle.$$

At this point we remark that we assume the reader is familiar with Sobolev norms and embedding theorems, and the method of weak equations on a manifold. We will use the notation $L^p_k(X, E)$ for the L^p norms summed to the kth derivative. To avoid clutter we shall assume our sections live in the correct space, taking care to distinguish them only in a "bootstrapping" or similar argument. Note that the above inequality applies for $u \in L^2_{2,loc}(X, E)$.

1.3 Spin

After setting up the fundamentals of gauge theory and Dirac operators, we now come to another crucial constituent of Seiberg-Witten theory, which is the structure of *spin.* As mentioned above, this could be an entire subject of its own, the geometry of spin and spinor bundles having lead to many profound results quite apart from Seiberg-Witten theory. (The Atiyah-Singer index theorem, numerous theorems of vanishing of betti numbers, results about manifolds of positive scalar curvature, and the Positive Mass Conjecture of general relativity, to name a few.) However, we will not need much of this theory; in particular we can restrict most of our discussion to 4-manifolds and the group $Spin(4)$ and its representations.

Of chief importance for us is the intimate relationship between spinor bundles and Dirac operators. When Paul Dirac originally searched for a coherent theory of the electron, he needed an operator to be (1) Lorentz invariant, (2) causally sensible, and (3) in agreement with the Klein-Gordon wave equation. The first requires treating time on equal footing with space, the second requires the time (thus space) component of the operator to be *first order,* and the third entails (roughly) squaring to a Laplacian. This is all wrapped into our definition above of the Dirac operator.

We can then ask: on which bundles do Dirac operators live? We will see in this section that the natural answer is the spinor bundles associated to a principle $Spin(n)$ bundle.

We begin with the Clifford structure defined by a Dirac operator. Let $\rho : T^*X \to \operatorname{End}(E)$ be the principle symbol of a Dirac operator $D : C^\infty(X, E) \to C^\infty(X, E)$. So by the composition law and Eqn. 1.3 we have at each $x \in X$

$$\rho(\xi)^2 = -|\xi|^2 \mathbf{1}_{E_x}. \tag{1.4}$$

This immediately gives us an algebra in $\operatorname{End}(E_x)$ which is generated by $\{\rho(e_i)\}$ for e_i an orthonormal basis of $T_x^*X =: V$, and the relation above is easily rewritten:

$$\{\rho(u), \rho(v)\} = -2\langle u, v\rangle.$$

The defining relation for the *Clifford algebra* $Cl(V)$ is just

$$\{e_i, e_j\} = -2\delta_{ij},$$

so the principle symbol gives a representation of $Cl(V)$, called *Clifford multiplication.* The principle symbol is actually a *bundle map* $T^*X \to \mathrm{End}(E)$ so in fact we have even more:

$$\{\rho(\lambda), \rho(\delta)\} = -2g(\lambda, \delta)\mathbf{1}_E \qquad \forall \lambda, \delta \in \Omega^1(X). \tag{1.5}$$

This representation of the Clifford bundle of T^*X is called a *Clifford structure*. We can also go the other direction: given a Clifford structure $\rho : T^*X \to \mathrm{End}(E)$ obeying 1.4, we obtain a Dirac operator from any connection by composing with ρ. (We would also require that ρ be self-adjoint and compatible with the connection.) In some sense this direction is more practical because one must first understand which bundles $E \to X$ permit such a structure, though in the end for us they are really just homes for the Dirac operators.

Next we discuss what $E \to X$ should in fact be; clearly it must involve representations of $Cl(V)$. But to see which ones we need to define some objects inside $Cl(V)$:

Definition 1.14. Let $Cl_0(V) \subset Cl(V)$ be the "even" subalgebra generated by products of an even number of e_i, and $Cl_1(V)$ the "odd" module over $Cl_0(V)$. Let $Pin(V) \subset Cl^\times(V)$ denote the multiplicative group generated by $v \in V$ with $\|v\|^2 = 1$, and set $Spin(V) := Pin(V) \cap Cl_0(V)$, the even part. We write $Spin(n)$ for $Spin(\mathbb{R}^n)$ etc.

The group $Spin(V)$ acts by conjugation on $Cl(V)$, and one can show by writing it as an even number of reflections that $V \subset Cl(V)$ is preserved with its orientation, so we have a map

$$\pi : Spin(V) \to SO(V).$$

One then finds the short exact sequence:

$$0 \to \mathbb{Z}_2 \to Spin(V) \to SO(V) \to 1, \tag{1.6}$$

so that $Spin(n)$ is the double cover of $SO(n)$, and the universal cover when $n \geq 3$ (so that $\pi_1(SO(n)) = \mathbb{Z}_2$). Now it turns out that $Spin(n)$ is exactly the Lie group we need, whose associated vector bundles support a Clifford representation and thereby a Dirac operator. The *reason* is that $Pin(n)$ generates $Cl(n)$, in the sense that $Pin(n)$ contains an $\mathbb{R}$-basis of $Cl(n)$. We get a short exact sequence similar to the above with $Pin(n)$ and $O(n)$ instead of $Spin(n)$ and $SO(n)$, but because $SO(n)$ is the structure group of the frame bundle of T^*X, we need to reduce from $Pin(n)$ to $Spin(n)$.

It happens that the existence of a principle $Spin(n)$ bundle as the lift of the principle $SO(n)$ frame bundle imposes a somewhat difficult topological constraint on X, so we generalize to consider the double cover of $SO(V) \times U(1)$, which can be described as follows:

Definition 1.15. Let

$$Spin^c(V) := (Spin(V) \times U(1))/\mathbb{Z}_2, \tag{1.7}$$

where the action of $\mathbb{Z}_2$ on $Spin(V)$ flips the sheet in the double cover, and on $U(1)$ flips the sign.

Since π is the same on both sheets, it induces a map $\pi : Spin^c(V) \to SO(V)$. Squaring the projection to $U(1)$ gives another well-defined map $Spin^c(V) \to U(1)$ which we will use (below) to define the line bundle of a $Spin^c$ structure. One can show by checking these properties and uniqueness that $Spin(3) \cong SU(2)$ and $Spin(4) \cong SU(2) \times SU(2)$.

Next we describe the representations that will be used for the bundle $E \to X$, and we restrict to the case $n = 4$. We consider the complexification $Cl(\mathbb{R}^4) \otimes \mathbb{C}$, which has as its (complex) volume element $\omega_{\mathbb{C}} = -e_1e_2e_3e_4$, satisfying $\omega_{\mathbb{C}}^2 = 1$. Left multiplication induces a splitting into eigenspaces of ± 1 :

$$Cl(\mathbb{R}^4) \otimes \mathbb{C} = (Cl(\mathbb{R}^4) \otimes \mathbb{C})^+ \oplus (Cl(\mathbb{R}^4) \otimes \mathbb{C})^-,$$

which descends to a similar splitting of $Cl_0(\mathbb{R}^4) \otimes \mathbb{C}$.

We then have representations as follows:

Proposition 1.16. *$Cl(\mathbb{R}^4) \otimes \mathbb{C} \cong \mathbb{C}[4]$, where $\mathbb{C}[4]$ is the 4×4 matrix algebra with complex entries. In particular, $Cl(\mathbb{R}^4)$ has precisely one complex representation, $W = \mathbb{C}^4$ (up to isomorphism). The action gives an isomorphism*

$$Cl(\mathbb{R}^4) \otimes \mathbb{C} \cong \mathrm{End}(W) = W \otimes W^*.$$

Under this identification, the action of $\omega_{\mathbb{C}}$ also splits

$$W = W^+ \oplus W^-$$

and gives isomorphisms

$$(Cl_0(\mathbb{R}^4) \otimes \mathbb{C})^{\pm} \cong \mathrm{End}(W^{\pm}) \tag{1.8}$$

and

$$(Cl_1(\mathbb{R}^4) \otimes \mathbb{C})^{\pm} \cong \mathrm{Hom}(W^{\mp}, W^{\pm}). \tag{1.9}$$

This pair of inequivalent, irreducible representations of $Cl_0(\mathbb{R}^4) \otimes \mathbb{C}$ induces a pair of inequivalent, irreducible representations

$$\Delta^{\pm} : \mathrm{Spin}(\mathbb{R}^4) \to \mathrm{Aut}(W^{\pm}).$$

Finally, these extend uniquely to representations of $Spin^c$.

Proof. The proofs are not hard, but to save space we refer the reader to [8] or [9]. □

The Δ^+ and Δ^- are called "positive" and "negative" spinor representations, and elements in W^+ and W^- are said to have "spin up" and "spin down." Also, most of this generalizes to higher (even) dimensions, but there is one fact peculiar to $n = 4$, and that is a correspondence with the *Hodge* splitting of $\Lambda^2(\mathbb{R}^4)$ into self-dual and anti-self-dual forms. Under the natural identification $Cl(\mathbb{R}^4) \otimes \mathbb{C} \cong \Lambda^*(\mathbb{R}^4) \otimes \mathbb{C}$, the self-dual forms $\Lambda^2_+(\mathbb{R}^4)$ only act on W^+ and $\Lambda^2_-(\mathbb{R}^4)$ only on W^-. This leads to the interesting fact (one can show) that the Seiberg-Witten equations are *overdetermined* in dimensions $n > 4$.

Definition 1.17. A *Spin* (resp. *Spin*^{c}) *structure* $\sigma : P \to X$ on the Riemannian manifold (X, g) is a principle *Spin* (resp. *Spin*^{c}) bundle over X lifted from the principle $SO(n)$ bundle associated to the orthonormal frame bundle. The set of structures is denoted $Spin(X)$ (resp. $Spin^c(X)$).

Remark 1.18. Note two facts about this definition: First, the lifting is itself important to a *Spin* structure; one may have distinct *Spin* structures that are equivalent as principle *Spin* bundles: as an example, one may check that there are two *Spin* structures over S^1, but of course there is just one *Spin* principle bundle (take a cross section).

Second, the metric g is necessary to specify a *Spin* or *Spin*^{c} structure. However, the data of a *Spin* or *Spin*^{c} structure induces another structure for every other metric. This is because the $SO(n)$ frame bundle is a *deformation retract* of the $GL(n)^+$ bundle of oriented bases. Indeed, one can define a notion of *spin equivalence* which is a diffeomorphism $X \to X$ that preserves orientation and *Spin* structures; one can then show that the spin class is *independent* of the choice of g. This will be crucial to showing that the Seiberg-Witten invariant is independent of g, since we must vary g, implicitly assuming this equivalence. See [8, p. 82, 90] for more details.

We can finally define the spinor bundle which we have been calling $E \to X$ above:

Definition 1.19. Let $Fr \to X$ be the tangent bundle TX with structure group reduced to $SO(4)$ by the metric, and let $\sigma : P \to X$ be a $Spin^c(4)$ structure. The representation $\Delta^{\pm} : \mathrm{Spin}^c(4) \to W^{\pm}$ gives the associated *spinor bundles*, also written $W^{\pm}$:

$$(W^{\pm} \to X) \quad = \quad P \times_{\Delta^{\pm}} W^{\pm}.$$

Now it is certainly not true that *Spin* and $Spin^c$ structures automatically exist. In fact it is not hard to describe when they exist and what they are. We give a sketch here (following [11]) for $Spin^c$ structures in terms of gluing data; a very similar method works for *Spin* structures.

Looking at our definition 1.7 of $Spin^c$, it is clear that a $Spin^c$ structure σ is presented by gluing data as a (good) cover $\{U_\alpha\}$, and a collection of maps:

$$g_{\alpha\beta} : U_{\alpha\beta} \to SO(n)$$

defining $Fr \to X$, and

$$h_{\alpha\beta} : U_{\alpha\beta} \to Spin(n),\ z_{\alpha\beta} : U_{\alpha\beta} \to U(1),$$

such that for $\pi : Spin(n) \to SO(n)$,

$$\pi(h_{\alpha\beta}) = g_{\alpha\beta},$$

and the overlap condition removes $\mathbb{Z}_2$ in parallel:

$$(w_{\alpha\beta\gamma}, \xi_{\alpha\beta\gamma}) := (h_{\alpha\beta}h_{\beta\gamma}h_{\gamma\alpha}, z_{\alpha\beta}z_{\beta\gamma}z_{\gamma\alpha}) \in \{(-1,-1),(1,1)\}. \tag{1.10}$$

As mentioned above, taking the square $\lambda_{\alpha\beta} = z_{\alpha\beta}^2$ gives a collection satisfying the cocycle condition, which thus determines the associated line bundle, and for which we will write $\det(\sigma) \to X$.

The key step is then to show that for *any* lift $h_{\alpha\beta}$ (not necessarily obeying 1.10), $w_{\alpha\beta\gamma}$ defines a Čech 2-cocycle which is independent of the lift, and without the restriction 1.10 such a lift is always possible. This cocycle is the second Steifel-Whitney class $w_2(X)$. (This is how one shows that *Spin* structures exist if and only if $w_2(X) = 0$.) On the other hand, using the exponential sheaf sequence for the line bundle $\det(\sigma) \to X$ and the induced long exact sequence on Čech cohomology to get the topological first Chern class of $\det(\sigma)$, the relation 1.10 is exactly the statement

$$c_1^{top}(\det\sigma) \equiv w_2(X) \mod 2.$$

One can show that this condition is sufficient. Furthermore, if we define

$$\mathcal{L}_X = \{\beta \in H^2(X;\mathbb{Z}) \mid \beta \equiv w_2(X) \bmod 2\},$$

then one can prove:

Proposition 1.20. *If $H_1(X;\mathbb{Z})$ has no 2-torsion (e.g., X is simply connected) then taking the determinant line bundle* $\det(\sigma)$ *gives a bijection between the set of $Spin^c$ structures $Spin^c(X)$ and $\mathcal{L}_X$.*

There is also a natural action of the set of line bundles $\mathrm{Pic}^\infty(X)$ on $Spin^c(X)$. Given a line bundle by gluing data $\tilde{\lambda}_{\alpha\beta} : U_{\alpha\beta} \to U(1)$, one can multiply the $U(1)$ part to get a new $Spin^c$ structure with data $(h_{\alpha\beta}, z_{\alpha\beta}\tilde{\lambda}_{\alpha\beta})$. (Note that the new $\det(\tilde{\sigma})$ comes from squaring the $U(1)$ part, so its c_1^{top} changes by an even number.) This action is free and transitive, and c_1^{top} gives an isomorphism $\mathrm{Pic}^\infty \cong H^2(X;\mathbb{Z})$, so in the end we also have:

Proposition 1.21. *$Spin^c(X)$ is an $H^2(X;\mathbb{Z})$-torsor; an affine space modeled on $H^2(X;\mathbb{Z})$.*

We finally state an important fact:

Theorem 1.22. *Every oriented smooth 4-manifold X has at least one $Spin^c$ structure.*

Proof. See [8] or [14] for the compact case; it has been proven by Teichner in [17] for the non-compact case as well. □

Later in the essay we will need to understand the induced $Spin^c(Y)$ from the set $Spin^c(\mathbb{R}\times Y)$ with $\dim(Y) = 3$, and the induced $Spin^c(X)$ from an almost complex structure J on X. We follow [11].

For the first, note that as groups $Spin(3) = SU(2)$ and $Spin(4) = SU(2)\times SU(2)$, so there is a diagonal inclusion $Spin(3) \hookrightarrow Spin(4)$. One easily checks that this commutes with projection to $SO(n)$, and yields a commutative diagram

$$\begin{tikzcd} Spin^c(3) \arrow[r, hook] \arrow[d] & Spin^c(4) \arrow[d] \\ SO(3) \arrow[r, hook] & SO(4). \end{tikzcd} \tag{1.11}$$

On the cylinder $\mathbb{R}\times Y$ over a 3-manifold Y, the tangent bundle can be written as the product $\mathbb{R}\times TY$ and the $SO(4)$ frame bundle actually reduces to an $SO(3)$ bundle. A $Spin^c(3)$ structure $\sigma : P \to Y$ can be trivially extended to a $Spin^c(4)$ structure $\hat{\sigma} : \mathbb{R}\times P \to \mathbb{R}\times Y$, and we have an isomorphism of $Spin^c$ structures

$$Spin^c(Y) \cong Spin^c(\mathbb{R}\times Y).$$

For the second, let (X, J, g) be an almost complex manifold of real dimension 4 with J and g compatible, so the $SO(4)$ frame bundle reduces to a $U(2, J)$ principle bundle (where unitarity means orthogonal and commuting with J). The main ingredient is:

Lemma 1.23. *The map $U(2, J) \to SO(4)$ factors through a map $\xi : U(2, J) \to Spin^c(4)$:*

$$\begin{tikzcd} U(2,J) \arrow[r, "\xi"] \arrow[dr] & Spin^c(4) \arrow[d] \\ & SO(4). \end{tikzcd} \tag{1.12}$$

Main idea of proof (cf. [11]): Take a path γ between 1 and any $\omega \in U(2,J)$, considered by the inclusion in $SO(4)$. Lift this to $Spin(4)$ on the one hand, and lift $\det(\gamma)$ by the double cover $U(1) \to U(1),\ z \mapsto z^2$ on the other. Pair the results and compare to 1.7. $\square$

Then gluing data $h_{\alpha\beta} : U_{\alpha\beta} \to U(2,J)$ defining the frame bundle gives gluing data $\xi(h_{\alpha\beta}) : U_{\alpha\beta} \to Spin^c(4)$ which defines a $Spin^c$ structure σ. By the construction of the lemma, we also have the convenient fact that $\det(\sigma)$ is precisely the (dual) of the canonical line bundle of (X,J,g).

The last bit of background information we will need is the definition of the Dirac operator of a $Spin^c$ structure σ. As mentioned above, we need a map $\rho : T^*X \to \mathrm{End}(W)$ and a connection on W; composition gives the Dirac operator. The first part is easy: let us be given a $Spin^c$ structure $\sigma : P \to X$ and associated spinor bundle $W \to X$. We have mentioned the identification $Cl(\mathbb{R}^4) \otimes \mathbb{C} \cong \Lambda^*(\mathbb{R}^4) \otimes \mathbb{C}$. This extends to the bundles, to yield

$$T^*X \otimes \mathbb{C} \subset Cl_1(TX) \otimes \mathbb{C}.$$

Then by 1.9 we have a linear map

$$Cl_1(TX) \otimes \mathbb{C} \to \mathrm{End}(W)$$

that switches W^+ and W^-. Putting this together is precisely our desired Clifford multiplication:

$$\rho : T^*X \otimes \mathbb{C} \to \mathrm{End}(W). \tag{1.13}$$

Lastly we need a connection on W. The main point is that we need only specify a $U(1)$ connection A on the determinant line bundle $\det(\sigma)$ of a $Spin^c$ structure σ. We have seen that connections on principle bundles yield connections (covariant derivatives) on their associated vector bundles. There is a connection ω on the principle $SO(4)$ bundle which underlies the Levi-Civita connection on the associated frame bundle. Then $\omega \oplus A$ is a connection on a principle $SO(4) \oplus U(1)$ bundle. Now we have noted one possible definition of $Spin^c$ as the double cover of $SO(4) \oplus U(1)$; since this cover is finite there is an induced connection on our $Spin^c$ structure σ given by pulling back along the covering projection. (Or thinking of H_p, pushing forward to *both sheets* by the differential of the inverse of the projection.) Thus we will write ∇_A for the covariant derivative on $W \to X$ arising from a $U(1)$ connection A.

One can check that $\tilde{\nabla}_A(\omega_{\mathbb{C}}) = 0$ (for $\tilde{\nabla}_A$ induced on $Cl(TX)$) so ∇_A respects the splitting $W = W^+ \oplus W^-$. The Dirac operator we finally obtain is

$$D_A : \Gamma(W^+) \to \Gamma(W^-) \quad \text{by} \quad D_A = \rho \circ \nabla_A,$$

where we think of the connection as a map

$$\nabla_A : \Gamma(W^+) \to \Gamma(T^*X \otimes W^+).$$

In local coordinates at x with $SO(4)$ frame $\{e_1 \dots e_4\}$ of T_xX, and a "$\cdot$" for the Clifford multiplication, this is written

$$\nabla_A(\psi)(x) = \sum_{i=1}^{4} e_i \cdot \nabla_{Ae_i}(\psi)(x).$$

The Weitzenböck presentation for this Dirac operator is the following:

Proposition 1.24 (Lichnerowicz)**.** *Let (X, g) be a Riemannian manifold with $Spin^c$ structure σ and a connection A on $\det(\sigma)$. Let D_A be the associated Dirac operator as above. Then the Weitzenböck presentation is*

$$D_A^* D_A = \nabla_A^* \nabla_A + \frac{s}{4} + \frac{1}{2}\rho(F_A^+). \tag{1.14}$$

In this formula s is the scalar curvature, or the trace of the Ricci curvature, of (X, g). The identification $Cl(TX) \otimes \mathbb{C} \cong \Lambda^*(TX) \otimes \mathbb{C}$ that we used to define ρ extends to

$$\rho : \Lambda^2(T^*X) \otimes \mathbb{C} \to \mathrm{End}(W),$$

where on 2-forms it preserves $W = W^+ \oplus W^-$. Also F_A^+ is just the self-dual part of the curvature of the connection A. Observe that because D_A^* takes $\Gamma(W^-)$ to $\Gamma(W^+)$ and the 2-form F_A^+ corresponds to something in Cl_0, this makes sense as a map $\Gamma(W^+) \to \Gamma(W^+)$.

2 Seiberg-Witten theory

In this section we aim to provide an overview of Seiberg-Witten theory. Our main goal is to give background for the results in §3, so we give full detail for the proof of compactness, and much less for most of the others. The first section presents the equations and describes the gauge group and its action. In the next we discuss the moduli space of solutions, and the last section is about constructing the Seiberg-Witten invariant from that space.

2.1 Equations

The (perturbed) Seiberg-Witten equations are:

$$SW(X, \sigma, g, \eta) = \begin{cases} D_A\psi = 0 \\ \rho(F_A^+ + i\eta^+) = \dfrac{1}{2}q(\psi), \end{cases} \tag{2.1}$$

and their explanation as follows. The data one initially provides is a connected, Riemannian 4-manifold (X, g), a *Spin*^{c} structure $\sigma : P \to X$ with corresponding spinor bundles $W_\sigma^\pm \to X$ and Clifford multiplication ρ. Now the closed *perturbation 2-form* $\eta \in \Omega^2(X)$ is used in the genericity argument to obtain a smooth manifold for the space of solutions. That process will be discussed in §2.3.2. Note that $U(1)$ is abelian with Lie algebra $i\mathbb{R}$ so the curvature $F_A = dA$ is a purely imaginary 2-form on X. The *variables* in the equation are pairs (ψ, A) called *configurations* where $\psi \in \Gamma(W_\sigma^+)$ and A is a $U(1)$ connection on the determinant line bundle $\det(\sigma)$, and *solutions* to these equations are called *Seiberg-Witten monopoles.*

The last object in need of explanation is $q(\psi) : \Gamma(W_\sigma^+) \to \mathrm{End}(W_\sigma^+)$. This is given by

$$q(\psi) := \psi \otimes \psi^* - \frac{1}{2}|\psi|^2\mathbf{1},$$

and is understood as follows: In a (complex) basis for W^+, $\psi = (\psi_1; \psi_2)$, we have

$$\psi \otimes \psi^* = \begin{pmatrix} |\psi_1|^2 & \psi_1\overline{\psi}_2 \\ \overline{\psi}_1\psi_2 & |\psi_2|^2 \end{pmatrix},$$

whereby $q(\psi)$ is a traceless, Hermitian endomorphism. Then to make sense of the equation we show that $\rho^{-1}(q(\psi))$ is a purely imaginary self-dual 2-form, in two lemmas; the first makes precise our comment above about dimension 4:

Lemma 2.1. *For a vector space V of dimension 4, the (complexified) Clifford–exterior algebra correspondence restricts to:*

$$(Cl_0(V) \otimes \mathbb{C})^+ \cong \mathbb{C}(\frac{1+\omega_{\mathbb{C}}}{2}) \oplus (\Lambda_+^2(V) \otimes \mathbb{C}). \tag{2.2}$$

This gives a correspondence between $\Lambda_+^2(TX) \otimes \mathbb{C}$ and the traceless endomorphisms of W_σ^+. We can then prove:

Lemma 2.2. *Under this correspondence, $q(\psi)$ is a purely imaginary self-dual 2-form.*

Proof sketch. We follow Morgan [9, Lem. 4.1.1]. The *real* (pre-complexified) subspace $Cl_0^+(T_xX)$ at a point is given by complex matrices of the form

$$\begin{pmatrix} \alpha & -\overline{\beta} \\ \beta & \overline{\alpha} \end{pmatrix},$$

that is, matrices λ satisfying $\lambda^* + \lambda = \mathrm{Tr}(\lambda)\mathbf{1} \in \mathbb{R}$. We know $\mathrm{Tr}(q(\psi)) = 0$. So to show $q(\psi)$ is imaginary it suffices to show $iq(\psi)$ is real, which then just means

$$(iq(\psi))^* + iq(\psi) = 0;$$

one rearranges this to

$$q(\psi) = q(\psi)^\dagger,$$

which is clearly true from the basis representation of $q(\psi)$. $\square$

The equations now make sense. We write $\mathcal{C}_\sigma(X)$ for the space of configurations, which is just the space of pairs (ψ, A) of sections of W^+ and $U(1)$ connections on $\det(\sigma)$. We might begin by taking C^∞ sections, but as we will need the standard PDE package for elliptic equations, we work instead with Sobolev spaces and require L^2_2 sections and connections for configurations.

2.2 Changes of gauge

The configuration space contains the enormous symmetry of the gauge group, which as mentioned above is

$$\mathcal{G} = \Omega^0(X, S^1).$$

We can write explicitly the formula for a gauge change $\gamma : X \to S^1$:

$$\gamma \cdot (\psi, A) = (\gamma\psi, A - 2(d\gamma)\gamma^{-1}).$$

Note an extra factor of 2 from the squaring involved in defining $\det(\sigma)$.

The set of solutions of 2.1 is invariant under this action: F_A^+ doesn't change and $q(e^{i\theta}\psi) = |e^{i\theta}|^2 q(\psi) = q(\psi)$. Also note that in order to keep the configurations in L^2_2, we must use gauge changes γ in L^2_3. The action is clearly free when $\psi \neq 0$. When $\psi \equiv 0$, the stabilizer $\mathrm{Stab}(\psi, A)$ is the set of maps γ with $-2(d\gamma)\gamma^{-1} = 0$; that is, $d\gamma = 0$ and $\mathrm{Stab}(\psi, \mathrm{A}) \cong S^1$ is the constant maps. Configurations (ψ, A) are called *reducible* if $\psi \equiv 0$; otherwise they are *irreducible.*

We intend to remove the large symmetry of $\mathcal{G}$ from the space $\mathcal{C}_\sigma$ and eventually the set $\mathcal{Z}_\sigma$ of solutions of 2.1. Thus we would like to take the quotients

$$\mathcal{M}_\sigma := \mathcal{Z}_\sigma/\mathcal{G} \subset \mathcal{C}_\sigma/\mathcal{G} =: \mathcal{B}_\sigma.$$

To define the SW invariant we need to show that these quotients make sense; first of all they should be Hausdorff topological spaces, and ultimately we would like smooth manifolds. A genericity argument can be used on the perturbation η to produce smooth manifolds for $\mathcal{M}_\sigma$; but we need to know that $\mathcal{B}_\sigma$ is a smooth (Hilbert) manifold in the first place.

In the finite dimensional realm, quotients of manifolds by Lie group actions yield smooth manifolds when the action is free and proper. This partly carries over in that $\mathcal{B}_\sigma$ will be locally a smooth manifold away from reducible configurations where the action isn't free, and reducible configurations may become singularities in the quotient. A convenient framework to deal with this problem is given by artificially forcing the action to be free by introducing a smaller gauge group:

Definition 2.3. Fix some point $* \in X$. The *reduced* gauge group is the set

$$\mathcal{G}^0 = \{\gamma \in \mathcal{G} \mid \gamma(*) = 1\}.$$

Then the *unreduced* quotient space is the quotient

$$\mathcal{B}^0 = \mathcal{C}_\sigma / \mathcal{G}^0$$

by the smaller *reduced* gauge group. Given the moduli space of solutions $\mathcal{M}_\sigma$, the *unreduced* moduli space $\mathcal{M}^0_\sigma$ is defined similarly.

The group of constant maps to S^1 remains a symmetry, and gives a principle S^1 bundle $\mathcal{M}^0_\sigma \to \mathcal{M}_\sigma$ whenever $\mathcal{M}_\sigma$ is a smooth manifold. The line bundle associated by the tautological representation of S^1 on $\mathbb{C}$ is called the *Seiberg-Witten line bundle*. The first Chern class of this bundle will be denoted μ.

The action of $\mathcal{G}^0$ is now free, and one can prove that $\mathcal{B}^0$ is always a smooth (Hilbert) manifold, and furthermore by the standard argument (see below), for generic perturbation η the unreduced moduli space $\mathcal{M}^0_\sigma$ will necessarily be a smooth (finite dimensional) manifold as well. In the end we will in fact want to deal with the fully reduced $\mathcal{B}_\sigma$ and $\mathcal{M}_\sigma$, and so one shows that the complement of the set of orbits of reducible configurations, written $\mathcal{B}^*_\sigma$, is a Hilbert manifold, and $\mathcal{M}_\sigma$ is a smooth finite dimensional manifold in a neighborhood of any irreducible solution. (The set of such is called $\mathcal{M}^*_\sigma$.)

To prove the facts about $\mathcal{B}_\sigma$ one constructs local "slice" coordinates for the action of $\mathcal{G}_\sigma$ as follows (more details can be found in [9, p. 64] or [11, Prop. 2.2.7]). The tangent space $T\mathcal{G}$ at $\mathbf{1}$ is the set of functions $L^2_m(X, i\mathbb{R})$ with the exponential map literally the exponential $if \mapsto e^{if}$. Write

$$\mathcal{L}_{(\psi,A)} : if \mapsto (\gamma\psi, -2idf)$$

for the differential of the action at the point (ψ, A). (So $\ker \mathcal{L}_{(\psi,A)}$ is 0 or $i\mathbb{R}$ if (ψ, A) is irreducible or reducible, respectively.)

The *linear slice* $\mathcal{S}_{(\psi,A)}$ at a point is the set of tangent vectors $(\dot{\psi}, i\dot{a}) \in T\mathcal{C}_\sigma$ orthogonal to the action of $\mathcal{L}_{(\psi,A)}$, that is, such that

$$\left\langle \mathcal{L}_{(\psi,A)}(if), (\dot{\psi}, i\dot{a}) \right\rangle = 0$$

for all $if \in T\mathcal{G}$. This is (by non-degeneracy) the same as

$$\mathcal{S}_{(\psi,A)} = \ker \mathcal{L}^*_{(\psi,A)}$$

for the formal adjoint

$$\mathcal{L}^*_{(\psi,A)} : T\mathcal{C}_\sigma \to T\mathcal{G}.$$

Plugging in the definition and rearranging we find that

$$\mathcal{L}^*_{(\psi,A)} : (\dot\psi, i\dot a) \mapsto -2id^*\dot a - i\,\mathrm{Im}\langle\psi,\dot\psi\rangle_{L^2_m}. \tag{2.3}$$

Intuitively the linear slice is the set of tangent vectors pointing "across orbits" of the gauge group action. Then "slice coordinates" centered at (ψ, A) are pairs (s,γ) for $s \in \mathcal{S}_{(\psi,A)}$ and $\gamma \in \mathcal{G}$; the first moves across orbits and the second along orbits. So to show $\mathcal{B}^*_\sigma$ is a manifold one proves:

Proposition 2.4. *For any irreducible point (ψ, A) there is a neighborhood $0 \in U \subset \mathcal{S}_{(\psi,A)}$ and diffeomorphism*

$$\mathcal{F} : U \times \mathcal{G} \to V,$$

where $V \subset \mathcal{C}_\sigma$ is an (irreducible) neighborhood of the orbit of (ψ, A).

The proof uses a trivial "exponential map"

$$\mathcal{F} : \Big((\dot\psi, i\dot a),\gamma\Big) \mapsto \gamma\cdot\Big((\psi,A) + (\dot\psi, i\dot a)\Big) = \Big(\gamma(\psi+\dot\psi), A + i\dot a - 2d\gamma/\gamma\Big).$$

One shows that the differential of this map is an isomorphism, applies the inverse function theorem, and argues that the result can be extended from a small neighborhood in $\mathcal{G}$ to all of $\mathcal{G}$, giving a map to the neighborhood of the whole orbit. (One can also prove a similar statement around reducible points by consistently removing the stabilizer.) We will have more to say about $\mathcal{M}^*_\sigma$ in the next section.

We should say a word about the topologies involved. On $\mathcal{C}_\sigma$ we start with the topology induced by the norm

$$\|(\psi_1, A_1) - (\psi_2, A_2)\|^2 = \int_X \Big(|\psi_1-\psi_2|^2 + |A_1 - A_2|^2\Big)\, dv_g.$$

(The analogous norms $\|\cdot\|_{L^2_k}$ are defined for Sobolev spaces.) Then the natural topology on the quotient is determined by the metric

$$\delta([C_1],[C_2]) = \inf_{p,q} \|\gamma_p\cdot C_1 - \gamma_q\cdot C_2\|,$$

where γ_p, γ_q range over $\mathcal{G}$ for $\mathcal{B}_\sigma$ and $\mathcal{G}^0$ for $\mathcal{B}^0_\sigma$. The moduli space inherits the subspace topology from $\mathcal{B}_\sigma$. The Lie group $\mathcal{G}$ is not compact, so we don't know *a priori* that the quotient will even be Hausdorff (though this does follow from the slice charts above). But it is not exceedingly difficult to show even more:

Proposition 2.5. *Using any of the Sobolev norms $\|\cdot\|_{L^2_k}$ with $k \geq 2$, the pair $(\mathcal{B}_\sigma, \delta)$ is a metric* *space.*

Proof. See [11, Prop. 2.2.2]. □

Then $\mathcal{M}_\sigma$ inherits the structure of a metric space as well. In the next section we explain why this metric space is always compact.

2.3 Moduli space

In this section we prove that the moduli space is always compact, and sketch why and when it is a smooth manifold.

2.3.1 Curvature and compactness

Perhaps the most significant fact about Seiberg-Witten theory is the following:

Proposition 2.6. *For a fixed perturbation* η*, the metric space* $(\mathcal{M}_\sigma(g,\eta),\delta)$ *is compact.*

Before we prove this it is useful to note two things. First, one can show that any solution to the Seiberg-Witten equations is gauge equivalent to a smooth solution. The proof of this is not difficult; see [11, Prop. 2.1.11] for full details. One way to think about it ([9, p. 77]) is that the smooth configurations are dense in the L^2_2 configurations, so the orbit of an L^2_2 configuration must pass through a slice which itself passes through a smooth solution. Then the SW equations can be written with smooth coefficients, and since they are elliptic (after gauge fixing), all solutions in such a slice are themselves smooth.

The second note is that this proposition can be proved with the same reasoning for L^p_k sections and connections for $k \geq 2$, but for clarity we only consider the case $k = 1$.

Proof of Prop. 2.6. We present the proof in three steps. The first step is to apply the Lichnerowicz formula to obtain absolute upper bounds on $|\psi|^2$ and $|F_A^+|$. This may well be the most important calculation in the theory. The second step is to fix representative gauges for the orbits in $\mathcal{M}_\sigma$ in order to find the relevant estimates; we will fix the Coulomb gauge. The last step is to find those estimates; namely to use the ellipticity of the gauge-fixed equations to obtain bounds on higher derivatives. We follow [11, Prop. 2.2.2] rather closely. See [9, §5.3] for an alternative.

Step 1.

Lemma 2.7. *Let* $C = (\psi, A)$ *be a solution to the Seiberg-Witten equations, and by gauge invariance of this result we may assume without loss of generality that* C *is smooth. Let* $\kappa = \min s(x)$ *be the smallest scalar curvature of the compact* X*. We then have the absolute bound:*

$$\|\psi\|^2_\infty \leq \max(0, -2\kappa + 4\|\rho(\eta^+)\|_\infty). \tag{2.4}$$

Proof. Recall the Kato inequality of Prop. 1.13, which gives the pointwise bound

$$\Delta_X |\psi|^2(x) \leq 2\,\mathrm{Re}\langle \nabla_A^* \nabla_A \psi(x), \psi(x)\rangle. \tag{2.5}$$

Since we assume ψ is a smooth solution, we also know that at any point x_0 where $\psi(x)$ is maximized,

$$0 \leq \Delta_X |\psi|^2(x_0) \text{ and } D_A\psi(x_0) = 0.$$

The Lichnerowicz formula 1.14 for the Weitzenböck presentation of the Dirac operator then gives

$$\nabla_A^* \nabla_A \psi = -\frac{1}{4} s\psi - \frac{1}{2}\rho(F_A^+)\psi,$$

and plugging this and the inequality into 2.5 yields

$$0 \leq -\frac{1}{2} s(x_0)|\psi(x_0)|^2 - \langle \rho(F_A^+)\psi(x_0), \psi(x_0)\rangle.$$

The second of the Seiberg-Witten equations lets us write $\rho(F_A^+) = \frac{1}{2}q(\psi) - \rho(i\eta^+)$, and by the definition of $q(\psi)$ we have

$$q(\psi)\psi = \langle \psi, \psi\rangle\psi - \frac{1}{2}|\psi|^2\psi$$

so

$$\langle \rho(F_A^+)\psi, \psi\rangle = \frac{1}{4}|\psi|^4 - \langle i\rho(\eta^+)\psi, \psi\rangle$$

and at x_0

$$\begin{aligned} 0 &\leq -\frac{1}{2}s|\psi|^2 - \frac{1}{4}|\psi|^4 + \langle i\rho(\eta^+)\psi, \psi\rangle \\ &\leq -\frac{1}{2}s|\psi|^2 - \frac{1}{4}|\psi|^4 + \|\rho(\eta^+)\|_\infty |\psi|^2. \end{aligned}$$

If $\psi(x_0) = 0$ the proposition is proved. If $\psi(x_0) \neq 0$ we can divide by $\frac{1}{4}|\psi(x_0)|^2$ and rearrange to find:

$$|\psi(x_0)|^2 \leq -2s(x_0) + 4\|\rho(\eta^+)\|_\infty. \tag{2.6}$$

The result follows.

∎

A bound on $|F_A^+|$ follows from this and the second SW equation.

Step 2.

To use these bounds to obtain compactness we need to fix a gauge. Let us start with a sequence

$C_n = (\psi_n, A_n)$ of L^2_2 solutions, and a fixed connection A_0. We can write the Hodge decomposition of the difference between A_n and A_0 :

$$A_n = A_0 + i\,(h_n + 2d\alpha_n + d^*\beta_n) \tag{2.7}$$

with real forms, h_n harmonic, and the 2 for convenience. Consider next the integral lattice $\mathcal{H}^1(X;\mathbb{Z}) \subset \mathcal{H}^1(X,g)$ in the space of harmonic forms. Let $\chi_n \in 4\pi\mathcal{H}^1(X;\mathbb{Z})$ be a 4π lattice point minimizing the distance to h_n. Note that this distance is bounded by the lattice size, that is,

$$\|\chi_n - h_n\|_{L^2} \leq N_g$$

for some $N_g \in \mathbb{R}_+$ depending on the metric.

We then define our change of gauge as follows. Fix $x_0 \in X$, and for any path Γ from x_0 to $x \in X$ set

$$f^n_\Gamma(x) = \int_\Gamma \chi_n.$$

For two such paths Γ_1, Γ_2, since χ_n is in $4\pi\mathcal{H}^1(X;\mathbb{Z})$ we have

$$f^n_{\Gamma_1}(x) - f^n_{\Gamma_2}(x) = \oint \chi_n \in 4\pi\mathbb{Z}.$$

Thus there is a well-defined function

$$\gamma_n := \exp(\frac{i}{2}f_n) : X \to S^1$$

where any path Γ can be used to compute γ_n. We also see that $d\gamma_n = \frac{i}{2}\chi_n\gamma_n$. Our change of gauge is then given by the composition

$$\begin{aligned}
C'_n &= \exp(i\alpha_n)\cdot\gamma_n\cdot C_n \\
&= (\psi'_n, A_0 + i(h_n + 2d\alpha_n + d^*\beta_n) - 2d\gamma_n/\gamma_n - 2d\alpha_n) \\
&= (\psi'_n, A_0 + i(h_n - \chi_n) + id^*\beta_n),
\end{aligned}$$

and if $A'_n = A_0 + a'_n$ we have $d^*a'_n = 0$ as well as a bound on the harmonic part $i(h_n - \chi_n)$, written $\|\Pi(a'_n)\| \leq N_g$ for the harmonic projection Π. From here on we use C'_n and remove the primes.

We can include these equations to obtain the Coulomb gauge Seiberg-Witten equations on our sequence C_n as follows. By writing the Dirac operator in coordinates, for a change of connection $A_n = A_0 + ia_n$ one can check the formula

$$D_{A_n} = D_{A_0} + \frac{1}{2}\rho(ia_n).$$

Implicitly assuming the correspondence ρ, the second equation becomes

$$\begin{aligned} F^+_{A_n} + i\eta^+ &= \frac{1}{2}q(\psi_n) \\ F^+_{A_0} + id^+a_n + i\eta^+ &= \frac{1}{2}q(\psi_n) \\ i(d^+ + d^*)a_n &= \frac{1}{2}q(\psi_n) - i\eta^+ - F^+_{A_0} \end{aligned}$$

where in the second step we used the general fact that $F_A = dA$ and in the third line we added $d^*a_n = 0$. So we end up with the Coulomb gauge Seiberg-Witten equations:

$$CGSW = \begin{cases} D_{A_0}\psi_n = -\frac{1}{2}\rho(ia_n)\psi_n \\ i(d^+ + d^*)a_n = \frac{1}{2}q(\psi_n) - i\eta^+ - F^+_{A_0} \\ \|\Pi a_n\|_{L^2} \leq N_g. \end{cases} \tag{2.8}$$

<u>Step 3.</u>

In this last step we summarize how to find L^p_2 bounds on ψ_n and a_n. This implies a subsequence weakly convergent in the Hilbert space L^2_2, and the compact Sobolev embedding $L^p_2 \subset\subset L^q_1$ for appropriate q gives a strongly convergent subsequence in L^q_1. One can then use a standard bootstrapping process to get convergence to a C^∞ solution of the Seiberg-Witten equations.

We need to note three facts at the outset. First, under the map q (assuming ρ^{-1}) we lose at most one derivative of regularity:

$$q : L^2_k(W^+) \to L^2_{k-1}(i\Lambda^2_+ T^*X).$$

This is not trivial but not difficult; details can be found in [11, Lem. 2.1.6]. Second, the operators D_{A_0} and $(d^+ + d^*)$ are elliptic. For any elliptic operator $D : L^2(E_0) \to L^2(E_1)$ with projection $P : L^2(E_0) \to \ker D$ we know two important facts:

(1) If $u \in L^p, v \in L^p_m$ and $Du = v$, then $u \in L^p_{m+k}$ where k is the order of D.

(2) If $u \in L^p_{m+k}$, then there is a constant C such that $\|u - Pu\|_{L^p_{m+k}} \leq C\|Du\|_{L^p_m}$.

Third, $\ker(d^+ + d^*) = \mathcal{H}^1(X, g)$. This is not hard to show. Given the Hodge decomposition for any candidate form $\omega = h + d\alpha + d^*\beta$, the standard argument for $d + d^*$ works except that we must check that $(dd^*\beta)^+ = 0$ implies $d^*\beta = 0$. For this, just write $(dd^*\beta)^+ = \frac{1}{2}(1+*)(dd^*\beta)$, pair with β, and use the fact that (up to sign) $*$ is self-adjoint (and we can assume β is closed).

The rest of the argument runs as follows. Using the right side of $CGSW(2)$ we have an L_0^∞ bound on $(d^+ + d^*)a_n$. Assuming X is compact, this gives L_0^p bounds for all p. Then using point (2) on elliptic operators (here $k = 1$) we get an L_1^p bound on $a_n - \Pi a_n$. Now since $\mathcal{H}^1$ has finite dimension, we don't distinguish its Sobolev spaces and $CGSW(3)$ gives by the triangle inequality with the previous statement an L_1^p bound on a_n for all p.

Next we have by Sobolev multiplication and our L_0^∞ bound on ψ_n and this L_1^p bound on a_n an L_0^∞ bound on $CGSW(1) = -\frac{1}{2}\rho(ia_n)\psi_n$; using the ellipticity of D_{A_0} gives an L_1^p bound on ψ_n, which plugging again into $CGSW(1)$ and applying ellipticity gives an L_2^p bound on ψ_n for all p.

Losing one derivative as mentioned, $CGSW(2)$ now gives an L_1^p bound on $(d^+ + d^*)a_n$, so repeating the argument above with Πa_n and ellipticity gives the desired L_2^p bound on a_n.

□

2.3.2 Generic smoothness and dimension of the moduli space

In this section we outline the method used to show that $\mathcal{M}_\sigma$ is generically a smooth manifold. The only complicating factor in this process is the possibility of reducible solutions, at which the quotient $\mathcal{B}_\sigma$ is itself not even smooth. There is a simple way to understand when reducible solutions can occur based on the harmonic part of the perturbation $i\eta^+$ which will be discussed in §2.4.2; this must be considered to define the invariant we are after, but for now we treat only the irreducible part $\mathcal{M}_\sigma^*$.

The method is the following. We introduce the irreducible *parametrized moduli space*

$$\tilde{\mathcal{M}}_\sigma^* = \bigcup_{\eta^+ \in \Omega^+} \mathcal{M}_{\sigma,\eta^+}^* \times \{\eta^+\},$$

which carries each moduli space tied to its perturbation value.

The strategy is to show that this larger space is a smooth manifold, then to apply the Sard-Smale theorem (the version of Sard's theorem appropriate to Hilbert manifolds) to the projection

$$\pi : \tilde{\mathcal{M}}_\sigma^* \to \Omega^+,$$

for which each preimage is a moduli space:

$$\mathcal{M}_{\sigma,\eta^+}^* \cong \pi^{-1}(\eta^+).$$

To show smoothness of $\tilde{\mathcal{M}}_\sigma^*$ we think of it as the zero locus of the (parametrized) Seiberg-Witten functional

$$\mathcal{F} : \mathcal{C}_\sigma \times \Omega^+ \to L^2_3 \left(W^- \oplus (\Lambda^2_+ T^* X \otimes i\mathbb{R}) \right)$$

given by

$$\mathcal{F}(\psi, A, \eta^+) = \left(D_A \psi, \rho(F_A^+ + i\eta^+) - \frac{1}{2} q(\psi) \right),$$

modulo the group of changes of gauge. (Note that because the equations are invariant, one sometimes switches between removing $\mathcal{G}$ before or after applying $\mathcal{F}$.)

We then establish:

Lemma 2.8. *At an irreducible zero of $\mathcal{F}$, the differential $D\mathcal{F}$ is surjective.*

Proof. Computing the differential is straightforward:

Let $(\dot\psi, i\dot a, i\delta^+)$ give a direction; then evaluate the derivative:

$$D\mathcal{F}(\dot\psi, i\dot a, i\delta^+) = \frac{d}{dt} \left(\mathcal{F}(\psi + t\dot\psi, A + ti\dot a, i\eta^+ + ti\delta^+) \right) \Big|_{t=0}.$$

Using the formula $D_{A+ia} = D_A + \frac{1}{2}\rho(ia)$ mentioned above, one finds for the first factor of the image:

$$\begin{aligned} \frac{d}{dt}\ D_{A+ti\dot a}(\psi + t\dot\psi)\Big|_{t=0} &= \frac{d}{dt} \left(D_A\psi + tD_A\dot\psi + \frac{t}{2}\rho(i\dot a)\psi + \frac{t^2}{2}\rho(i\dot a)\dot\psi \right) \Bigg|_{t=0} \\ &= D_A\dot\psi + \frac{1}{2}\rho(i\dot a)\psi. \end{aligned}$$

Then since $F^+_{A+i\dot a} = F^+_A + d^+(i\dot a)$, noting the linearity of ρ, a similar calculation gives for the second factor of the image:

$$\rho(d^+(i\dot a) + i\delta^+) - \frac{1}{2}\dot q(\psi, \dot\psi).$$

And we can give an expression for the derivative $\dot q$:

$$\begin{aligned} \dot q(\psi, \dot\psi) &= \frac{d}{dt}\ q(\psi + t\dot\psi)\Big|_{t=0} \\ &= \frac{d}{dt}\ (\psi + t\dot\psi) \otimes (\psi + t\dot\psi)^* - \frac{1}{2}\|\psi + t\dot\psi\|^2 \mathbf{1} \Big|_{t=0}; \end{aligned}$$

using the polarization identity:

$$\|\psi + t\dot\psi\|^2 = 2\,\mathrm{Re}\langle\psi, t\dot\psi\rangle + \langle\psi,\psi\rangle + \langle t\dot\psi, t\dot\psi\rangle,$$

so

$$\dot{q}(\psi, \dot{\psi}) = \psi \otimes \dot{\psi}^* + \dot{\psi} \otimes \psi^* - \mathrm{Re}\langle \psi, \dot{\psi} \rangle \mathbf{1}.$$

The differential is:

$$D\mathcal{F}(\dot{\psi}, i\dot{a}, i\delta^+) = \begin{pmatrix} D_A\dot{\psi} + \frac{1}{2}\rho(i\dot{a})\psi \\ \rho(d^+(i\dot{a})) + \rho(i\delta^+) - \frac{1}{2}\dot{q}(\psi, \dot{\psi}) \end{pmatrix}. \tag{2.9}$$

To show this is surjective we note that by varying $i\delta^+$ one can cover the second factor without further thought. For the first factor, assume $\phi \neq 0$ is orthogonal to $D_A\dot{\psi} + \frac{1}{2}\rho(i\dot{a})\psi$ for all $(\dot{\psi}, i\dot{a})$. By letting $i\dot{a} = 0$ and varying $\dot{\psi}$ we see that $D_A^*\phi = 0$ so

$$\int_X \left\langle \frac{1}{2}\rho(i\dot{a})\psi, \phi \right\rangle = 0 \qquad \forall\, \dot{a}.$$

But it is easy to construct a form $\dot{a}$ with $\frac{1}{2}\rho(i\dot{a})\psi = \phi$ in a small neighborhood of a point x_0 on which both ϕ and ψ are bounded away from zero, and which is quickly zero outside. (Such a neighborhood exists because they are nontrivial solutions to the Dirac and adjoint-Dirac equations, so they cannot vanish on any open set.) This is a contradiction. □

Since this differential is surjective, the implicit function theorem establishes that $\tilde{\mathcal{M}}_\sigma^*$ is a smooth manifold. We then consider the projection to Ω^+. This requires the basic Fredholm idea and the theorem of Sard-Smale; proofs can be found in many places. We are indebted here to the discussion in [11, pp. 110-12]:

Definition 2.9. A smooth map F between Hilbert manifolds is *Fredholm* if its differential is a Fredholm operator. For connected domain, the index of this operator is constant and denoted $\mathrm{ind}(F)$.

Theorem 2.10 (Sard-Smale)**.** *Let $F : M \to N$ be a Fredholm map with M connected and M and N Hilbert manifolds. Then generic $n \in N$ are regular values at which $F^{-1}(n)$ is a smooth manifold of dimension* $\mathrm{ind}(F)$ *(empty for* $\mathrm{ind}(F) < 0$*).*

To complete our argument, then, we need to show that the projection is Fredholm; its index will also give us the dimension of $\mathcal{M}_\sigma^*$. To do that we cite one more very general lemma, useful in many similar situations:

Lemma 2.11. *Let $F : \Lambda \times X \to Y$ be a smooth map between Hilbert manifolds (Λ and X connected) with regular value y_0. Suppose $F_\lambda : x \mapsto F(\lambda, x)$ is Fredholm at each λ. Then the projection $\pi : F^{-1}(y_0) \to \Lambda$ is Fredholm with the same index.*

Putting these together, we conclude with the following result in our case:

Proposition 2.12. *The Seiberg-Witten differential $D\mathcal{F}_{\eta^+}$ is Fredholm for each η^+, and its index is given by*

$$d(\sigma) = \frac{1}{4}\left(c_1(\sigma)^2 - (2\chi + 3\tau)\right), \tag{2.10}$$

where $\tau = b_2^+ - b_2^-$ is the signature and χ the Euler characteristic of X.

Proof idea. We do not have space for a full proof; see the treatments in [9] or [11]. The main idea is that we add to the differential 2.9 (with $\delta = 0$ for evaluation at η^+) another row for the Coulomb condition 2.3, and then remove the zeroth order term

$$\begin{pmatrix} \frac{1}{2}\rho(i\dot{a})\psi \\ -\frac{1}{2}\dot{q}(\psi, \dot{\psi}) \\ -i\,\mathrm{Im}\langle\psi, \dot{\psi}\rangle \end{pmatrix}$$

by a homotopy (multiply by $-t$ for $t \in [0,1]$ and add) which preserves the Fredholm property and index. The remaining operator is

$$\begin{pmatrix} D_A \\ \rho(d^+) \\ -2d^* \end{pmatrix}$$

and the formula comes from the Atiyah-Singer index theorem. □

Remark 2.13. We are somewhat hiding the gauge group here; to deal with it explicitly one constructs an *elliptic complex* called the *deformation complex* which incorporates the linearized gauge changes and linearized Seiberg-Witten functional on configuration space, letting one "linearize" in a gauge invariant fashion (that is, on the quotient space). One shows first that the linearized SW equations (with fixed gauge) are Fredholm because they are elliptic, then that the deformation complex is a Fredholm complex, giving the appropriate relations on the dimensions of domains and codomains.

Putting this into the above lemma and applying the Sard-Smale theorem shows that for generic η^+, the preimage

$$\mathcal{M}^*_{\sigma,\eta^+} = \pi^{-1}(\eta^+)$$

is a smooth manifold of dimension $d(\sigma)$.

2.4 Invariant

We are now in a position to define the Seiberg-Witten invariant. In the first section of this part we give the definition and explain why the invariant is independent of metric and perturbation. We have mentioned the issue of reducible solutions; this will be taken up in §2.4.2 where we describe the context under which all solutions are irreducible; until then we assume they are. We also define a very simple version of the invariant which will be sufficient for the proof of the Thom Conjecture in the final part of this essay.

2.4.1 Definition and basic properties

Let us be given any closed Riemannian 4-manifold (X, g), and fix a $Spin^c$ structure $\sigma : P \to X$. Then for generic η^+ the irreducible part $\mathcal{M}^*_\sigma$ of the moduli space is a smooth manifold of dimension $d(\sigma)$. We assume for now that there are no reducible solutions, so $\mathcal{M}_\sigma \subset \mathcal{B}^*_\sigma$. We shall see that for this to be true generically it is enough that $b_2^+ > 0$, and that if furthermore $b_2^+ > 1$ there are no reducibles for a generic *path* of metrics.

Recall the 2-form μ defined above as the Chern class of the Seiberg-Witten line bundle arising from the unreduced moduli space $\mathcal{M}^0_\sigma$. We define the Seiberg-Witten invariant to be

$$SW_\sigma(g, \eta^+) = \int_{\mathcal{M}_\sigma(g,\eta^+)} \mu^{d(\sigma)/2} \in \mathbb{Z} \tag{2.11}$$

when $d(\sigma)$ is positive even, and $SW_\sigma(g, \eta^+) = 0$ when $d(\sigma)$ is odd or negative. One can show (with some work) that an orientation on $H^0(X;\mathbb{R}) \oplus H^1(X;\mathbb{R}) \oplus H^2_+(X;\mathbb{R})$ is sufficient to provide an orientation on $\mathcal{M}_\sigma(g, \eta^+)$. When $d(\sigma) = 0$, the smooth oriented manifold $\mathcal{M}_\sigma(g, \eta^+)$ amounts to a signed collection of separated points, and we define $SW_\sigma(g, \eta^+)$ to be the signed sum in that case.

The task is to show that these quantities are independent of the choice of g and η^+.

Proposition 2.14. *Assume $b_2^+(X) > 1$ so that one may completely avoid reducible solutions. Then $SW(\sigma)$ does not depend on g and η^+.*

Proof idea. See [11, pp. 141ff.]. The goal is to find a path (g_s, η_s^+) joining (g_0, η_0^+) to (g_1, η_1^+) so that each $\mathcal{M}_\sigma(g_s, \eta_s^+)$ is a smooth manifold and furthermore the union

$$\mathcal{M}_\sigma(g_s, \eta_s^+) \times \{s\} \subset \mathcal{B}^*_\sigma \times I$$

is a smooth cobordism between $\mathcal{M}_\sigma(g_0, \eta_0^+)$ and $\mathcal{M}_\sigma(g_1, \eta_1^+)$. This will show that the homology classes $[\mathcal{M}_\sigma(g_0, \eta_0^+)]$ and $[\mathcal{M}_\sigma(g_1, \eta_1^+)]$ within $\mathcal{B}^*_\sigma$ are the same, from which the result follows by Stokes's theorem and the Chern-Weil theorm applied to μ.

To find such a path we start with a smooth path g_s connecting the metrics, and apply the Sard-Smale theorem again to show that generic paths joining η_0^+ and η_1^+ for this path of metrics give smooth cobordisms. Here is the setup in a little more detail: Write $\mathcal{E}$ for the space of *all* paths (in the appropriate L^2_k space; at least $k \geq 4$) of perturbations η_s^+ that agree with η_0^+ and η_1^+ near the end points (and that don't yield any reducible solutions). Then extend the configuration space to

$$\tilde{\mathcal{C}}_\sigma := \mathcal{C}_\sigma \times I,$$

so each path $\tilde{\eta}^+$ gives a functional on $\tilde{\mathcal{C}}_\sigma$:

$$\tilde{\mathcal{F}}_{\tilde{\eta}^+} := \mathcal{F}_{g(s),\eta_s^+}.$$

The potential cobordism is just

$$\tilde{\mathcal{M}}_\sigma = \tilde{\mathcal{F}}_{\tilde{\eta}^+}^{-1}(0)/\mathcal{G}_\sigma,$$

So we take the master space, this time parametrized by *paths* $\tilde{\eta}^+$:

$$\tilde{\mathscr{M}}_\sigma = \bigcup_{\tilde{\eta}^+ \in \mathcal{E}} \tilde{\mathcal{M}}_\sigma \times \{\tilde{\eta}^+\}.$$

Then the actual work of the proof is to show that $D\tilde{\mathcal{F}}_{[\cdot]}$ mapping from $T\tilde{\mathscr{M}}_\sigma$ is surjective at 0, and each $\tilde{\mathcal{F}}_{\tilde{\eta}^+}$ is Fredholm of appropriate index. (The proofs are almost the same as for $D\mathcal{F}$.) Applying Sard-Smale and our lemma about projections then gives the result that generic paths induce smooth cobordisms. □

Remark 2.15. One important fact that we have been tacitly assuming is that we can vary the metric without disturbing the *Spin*^{c} structure σ. This is not possible *a priori* as σ is defined with reference to the metric (in the reduction of the frame bundle to an $SO(n)$ bundle). But fortunately, as mentioned in Remark 1.18, we can find a (unique) equivalent *Spin*^{c} structure for every other metric given σ for g_0; in particular we are assuming that we have chosen a *path* of equivalent *Spin*^{c} structures induced by the metrics g_s.

2.4.2 The wall

We now face directly the problem of reducible solutions. Recall that a reducible solution satisfies $\psi \equiv 0$ and $F_A^+ + i\eta^+ = 0$. It will be convenient to describe when this happens in the context of the vector space of harmonic representatives $\mathcal{H}^2(X;\mathbb{R})$ of $H^2(X;\mathbb{R})$. Denote the projection again by Π. Let us write $c_1(\sigma)$ for $c_1(\det\sigma)$. Recall that

$$c_1(\sigma) = \frac{i}{2\pi}F_A \qquad \text{so} \qquad 2\pi\Pi(c_1(\sigma)) = i\Pi(F_A).$$

We also have the very general fact:

Lemma 2.16. *The harmonic projection* Π *commutes with the projection* $\frac{1}{2}(1+*)$ *to self-dual forms.*

Proof. Write the Hodge decomposition for an arbitrary 2-form:

$$\eta = \gamma + d\alpha + d^*\beta.$$

Then $\Pi(\eta)^+ = \gamma^+$. On the other hand, we can take the self-dual part:

$$\begin{aligned}\eta^+ &= \gamma^+ + \frac{1}{2}(d\alpha + *d\alpha) + \frac{1}{2}(d^*\beta + *d^*\beta) \\ &= \gamma^+ + \frac{1}{2}(d\alpha - *d*(*\alpha)) + \frac{1}{2}(d^*\beta + **d*\beta) \\ &= \gamma^+ + \frac{1}{2}(d\alpha - d^*(*\alpha)) + \frac{1}{2}(d^*\beta + d(*\beta)).\end{aligned}$$

Since γ^+ is itself harmonic, applying the Hodge decomposition theorem again shows that $\Pi(\eta^+) = \gamma^+ = \Pi(\eta)^+$. □

Characterizing the presence of reducibles is now easy. Any reducible solution gives a connection A satisfying

$$2\pi\Pi\left(c_1(\sigma)^+\right) = \Pi(\eta^+). \tag{2.12}$$

Alternatively, starting with a perturbation that satisfies 2.12 we can construct a reducible solution: Since η is closed, its Hodge decomposition is

$$\eta = \Pi(\eta) + d\alpha.$$

Choose a connection A so that $id\alpha = \Pi(F_A) - F_A$. (This is possible: the connections generating $\Pi(F_A)$ and F_A differ by an imaginary 1-form, say ida. Add $i(\alpha - a)$ to A.) Then applying the lemma twice and our assumption 2.12 yields:

$$\Pi(F_A)^+ = \Pi(F_A^+) = -2\pi i\Pi\left(c_1(\sigma)^+\right) = -i\Pi(\eta^+) = -i\Pi(\eta)^+$$

so

$$\Pi(F_A)^+ + i\Pi(\eta)^+ = 0.$$

Plugging in the formulas for $\Pi(F_A)$ and α gives:

$$\Pi(F_A)^+ = (F_A + id\alpha)^+ = F_A^+ + \left(i\eta - i\Pi(\eta)\right)^+$$

so

$$\Pi(F_A)^+ + i\Pi(\eta)^+ = F_A^+ + i\eta^+ = 0.$$

Thus $(0, A)$ is a reducible solution.

We have shown the following:

Proposition 2.17. *There are reducible solutions for (σ, g, η^+) if and only if*

$$2\pi\Pi\left(c_1(\sigma)^+\right) = \Pi(\eta^+).$$

This explains the conditions on b_2^+ mentioned above. If $b_2^+ > 0$, there is “room” in the harmonic part of Ω_2^+ to avoid this equality for generic η^+. However, if $b_2^+ = 1$, it may happen that for a *path* of metrics and corresponding path of $2\pi\Pi\left(c_1(\sigma)^+\right)$ one cannot avoiding hitting the “wall” defined by $\Pi(\eta^+)$. If $b_2^+ > 1$ then we can always find generic paths avoiding this equality, hence Prop. 2.14 for that case.

Once we restrict to the case $b_2^+ = 1$, we can clean up the presentation somewhat by giving a basis ω_g for the line $\mathcal{H}_+^1(X;\mathbb{R})$. Then using the Hodge inner product on 2-forms (which is equivalent to the cup product since ω_g is self-dual), the condition 2.12 is equivalent to the condition

$$\langle \eta, \omega_g \rangle = 2\pi \langle c_1(\sigma), \omega_g \rangle.$$

This condition defines a hyperplane in the space of perturbations, which is called the “wall” separating the two “chambers” of that space associated to a given metric (called the positive chamber when the left side is greater, negative when the right is greater). Suppose we continuously vary the metric and thereby ω_g. As long as $c_1(\sigma) \smile [\omega_{g(s)}] \neq 0$ (in which case we call $g(s)$ *good*) we can find (sufficiently small) perturbations in one chamber to get a smooth manifold, and show that the invariant is independent of the metric in that chamber. But as soon as our path $g(s)$ takes $[\omega_{g(s)}]$ orthogonal to $c_1(\sigma)$, which is to say, as soon as $c_1(\sigma)$ becomes *anti-self-dual,* the argument fails and we must examine the reducible solutions where the moduli space is singular.

We have mentioned that the unreduced moduli space $\mathcal{M}_\sigma^0$ and thereby the unreduced cobordism $\tilde{\mathcal{M}}_\sigma^0$ is always smooth; in this sense the S^1 stabilizer in $\mathcal{G}_\sigma$ at a reducible solution really is the obstruction to the reduced cobordism $\tilde{\mathcal{M}}_\sigma$ being smooth. One finds, then, that a (punctured) neighborhood of a reducible solution $[(0, A)]$ on $\tilde{\mathcal{M}}_\sigma^0$ looks like $\mathbb{C}^{d(\sigma)/2+1} - \{0\}$, and therefore after cutting out the singular point, a neighborhood of $[(0, a)] \in \tilde{\mathcal{M}}_\sigma$ looks like

$$\left(\mathbb{C}^{d(\sigma)/2+1} - \{0\}\right)/S^1 = \mathbb{CP}^{d(\sigma)/2} \times \mathbb{R}_+;$$

that is, a cone on $\mathbb{CP}^{d(\sigma)/2}$.

It is easy to show that only one reducible solution appears from a smooth path that crosses the wall in just one place. So after we cut out the singular reducible point (and labor somewhat to keep track of orientations), we find the boundary of the cobordism is

$$\partial \tilde{\mathcal{M}}_\sigma^0 = \mathcal{M}_\sigma(g_1, \eta_1^+) - \mathcal{M}_\sigma(g_0, \eta_0^+) + \mathbb{CP}^{d(\sigma)/2}.$$

One also knows that μ itself comes from the same S^1 action that we are removing, and can use this to show

$$\int_{\mathbb{CP}^{d(\sigma)/2}} \mu^{d(\sigma)/2} = (-1)^{d(\sigma)/2}.$$

Putting everything together gives the general *wall crossing formula:*

Proposition 2.18. *Let X be a compact, oriented, simply connected smooth 4-manifold with $b_2^+(X) = 1$, and σ a Spin^c structure with $d(\sigma)$ even. Write SW_σ^+ (resp. SW_σ^-) for the SW invariant with a metric g such that a small perturbation is in the positive (resp. negative) chamber. Then*

$$SW_\sigma^+ - SW_\sigma^- = (-1)^{d(\sigma)/2}. \tag{2.13}$$

Proof. The proof follows the general method sketched above, but some details do require attention. It can be found in [9] or [11]; the latter nicely separates the case $d(\sigma) \neq 0$ from $d(\sigma) = 0$. □

The setting for the proof of the Thom Conjecture is the manifold $X = \mathbb{CP}^2 \# n\overline{\mathbb{CP}}^2$. Unfortunately this satisfies $b_2^+(X) = 1$. It is a rather pleasant fact, however, that we will only need a very simple version of the SW invariants, namely the mod 2 count of solutions when $d(\sigma) = 0$. Let us denote this number by $\bar{n}_\sigma(g)$. Since $H_1(X; \mathbb{Z}) = 0$, the proposition just stated immediately yields:

Corollary 2.19. *Let us be given an orientation on $\mathcal{H}_+^2(\mathbb{CP}^2 \# n\overline{\mathbb{CP}}^2; \mathbb{R})$ and assume that any ω_g is positive with respect to this orientation. If g_0 is such that $c_1(\sigma) \smile [\omega_{g_0}] > 0$, and g_1 such that $c_1(\sigma) \smile [\omega_{g_1}] < 0$, then*

$$\bar{n}_\sigma(g_0) \equiv \bar{n}_\sigma(g_1) + 1 \qquad (\textit{mod } 2).$$

Proof. The conditions on g_0 and g_1 put small perturbations in opposite chambers, so this follows from formula 2.13. Note that it is quite possible to prove this simplified relation without the full power of the proposition; a direct proof in this case is given in [6]. □

Corollary 2.20. *If g is a Riemannian metric on $\mathbb{CP}^2 \# n\overline{\mathbb{CP}}^2$ with $c_1(\sigma) \smile [\omega_g] < 0$, then $\mathcal{M}_\sigma(g)$ is not empty for the Spin^c structure σ induced by the almost complex structure on $\mathbb{CP}^2 \# n\overline{\mathbb{CP}}^2$.*

Proof. The main bound of Lemma 2.7 used to prove compactness also immediately shows that if a manifold has positive scalar curvature, then for sufficiently small perturbation, all solutions are reducible. If all solutions are also irreducible, there must not be any solutions. Hitchin shows in [4] that there is a metric of postive scalar curvature on $\mathbb{CP}^2 \# n\overline{\mathbb{CP}}^2$, and this also satisfies $c_1(\sigma) \smile [\omega_g] > 0$. Then simply applying the above corollary gives this result. □

3 The Thom Conjecture

In this final part we seek to explain how Kronheimer and Mrowka in [6] confirmed the Thom Conjecture. There are four key ingredients to the proof. The most important idea, beyond use of the Seiberg-Witten invariants in the first place, is the method of "pulling apart" $\mathbb{CP}^2 \# n\overline{\mathbb{CP}}^2$ along a neighborhood of the embedded curve Σ. This generates an arbitrarily long cylinder with Σ in a cross section, lengthening which is called "stretching the neck." The result follows from three more facts. First, for long enough neck, $\mathcal{M}_\sigma$ is not empty: there are SW solutions. Second, one can extract from these solutions a translation invariant solution on the cylinder that we are stretching. Third, using our absolute bounds on ψ and F_A^+ for this solution, one obtains the desired relation between the genus and homology class of Σ.

Our first task is to set up the geometry. Write H for the Poincaré dual of a copy of $\mathbb{CP}^1$ inside $\mathbb{CP}^2$, which is a sphere and generates $H_2(\mathbb{CP}^2;\mathbb{Z}) \cong H^2(\mathbb{CP}^2;\mathbb{Z}) \cong \mathbb{Z}$. An embedded degree d algebraic curve has class dH, so let us be given an arbitrary smooth representative $\Sigma \hookrightarrow \mathbb{CP}^2$ of that class, and suppose that it has genus g. Now our strategy is to stretch $\mathbb{CP}^2$ in a neighborhood of Σ so that we get a cylinder with Σ living in the cross section, and monopoles on this cylinder will restrict the genus of Σ. But unfortunately the Euler class of the normal bundle to Σ is given by $\Sigma \cdot \Sigma = d^2 H \cdot H = d^2 \neq 0$, so the normal bundle is not trivial and a tubular neighborhood of Σ can't be written as a product for us to perform this stretching.

The situation is remedied by the following trick. We take instead the *blow up* $X = \mathbb{CP}^2 \# d^2 \overline{\mathbb{CP}}^2$ in exactly d^2 points. Each copy of $\overline{\mathbb{CP}}^2$ has its own generator E_i, this time because of the opposite orientation $E_i \cdot E_i = -1$. Together the $\{H, E_i\}$ generate $H^2(X;\mathbb{Z}) = \mathbb{Z}^{1+d^2}$, and it follows that the intersection form (or the wedge-integral form on Poincaré duals) has signature $(1, d^2)$. (Of course this means $b_2^+ = 1$ and we will need to employ the wall crossing formula eventually.) The trick is that we can join $\Sigma \hookrightarrow \mathbb{CP}^2$ by thin tubes to spheres dual to the $-E_i$ producing a surface $\tilde{\Sigma}$ with $[\tilde{\Sigma}] = dH - \sum E_i$, so that (by orthogonality)

$$\tilde{\Sigma} \cdot \tilde{\Sigma} = d^2 H \cdot H + \sum E_i \cdot E_i = d^2 - d^2 = 0$$

and $\tilde{\Sigma}$ has the *same genus* (since we've only attached spheres). Our task is then to show that the genus g of $\tilde{\Sigma}$ is more than $\frac{1}{2}(d-1)(d-2)$. This is already known when $d \leq 3$, so we assume $d > 3$.

Since $\tilde{\Sigma} \cdot \tilde{\Sigma} = 0$ and in particular the Euler class of *its* normal bundle is 0, we can show that the normal bundle must itself be trivial. One first finds that since the dimension of a Euclidean fiber is the same as that of the base, the Euler class is precisely the obstruction to finding a non-vanishing section. The $U(1)$ structure group on the fibers (from the embedding and orientations) gives another section, showing triviality. Then by the tubular neighborhood theorem, there is a neighborhood of $\tilde{\Sigma}$ diffeomorphic to $D^2 \times \tilde{\Sigma}$. We denote its boundary by $Y := S^1 \times \tilde{\Sigma}$, and a neighborhood of this boundary is $[-\epsilon, \epsilon] \times Y$.

Now we can see what it means to "stretch" X away from $\tilde{\Sigma}$. It means to supply metrics g_R on X so that the first component of (the restriction to) this neighborhood product becomes very long, while the second component isn't changed; it will be isometric to $[-R, R] \times Y =: \hat{Y}$ with a product metric and axis length $2R$. Also the metric should not change in X outside $\hat{Y}$, so that the scalar curvature s_R of g_R is bounded uniformly in R. We call the resulting Riemannian manifold

(X_R, g_R). The hypersurface Y divides X into two regions: the tubular neighborhood $D^2 \times \tilde{\Sigma}$ and its exterior. Let X_- denote that part of X within $D^2 \times \tilde{\Sigma}$ by more than ϵ, and X_+ that part outside $D^2 \times \tilde{\Sigma}$ by more than ϵ; so in the end X_R has three regions, glued in order: $X_R = X_- \cup \hat{Y} \cup X_+$. Finally, since in each cross section of the tube $\hat{Y}$, we know that $\tilde{\Sigma}$ is a compact surface with genus more than 1, we can ask that its scalar curvature (which is twice the Gaussian curvature) be a single (negative) constant s there for each g_R. The scalar curvature on the cylinder $\hat{Y}$ with product metric is of course the same value. Then the Gauss-Bonnet theorem tells us that

$$\frac{1}{4\pi} \int_{\tilde{\Sigma}} s = 2 - 2g = \frac{1}{4\pi} s \operatorname{Area}(\tilde{\Sigma}),$$

and rescaling the Y part of g_R to get unit area, we have the formula:

$$s = -4\pi(2g - 2). \tag{3.1}$$

We should mention here that for the *Spin*^{c} structure σ induced by the almost complex structure on X, one can show that $d(\sigma) = 0$ and $c_1(\sigma) = 3H - \sum E_i$.

The last piece of setting up that we shall need is the existence of monopoles on X_R. Now the quadratic form Q on $H^2(X; \mathbb{R})$ with signature $(1, d^2)$ splits $H^2(X; \mathbb{R})$ into a *light cone* where self-intersection is positive; the generator H sets the *forward* direction. The metric g_R defines the linear harmonic subspace $\mathcal{H}^2_+(X; \mathbb{R})$, and we set ω_{g_R} to be the unique (forward) vector in that space satisfying $\omega_{g_R} \cdot H = 1$.

Lemma 3.1. *For $R >> 1$, there are monopoles on (X_R, g_R).*

Proof. We follow [11, p. 171] and [6, Lem. 10]. By Corollary 2.20, it suffices to show that $c_1(\sigma) \smile [\omega_g] < 0$ for $R >> 1$. Assume we have chosen the harmonic representative of H. Then writing $H = h_0\omega_0 + \sum h_i\omega_i$ in an orthonormal basis such that $\omega_{g_R} = x_0\omega_0$ and $x_0 > 0$, the two conditions $H \cdot H = 1$ (so $h_0 \geq 1$) and $\omega_{g_R} \cdot H = x_0 h_0 = 1$ imply $\|\omega_{g_R}\|_{L_2} = x_0 \leq 1$.

Let us identify compact regions on $X_- \cup \big([-R, R] \times Y\big)$ with compact regions on

$$M_- := X_- \cup \big([0, \infty) \times Y\big)$$

by associating $\{-R\} \times Y \longleftrightarrow \{0\} \times Y$. (Going the reverse direction requires letting $2R$ grow past the chosen region.) The above bound gives, for every sequence $\omega_{g_{R_i}}$ with $R_i \to \infty$, a subsequence converging (on compact regions) to an L_2 harmonic form ω_- on M_-. One can show (cf. [1]) that there are no nonzero L_2 harmonic forms on the infinite M_-, so $\omega_- = 0$. Then $\omega_{g_{R_i}}\Big|_{M_-} \to 0$. (Suppose this has a subsequence bounded away from zero. Then it has another subsequence converging away from zero, a contradiction.)

Now we just plug into the formula:

$$c_1(\sigma) \smile [\omega_{g_{R_i}}] = (3H - \sum E_i) \cdot \omega_{g_{R_i}} = \Big(3H - dH + (dH - \sum E_i)\Big) \cdot \omega_{g_{R_i}}$$
$$= \Big((3-d)H + [\tilde{\Sigma}]\Big) \cdot \omega_{g_{R_i}} = 3 - d + \int_{\tilde{\Sigma}} \omega_{g_{R_i}}.$$

For large R_i the integral is small and we are assuming $3 - d < 0$.

□

3.1 Monopoles on tubes: gradient flow

To see why we get translation-invariant solutions on $\mathbb{R} \times Y$ by stretching $\hat{Y}$, we must first see what monopoles on the cylinder $\mathbb{R} \times Y$ look like. We largely follow [11, §2.4.1].

We have seen that a $Spin^c(4)$ structure $\hat{\sigma}$ on $\mathbb{R} \times Y$ induces a $Spin^c(3)$ structure σ on Y. Throughout this section we will adopt the convention that a hat ˆ indicates something on $\mathbb{R} \times Y$, and no hat something on Y. We will use t for the $\mathbb{R}$-coordinate. The image of dt under Clifford multiplication, $\hat{\rho}(dt)$, gives an isomorphism

$$W_3 := \hat{W}^+_\sigma \cong \hat{W}^-_\sigma.$$

This is also naturally isomorphic to the spinor bundle of σ. Also note that the line bundle $\det(\hat{\sigma})$ pulls back from $\det(\sigma)$. We have a new Clifford multiplication on W_3 defined by the composition

$$\rho := \hat{\rho}(dt)\hat{\rho}([\cdot]),$$

which, recalling again Eqn. 1.9, defines a map $\rho : T^*Y \to \mathrm{End}(W_3)$. Together with A on Y, we get an induced Dirac operator also written D_A on Y which is formally self-adjoint.

A connection $\hat{A}$ on $\det(\hat{\sigma})$ is called *temporal* if it has no dt component; so we can write it as the pull-back of

$$A(t) = A_0 + ia_t$$

where A_0 is a connection on $\det(\sigma)$ and a_t is a path of 1-forms on Y, and the i factor puts it in $\mathrm{Lie}(U(1)) = i\mathbb{R}$. We can always change gauge to get a temporal connection: start with an arbitrary connection in coordinates

$$\hat{A} = \hat{A}_0 + i\left(f_t dt + a_t\right),$$

where f_t is just the idt coefficient of the 1-form obtained by subtracting $\hat{A}_0$ and is a path of functions on Y. Choose a change of gauge $\gamma : \mathbb{R} \times Y \to S^1$ defined by $\gamma(t, y) = \exp(ig_t(y))$ with

$$2g_t(y) = \int_{s=-\infty}^{t} f_s(y) ds.$$

Then

$$-2\hat{d}\gamma/\gamma = -2i\hat{d}g(t, y) = -2if_t dt - 2idg_t(y)$$

so the gauge change removes the dt part.

Claim 3.2. *The Seiberg-Witten equations on $\mathbb{R} \times Y$ in temporal gauge (omitting the perturbation) are:*

$$SW_X(\sigma) = \begin{cases} \dfrac{d\psi}{dt} = D_A\psi \\ \dfrac{dA}{dt} = \rho^{-1}(\dfrac{1}{2}q(\psi)) - *F_A. \end{cases} \tag{3.2}$$

Proof of claim. We note again A and F_A without hats live on Y.

Both equations are straightforward computations. We roughly follow [11].

For the first, the full Dirac operator on $\mathbb{R} \times Y$ looks like

$$\hat{D}_{\hat{A}} = \hat{\rho} \circ \hat{\nabla}^{\hat{A}} = \hat{\rho}(dt)\partial_t + \sum \hat{\rho}(e_i)\nabla^A_{e_i}$$

and because the Clifford rule gives $\hat{\rho}(dt)^2 = -1$ and by definition $\rho(e_i) = \hat{\rho}(dt)\hat{\rho}(e_i)$, we can write

$$\hat{D}_{\hat{A}} = \hat{\rho}(dt)\left(\partial_t - \sum \rho(e_i)\nabla^A_{e_i}\right) = \hat{\rho}(dt)\left(\partial_t - D_A\right).$$

The first equation above then follows from the first Seiberg-Witten equation on $\mathbb{R} \times Y$.

To derive the second equation, we need to compute $\hat{F}^+_{\hat{A}} = \hat{d}^+(ia(t))$, where $a(t)$ has no dt part. We will need the very general fact that for any 2-form $\hat{\eta}$, $\hat{\rho}(\hat{*}\hat{\eta}) = \hat{\rho}(\hat{\eta})$, as well as the observation that when α contains no dt, $\hat{*}\alpha = dt \wedge *\alpha$. We write $\dot{a}$ for $a'(t)$. Then:

$$\begin{aligned} 2\hat{F}^+_{\hat{A}} &= 2\hat{d}^+(i\dot{a}) = (1 + \hat{*})\hat{d}(i\dot{a}) \\ &= (1 + \hat{*})(dt \wedge i\dot{a} + da) \\ &= dt \wedge i\dot{a} + da + *i\dot{a} + dt \wedge (*da) \\ &= dt \wedge (i\dot{a} + *da) + *(i\dot{a} + *da) \\ &= dt \wedge (i\dot{a} + *F_A) + *(i\dot{a} + *F_A). \end{aligned}$$

Also

$$\hat{\rho}\left(*(i\dot{a}+*F_A)\right)=\hat{\rho}\left(\hat{*}*(i\dot{a}+*F_A)\right)=\hat{\rho}\left(dt\wedge(i\dot{a}+*F_A)\right),$$

so

$$\hat{\rho}\left(\hat{F}^+_{\hat{A}}\right)=\hat{\rho}\left(dt\wedge(i\dot{a}+*F_A)\right)=\rho\left(i\dot{a}+*F_A\right).$$

The second equation above then follows from the second Seiberg-Witten equation.

□

The wonderful thing about these equations is that they are precisely the gradient flow of the *topological energy functional:*

Claim 3.3. *Equations 3.2 are the gradient flow equations for (half) the functional*

$$\mathcal{E}(\psi,A)=\int(A-A_0)\wedge F_A+\int\langle\psi,D_A\psi\rangle \tag{3.3}$$

where A is a connection on $\det(\sigma)$ and $\psi\in\Gamma(W_3)$, and we take the real part of the inner product on W_3.

Proof of claim. Again this is a computation. By definition a vector $w\in H$ in a Hilbert space satisfies the functional gradient flow equation $\nabla\mathcal{E}=w$ at a point f when, for all $v\in H$,

$$\mathcal{E}'[f](v)=\langle w,v\rangle. \tag{3.4}$$

So we take an arbitrary direction in configuration space $v=(\dot{\psi},i\dot{a})\in\mathcal{C}_3$ and evaluate the left side of 3.4 at a point (ψ,A) :

$$\begin{aligned}\mathcal{E}'[\psi,A](\dot{\psi},i\dot{a}) &= \frac{d}{dt}\mathcal{E}(\psi+t\dot{\psi},A+ti\dot{a})\Big|_{t=0}\\ &= \frac{d}{dt}\int(A+ti\dot{a}-A_0)\wedge F_{A+ti\dot{a}}\Big|_{t=0}+\frac{d}{dt}\int\langle\psi+t\dot{\psi},D_{A+ti\dot{a}}(\psi+t\dot{\psi})\rangle\Big|_{t=0}\\ &= (1)+(2).\end{aligned}$$

We also know $F_{A+ti\dot{a}}=F_A+tid\dot{a}$ and $D_{A+ti\dot{a}}=D_A+\frac{t}{2}\rho(i\dot{a})$. Only terms with one factor of t survive:

$$
\begin{aligned}
(1) &= \int i\dot{a} \wedge F_A + \int (A - A_0) \wedge i d\dot{a} = \int i\dot{a} \wedge F_A + \int i\dot{a} \wedge d(A - A_0) \\
&= 2\int i\dot{a} \wedge F_A = -2\int \langle i\dot{a}, *F_A \rangle dv_g
\end{aligned}
$$

where in the second equality we used Stokes's Law, and for the sign in the last, note that F_A is imaginary and the Hodge dual satisfies $\alpha \wedge *\bar{\beta} = \langle \alpha, \beta \rangle dv_g$.

And the other part:

$$
\begin{aligned}
(2) = &\frac{d}{dt} \int \langle \psi + t\dot{\psi}, D_A\psi + tD_A\dot{\psi} + \frac{t}{2}\rho(i\dot{a})\psi + \frac{t^2}{2}\rho(i\dot{a})\dot{\psi} \rangle \Big|_{t=0} \\
= &\int \langle \dot{\psi}, D_A\psi \rangle + \langle \psi, \frac{1}{2}\rho(i\dot{a})\psi \rangle + \langle \psi, D_A\dot{\psi} \rangle.
\end{aligned}
$$

From the definition of $q(\psi)$ one has for any traceless endomorphism (not just ρ) the identity $\langle \psi, \rho(i\dot{a})\psi \rangle = \langle \rho(i\dot{a}), q(\psi) \rangle$. By the Clifford multiplication rule $|\rho(\alpha)|^2 = 2|\alpha|^2$, so the polarization identity on the inner product gives another general identity:

$$
\langle \rho(\alpha), \rho(\beta) \rangle = 2\langle \alpha, \beta \rangle.
$$

Applying this with $\alpha = i\dot{a}$ and $\beta = \rho^{-1}(q(\psi))$ leads to

$$
\langle \psi, \frac{1}{2}\rho(i\dot{a})\psi \rangle = \langle i\dot{a}, \rho^{-1}(q(\psi)).
$$

Also since D_A is now self-adjoint and we are taking the real part, both D_A terms in (2) are the same. We finally have

$$
\frac{1}{2}\mathcal{E}'[\psi, A](\dot{\psi}, i\dot{a}) = \int \langle \rho^{-1}(q(\psi)) - *F_A, i\dot{a} \rangle + \int \langle D_A\psi, \dot{\psi} \rangle.
$$

By varying one of A or ψ at a time and comparing to 3.4, we see that 3.2 are the gradient flow equations.

□

We need lastly to know how this functional behaves under a change of gauge. Unfortunately it is not gauge invariant, but it is not difficult to see what happens. The integral on the right doesn't change, and the one on the left is changed by the term

$$\int 2\frac{d\gamma}{\gamma} \wedge F_A = 2 \cdot 4\pi^2 \int \frac{d\ln\gamma}{2\pi i} \wedge \frac{i}{2\pi}F_A$$
$$= 8\pi^2 \int_Y \deg\gamma \wedge c_1(\sigma) = 8\pi^2 c_1(\sigma) \smile [\deg\gamma]$$

where $[\deg\gamma] \in H^1(Y;\mathbb{Z})$. So in particular, $\mathcal{E}$ is invariant under a gauge transformation γ if the degree of γ on a cycle dual to $c_1(\sigma)$ is 0.

In our situation each gauge transformation γ is actually a map from X. A cycle ℓ dual to $c_1(\sigma)$ in Y is actually the boundary of the two halves (corresponding to $X_\pm$) of Γ dual to $c_1(\sigma)$ in X. The degree of γ on ℓ is then 0 since γ extends to a surface with boundary ℓ. So we know $\mathcal{E}$ actually *is* invariant under the gauge transformations that are relevant in our particular problem.

3.2 Stretching the neck

We now come to the key argument of the proof. We are trying to find a translation invariant solution on the cylinder $\hat{Y}$. Such a solution is precisely a fixed point of the gradient flow equations: a solution of 3.2 with $\frac{d\psi}{dt} = \frac{dA}{dt} = 0$. (Such an object is called a monopole on Y.) Since 3.2 are the gradient flow of a functional, the functional must decrease monotonically along the cylinder, and the difference in value of the functional at two points on the cylinder measures the "translation variance" of a solution. So our method is first to find an absolute bound (independent of R) on the change $\Delta\mathcal{E}$ along $[-R,R] \times Y$ and then to use that to find a region where $\Delta\mathcal{E}$ is small, and finally to construct a sequence of these region/solution pairs to obtain (by compactness) a solution with $\Delta\mathcal{E} = 0$.

Let $R_i = i \in \mathbb{N}$ so that by Lemma 3.1 there is a solution (ψ_i, A_i) on the stretched manifold X_{R_i}; we can assume A_i is in temporal gauge on the cylinder part. We write

$$\Delta\mathcal{E}(\hat{Y}_{R_i}) = \mathcal{E}(\psi_i(R_i), A_i(R_i)) - \mathcal{E}(\psi_i(-R_i), A_i(-R_i))$$

for the change in $\mathcal{E}$ along the cylinder part where $(\psi_i(R_i), A_i(R_i))$ are the restrictions to slices $\{R_i\} \times Y$.

Now the metric on $[-R_i, R_i] \times Y$ is a product with scalar curvature equal to the scalar curvature s on Y, so that (as we have noted) the scalar curvature s_{R_i} on X_{R_i} is bounded independent of R_i. This puts an absolute bound on the solutions (ψ_i, A_i) themselves (in Coulomb gauge; note that this is compatible with temporal gauge), also implying a bound B on $\Delta\mathcal{E}(\hat{Y}_{R_i})$.

Lastly, by the pigeonhole principle we can find an interval $[k_i, k_i + 1]$ over which $\mathcal{E}$ changes by less than $B/2i$. Placing this solution on $[0,1] \times Y$ for each i, we get a sequence of solutions there for which $\Delta\mathcal{E}$ becomes arbitrarily small, and by the main compactness theorem these converge to a solution with $\Delta\mathcal{E} = 0$, which we extend trivially to $\mathbb{R} \times Y$.

So we have:

Lemma 3.4. *Given that for $R >> 1$ there are monopoles on (X_R, g_R), there is a translation invariant monopole (in temporal gauge) on $\mathbb{R} \times Y$.*

3.3 Bounding the genus

Lemma 3.5. *A translation invariant monopole (in temporal gauge) on $\mathbb{R} \times Y$ implies the bound*

$$|c_1(\sigma)[\tilde{\Sigma}]| \leq 2g - 2. \tag{3.5}$$

Proof. The fundamental curvature bound $|\psi|^2 \leq -2 \min s(x)$ together with Eqn. 3.1 gives

$$|\psi|^2 \leq 16\pi(g-1).$$

By the Clifford rule, $|\rho(*F_A)| = \sqrt{2}|*F_A| = \sqrt{2}|F_A|$, and we know $|q(\psi)|^2 = \frac{1}{2}|\psi|^4$. Combining these with the second SW equation on cylinders yields

$$|F_A| \leq 2\pi(2g-2).$$

The result follows by integrating:

$$\left|c_1(\sigma)[\tilde{\Sigma}]\right| = \left|\frac{i}{2\pi}\int_{\tilde{\Sigma}} F_A\right| \leq \frac{1}{2\pi}\sup|F_A| \leq 2g-2.$$

□

The Thom Conjecture is proved by observing that

$$3d - d^2 = (3H - \sum E_i)(dH - \sum E_i) = c_1(\sigma)[\tilde{\Sigma}] \geq -(2g-2)$$

so

$$g \geq \frac{d^2 - 3d + 2}{2} = \frac{(d-1)(d-2)}{2}.$$

Remark 3.6. It is worth reflecting on the use of Seiberg-Witten theory in this proof. Of course the simplified invariant plays a key role; but its invariant nature is not so significant: it is in fact a Seiberg-Witten monopole that we use, the invariant only finds one. The absolute bounds on ψ and F_A^+ play a crucial role in multiple ways: They are firstly necessary to show the moduli space is compact, as we have seen, and therefore the number of solutions when $d(\sigma) = 0$ must be finite. This is necessary to get $\bar{n}_\sigma \equiv 1$ in our case. (Though conceivably an infinity of solutions could be made to work in our proof, this invariant wouldn't exist.) Compactness is used again in its other sense: the convergent subsequence property is crucial for finding the translation invariant solution on the cylinder from monopoles on the stretched spaces. Understanding the nature of translation invariant solutions itself requires nontrivial facts about the geometric background of Seiberg-Witten theory. Finally, the absolute bounds are used yet again for the crucial relation we sought between the genus and class of the embedded surface.

4 References

[1] M. Atiyah, V. Patodi, I. Singer, *Spectral asymmetry and Riemannian geometry I*, Math. Proc. Cambridge Philos. Soc. **77** (1975), 43-69.

[2] N. Berline, E. Getzler, M. Vergne, *Heat Kernels and Dirac Operators*, Springer Verlag, 1992.

[3] S. Donaldson, *Lecture Notes for course: "Geometric Analysis"* (2008). http://wwwf.imperial.ac.uk/ skdona/GEOMETRICANALYSIS.PDF

[4] N. Hitchen, *On the curvature of rational surfaces*, Diff. Geom., Proc. Sympos. Pure Math., **27**, Amer. Math. Soc. (1975) 65-80.

[5] M. Hutchings and C. Taubes, *An introduction to the Seiberg-Witten equations on symplectic manifolds*, Symp. Geom. & Top. Amer. Math. Soc. (1999) 103-142.

[6] P. Kronheimer and T. Mrowka, *The Genus of Embedded Surfaces in the Projective Plane*, Math. Res. Lett. **1** (1994) 797-808.

[7] P. Kronheimer and T. Mrowka, *Monopoles and Three-Manifolds*, Cambridge University Press, 2007.

[8] H. Lawson, and M. Michelsohn, *Spin Geometry*, Princeton University Press, 1989.

[9] J. Morgan, *The Seiberg-Witten Equations and Applications to the Topology of Smooth 4-Manifolds*, Princeton University Press, 1996.

[10] M. Nakahara, *Geometry, Topology and Physics*, Taylor and Francis, 2003.

[11] L. Nicolaescu, *Notes on Seiberg-Witten Theory*, Graduate Studies in Mathematics **28**, American Mathematical Society, 2000.

[12] J. Figueroa-O'Farrill, *Lecture notes on Gauge Theory* (2006). http://empg.maths.ed.ac.uk/Activities/GT/

[13] Z. Szabó, *Notes on Seiberg-Witten Invariants* in *Geometry and Physics (Lecture Notes in Pure and Applied Mathematics)*, CRC Press 1996.

[14] D. Salamon, *Spin Geometry and Seiberg-Witten Invariants* (1999). https://people.math.ethz.ch/ salamon/PREPRINTS/witsei.pdf

[15] N. Seiberg and E. Witten, *Electric-magnetic duality, monopole condensation, and confinement in $N = 2$ supersymmetric Yang-Mills theory*, Nucl. Phys. B **426** (1994) 19-52.

[16] T. Tao, *What is a gauge?* (2008). https://terrytao.wordpress.com/2008/09/27/what-is-a-gauge/

[17] P. Teichner, *All oriented 4-manifolds have spin-c structures* (1994). http://people.mpim-bonn.mpg.de/teichner/Papers/spin.pdf

[18] V. Varadarajan *Vector Bundles and Connections in Physics and Mathematics: Some Historical Remarks*, Trends Math. Birkhuser, (2003) 502-541.

[19] E. Witten, *Monopoles and four-manifolds*, Math. Res. Lett. **1** (1994) 769-796.