COMPACTNESS RESULTS FOR NECK-STRETCHING LIMITS OF INSTANTONS

DAVID L. DUNCAN

Abstract. We prove that, under a suitable degeneration of the metric, instantons on a cylinder converge to holomorphic curves. This is part of a program for proving the quilted Atiyah-Floer conjecture.

Contents

1. Introduction

In 1988, Atiyah [1] gave a heuristic argument suggesting that the instanton Floer group of a homology 3-sphere S should agree with the Lagrangian intersection Floer group of S. The former Floer group is defined by counting instantons of index one, while the latter counts certain index-one holomorphic curves arising from a handlebody decomposition of S. Atiyah observed that if one chooses a metric on S in which the handlebodies are separated by a very long 'neck', then the instanton equation formally converges to the

holomorphic curve equation as the neck becomes longer. For such a metric, one therefore expects that the number of instantons should agree with the number of holomorphic curves. In particular, the two Floer groups should be isomorphic. The resulting conjecture became known as the *Atiyah-Floer conjecture*.

At the time of writing, it is not clear how the Atiyah-Floer conjecture should be interpreted. The issue is due to the presence of reducible connections. These lead to singularities in the relevant symplectic moduli spaces, and the current technology is not equipped to define the Lagrangian intersection Floer group in the presence of such singularities. (However, see [23, 25, 33] for various approaches to resolving this matter.)

Nevertheless, if one replaces the homology 3-sphere S with a 3-manifold Y *with positive first Betti number*, then it is possible to avoid these reducible connections. Indeed, K. Wehrheim and C. Woodward [39] described how to associate to Y a well-defined Lagrangian intersection Floer group $HF_{\mathrm{symp}}(Y)$; their theory of holomorphic *quilts* was the key ingredient in showing that $HF_{\mathrm{symp}}(Y)$ is independent of auxiliary choices. On the other hand, there is still a well-defined instanton Floer group $HF_{\mathrm{inst}}(Y)$ [16, 4]. Atiyah's heuristic carries over to the positive Betti number setting as well, and so there is a well-posed variant of the Atiyah-Floer conjecture for manifolds Y with $b_1(Y) > 0$. We refer to this variant as the *quilted Atiyah-Floer conjecture*.

Quilted Atiyah-Floer Conjecture. *Suppose Y is a closed, connected, oriented 3-manifold with positive first Betti number. Then there is a natural group isomorphism*

$$HF_{\mathrm{inst}}(Y) \cong HF_{\mathrm{symp}}(Y). \tag{1}$$

Strictly speaking, the construction of each Floer group appearing in (1) depends on the choice of admissible bundle over Y (this is the bundle Q introduced below), so really there is a different conjecture for each admissible bundle isomorphism type. Moreover, it is possible to equip each Floer group with a relative $\mathbb{Z}_4$-grading, but this too depends on a choice. It turns out that the choice of grading is pinned down for each Floer group through the choice of a preferred element of $H^1(Y)$ (this is the cohomology class $[f]$ below). Then there is an extension of the quilted Atiyah-Floer conjecture stating that the isomorphism (1) preserves the relative gradings induced from this choice. See [12] for more details.

In light of Atiyah's heuristic, the heart of the quilted Atiyah-Floer conjecture is in understanding the relationship between instantons and holomorphic curves. In this paper we show that, for a suitably chosen metric g_ϵ on Y, each index-one instanton is close to an index-one holomorphic curve. The parameter $\epsilon > 0$ appearing in the metric g_ϵ is inversely proportional to the length of the 'neck' on Y; see Remark 1.3 (b). Our result establishes

roughly one-half of Atiyah's heuristic for the quilted Atiyah-Floer conjecture, the other half being that each holomorphic curve is close to a unique instanton.

To describe our result in more detail, we suppose from now on that Y is a closed, connected, and oriented 3-manifold with $b_1(Y) > 0$. Fix a preferred class $[f]$ in $H^1(Y) \cong \mathbb{Z}^{b_1(Y)}$ that is primitive. Then $[f]$ can be represented by a Morse function

$$f : Y \longrightarrow S^1$$

with connected, nonempty fibers; see [19, Theorem 1.3]. Now fix a section $\gamma : S^1 \to Y$ of f. This determines a principal SO(3)-bundle $Q \to Y$ by requiring that the Stiefel-Whitney class $w_2(Q) \in H^2(Y, \mathbb{Z}_2)$ is Poincaré dual to the cohomology class $[\gamma] \in H_1(Y, \mathbb{Z}_2)$. Then Q is uniquely determined, up to isomorphism, by this condition. This specific choice of Q is *admissible* in the sense of Braam-Donaldson [4].

We will be interested in connections A on the bundle $\mathbb{R} \times Q$ that are anti-self dual instantons relative to the metric $ds^2 + g_\epsilon$ on $\mathbb{R} \times Y$. We will call these connections *ϵ-ASD*. The metric g_ϵ will be constructed explicitly in Section 2. Its relevant property is that it degenerates, as ϵ goes to zero, to a symmetric 2-form that *vanishes* along the fibers of f; see also (3). Just as in Atiyah's heuristic, this specific degeneration is tailored so that, as ϵ goes to zero, the ϵ-ASD equation formally recovers the holomorphic curve equation for curves mapping into the space of flat connections on the fibers of f. We say a connection is a *holomorphic curve representative* if it satisfies this holomorphic curve equation; see Section 2.4 for a more precise definition.

We express our main result in terms of a compactness theorem. It is similar to Uhlenbeck's theorem, but relative to the degenerating metrics g_ϵ as opposed to a single fixed metric. A precise statement of our main result is given in Theorem 2.3. The statement is rather technical, so we include a less cumbersome version here.

Main Compactness Theorem. *Fix a sequence of positive numbers ϵ_ν converging to zero, and suppose that A_ν is an ϵ_ν-ASD connection on $\mathbb{R} \times Q$. Assume further that the energies of the A_ν are uniformly bounded. Then there is a holomorphic curve representative A_∞ so that, after passing to a subsequence, the A_ν converge (in the sense of Theorem 2.3) to A_∞.*

The use of the term 'compactness' is explained, in part, by the following picture. Let $\widehat{M}_\epsilon$ denote the moduli space of index-one ϵ-ASD connections, and let $\widehat{M}_0$ denote the moduli space of index-one holomorphic curves. The hat indicates that we are quotienting by the natural $\mathbb{R}$-action, so these are expected to be zero-dimensional spaces. Assume, for simplicity, that all moduli spaces are smooth and of the expected dimension; this excludes bubbling phenomena. Then our compactness theorem implies that the union

$$\widehat{M}_{0<\epsilon\leq 1} := \bigcup_{\epsilon\in(0,1]} \widehat{M}_\epsilon$$

over all $0 < \epsilon \leq 1$ has a compactification

$$\overline{\widehat{M}_{0<\epsilon\leq 1}}$$

in terms of holomorphic curves. That is, there is a natural inclusion

$$\overline{\widehat{M}_{0<\epsilon\leq 1}} \subseteq \widehat{M}_0 \bigcup \widehat{M}_{0<\epsilon\leq 1}. \tag{2}$$

Remark 1.1. *Though our theorem only proves that (2) is an inclusion, Atiyah's heuristic suggests that (2) is actually an equality. This is illustrated in Figure 1. Moreover, it is likely that the compactification of $\widehat{M}_{0<\epsilon\leq 1}$ is a 1-manifold with boundary $\widehat{M}_0 \cup \widehat{M}_1$; see Remark 1.3 (d). If one could show this, then the quilted Atiyah-Floer conjecture would follow because the two Floer groups depend only on the oriented cobordism types of their respective moduli spaces $\widehat{M}_\epsilon$ (for any $0 < \epsilon \leq 1$) and $\widehat{M}_0$. We leave a further investigation of this to future work.*

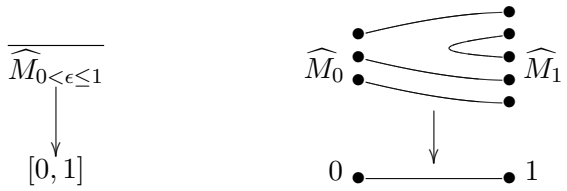

FIGURE 1. The space $\widehat{M}_{0<\epsilon\leq 1}$ has a natural projection to the interval $(0, 1]$. This projection extends to the compactification of each space, and is illustrated on the left above. One expects that the compactification of $\widehat{M}_{0<\epsilon\leq 1}$ is a smooth 1-manifold with boundary $\widehat{M}_0 \cup \widehat{M}_1$. This is illustrated on the right above, and corresponds to (2) being an equality.

Now we discuss the underlying geometry of, and proof strategy for our compactness theorem. We begin with a particularly simple special case that captures much of the spirit of our overall approach.

Example 1.2. *Consider the case where $Y = S^1 \times \Sigma$ is a product, with Σ a surface. Assume f is the projection and γ is the inclusion of a fiber. Then the bundle Q has the form $Q = S^1 \times P$, where $P \to \Sigma$ is the non-trivial* SO(3)*-bundle. The metric g_ϵ should be taken to be of the form*

$$g_\epsilon = dt^2 + \epsilon^2 g_\Sigma, \tag{3}$$

where t is the coordinate variable on S^1, and g_Σ is a fixed metric on Σ.

Consider the quotient space

$$M(P) = \mathcal{A}_{\text{flat}}(P)/\mathcal{G}_0(P)$$

where $\mathcal{A}_{\text{flat}}(P)$ is the space of flat connections on P, and $\mathcal{G}_0(P)$ is the identity component of the gauge group. The non-triviality of the bundle P implies

that $M(P)$ is a smooth symplectic manifold. The metric $\epsilon^2 g_\Sigma$ induces a complex structure J on $M(P)$, and J is independent of ϵ.

Any function $v : \mathbb{R} \times S^1 \to M(P)$ lifts to a map $\alpha_\infty : \mathbb{R} \times S^1 \to \mathcal{A}_{\text{flat}}(P)$, and this lift is unique up to the action of the group $\text{Maps}(\mathbb{R} \times S^1, \mathcal{G}_0(P))$. Each lift α_∞ has a canonical extension to a connection A_∞ on the 4-manifold $\mathbb{R} \times S^1 \times \Sigma$. This extension has the property that, for each $(s,t) \in \mathbb{R} \times S^1$, the restriction

$$A_\infty|_{\{(s,t)\}\times\Sigma} = \alpha_\infty(s,t)$$

recovers α_∞. We call the connection A_∞ a representative *of v. If v is J-holomorphic, then we say that A_∞ is a* holomorphic curve *representative.*

Now suppose A is ϵ-ASD. Notice that the metric in (3) shrinks the volume of the fibers of f to zero. The proof of our compactness theorem will show that, when $\epsilon > 0$ is sufficiently small (and when there is no bubbling), each restriction

$$\alpha(s,t) := A|_{\{(s,t)\}\times\Sigma}$$

is a connection on Σ that has small curvature controlled by ϵ^2; in particular, it is in the stable range when ϵ is small. Motivated by Donaldson [6], our strategy for proving the theorem is to use an analytic Narasimhan-Seshadri correspondence on the surface Σ to map $\alpha(s,t)$ to a nearby flat connection. We denote this nearby flat connection by $\text{NS}(\alpha(s,t))$. This correspondence preserves the equations in the following sense: if A is an instanton, then allowing (s,t) to vary, we obtain a map

$$\mathbb{R} \times S^1 \to \{\textit{flat connections on } \Sigma\}, \quad (s,t) \mapsto \text{NS}(\alpha(s,t))$$

that is holomorphic *relative to the complex structure on the space of all connections on Σ. Then the convergence result for the case where $Y = S^1 \times \Sigma$ essentially follows from the standard Gromov compactness theorem for holomorphic curves.*

The discussion from Example 1.2 carries over, with little significant change, to the case where f has no critical points (i.e., where Y is a mapping torus). However, the more general case, where f has critical points, is significantly more difficult both from a geometric and an analytic perspective. First we describe the basic geometric set-up in the general case.

Just as in the model case where Y is a product $S^1 \times \Sigma$, in the general case we will construct the metrics g_ϵ so they shrink to zero in directions tangent to the regular fibers of f. However, the critical fibers require special treatment. In a neighborhood of a critical fiber, we define the metric by shrinking in all three directions; see Section 2. The point with this particular rescaling of the metric is that we expect the limiting holomorphic curve representative A_∞ to now have Lagrangian boundary conditions coming from the Morse critical points. Indeed, suppose A is ϵ-ASD. When $t \in S^1$ is not near a critical value of f, we can construct the flat connection $\text{NS}(\alpha(s,t))$ as in Example 1.2, and this continues to be holomorphic as (s,t) varies ($s \in \mathbb{R}$ is arbitrary, but t should be bounded away from the critical values of f). On

the other hand, any critical point $p \in Y$ of f determines a boundary $\mathbb{R} \times \{p\}$ for the domain of $\mathrm{NS}(\alpha(s,t))$. The instanton equation for A implies that, for t near the critical value, the value of $\mathrm{NS}(\alpha(s,t))$ is *almost* on the Lagrangian submanifold associated to p. In the small ϵ limit, these 'almost' Lagrangian boundary conditions become exact Lagrangian boundary conditions.

This brings us to the analytic difficulties that arise in the presence of critical points. These difficulties stem from the observation that the standard Gromov compactness theorem we used in Example 1.2 breaks down when one does not have Lagrangian boundary conditions on the nose. We get around this using the following two ingredients. First, we establish several $\mathcal{C}^1$-estimates for the map NS (see Proposition 3.2), and then $W^{2,2}$-estimates for instantons (see Theorem 4.1). These allow us to control the behavior of the holomorphic curves $\mathrm{NS}(\alpha(s,t))$ near the boundary. The second ingredient is provided by the Yang-Mills heat flow on 3-manifolds with boundary. This gives us candidates for what the boundary conditions might be if they were Lagrangian. Combining this with the $\mathcal{C}^1$- and $W^{2,2}$-estimates allows us to reprove a version of Gromov's compactness theorem for *almost* Lagrangian boundary conditions, and this is enough to establish the convergence in our compactness theorem.

Now we describe the sense of convergence claimed in our compactness theorem. On the one hand, the convergence is reminiscent of Uhlenbeck convergence on cylinders in the sense that bubbles and broken trajectories can form in the limit, and we only expect convergence modulo the action of the gauge group. The broken trajectories consist of holomorphic curves, but the bubbles can be either holomorphic curves or instantons. On the other hand, we only claim a certain type of $\mathcal{C}^0 \cap W^{1,2}$-convergence, as opposed to the $\mathcal{C}^\infty$-convergence of Uhlenbeck's theorem. This type of weaker convergence is to be expected for two reasons. The first is due to the degeneration of the metrics g_ϵ. The consequence being that we lose much of our uniform control when taking certain derivatives; we discuss a related issue in the next paragraph. The second reason occurs only when f has critical points. In this case the holomorphic curve representatives are typically not smooth (however, they are of Sobolev class $W^{1,2}$), and so we cannot expect that the instantons converge in the $\mathcal{C}^\infty$-topology; see Remark 2.2 (b).

In order to discuss convergence, we need to specify a metric on Y with which to define the relevant Sobolev spaces. For the most part, we use Sobolev norms defined using the *fixed* metric g_1 on Y (taking $\epsilon = 1$). This allows us to appeal to the standard Sobolev embedding and compactness results, which are used at the crux of the proof (the ϵ-dependent metrics are degenerating in ϵ, so we are prohibited from a direct use of Sobolev embedding relative to an ϵ-dependent norm). The downside is that this fixed metric does not interact well with the ϵ-ASD equation, which is defined using g_ϵ. To deal with this, we need a mechanism for translating between the analytically necessary fixed metric, and the geometrically natural metrics

coming from the g_ϵ. This mechanism is Theorem 4.1, the proof of which establishes elliptic estimates that are standard for the instanton equations, except we keep track of the ϵ-dependence in the constants.

Remark 1.3. *(a) As discussed above, the main technical difficulty in the proof of our compactness result arises due to the presence of critical points of f. To help streamline the exposition, we work almost exclusively under the assumption that f has at least one critical point. Note that this is a rather mild assumption since critical points can always be created by homotoping f; see* [27].

(b) Conformally rescaling our ϵ-dependent metric $ds^2 + g_\epsilon$ by ϵ^{-2}, one obtains the 'neck-stretching metrics' described by Atiyah [1] *(the 'neck' consists of the regular fibers of f). However, we prefer to work with metrics of the form $ds^2 + g_\epsilon$, since they keep the size of $\mathbb{R} \times S^1$ bounded. Due to the conformal invariance of the instanton and holomorphic curve equations, this is equivalent to Atiyah's set-up, and so is just a matter of taste.*

(c) Though we have been discussing the structure group $\mathrm{SO}(3) = \mathrm{PU}(2)$, in the sequel we work entirely with any projective unitary group $\mathrm{PU}(r)$ for $r \geq 2$. Ultimately, this is allowable because every closed surface has a $\mathrm{PU}(r)$-bundle that does not admit any reducible flat connections.

More generally, one could work with an arbitrary compact simple Lie group G, if one could ensure that (the identity-component of) the gauge group acts freely on the space of flat connections on the fibers of f. The only significant changes would be in the values of various constants.

(d) S. Dostoglou and D. Salamon proved the quilted Atiyah-Floer conjecture (1) in the special case where Y is a mapping torus [9, 10, 11]. *They do this using an implicit function theorem argument that shows $\widehat{M}_{0<\epsilon\leq 1}$ has a compactification as a 1-manifold with boundary $\widehat{M}_0 \cup \widehat{M}_1$, as described in Remark 1.1. This mapping torus case corresponds, in our set-up, to the case where the Morse function f is homotopic to a function with no critical points.*

When f has no critical points, our proof of the compactness theorem simplifies drastically, and can be viewed as a repackaging of much of the Dostoglou-Salamon analysis (e.g., Sections 8 and 9 of [10], *and Remark 4.8 below). On the other hand, the Dostoglou-Salamon implicit function theorem that was successful in the mapping torus case does not directly extend to the more general case considered here.*

(e) Upon completion of this project, the author learned of an earlier work [30] *wherein T. Nishinou proves a variant of our compactness theorem with $\mathbb{R} \times Y$ replaced by a product of surfaces. From an analytic standpoint, this is analogous to the case where Y has no critical points.*

(f) Recently M. Lipyanskiy [24] *has proven a compactness result for the quilted Atiyah-Floer conjecture. He considers sequences of pairs of holomorphic curves and instantons (relative to a fixed metric) that are matched via certain seam conditions.*

Acknowledgments: The author is grateful to his thesis advisor Chris Woodward for his insight and valuable suggestions. In addition, the author benefited greatly from discussions with Katrin Wehrheim, via Chris Woodward, of her unpublished work [37] on the Atiyah-Floer conjecture. This unpublished work analyzes various bubbles that appear in the limit of instantons with Lagrangian boundary conditions with degenerating metrics, and it outlines the remaining problems in such an approach. The present paper takes a different approach that avoids instantons with Lagrangian boundary conditions. The author would also like to thank an anonymous referee for suggestions that helped simplify the exposition and some of the proofs. This work was partially supported by NSF grants DMS 0904358 and DMS 1207194.

2. Set-up and statement of the main results

Let $f : Y \to S^1$ be a Morse function as in the introduction. By homotoping f, we can assume that the number N of critical points of f is positive (see Remark 1.3 (a)), and that these critical points have distinct critical values. This implies that $N > 0$ is even, and also allows us to view Y as a composition of cobordisms:

$$Y_{01} \cup_{\Sigma_1} (I \times \Sigma_1) \cup_{\Sigma_1} Y_{12} \cup_{\Sigma_2} \ldots \cup_{\Sigma_{N-1}} Y_{(N-1)0} \cup_{\Sigma_0} (I \times \Sigma_0) \cup_{\Sigma_0} \tag{4}$$

where $I := [0, 1]$ is the unit interval, each $\Sigma_j \subset Y$ is a fixed regular fiber of f, and each $Y_{j(j+1)} \subset Y$ is a cobordism from Σ_j to Σ_{j+1} such that $f|_{Y_{j(j+1)}}$ has exactly one critical point. Here and below the indices j are taken modulo N. Note also that (4) is *cyclic* in the sense that the cobordism $I \times \Sigma_0$ on the right is glued to the cobordism Y_{01} on the left, reflecting the fact that f maps to the circle.

We set

$$Y_\bullet := \sqcup_j Y_{j(j+1)}, \quad \text{and} \quad \Sigma_\bullet := \sqcup_j \Sigma_j,$$

and refer to the components of the boundary $\partial Y_\bullet$ as the *seams*.

Fix a metric g on Y, and assume this has been chosen to have the form

$$g|_{I \times \Sigma_\bullet} = dt^2 + g_\Sigma,$$

where g_Σ is a metric on $\Sigma_\bullet$ that is independent of $t \in I$. We call g the *fixed metric*. For $\epsilon > 0$ we define a new metric g_ϵ by setting

$$g_\epsilon|_{Y_\bullet} := \epsilon^2 g|_{Y_\bullet}, \quad g_\epsilon|_{I \times \Sigma_\bullet} := dt^2 + \epsilon^2 g_\Sigma.$$

See Figure 3. We call g_ϵ the *ϵ-dependent metric*.

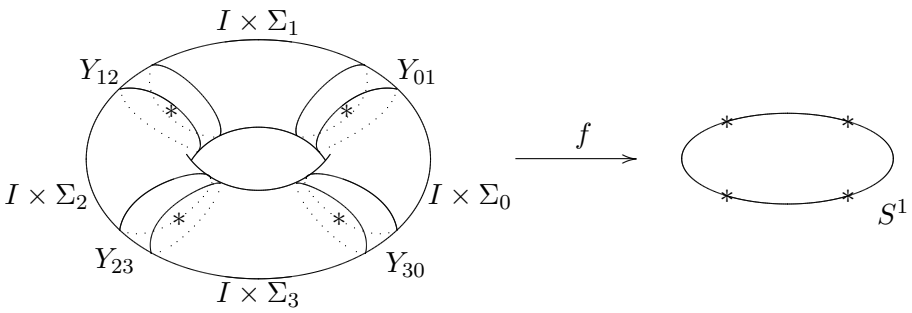

FIGURE 2. *Here is an illustration of $f : Y \to S^1$ with $N = 4$ critical points. Each $Y_{i(i+1)}$ has a unique critical point indicated by a star. The corresponding critical value in S^1 is also indicated by a star.*

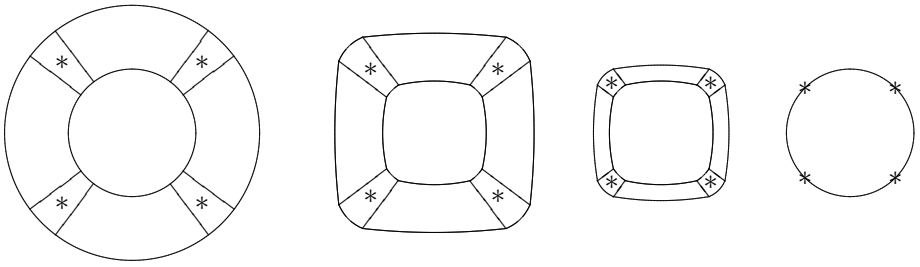

FIGURE 3. *Pictured above are four copies of the manifold Y, viewed from the 'top' relative to the illustration in Figure 2. Moving from left to right, the different copies represent the metric g_ϵ as ϵ decreases to zero. Notice that the volumes of the Σ_i and the $Y_{i(i+1)}$ are going to zero. However, the length in the I-direction (the 'neck') of each $I \times \Sigma_i$ is remaining fixed. In the picture on the far right, the $Y_{i(i+1)}$ have collapsed entirely to the critical points of f.*

2.1. **ϵ-dependent smooth structures.** Let $\mathcal{S}_1$ denote the smooth structure on Y, i.e., the smooth structure in which g and f are smooth. We call this the *standard smooth structure*. When $\epsilon \neq 1$ the metric g_ϵ is not smooth relative to $\mathcal{S}_1$. To see this, set $V = \nabla f / |\nabla f|$, where the norm and gradient are taken with respect to the fixed metric $g = g_1$. Then away from the critical points of f, the vector field V is smooth on $(Y, \mathcal{S}_1)$, but

$$g_\epsilon(V, V) = \begin{cases} 1 & \text{on } I \times \Sigma_\bullet \\ \epsilon^2 & \text{on } Y_\bullet \setminus \{\text{critical points}\} \end{cases}$$

is not even continuous, so g_ϵ cannot be continuous on $(Y, \mathcal{S}_1)$.

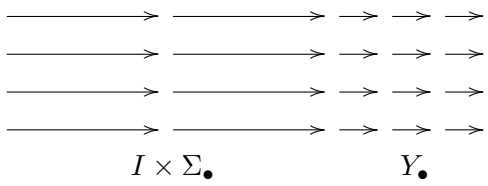

FIGURE 4. *The arrows in the figure represent the vector field V measured relative to the metric g_ϵ. Note the discontinuity at the seam.*

Observe that g_ϵ only fails to be smooth on $(Y, \mathcal{S}_1)$ at the seams $\partial Y_\bullet$. Furthermore, even at the seams, the metric g_ϵ is smooth in directions *parallel* to the seams. So the discontinuity illustrated in Figure 4 is the only thing that goes wrong.

The above observations imply that we need to be a little careful when dealing with g_ϵ. Fortunately, there is a *different* smooth structure in which g_ϵ is smooth. We call this the *ϵ-dependent smooth structure*, and denote it by $\mathcal{S}_\epsilon$. We will say a function, tensor, connection, etc. is *ϵ-smooth* if it is smooth with respect to the ϵ-dependent smooth structure.

The existence of $\mathcal{S}_\epsilon$ can be seen as follows. View Y as the *topological* manifold in (4). Following Milnor [27, Theorem 1.4], any choice of collar neighborhoods of the seams determines a smooth structure on Y. In this language, the smooth structure $\mathcal{S}_1$ arises by choosing collar neighborhoods of the seams determined by the time-δ gradient flow of f, and then using the identity to glue these neighborhoods on the overlap. Here $\delta > 0$ is small, but fixed. On the other hand, the smooth structure $\mathcal{S}_\epsilon$ arises by taking the time-δ gradient flow on the $Y_{i(i+1)}$ side of the seam $\{1\} \times \Sigma_i$, but the time-$\delta\epsilon$ gradient flow on the $[0,1] \times \Sigma_i$ side of the seam, and then gluing using the gluing map $(t, \sigma) \mapsto (\epsilon t, \sigma)$.

Remark 2.1. *(a) In dimension three, the topological category agrees with the smooth category, so it also follows that the smooth manifold $(Y, \mathcal{S}_\epsilon)$ is actually* diffeomorphic *to $(Y, \mathcal{S}_1)$. This diffeomorphism can be realized concretely by fixing isotopies of the collar neighborhoods above, since these isotopies determine a diffeomorphism between the two smooth structures* [27].

(b) Fix $1 \leq p \leq \infty$ and let $W^{1,p}(Y, \mathcal{S}_\epsilon)$ denote the Sobolev space associated to the ϵ-dependent smooth structure (as a vector space, this is independent of the choice of connection used to define the derivative). Every ϵ-smooth function *h on Y is in the Sobolev space $W^{1,p}(Y, \mathcal{S}_1)$ associated to the standard smooth structure. That is, the identity map on Y determines a pullback map of the form $\mathrm{Id}^* : C^\infty(Y, \mathcal{S}_\epsilon) \longrightarrow W^{1,p}(Y, \mathcal{S}_1)$. This can be seen as follows: The underlying topologies on $(Y, \mathcal{S}_1)$ and $(Y, \mathcal{S}_\epsilon)$ are identical, so the ϵ-smooth function h is continuous on $(Y, \mathcal{S}_1)$. Moreover, on the complement of the seams, h is 1-smooth with bounded derivative. This implies $\mathrm{Id}^* h$ is of Sobolev class $W^{1,p}(Y, \mathcal{S}_1)$.*

However, in general, an ϵ-smooth tensor *or* connection *will only be in $L^p(Y, \mathcal{S}_1)$ with respect to the standard smooth structure. This is because any component transverse to the seam will have a jump discontinuity as in Figure 4 (unless that component vanishes), and so taking a derivative will introduce a delta function.*

2.2. **Principal PU(r)-bundles.** As mentioned in Remark 1.3 (c), we will be working with a principal PU(r)-bundle over Y. Before specifying the bundle, we review several topological facts about PU(r)-bundles.

On manifolds X of dimension at most 4, the principal PU(r)-bundles $P \to X$ are classified, up to bundle isomorphism, by two characteristic classes $t_2(P) \in H^2(X, \mathbb{Z}_r)$ and $q_4(P) \in H^4(X, \mathbb{Z})$. These generalize the 2nd Stiefel-Whitney class and 1st Pontryagin class, respectively, to the group PU(r); see [40].

We equip the Lie algebra $\mathfrak{pu}(r)$ with the inner product given by $\langle \mu, \nu \rangle := -\frac{1}{4\pi^2}\mathrm{tr}(\mu \cdot \nu)$, where the trace is the one induced by the inclusion $\mathfrak{pu}(r) \cong \mathfrak{su}(r) \subset \mathrm{End}(\mathbb{C}^r)$. When X is a closed, oriented 4-manifold, the class q_4 can be expressed via the Chern-Weil formula

$$q_4(P) = -r \int_X \langle F_A \wedge F_A \rangle \tag{5}$$

which holds for any connection A on P; see [13]. Here F_A is the curvature of A and the notation $\langle \cdot \wedge \cdot \rangle$ combines the wedge on forms with the inner product on $\mathfrak{pu}(r)$. If P is the reduction of an SU(r)-bundle $P' \to X$, then $2rc_2(P') = q_4(P)$ and (5) recovers the usual Chern-Weil formula for complex vector bundles.

Now assume $\dim(X) \leq 3$, and let $\mathcal{G}(P)$ denote the group of gauge transformations on P. There are maps

$$\eta : \mathcal{G}(P) \longrightarrow H^1(X, \mathbb{Z}_r), \quad \deg : \mathcal{G}(P) \longrightarrow H^3(X, \mathbb{Z})$$

called the *parity* and *degree*. These detect the connected components of $\mathcal{G}(P)$ in the sense that u can be connected to u' by a path if and only if $\eta(u) = \eta(u')$ and $\deg(u) = \deg(u')$. In particular, a gauge transformation $u \in \mathcal{G}(P)$ lies in the identity component if and only if $\eta(u) = 0$ and $\deg(u) = 0$. See [13].

Now we are at a place to define the bundle of the 3-manifold Y. Fix a generator $d \in \mathbb{Z}_r$. Then we take $Q \to Y$ to be the principal PU(r)-bundle with the property that the characteristic class $t_2(Q)$ is Poincaré dual to $d\,[\gamma] \in H_1(Y, \mathbb{Z}_r)$, where $d \in \mathbb{Z}_r$ is a generator, and $\gamma : S^1 \to Y$ is a section of $f : Y \to S^1$. When $r = 2$, we have PU(2) = SO(3), and this is exactly the bundle Q from the introduction.

Set

$$Q_{j(j+1)} := Q|_{Y_{j(j+1)}}, \quad Q_\bullet := \sqcup_j Q_{j(j+1)}, \quad P_j := Q|_{\{0\}\times\Sigma_j}, \quad P_\bullet := \sqcup_j P_j.$$

The assumption on $t_2(Q)$ ensures

$$t_2(Q_{j(j+1)})\,[\Sigma_j] = t_2(P_j)\,[\Sigma_j] = d,$$

which will be useful in establishing smooth moduli spaces in Section 2.4. Here $[\Sigma_j] \in H^2(\Sigma_j, \mathbb{Z}_r) = H^2(Y_{j(j+1)}, \mathbb{Z}_r)$ is the obvious homology class.

In [13] we show that the fibers $\Sigma_\bullet \subset Y$ determine a connected component in $\mathcal{G}(Q)$ consisting of degree 1 gauge transformations. We let

$$\mathcal{G}_\Sigma \subset \mathcal{G}(Q)$$

denote the subgroup generated by this component. This subgroup can be equivalently described as the set of gauge transformations that restrict on each fiber of f to an identity component gauge transformation.

For each $\epsilon > 0$, there is a smooth structure on Q for which the map $Q \to Y$ is smooth relative to the ϵ-dependent smooth structure $\mathcal{S}_\epsilon$ on Y. When Y is equipped with $\mathcal{S}_\epsilon$, we will always assume that Q is equipped with the smooth structure just described.

2.3. **Gauge theory.** Here we examine the instanton equation on $\mathbb{R} \times Y$. We begin by discussing the notation we will be using; see [7, Chapter 2] for more details.

Let G be a Lie group with Lie algebra $\mathfrak{g}$. Assume $\mathfrak{g}$ is equipped with an Ad-invariant inner product $\langle \cdot, \cdot \rangle$. Suppose X is an oriented manifold, possibly with boundary, and fix a principal G-bundle $P \to X$. We will write $\mathcal{A}(P)$ for the space of connections on P. Given $A \in \mathcal{A}(P)$, we will denote the associated covariant derivative and curvature by

$$d_A : \Omega^k(X, P(\mathfrak{g})) \longrightarrow \Omega^{k+1}(X, P(\mathfrak{g})), \quad \text{and} \quad F_A \in \Omega^2(X, P(\mathfrak{g})).$$

Here $\Omega^k(X, P(\mathfrak{g}))$ is the space of k-forms on X with values in the associated adjoint bundle $P(\mathfrak{g})$. The space $\mathcal{A}(P)$ is an affine space modeled on $\Omega^1(X, P(\mathfrak{g}))$, and we use additive notation to denote the affine action. In particular, the covariant derivative and curvature satisfy

$$d_{A+\xi} = d_A + [\xi \wedge \cdot\,], \quad F_{A+\xi} = F_A + d_A\xi + \frac{1}{2}\,[\xi \wedge \xi\,],$$

where ξ is a 1-form and $[\cdot \wedge \cdot]$ combines the wedge on forms with the Lie bracket on $\mathfrak{g}$. We say that a connection A is *irreducible* if d_A is injective on 0-forms. If X has a metric, then the formal adjoint of d_A is

$$d_A^* := -(-1)^{(n-1)(k-1)} * d_A * : \Omega^k(X, P(\mathfrak{g})) \longrightarrow \Omega^{k-1}(X, P(\mathfrak{g})),$$

where $*$ is the Hodge star, and n is the dimension of X.

Denote by $\mathcal{A}_{\text{flat}}(P) \subset \mathcal{A}(P)$ the subspace of flat connections. If A is flat, then $d_A \circ d_A = 0$, and we can form the *harmonic spaces*

$$H_A^k := \ker\left(d_A|_{\Omega^k}\right) / \text{im}\ \left(d_A|_{\Omega^{k-1}}\right).$$

These are finite-dimensional when X is compact. We will say a flat connection A is *non-degenerate* if $H_A^1 = 0$.

Suppose X is a 4-manifold. We will say a connection A is *ASD* or an *instanton* if

$$F_A + *F_A = 0.$$

Note that the metric (and orientation) of X is encoded in this equation via the Hodge star.

The space $\Omega^0(X, P(\mathfrak{g}))$ can be naturally identified with the Lie algebra of the gauge group $\mathcal{G}(P)$. The exponential map is given by $\xi \mapsto \exp(\xi)$, where exp is the exponential map for the finite-dimensional group G. We will denote the identity component of the gauge group by $\mathcal{G}_0(P)$. The gauge group acts (on the right) on the space of connections by pullback, and the curvature transforms according $F_{u^*A} = \mathrm{Ad}(u^{-1})F_A$. Consequently, the gauge group restricts to an action on the flat connections, as well as on the space of instantons.

Notation Convention: When the base manifold has dimension 4 we will use capital letters A, U to denote connections and gauge transformations. In dimension 3 we will use lower case letters a, u, and in dimension 2 we will use lower case Greek letter α, μ to denote connections and gauge transformations.

Now we consider $\mathbb{R} \times Y$, where Y is the 3-manifold from the introduction. Let $Q \to Y$ be as in Section 2.2, and consider the bundle $\mathbb{R} \times Q \to \mathbb{R} \times Y$. We want to express connections A on $\mathbb{R} \times Q$ in certain local coordinates. We first note that we can write

$$A = a + p\, ds,$$

where a is a path of connections on Y, p is a path of 0-forms on Y (with values in $Q(\mathfrak{g})$), and we are using s to denote the coordinate variable on $\mathbb{R}$. This decomposition is unique; for example, the path a is uniquely determined by restricting

$$a(s) = A|_{\{s\} \times Y}.$$

Recall the decomposition $Y = Y_\bullet \cup (I \times \Sigma_\bullet)$. We will primarily use the above coordinate description over $\mathbb{R} \times Y_\bullet$. Over the region $\mathbb{R} \times I \times \Sigma_\bullet$, we can decompose even further. Using coordinates (s, t) for the strip $\mathbb{R} \times I$, we can write

$$A|_{\mathbb{R} \times I \times \Sigma} = \alpha + \phi\, ds + \psi\, dt,$$

where α is a map from the strip into the space of connections on $\Sigma_\bullet$, and ϕ, ψ are maps from the strip into the space of 0-forms on $\Sigma_\bullet$. Once again, this decomposition is unique. The relation between these coordinates and the above is given by

$$a|_{\mathbb{R} \times I \times \Sigma_\bullet} = \alpha + \psi\, dt, \quad p|_{\mathbb{R} \times I \times \Sigma_\bullet} = \phi.$$

Equip Y with the ϵ-dependent metric g_ϵ (and hence the ϵ-dependent smooth structure). We say that a connection on $\mathbb{R} \times Y$ is *ϵ-ASD* if it is

an instanton relative to the metric $ds^2 + g_\epsilon$. In the local coordinates over $\mathbb{R} \times Y_\bullet$, the ϵ-ASD equation takes the form

$$\partial_s a(s) - d_{a(s)} p(s) + \epsilon^{-1} *_Y F_{a(s)} = 0,$$

where $*_Y$ is the Hodge star coming from the fixed metric g on Y. Over $\mathbb{R} \times I \times \Sigma_\bullet$, the ϵ-ASD equation can be written as

$$\begin{aligned} \partial_s \alpha - d_\alpha \phi + *_\Sigma (\partial_t \alpha - d_\alpha \psi) &= 0 \\ \epsilon^2 (\partial_s \psi - \partial_t \phi - [\psi, \phi]) + *_\Sigma F_\alpha &= 0 \end{aligned}$$

where $*_\Sigma$ is the Hodge star associated to the metric g_Σ.

We define the *ϵ-energy* of a connection A on $\mathbb{R} \times Q \to \mathbb{R} \times Y$ by

$$E^{\mathrm{inst}}_\epsilon(A) := \frac{1}{2} \int_Z \langle F_A \wedge *_\epsilon F_A \rangle.$$

This is well-defined for any connection on $\mathbb{R} \times Q$ that is $W^{1,2}$ with respect to the ϵ-smooth structure. If A is ϵ-ASD, then the ϵ-energy is finite if and only if *A limits to flat connections at $\pm\infty$* in the sense that there are flat connections $a^\pm \in \mathcal{A}_{\mathrm{flat}}(Q)$ such that $\lim_{s \to \pm\infty} a(s) = a^\pm$, and $\lim_{s \to \pm\infty} p(s) = 0$. When this is the case, $E^{\mathrm{inst}}_\epsilon(A) = \mathcal{CS}(a^+) - \mathcal{CS}(a^-)$, where $\mathcal{CS}$ is the Chern-Simons functional. In particular, the energy of A depends only on the limits $a^\pm$, and is independent of ϵ.

2.4. **Symplectic geometry.** The goal of this section is to define the term *holomorphic curve representative*. We begin with some generalities, and then we specialize to the situation directly relevant to this paper.

Suppose Σ is a closed, connected, oriented surface, and $P \to \Sigma$ is a principal $\mathrm{PU}(r)$-bundle with $t_2(P)[\Sigma] \in \mathbb{Z}_r$ a generator. This implies all flat connections on P are irreducible, and that $\mathcal{G}_0(P)$ acts freely on $\mathcal{A}_{\mathrm{flat}}(P)$. Moreover, the space

$$M(P) := \mathcal{A}_{\mathrm{flat}}(P)/\mathcal{G}_0(P)$$

is a compact, simply-connected, smooth, symplectic manifold. The tangent space at $[\alpha] \in M(P)$ is canonically identified with the harmonic space H^1_α, for any choice of representative $\alpha \in [\alpha]$. Fixing a metric on Σ, the Hodge theorem [36, Theorem 6.8] allows us to identify H^1_α with the kernel

$$\ker (d_\alpha \oplus d^*_\alpha) \subset \Omega^1(\Sigma).$$

We use the term *harmonic 1-form representatives* to refer to the forms in this kernel. The Hodge star $*_\Sigma$ on Σ determines a complex structure on $\mathcal{A}(P)$. Moreover, this restricts to the space of harmonic 1-form representatives, and hence defines an (integrable) complex structure on $M(P)$ that is compatible with the symplectic form. See [2] for more details regarding these assertions.

Now suppose Y_{ij} is an oriented, connected cobordism between closed, connected, oriented, nonempty surfaces Σ_i and Σ_j. Then we can find a $\mathrm{PU}(r)$-bundle $Q_{ij} \to Y_{ij}$ with $t_2(Q_{ij})[\Sigma_i] \in \mathbb{Z}_r$ a generator. This implies the flat connections on Q_{ij} are irreducible, and the quotient $\mathcal{A}_{\mathrm{flat}}(Q_{ij})/\mathcal{G}_0(Q_{ij})$

is a finite-dimensional, simply-connected, smooth manifold [39, Theorem 3.4.3]. Restricting to the two boundary components induces an embedding

$$\mathcal{A}_{\text{flat}}(Q_{ij})/\mathcal{G}_0(Q_{ij}) \hookrightarrow M(Q_{ij}|_{\Sigma_i}) \times M(Q_{ij}|_{\Sigma_j}),$$

and we let $L(Q_{ij})$ denote the image. If Y_{ij} admits a Morse function with only index 1 critical points, then

$$L(Q_{ij}) \subset M(Q_{ij}|_{\Sigma_i})^- \times M(Q_{ij}|_{\Sigma_j})$$

is a smooth Lagrangian submanifold, where the superscript in $M(Q_{ij}|_{\Sigma_i})^-$ means that we have replaced the symplectic form with its negative (this is necessary for $L(Q_{ij})$ to be Lagrangian). We refer the reader to [1] or [3, Chapter 3] for a general description of Lagrangians arising in this way.

As in the case of surfaces, the tangent space to $T_{[a]}L(Q_{ij})$ can be identified with the harmonic space H^1_a. This, in turn, can be identified with a space of 1-form representatives. Indeed,

$$H^1_a \cong \ker(d_a \oplus d_a^* \oplus \partial *_Y) \subset \Omega^1(Y_{ij}, Q_{ij}(\mathfrak{g})), \tag{6}$$

where $\partial *_Y$ is the map sending a 1-form w on Y_{ij} to the 2-form $(*_Y w)|_{\partial Y_{ij}}$ on ∂Y_{ij}. The proof of (6) follows just as in the case for closed manifolds [36, Theorem 6.8], with this boundary condition being used to kill off the boundary term that appears during integration by parts.

Now let $Q \to Y$ be as in Section 2.2. By the above observations, restricting to each of the two boundary components of $Y_{j(j+1)}$ determines a Lagrangian embedding

$$L(Q_{j(j+1)}) \hookrightarrow M(P_j)^- \times M(P_{j+1}). \tag{7}$$

Then we set

$$\begin{aligned} M &:= M(P_0)^- \times M(P_1) \times M(P_2)^- \times \ldots \times M(P_{N-1}) \\ L_{(0)} &:= L(Q_{01}) \times L(Q_{23}) \times \ldots \times L(Q_{(N-2)(N-1)}) \\ L_{(1)} &:= L(Q_{12}) \times L(Q_{34}) \times \ldots \times L(Q_{(N-1)N}). \end{aligned} \tag{8}$$

The maps (7) piece together to define Lagrangian embeddings

$$L_{(0)}, L_{(1)} \hookrightarrow M.$$

The Hodge star $*_\Sigma$ from g_Σ determines an almost complex structure J on M by declaring the $M(P_j)$-component of J to be $*_\Sigma$ when j is odd, and to be the negative $-*_\Sigma$ when j is even; this condition makes J compatible with the symplectic form on M specified by the superscripts in (8). The intersection points $L_{(0)} \cap L_{(1)}$ can be canonically identified with the space $\mathcal{A}_{\text{flat}}(Q)/\mathcal{G}_\Sigma$ of flat connections on Y modulo gauge; see [12].

The quilted Floer group $HF_{\text{symp}}(Y)$ from the introduction was defined in [39], using Floer's Lagrangian intersection homology group [17] applied to $M, L_{(0)}, L_{(1)}$. The group is defined by counting holomorphic maps $\mathbb{R} \times I \to M$ with boundary conditions in the Lagrangians.

Fix a smooth map

$$v : (\mathbb{R} \times I, \mathbb{R} \times \{0\}, \mathbb{R} \times \{1\}) \longrightarrow (M, L_{(0)}, L_{(1)}).$$

Our goal now is to associate to this a connection $A(v)$ on the bundle $\mathbb{R} \times Q$. To define this connection, first write the components of v as $v = (v_0, \ldots, v_{N-1})$, so v_j maps into $M(P_j) = \mathcal{A}_{\text{flat}}(P_j)/\mathcal{G}_0(P_j)$. Fix a smooth lift

$$\alpha_j : \mathbb{R} \times I \longrightarrow \mathcal{A}_{\text{flat}}(P_j)$$

of v_j with the property that $\alpha_j(s,t) = v_j(s,t)$ if j is odd, and $\alpha_j(s,t) = v_j(s,1-t)$ if j is even. The sign convention is such that v is J-holomorphic in M if and only if α_j is $*_\Sigma$-holomorphic in $\mathcal{A}(P_j)$ for each j. The α_j piece together, in the obvious way, to form a map

$$\alpha : \mathbb{R} \times I \longrightarrow \mathcal{A}_{\text{flat}}(P_\bullet) = \mathcal{A}_{\text{flat}}(P_0) \times \ldots \times \mathcal{A}_{\text{flat}}(P_{N-1}).$$

Since $\alpha(s,t)$ is flat for each (s,t), there are unique maps ϕ, ψ from the strip $\mathbb{R} \times I$ into the space of 0-forms on $\Sigma_\bullet$, with the property that the 1-forms

$$\partial_s \alpha - d_\alpha \phi, \quad \text{and} \quad \partial_t \alpha - d_\alpha \psi$$

each lie in the harmonic space $\ker(d_\alpha \oplus d_\alpha^*)$. Existence and uniqueness follow from the irreducibility of flat connections on the bundle $P_\bullet \to \Sigma_\bullet$, which says that the operator

$$d_\alpha^* d_\alpha : \Omega^0(\Sigma_\bullet, P_\bullet(\mathfrak{g})) \to \Omega^0(\Sigma_\bullet, P_\bullet(\mathfrak{g}))$$

is invertible. To avoid the complications at the seams described in Remark 2.1 (b), we assume that ψ satisfies

$$\psi(s,0) = \psi(s,1) = 0, \quad \forall s \in \mathbb{R}. \tag{9}$$

This can always be arranged by choosing a suitable lift α (this is just temporal gauge for the connection $\alpha + \psi\, dt$ near the boundary of $I \times \Sigma_\bullet$). Then we define the connection $A(v)$ on the cylinder $\mathbb{R} \times I \times \Sigma_\bullet$ by setting

$$A(v)|_{\mathbb{R} \times I \times \Sigma_\bullet} := \alpha + \phi\, ds + \psi\, dt.$$

Now we define $A(v)$ on $\mathbb{R} \times Y_\bullet$. Unraveling the definition of the Lagrangians $L_{(0)}, L_{(1)}$, one finds that the Lagrangian boundary conditions of v are equivalent to the existence of a map $a : \mathbb{R} \to \mathcal{A}_{\text{flat}}(Q_\bullet)$ such that

$$a(s)|_{\partial Y_\bullet} = \alpha(s, \cdot)|_{\partial(I \times \Sigma_\bullet)}. \tag{10}$$

Indeed, the Lagrangian $L_{(0)}$ (resp. $L_{(1)}$) determines the gauge class of a on $Y_{j(j+1)}$ for j even (resp. j odd). Moreover, given α, the condition (10) determines a uniquely, modulo the action of the gauge transformations that restrict to the identity on $\partial Y_\bullet$. This is because flat connections are determined (modulo gauge) by their induced representations on the fundamental group, and for each j the fundamental groups $\pi_1(\Sigma_j), \pi_1(\Sigma_{j+1})$ generate $\pi_1(Y_{j(j+1)})$.

Note that (10) does not determine the component of a in the direction transverse to the seam. To match up with (9), we assume that the lift a has been chosen so that this transverse component vanishes:

$$\iota_{\partial_n} a|_{\partial Y_\bullet} = 0, \tag{11}$$

where ∂_n is a unit normal to $\partial Y_\bullet$. This can always be arranged by applying a suitable gauge transformation to a (as with $\alpha + \psi\, dt$ on $I \times \Sigma_\bullet$, this can be viewed as a temporal gauge for a near $\partial Y_\bullet$). We can also assume this gauge transformation restricts to the identity on the boundary, so (10) continues to hold.

Fix a. Then irreducibility for the bundle $Q_\bullet \to Y_\bullet$ implies that there is a unique path p into the space of 0-forms on $Y_\bullet$ with the property that

$$\partial_s a - d_a p$$

lies in the space $\ker(d_a \oplus d_a^* \oplus \partial *_Y)$ of harmonic 1-form representatives. Then we set

$$A(v)|_{\mathbb{R}\times Y_\bullet} := a + p\, ds.$$

The boundary conditions (9-11), as well as the uniqueness of ϕ, ψ and p, imply that $A(v)$ is *continuous*, even at the seams $\mathbb{R} \times \partial Y_\bullet$ and relative to the ϵ-smooth structure on Y for all $\epsilon > 0$. It is not hard to see that each restriction $A(v)|_{\mathbb{R}\times I\times\Sigma_\bullet}$ and $A_v|_{\mathbb{R}\times Y_\bullet}$ is locally of Sobolev class $W^{1,p}$ for any $1 \le p \le \infty$. It therefore follows that $A(v)$ is in $W^{1,p}_{loc}$ as well, relative to all ϵ-smooth structures. We will say a connection is a *representative* if it is of the form $A(v)$ for some v, and we say that it is a *holomorphic curve representative* if v is holomorphic relative to the complex structure J from above. Writing a connection A in components in the usual way, it follows that a representative A is a holomorphic curve representative if and only if the following hold

$$\begin{aligned} \partial_s\alpha - d_\alpha\phi + *_\Sigma(\partial_t\alpha - d_\alpha\psi) &= 0 \quad \text{on } \mathbb{R}\times I\times\Sigma_\bullet \\ F_\alpha &= 0 \quad \text{on } \mathbb{R}\times I\times\Sigma_\bullet \\ \\ F_a &= 0 \quad \text{on } \mathbb{R}\times Y_\bullet. \end{aligned}$$

Remark 2.2. *(a) Given two representatives $A(v)$ and $A(v)'$ of v, there is a unique gauge transformation on $\mathbb{R}\times Q$ taking $A(v)$ to $A(v)'$. Moreover, this gauge transformation lies in (a suitable Sobolev completion of) the space of maps from $\mathbb{R}$ into the subgroup $\mathcal{G}_\Sigma$ from Section 2.2.*

(b) Though all representatives are in the space $W^{1,p}$ (relative to any ϵ-smooth structure), they typically do not have much higher regularity. Indeed, suppose $A(v)$ is a representative. Using the gradient flow of the circle-valued function f, we can extend the decomposition $a = \alpha + \psi\, dt$ from $I\times\Sigma_\bullet$ into a small neighborhood of the boundary in $Y_\bullet$. Then since a is flat on the $Y_\bullet$ side of the seam we have $\partial_t\alpha - d_\alpha\psi = 0$. On the $I\times\Sigma_\bullet$ side we have $\partial_t\alpha - d_\alpha\psi = \partial_t v$. Consequently, when $\partial_t v(s,0) \neq 0$ or $\partial_t v(s,1) \neq 0$, then A_v

is not even $\mathcal{C}^1$. Note that holomorphic curves typically have $\partial_t v(s,0) \neq 0$, unless they are constant or sphere bubbles.

We define the *energy* of a representative A by the formula

$$E^{\mathrm{symp}}(A) := \frac{1}{2}\int_{\mathbb{R}\times I\times \Sigma_\bullet} |\partial_s \alpha - d_\alpha \phi|^2 + |\partial_t \alpha - d_\alpha \psi|^2 \,\mathrm{dvol}.$$

This is concocted so that it recovers the energy of the curve that A represents. If A is a holomorphic curve representative, then the energy $E^{\mathrm{symp}}(A)$ is finite if and only if A limits to flat connections $a^\pm$ at $\pm\infty$; see [12]. In this case we have

$$E^{\mathrm{symp}}(A) = \mathcal{CS}(a^+) - \mathcal{CS}(a^-)$$

and so the energy is again a topological quantity depending only on the limiting connections. (The symplectic action functional for $(M, L_{(0)}, L_{(1)})$ is given by the Chern-Simons functional of representatives.)

2.5. **Statement of the main results.** Let $Q \to Y$ be as in section 2.2, and consider the induced bundle $\mathbb{R}\times Q \to \mathbb{R}\times Y$ over the cylinder. For $A \in \mathcal{A}(\mathbb{R}\times Q)$ and $s_0 \in \mathbb{R}$, let $\tau_{s_0}^* A \in \mathcal{A}(\mathbb{R}\times Q)$ be the connection defined by translating in the $\mathbb{R}$-direction by s_0. Given $(s,t) \in \mathbb{R}\times I$, we will use

$$\iota_{(s,t)} : \Sigma_\bullet \hookrightarrow \{(s,t)\}\times\Sigma_\bullet \subset \mathbb{R}\times Y$$

to denote the inclusion. In particular, the pullback $\alpha(s,t) := \iota_{(s,t)}^* A$ can be viewed as an $\mathbb{R}\times I$-dependent connection on $P_\bullet \to \Sigma_\bullet$ (note that this pullback has no ds- or dt-component).

Given ϵ-smooth flat connections $a^\pm \in \mathcal{A}_{\mathrm{flat}}(Q)$, we will denote by

$$\mathcal{A}_\epsilon^{1,q}(a^-, a^+)$$

the space of connections on $\mathbb{R}\times Q$ that (i) are of Sobolev class $W^{1,q}$ relative to the ϵ-smooth structure, and (ii) limit to $a^\pm$ at $\pm\infty$. Note that the $a^\pm$ can be gauge transformed so they are ϵ-smooth for all $\epsilon > 0$; see Lemma 4.6.

Now we can state the main theorem.

Theorem 2.3. *(Main Compactness Theorem) Fix $2 < q < \infty$, and let $\mathbb{R}\times Q \to \mathbb{R}\times Y$ be as above. Assume all flat connections on Q are non-degenerate, and fix flat connections $a^\pm \in \mathcal{A}_{\mathrm{flat}}(Q)$ that are ϵ-smooth for all $\epsilon > 0$.*

Let $(\epsilon_\nu)_{\nu\in\mathbb{N}}$ be a sequence of positive numbers converging to 0. Assume that, for each ν, there is an ϵ_ν-smooth ϵ_ν-ASD connection $A_\nu \in \mathcal{A}_{\epsilon_\nu}^{1,q}(a^-, a^+)$. Then there is

(i) *a subsequence of the A_ν (still denoted A_ν);*
(ii) *a sequence of ϵ_ν-smooth gauge transformations U_ν on $\mathbb{R}\times Q$;*
(iii) *a finite sequence of flat connections*

$$\left\{a^0 = a^-, a^1, \ldots, a^{J-1}, a^J = a^+\right\} \subseteq \mathcal{A}_{\mathrm{flat}}(Q);$$

(iv) *for each* $j \in \{1, \dots, J\}$, *a holomorphic curve representative*

$$A^j \in \mathcal{A}_1^{1,q}(u_{j-1}^* a^{j-1}, u_j^* a^j),$$

for some $u_{j-1}, u_j \in \mathcal{G}_\Sigma$, *possibly depending on* A^j;

(v) *for each* j, *a finite set* $B_j \subset \mathbb{R} \times I$;

(vi) *and, for each* j, *a sequence* $s_\nu^j \in \mathbb{R}$

such that the following holds. For each $j \in \{1, \dots, J\}$, *the restrictions* $\iota_{(s,t)}^* U_\nu^* \tau_{s_\nu^j}^* A_\nu$ *converge to* $\iota_{(s,t)}^* A^j$ *in* $\mathcal{C}^0$ *on compact sets* $K \subset \mathbb{R} \times I \backslash B_j$:

$$\sup_{(s,t)\in K} \left\| \iota_{(s,t)}^* \left(U_\nu^* \tau_{s_\nu^j}^* A_\nu - A^j \right) \right\|_{\mathcal{C}^0(\Sigma_\bullet)} \xrightarrow{\nu} 0. \tag{12}$$

The gauge transformations U_ν *can be chosen so that they restrict to the identity component* $\mathcal{G}_0(P_i)$ *on each* $\{(s,t)\} \times \Sigma_i$. *Moreover, for each* $b \in B_j$ *there is a positive integer* $m_b > 0$ *such that for any* ν,

$$\sum_{j=1}^{J} E^{\mathrm{symp}}(A^j) = E_{\epsilon_\nu}^{\mathrm{inst}}(A_\nu) - \frac{1}{r} \sum_{\substack{1 \le j \le J \\ b \in B_j}} m_b. \tag{13}$$

The conclusion of the theorem states that the A_ν converge, modulo gauge and bubbling, to the broken holomorphic curve trajectory $(A^1, \dots, A^J)$. The set B_j denotes the bubbling set associated to the limiting connection A^j, and m_b/r is the energy associated to the bubble(s) that form at $b \in B_j$ (the particular value of the energy is an artifact of our choice of inner product on the Lie algebra $\mathfrak{pu}(r)$). Our proof will show that the following types of bubbles can form:

- *Holomorphic spheres in* $M(P_i)$. These can occur at any point in $\mathbb{R} \times I$.
- *Holomorphic disks in* $M(P_i)^- \times M(P_{i+1})$, *with Lagrangian boundary conditions in* $L(Q_{i(i+1)})$. These occur only on the boundary $\mathbb{R} \times \partial I$.
- *Instantons on* S^4. These can occur at any point in $\mathbb{R} \times I$.
- *Instantons on* $\mathbb{R} \times Y_{i(i+1)}^\infty$. Here $Y_{i(i+1)}^\infty$ is obtained from $Y_{i(i+1)}$ by attaching a cylindrical end. These bubbles only occur on the boundary $\mathbb{R} \times \partial I$, and are the instanton analogue of holomorphic disks.

The value of m_b is a multiple of $2r$ if an instanton on S^4 is the only bubble that forms at b (otherwise, m_b can be any positive integer). This is because all $\mathrm{PU}(r)$-bundles on S^4 are induced from $\mathrm{SU}(r)$-bundles, while the other types of bubbles are associated to bundles with $t_2 \neq 0$.

To simplify the exposition, in the statement of Theorem 2.3, we have assumed that all flat connections $a \in \mathcal{A}_{\mathrm{flat}}(Q)$ are non-degenerate. In general, this need not be the case. However, non-degeneracy can always be achieved by first performing a suitable perturbation to the defining equations (see [12]). Though we do not pursue this here, our results have direct extensions to the perturbed setting.

It will be useful to recast the conclusion of Theorem 2.3 in the coordinate notation of Sections 2.3 and 2.4. To do this, let $(s,t) \in \mathbb{R} \times I$ and set

$$\alpha_\nu(s,t) = \iota^*_{(s,t)} A_\nu, \quad \alpha^j(s,t) = \iota^*_{(s,t)} A^j, \quad \text{and} \quad \mu(s,t) = \iota^*_{(s,t)} U.$$

In this notation, the convergence statement (12) is that, for each j, the quantity

$$\sup_{(s,t)\in K} \|\mu(s,t)^* \alpha_\nu(s - s^j_\nu, t) - \alpha^j(s,t)\|_{\mathcal{C}_0(\Sigma_\bullet)}$$

converges to zero as ν goes to infinity.

It follows from the construction of Section 2.4 that a holomorphic curve in M is entirely determined by the value of the restrictions $\alpha(s,t) = \iota^*_{(s,t)} A$ of any representative A. It is essentially due to this that the conclusion of Theorem 2.3 only includes a statement about convergence on $\mathbb{R} \times I \times \Sigma_\bullet$. That being said, the A_ν do converge in a certain sense on the complementary region $\mathbb{R} \times Y_\bullet$. A precise statement to this effect appears in (15) and (17) of the following lemma. The lemma is a variation of Theorem 2.3 with additional hypotheses that will be used to exclude bubbling a priori.

Lemma 2.4. *(Compactness Lemma) Fix a real number $2 < q < \infty$. Let $B \subset \mathbb{R} \times I$ be finite, and set $S_0 := \mathbb{R} \times I \backslash B$.*

Let $(\epsilon_\nu)_{\nu\in\mathbb{N}}$ be any sequence of positive numbers (not necessarily converging to zero). Assume that, for each ν, there is an ϵ_ν-smooth ϵ_ν-ASD connection A_ν satisfying the following conditions:

(i) *For each compact $K \subset S_0$, the slice-wise curvatures on $\Sigma_\bullet$ converge to zero:*

$$\sup_{(s,t)\in K} \|F_{\alpha_\nu(s,t)}\|_{L^\infty(\Sigma_\bullet)} \xrightarrow{\nu} 0.$$

(ii) *For each compact $L \subset \mathbb{R}$ with $L \times \{0,1\} \cap B = \emptyset$, the slice-wise curvatures on $Y_\bullet$ converge to zero:*

$$\sup_{s\in L} \|F_{a_\nu(s)}\|_{L^\infty(Y_\bullet)} \xrightarrow{\nu} 0, \qquad a_\nu(s) := \iota^*_s A_\nu.$$

(iii) *For each compact $K \subset S_0$, there is some constant C with*

$$\sup_\nu \sup_{(s,t)\in K} \|\partial_s \alpha_\nu(s,t) - d_{\alpha_\nu(s,t)} \phi_\nu(s,t)\|_{L^2(\Sigma_\bullet)} \leq C.$$

(iv) *For each compact $L \subset \mathbb{R}$ with $L \times \{0,1\} \cap B = \emptyset$, there is some constant C with*

$$\sup_\nu \sup_{s\in L} \|\partial_s a_\nu(s) - d_{a_\nu(s)} p_\nu(s)\|_{L^2(Y_\bullet)} \leq C.$$

(v) *The energies are uniformly bounded: $\sup_\nu E^{\mathrm{inst}}_{\epsilon_\nu}(A_\nu) < \infty$.*

Then there is a subsequence of the connections (still denoted A_ν), a sequence of gauge transformations U_ν, and a holomorphic curve representative A_∞ such that:

$$\sup_{(s,t)\in K} \left\| \alpha_\infty(s,t) - \mu_\nu(s,t)^* \alpha_\nu(s,t) \right\|_{\mathcal{C}^0(\Sigma_\bullet)} \xrightarrow{\nu} 0, \tag{14}$$

$$\sup_{s\in L} \left\| a_\infty(s) - u_\nu(s)^* a_\nu(s) \right\|_{L^4(Y_\bullet)} \xrightarrow{\nu} 0, \tag{15}$$

$$\begin{aligned} \sup_{(s,t)\in K} \Big(& \left\| \partial_s \alpha_\infty(s,t) - d_{\alpha_\infty(s,t)} \phi_\infty(s,t) \right\|_{L^2(\Sigma_\bullet)} \\ & - \left\| \partial_s \alpha_\nu(s,t) - d_{\alpha_\nu(s,t)} \phi_\nu(s,t) \right\|_{L^2(\Sigma_\bullet)} \Big) \xrightarrow{\nu} 0, \end{aligned} \tag{16}$$

$$\begin{aligned} \sup_{s\in L} \Big(& \left\| \partial_s a_\infty(s) - d_{a_\infty(s)} p_\infty(s) \right\|_{L^2(Y_\bullet)} \\ & - \left\| \partial_s a_\nu(s) - d_{a_\nu(s)} p_\nu(s) \right\|_{L^2(Y_\bullet)} \Big) \xrightarrow{\nu} 0, \end{aligned} \tag{17}$$

for any compact $K \subseteq S_0$, and any compact $L \subset \mathbb{R}$ with $(L \times \{0,1\}) \cap B = \emptyset$.

The key technical point of this lemma is that the convergence in (16) holds even for compact sets K that *intersect* the boundary of $\mathbb{R} \times I$.

Theorem 2.3 and Lemma 2.4 are proved in Sections 4.3 and 4.2, respectively. The proof of Theorem 2.3 consists of repeated applications of Lemma 2.4, together with fairly standard rescaling arguments to isolate the bubbles. The proof of Lemma 2.4 relies on three tools. The first is the Narasimhan-Seshadri correspondence for surfaces. This is developed in Section 3, and is used to associate a holomorphic curve representatives on $\mathbb{R} \times (0,1) \times \Sigma_\bullet$ to each instanton A_ν. The second tool is the Yang-Mills heat flow on the components of $Y_\bullet$. This is developed in Section 4.1, and is used to determine *nearby* Lagrangian boundary conditions for the representatives. The last tool consists of various elliptic estimates for the ϵ-ASD operator. This is addressed in Theorem 4.1.

3. The Narasimhan-Seshadri correspondence

The goal of this section is to prove Theorem 3.1, below. It provides a method for associating flat connections to connections with small curvature. The idea is to use the well-known fact that quotienting the subset of small curvature connections by the action of the complexified gauge group recovers the moduli space of flat connections; this is the *Narasihman-Seshadri correspondence*. The details of this procedure were originally carried out by Narasimhan and Seshadri [29]. They worked with unitary bundles, and this allowed them to use algebraic techniques. Later, their techniques were extended to more general structure groups by Ramanathan in his thesis [32]. (See Kirwan's book [22] for a finite-dimensional version.)

In preparation for a boundary-value problem, we need to work in an analytic category. Consequently, we adopt an approach of Donaldson [6], and use an implicit function theorem argument to arrive at a Narasimhan-Seshadri correspondence in our setting. This allows us to establish several

$\mathcal{C}^1$-estimates that will be needed for our proof of the main theorem. This approach also appears in Fukaya [18] and Nishinou [30].

Let $P \to \Sigma$ be a principal bundle over a closed surface. We will use superscripts to denote Sobolev completions of various spaces. For example,

$$\mathcal{A}^{k,q}(P), \quad \text{and} \quad \mathcal{G}^{k,q}(P)$$

denote the $W^{k,q}$-Sobolev completions of the space of connections and the gauge group, respectively. We recall that $\mathcal{G}^{k,q}(P)$ is a smooth Banach manifold whenever $kq > 2$. When this is the case, the usual action of the gauge group on $\mathcal{A}(P)$ extends to a *smooth* action of $\mathcal{G}^{k,q}(P)$ on $\mathcal{A}^{k-1,q}(P)$; see [38, Appendix B] for a general treatment. We will use $\|\cdot\|_{W^{k,q}(\Sigma)}$ and $\|\cdot\|_{L^q(\Sigma)}$ to denote the obvious Sobolev and L^q-norms.

Now we state the main theorem of this section.

Theorem 3.1. *(Narasimhan-Seshadri Theorem) Suppose G is a compact, connected Lie group, Σ is a closed, oriented surface with metric, and $P \to \Sigma$ is a principal G-bundle such that all flat connections are irreducible. Then the following hold.*

(i) *For any $1 < q < \infty$, there are constants $C > 0$ and $\epsilon_0 > 0$, and a $\mathcal{G}^{2,q}(P)$-equivariant deformation retract*

$$\mathrm{NS}_P : \left\{ \alpha \in \mathcal{A}^{1,q}(P) \;\middle|\; \|F_\alpha\|_{L^q(\Sigma)} < \epsilon_0 \right\} \longrightarrow \mathcal{A}^{1,q}_{\mathrm{flat}}(P) \tag{18}$$

that is smooth with respect to the $W^{1,q}$-topology on the domain and codomain.

(ii) *The map NS_P is also smooth with respect to the L^p-topology on the domain and codomain, for any $2 < p \leq \infty$.*

(iii) *Suppose $2 < p \leq \infty$. Then there are constants $C > 0$ and $\epsilon_0 > 0$ such that*

$$\|\mathrm{NS}_P(\alpha) - \alpha\|_{\mathcal{C}^0(\Sigma)} \leq C\|F_\alpha\|_{L^p(\Sigma)} \tag{19}$$

for all $\alpha \in \mathcal{A}^{1,q}(P)$ with $\|F_\alpha\|_{L^p(\Sigma)} < \epsilon_0$.

The proof will be given in Section 3.2, after we review the complexified gauge group in Section 3.1. The basic idea of the proof is to show that for each α in the domain of NS_P, there is a 'purely imaginary' complex gauge transformation μ such that $\mu^*\alpha$ is a flat connection, and μ is unique provided it lies sufficiently close to the identity. We then define $\mathrm{NS}(\alpha) := \mu^*\alpha$. The estimate (19) shows that if α has small curvature, then the flat connection $\mathrm{NS}_P(\alpha)$ is close to α.

The next proposition establishes several additional properties of the map NS_P. To state it, we restrict to the case $G = \mathrm{PU}(r)$ since our proof relies on the *smoothness* of the moduli space $M(P) = \mathcal{A}_{\mathrm{flat}}(P)/\mathcal{G}_0(P)$, and not just irreducibility of flat connections. We recall that Uhlenbeck compactness

implies that each connection in $\mathcal{A}^{1,q}_{\text{flat}}(P)$ is gauge equivalent to a smooth connection. In particular, the natural inclusion $\mathcal{A}_{\text{flat}}(P) \hookrightarrow \mathcal{A}^{k-1,q}_{\text{flat}}(P)$ induces an identification $M(P) = \mathcal{A}^{k-1,q}_{\text{flat}}(P)/\mathcal{G}^{k,q}_0(P)$ for any k, q with $kq > 2$. Moreover, the projection

$$\Pi : \mathcal{A}^{1,q}_{\text{flat}}(P) \longrightarrow M(P)$$

is smooth in the given topologies. The composition $\Pi \circ \text{NS}_P$ is exactly the Narasimhan-Seshadri correspondence of [29].

Proposition 3.2. *Suppose Σ is a closed, connected, oriented Riemannian surface, and $P \to \Sigma$ is a principal $\text{PU}(r)$-bundle with $t_2(P)[\Sigma] \in \mathbb{Z}_r$ a generator. Let $1 < q < \infty$ and suppose α is in the domain of NS_P. Then the following hold.*

(i) *The linearization $D_\alpha(\Pi \circ \text{NS}_P)$ is complex-linear*

$$*D_\alpha(\Pi \circ \text{NS}_P) = D_\alpha(\Pi \circ \text{NS}_P)*,$$

where $*$ is the Hodge star on Σ. The kernel of $D_\alpha(\Pi \circ \text{NS}_P)$ is the space of exact/coexact 1-forms:

$$\ker D_\alpha(\Pi \circ \text{NS}_P) = \text{im}\, d_\alpha \oplus \text{im}\, d^*_\alpha.$$

(ii) *There is a constant C such that*

$$\left| D_\alpha(\Pi \circ \text{NS}_P)(\eta) \right|_{M(P)} \leq C\|\eta\|_{L^q(\Sigma)} \tag{20}$$

for all 1-forms $\eta \in W^{1,q}(T^*\Sigma \otimes P(\mathfrak{g}))$.

(iii) *There is a constant $\epsilon_0 > 0$ and a continuous function $f : \mathcal{A}^{0,2q}(P) \to \mathbb{R}^{\geq 0}$ such that for each $\alpha \in \mathcal{A}^{1,q}(P)$ with $\|F_\alpha\|_{L^{2q}(\Sigma)} < \epsilon_0$, the estimate*

$$\|\text{proj}_\alpha \eta - D_\alpha(\Pi \circ \text{NS}_P)\eta\|_{L^q(\Sigma)} \leq f(\alpha)\|\text{proj}_\alpha \eta\|_{L^q(\Sigma)} \tag{21}$$

holds for all $\eta \in L^q(T^*\Sigma \otimes P(\mathfrak{g}))$. Here proj_α is the projection defined in Lemma 3.5, below.

The function f can chosen so that $f(\alpha) \to 0$ as $\|F_\alpha\|_{L^{2q}(\Sigma)} \to 0$ in the following sense: for every $\epsilon > 0$, there is a $\delta > 0$ so that if α is a connection with $\|F_\alpha\|_{L^{2q}(\Sigma)} < \delta$, then $f(\alpha) < \epsilon$.

Since NS_P is smooth in the $W^{1,q}$-topologies, the estimate (20) would be obvious if the L^q-norm on the right were a $W^{1,q}$-norm. In our applications, we will use (20) with $q = 2$ in order to obtain energy estimates for various derivatives. The estimate (21) combines with (19) to show that, to first order, the map NS_P is the identity plus the L^2-orthogonal projection to the tangent space of flat connections.

We prove Proposition 3.2 in Section 3.4. The preparatory Section 3.3 addresses various elliptic properties for connections that are almost flat. In particular, it introduces the projection proj_α from Proposition 3.2 (iii).

3.1. **The complexified gauge group.** Here we review the construction of the complexified gauge group. We refer the reader to [7, Chapter 6] for an overview.

Let G be a compact, connected Lie group and fix a faithful representation $\rho : G \hookrightarrow \mathrm{U}(n)$ for some n. We identify G with its image in $\mathrm{U}(n)$.

Let Σ be a closed, connected, oriented surface, and fix a principal G-bundle $P \to \Sigma$. Using the standard representation of $\mathrm{U}(n)$ on $\mathbb{C}^n$, define the associated bundle $E := P \times_G \mathbb{C}^n$. Then E inherits a (real) inner product $\langle \cdot, \cdot \rangle$ as well as a complex structure J_E.

Fix a complex structure j_Σ on Σ. We denote by $\mathcal{C}(E)$ the space of Cauchy-Riemann operators, which are maps $\overline{D} : \Omega^0(\Sigma, E) \to \Omega^{0,1}(\Sigma, E)$ satisfying the Leibniz rule

$$\overline{D}(f\xi) = f\overline{D}\xi + (\overline{\partial} f)\xi$$

for $\xi \in \Omega^0(\Sigma, E)$ and $f : \Sigma \to \mathbb{C}$. Then $\mathcal{C}(E)$ can be identified with the space $\mathcal{A}(E)$ of Hermitian $\mathbb{C}$-linear covariant derivatives on E. This identification is given by

$$\mathcal{A}(E) \xrightarrow{\cong} \mathcal{C}(E), \quad D \longmapsto \frac{1}{2}\left(D + J_E D \circ j_\Sigma\right). \tag{22}$$

See [7, Lemma 2.1.54]. Let $\mathcal{G}(E)$ (resp. $\mathcal{G}(E)^{\mathbb{C}}$) denote the real (resp. complexified) gauge group of E. This is a Lie group with Lie algebra given by the space of sections of $P \times_G \mathfrak{u}(n)$ (resp. $P \times_G \mathrm{End}(\mathbb{C}^n)$). The exponential map is given pointwise by the matrix exponential map on $\mathrm{End}(\mathbb{C}^n)$. Our convention is that this association is given by

$$\xi \longmapsto \exp(\xi).$$

The space $\mathcal{G}(E)^{\mathbb{C}}$ acts on $\mathcal{C}(E)$ by the map

$$\mathcal{G}(E)^{\mathbb{C}} \times \mathcal{C}(E) \longrightarrow \mathcal{C}(E), \quad (\mu, \overline{D}) \longmapsto \mu^{-1} \circ \overline{D} \circ \mu. \tag{23}$$

We have made this a right action so that it is consistent with the action of the real gauge group on the space of connections.

We want to restrict to the subspace of $\mathcal{C}(E)$ that remembers the bundle P. To do this, let $P(\mathfrak{g})^{\mathbb{C}}$ denote the complexification of the vector bundle $P(\mathfrak{g})$. Then we have bundle inclusions

$$P(\mathfrak{g}) \subset P(\mathfrak{g})^{\mathbb{C}} \subset \mathrm{End}(E),$$

where $\mathrm{End}(E)$ is the bundle of complex linear automorphisms of E and the second inclusion is induced by the embedding ρ. Each connection $\alpha \in \mathcal{A}(P)$ induces a covariant derivative

$$d_{\alpha,\rho} : \Omega^k(\Sigma, E) \to \Omega^{k+1}(\Sigma, E),$$

and so we have a map $\mathcal{A}(P) \to \mathcal{A}(E)$. Furthermore, this map is an embedding of $\Omega^1(\Sigma, P(\mathfrak{g}))$-affine spaces, where $\Omega^1(\Sigma, P(\mathfrak{g}))$ acts on $\mathcal{A}(E)$ via

the inclusion $\Omega^1(\Sigma, P(\mathfrak{g})) \subseteq \Omega^1(\Sigma, \mathrm{End}(E))$. In particular, restricting to the image of $\mathcal{A}(P)$ in $\mathcal{A}(E)$, the map (22) takes the form

$$\mathcal{A}(P) \longrightarrow \mathcal{C}(E), \quad \alpha \longmapsto \overline{\partial}_\alpha := \frac{1}{2}\left(d_{\alpha,\rho} + J_E d_{\alpha,\rho} \circ j_\Sigma\right) \tag{24}$$

The image of (24) is the set of Cauchy Riemann operators that preserve the G-structure, and we denote this image by $\mathcal{C}(P)$. See [26, Appendix C] for the case when $G = \mathrm{U}(n)$. The space $\mathcal{C}(P)$ is an affine space modeled on the space $\Omega^{0,1}(\Sigma, P(\mathfrak{g})^{\mathbb{C}})$ of anti-linear 1-forms. Similarly, $\mathcal{A}(P)$ is an affine space modeled on $\Omega^1(\Sigma, P(\mathfrak{g}))$, and (24) intertwines these affine actions, where we identify $\Omega^1(\Sigma, P(\mathfrak{g}))$ with $\Omega^{0,1}(\Sigma, P(\mathfrak{g})^{\mathbb{C}})$ in the same way real 1-forms are identified with $(0,1)$-forms. To summarize, we have a commutative diagram

$$\begin{array}{ccc} \mathcal{A}(P) & \xrightarrow{\cong} & \mathcal{C}(P) \\ \downarrow & & \downarrow \\ \mathcal{A}(E) & \xrightarrow{\cong} & \mathcal{C}(E) \end{array}$$

where the vertical arrows are inclusions and everything is equivariant with respect to the action of $\Omega^1(\Sigma, P(\mathfrak{g}))$.

Let $\alpha \in \mathcal{A}(P)$ be a connection on P with curvature $F_\alpha \in \Omega^2(\Sigma, P(\mathfrak{g}))$. Consider the associated covariant derivative $d_{\alpha,\rho} \in \mathcal{A}(E)$ as well as its curvature $F_{\alpha,\rho} = d_{\alpha,\rho} \circ d_{\alpha,\rho} \in \Omega^2(\Sigma, \mathrm{End}(E))$. Since the representation ρ is faithful, we have pointwise estimates of the form

$$c|F_{\alpha,\rho}| \leq |F_\alpha| \leq C|F_{\alpha,\rho}|;$$

this allow us to discuss curvature bounds in terms of either F_α or $F_{\alpha,\rho}$.

In a similar vein, we define the *complexified gauge group* $\mathcal{G}(P)^{\mathbb{C}}$ to be the space of sections of the bundle $P \times_G G^{\mathbb{C}}$, where $G^{\mathbb{C}}$ is the complexification of the Lie group G. See [21] or [20] for more details. The Lie algebra of the complexified gauge group can be identified with the space $\Omega^0(\Sigma, P(\mathfrak{g})^{\mathbb{C}})$. We have the obvious inclusions

$$\begin{array}{ccc} \mathcal{G}(P) & \longrightarrow & \mathcal{G}(P)^{\mathbb{C}} \\ \downarrow & & \downarrow \\ \mathcal{G}(E) & \longrightarrow & \mathcal{G}(E)^{\mathbb{C}} \end{array}$$

We note that each element of $\mathcal{G}(P)^{\mathbb{C}}$ can be written uniquely as

$$\mu \exp(i\xi)$$

for some $\mu \in \mathcal{G}(P)$ and $\xi \in \Omega^0(\Sigma, P(\mathfrak{g}))$. This follows from the analogous statement for the finite dimensional case of $\mathrm{U}(n)^{\mathbb{C}}$.

The action (23) determines a (right) action of $\mathcal{G}(P)^{\mathbb{C}}$ on $\mathcal{A}(P)$. This follows because (i) the action of the real gauge group $\mathcal{G}(P) \subset \mathcal{G}(P)^{\mathbb{C}}$ preserves

the complex structure on $\mathcal{A}(P)$, and (ii) the induced action from (23) of the Lie *algebra* of $\mathcal{G}(P)^{\mathbb{C}}$ is

$$\Omega^0\left(\Sigma, P(\mathfrak{g})^{\mathbb{C}}\right) \times \mathcal{A}(P) \longrightarrow \mathcal{A}(P), \quad (\xi + i\zeta, \alpha) \longmapsto \alpha + d_\alpha \xi + *_\Sigma d_\alpha \zeta,$$

which plainly preserves $\mathcal{A}(P)$. We will write $\mu^*\alpha$ for the action of a complex gauge transformation μ on a connection α; this is consistent with the notation used for the action of the real gauge group. Explicitly, identifying $\mathcal{A}(P)$ with its image in $\mathcal{A}(E)$, the action on $\mathcal{A}(P)$ takes the form

$$d_{\mu^*\alpha,\rho} = \mu^{-1} \circ \overline{\partial}_\alpha \circ \mu + \mu^\dagger \circ \partial_\alpha \circ \left(\mu^\dagger\right)^{-1} = d_{\alpha,\rho} + \lambda - \lambda^\dagger.$$

where $\partial_\alpha := \frac{1}{2}(d_{\alpha,\rho} - J_E d_{\alpha,\rho} \circ j_\Sigma)$, and we have set

$$\lambda := \mu^{-1}\overline{\partial}_\alpha \mu.$$

The curvature transforms under $\mu \in \mathcal{G}(P)^{\mathbb{C}}$ by

$$F_{\mu^*\alpha,\rho} = F_{\alpha,\rho} + d_{\alpha,\rho}\left(\lambda - \lambda^\dagger\right) + \frac{1}{2}\left[\lambda \wedge \lambda\right]. \tag{25}$$

3.2. **Proof of the Narasimhan-Seshadri Theorem 3.1.** Fix a flat connection, and use this to define the $W^{1,q}$-norm on $\mathcal{A}^{1,q}(P)$. The specific flat connection chosen will not be relevant to our proof.

Suppose we can define NS_P on the set

$$\left\{\alpha \in \mathcal{A}^{1,q}(P) \,\middle|\, \mathrm{dist}_{W^{1,q}}\left(\alpha, \mathcal{A}^{1,q}_{\mathrm{flat}}(P)\right) < \epsilon_0\right\}, \tag{26}$$

for some $\epsilon_0 > 0$, and show that it satisfies the desired properties on this smaller domain. Then the $\mathcal{G}^{2,q}$-equivariance will imply that it extends uniquely to the flow-out by the real gauge group:

$$\left\{\mu^*\alpha \in \mathcal{A}^{1,q}(P) \,\middle|\, \mu \in \mathcal{G}^{2,q}(P), \;\; \mathrm{dist}_{W^{1,q}}\left(\alpha, \mathcal{A}^{1,q}_{\mathrm{flat}}(P)\right) < \epsilon_0\right\},$$

and continues to have the desired properties on this larger domain. The next claim shows that this flow-out contains a neighborhood of the form appearing in the domain in (18), thereby reducing the problem to defining NS_P on a set of the form (26).

Claim: For any $\widetilde{\epsilon}_0 > 0$, there is some $\epsilon_0 > 0$ with

$$\begin{aligned}&\left\{\alpha \in \mathcal{A}^{1,q}(P) \middle| \, \|F_\alpha\|_{L^q} < \epsilon_0\right\} \\ &\qquad \subseteq \left\{\mu^*\alpha \in \mathcal{A}^{1,q}(P) \,\middle|\, \mu \in \mathcal{G}^{2,q}(P), \;\; \mathrm{dist}_{W^{1,q}}\left(\alpha, \mathcal{A}^{1,q}_{\mathrm{flat}}(P)\right) < \widetilde{\epsilon}_0\right\}.\end{aligned}$$

This claim follows from a straight-forward patching argument, together with Uhlenbeck's gauge fixing theorem for coordinate balls; see [7, Theorem 2.3.7].

To define NS_P, it therefore suffices to show that for α sufficiently $W^{1,q}$-close to $\mathcal{A}_{\mathrm{flat}}(P)$, there is a unique $\Xi(\alpha) \in \Omega^0(\Sigma, P(\mathfrak{g}))$ close to 0, with $F_{\exp(i\Xi(\alpha))^*\alpha,\rho} = 0$. Once we have shown this, then we will define

$$\mathrm{NS}_P(\alpha) := \exp(i\Xi(\alpha))^*\alpha.$$

Consider the map $\mathcal{F}$ sending $(\alpha, \xi) \in \mathcal{A}(P) \times \Omega^0(\Sigma, P(\mathfrak{g}))$ to

$$\mathcal{F}(\alpha, \xi) := *F_{\exp(i\xi)^*\alpha,\rho}. \tag{27}$$

Then finding $\Xi(\alpha)$ is equivalent to solving for ξ in $\mathcal{F}(\alpha, \xi) = 0$; simply set $\Xi(\alpha) := \xi$. We will appeal to the implicit function theorem to solve for ξ, and so we need to pass to suitable Sobolev completions.

It follows from (25), and the Sobolev embedding and multiplication theorems, that $\mathcal{F}$ extends to a smooth map $\mathcal{A}^{1,q}(P) \times \mathrm{Lie}(\mathcal{G}(P))^{2,q} \to \mathrm{Lie}(\mathcal{G}(P))^{0,q}$, whenever $q > 1$. Suppose $\alpha_\flat$ is a flat connection. By (25), the linearization of $\mathcal{F}$ at $(\alpha_\flat, 0)$ in the direction of $(0, \xi)$ is

$$D_{(\alpha_\flat,0)}\mathcal{F}(0, \xi) = *id_{\alpha_\flat,\rho}\left(\overline{\partial}_{\alpha_\flat}\xi - \partial_{\alpha_\flat}\xi\right) = d^*_{\alpha_\flat,\rho}d_{\alpha_\flat,\rho}\xi.$$

where, in the second equality, we used the Kähler identities (see [7, p.213] and note that, since we are in dimension two, the Kähler operator Λ on $\Omega^{1,1} = \Omega^2$ is just the Hodge star). By assumption, all flat connections are irreducible, so Hodge theory for the Laplacian $d^*_{\alpha_\flat,\rho}d_{\alpha_\flat,\rho}$ tells us that this linearized operator

$$D_{(\alpha_\flat,0)}\mathcal{F}(0, \cdot) : \mathrm{Lie}(\mathcal{G}(Q))^{2,q} \longrightarrow \mathrm{Lie}(\mathcal{G}(Q))^{0,q}$$

is an isomorphism. Since $\alpha_\flat$ is flat, the pair $(\alpha_\flat, 0)$ is clearly a solution to $\mathcal{F}(\alpha, \xi) = 0$. It therefore follows by the implicit function theorem that there are $\epsilon_{\alpha_\flat}, \epsilon'_{\alpha_\flat} > 0$ such that, for any $\alpha \in \mathcal{A}^{1,q}$ with $\|\alpha - \alpha_\flat\|_{W^{1,q}} < \epsilon_{\alpha_\flat}$, there is a unique $\Xi = \Xi(\alpha) \in \mathrm{Lie}(\mathcal{G}(P))^{2,q}$ with $\|\Xi(\alpha)\|_{W^{2,q}} < \epsilon'_{\alpha_\flat}$ and $\mathcal{F}(\alpha, \Xi(\alpha)) = 0$. The implicit function theorem also implies that $\Xi(\alpha)$ varies smoothly in α in the $W^{1,q}$-topology. Moreover, by the uniqueness assertion, it follows that $\Xi(\alpha) = 0$ if α is flat.

We need to show that $\epsilon_{\alpha_\flat}$ and $\epsilon'_{\alpha_\flat}$ can be chosen to be independent of $\alpha_\flat \in \mathcal{A}^{1,q}_{\mathrm{flat}}(P)$. Since the moduli space $\mathcal{A}_{\mathrm{flat}}/\mathcal{G}$ of flat connections on P is compact, it suffices to show that $\epsilon_{\alpha_\flat} = \epsilon_{\mu^*\alpha_\flat}$, for all real gauge transformations $\mu \in \mathcal{G}^{2,q}(P)$, and likewise for $\epsilon'_{\alpha_\flat}$. Fix $\mu \in \mathcal{G}^{2,q}(P)$ and α a connection that is $W^{1,q}$-close to $\alpha_\flat$, then find $\Xi(\alpha)$ as above. These satisfy

$$\exp(i\Xi(\alpha))\mu = \mu\exp(i\mathrm{Ad}(\mu^{-1})\Xi(\alpha)).$$

Since the curvature is $\mathcal{G}^{2,q}(P)$-equivariant, we also have

$$0 = \mathrm{Ad}(\mu^{-1})F_{\exp(i\Xi)^*\alpha} = F_{(\exp(i\Xi)\mu)^*\alpha} = F_{\exp(i\mathrm{Ad}(\mu^{-1})\Xi)^*(\mu^*\alpha)}.$$

It follows that $\Xi(\mu^*\alpha) = \mathrm{Ad}(\mu^{-1})\Xi(\alpha)$ since $\Xi(\mu^*\alpha)$ is uniquely defined by $F_{\exp(i\Xi(\mu^*\alpha))^*(\mu^*\alpha)} = 0$. We therefore have $\epsilon_{\mu^*\alpha_\flat} = \epsilon_{\alpha_\flat}$ and $\epsilon'_{\mu^*\alpha_\flat} = \epsilon'_{\alpha_\flat}$, so we can take ϵ_0 to be the minimum of

$$\inf_{[\alpha_\flat]\in\mathcal{A}_{\text{flat}}/\mathcal{G}} \epsilon_{\alpha_\flat} > 0 \quad \text{and} \quad \inf_{[\alpha_\flat]\in\mathcal{A}_{\text{flat}}/\mathcal{G}} \epsilon'_{\alpha_\flat} > 0.$$

This argument also shows that NS_P is $\mathcal{G}^{2,q}(P)$-equivariant. This finishes the proof of (i) in the statement of Theorem 3.1.

To prove (ii), it suffices to show that the map $\alpha \mapsto \Xi(\alpha)$ extends to a map $\mathcal{A}^{0,p}(P) \to \mathrm{Lie}(\mathcal{G}(P))^{1,p}$ that is smooth with respect to the specified topologies. To see this, note that $\mathcal{F}$ from (27) is well-defined as a map

$$\mathcal{A}^{0,p}(P) \times \mathrm{Lie}(\mathcal{G}(P))^{1,p} \longrightarrow \mathrm{Lie}(\mathcal{G}(P))^{-1,p}.$$

and is smooth with respect to the specified topologies (the restriction to $p > 2$ is required so that Sobolev multiplication is well-defined). Then the implicit function theorem argument we gave above holds verbatim to show that for each α sufficiently L^p-close to $\mathcal{A}^{0,p}_{\text{flat}}(P)$, there is a unique $W^{1,p}$-small $\widetilde{\Xi}(\alpha) \in \mathrm{Lie}(\mathcal{G}(P))^{1,p}$ such that $\exp(i\widetilde{\Xi}(\alpha))^*\alpha$ is flat. Moreover, the assignment

$$\mathcal{A}^{0,p}_{\text{flat}}(P) \longrightarrow \mathrm{Lie}(\mathcal{G}(P))^{1,p}, \quad \alpha \longmapsto \widetilde{\Xi}(\alpha)$$

is smooth. The uniqueness of $\widetilde{\Xi}(\alpha)$ and $\Xi(\alpha)$ ensures that the former is indeed an extension of the latter. This proves (ii).

Remark 3.3. *Let $\Pi : \mathcal{A}^{1,q}_{\text{flat}}(P) \to \mathcal{A}^{1,q}_{\text{flat}}(P)/\mathcal{G}^{2,q}-0(P)$ denote the projection. The above proof shows that the composition $\Pi \circ \mathrm{NS}_P$ is invariant under a small neighborhood of $\mathcal{G}^{2,q}_0(P)$ in $\mathcal{G}^{2,q}(P)^{\mathbb{C}}$. Indeed, α and $\exp(i\xi)^*\alpha$ both map to the same flat connection under NS_P whenever they are both in the domain of NS_P.*

More generally, differentiating the identities

$$\mathrm{NS}_P(\exp(t\phi)^*\alpha) = \exp(t\phi)^*\mathrm{NS}_P(\alpha), \quad \mathrm{NS}_P(\exp(it\phi)^*\alpha) = \mathrm{NS}_P(\alpha)$$

at $t = 0$ gives

$$D_\alpha\mathrm{NS}_P(d_\alpha\phi) = d_{\mathrm{NS}_P(\alpha)}\phi, \quad D_\alpha\mathrm{NS}_P(*d_\alpha\phi) = 0,$$

where $D_\alpha\mathrm{NS}_P$ is the linearization of NS_P at α.

Now we prove (iii). When $p > 2$, a straightforward contradiction argument using Uhlenbeck compactness shows that there are constants $C_0, \epsilon_0 > 0$ so that if α is any connection with $\|F_\alpha\|_{L^p} < \epsilon_0$, then there is a flat connection $\alpha_\flat$ with

$$\|\alpha - \alpha_\flat\|_{\mathcal{C}^0(\Sigma)} \leq C_0\|F_\alpha\|_{L^p(\Sigma)}. \tag{28}$$

(See [10, Lemma 7.6] for a similar proof.) On the other hand, using the fact that NS_P is the identity on flat connections, we have

$$\begin{aligned}\|\alpha - \mathrm{NS}_P(\alpha)\|_{\mathcal{C}^0(\Sigma)} &\leq \|\alpha - \alpha_\flat\|_{\mathcal{C}^0(\Sigma)} + \|\mathrm{NS}_P(\alpha) - \mathrm{NS}_P(\alpha_\flat)\|_{\mathcal{C}^0(\Sigma)} \\ &\leq \|\alpha - \alpha_\flat\|_{\mathcal{C}^0(\Sigma)} + \|D_{\alpha_\flat}\mathrm{NS}_P(\alpha - \alpha_\flat)\|_{\mathcal{C}^0(\Sigma)} \\ &\leq (1 + C_1)\|\alpha - \alpha_\flat\|_{\mathcal{C}^0(\Sigma)},\end{aligned}$$

where we used (ii) from Theorem 3.1 (with $p = \infty$) and the mean value inequality for NS_P. The constant C_1 can be taken to be independent of the flat connection $\alpha_\flat$ because the moduli space of flat connections is compact. Combining this with (28) verifies (iii). □

3.3. **Analytic properties of almost flat connections.** In this section we establish several elliptic properties for connections with small curvature; these results are all standard for flat connections. Write $W^{k,q}(P(\mathfrak{g}))$ and $W^{k,q}(T^*\Sigma \otimes P(\mathfrak{g}))$ for the obvious Sobolev completions of $\Omega^0(\Sigma, P(\mathfrak{g}))$ and $\Omega^1(\Sigma, P(\mathfrak{g}))$. Fix a flat connection on P, and use this to define the Sobolev norms on these spaces; the results below are independent of the specific flat connection chosen.

The following lemma addresses elliptic regularity for the operator d_α on 0-forms.

Lemma 3.4. *Suppose G is a compact Lie group, Σ is a closed, oriented surface with metric, and $P \to \Sigma$ is a principal G-bundle such that all flat connections are irreducible. Let $1 < q < \infty$. Then there are constants $C > 0$ and $\delta_0 > 0$ so that the following holds for all $\alpha \in \mathcal{A}^{1,q}(P)$ with $\|F_\alpha\|_{L^q(\Sigma)} < \delta_0$.*

(i) *The covariant derivative $d_\alpha : W^{1,q}(P(\mathfrak{g})) \to W^{0,q}(T^*\Sigma \otimes P(\mathfrak{g}))$ is a Banach space isomorphism onto its image. Moreover, for all $f \in W^{1,q}(P(\mathfrak{g}))$ the following holds*

$$\|f\|_{W^{1,q}(\Sigma)} \leq C\|d_\alpha f\|_{L^q(\Sigma)}.$$

(ii) *The Laplacian $d_\alpha^* d_\alpha : W^{2,q}(P(\mathfrak{g})) \to W^{0,q}(P(\mathfrak{g}))$ is a Banach space isomorphism. Moreover, for all $f \in W^{2,q}(P(\mathfrak{g}))$ the following holds*

$$\|f\|_{W^{2,q}(\Sigma)} \leq C\|d_\alpha * d_\alpha f\|_{L^q(\Sigma)}.$$

Proof. This is basically the statement of [10, Lemma 7.6], but adjusted a little to suit our situation. We prove (ii), the proof of (i) is similar. The assumption that all flat connections $\alpha_\flat$ are irreducible implies that the kernel and cokernel of the elliptic operator $d_{\alpha_\flat}^* d_{\alpha_\flat} : W^{2,q}(P(\mathfrak{g})) \to L^q(P(\mathfrak{g}))$ are trivial. In particular, we have an estimate $\|f\|_{W^{2,q}} \leq C\|d_{\alpha_\flat}^* d_{\alpha_\flat} f\|_{L^q}$ for all $f \in W^{2,q}(P(\mathfrak{g}))$. The moduli space of flat connections is compact, so this constant C can be taken to be independent of $\alpha_\flat$. Hence the statement of the lemma holds when $\alpha = \alpha_\flat$ is flat.

To extend to arbitrary connections with small curvature, fix any $\alpha \in \mathcal{A}^{1,q}(P)$ with $\|F_\alpha\|_{L^q} < \delta_0$. By taking δ_0 sufficiently small, we can find a flat connection $\alpha_\flat$ so that (28) holds. Fix $f \in W^{2,q}(P(\mathfrak{g}))$. Then by the above discussion, and the relation $d_{\alpha_\flat} f = d_\alpha f + [\alpha_\flat - \alpha, f]$, we have

$$\begin{aligned}
\|f\|_{W^{2,q}} &\leq C\|d^*_{\alpha_\flat} d_{\alpha_\flat} f\|_{L^p} \\
&\leq C\Big\{ \|d^*_\alpha d_\alpha f\|_{L^q} \\
&\qquad + \|d^*_\alpha [\alpha - \alpha_\flat, f]\|_{L^q} + \|[\alpha - \alpha_\flat \wedge [*(\alpha - \alpha_\flat), f]]\|_{L^q} \Big\} \\
&\leq C\Big\{ \|d^*_\alpha d_\alpha f\|_{L^q} \\
&\qquad + C'\|f\|_{W^{2,q}} \Big(\|d^*_\alpha(\alpha - \alpha_\flat)\|_{L^q} \\
&\qquad\qquad + \|\alpha - \alpha_\flat\|_{L^q} + \|\alpha - \alpha_\flat\|^2_{L^{2q}} \Big) \Big\},
\end{aligned}$$

where we have used the embeddings $W^{2,q} \hookrightarrow W^{1,q}$ and $W^{2,q} \hookrightarrow L^\infty$ in the last line. By composing $\alpha_\flat$ with a suitable gauge transformation, we can assume that $\alpha_\flat$ is in Coulomb gauge relative to α:

$$d^*_\alpha(\alpha - \alpha_\flat) = 0,$$

while retaining the bound (28); see [38, Theorem 8.1]. We therefore have

$$\begin{aligned}
\|f\|_{W^{2,q}} &\leq C\|d^*_\alpha d_\alpha f\|_{L^q} + CC'\|f\|_{W^{2,q}}(\|\alpha - \alpha_\flat\|_{L^q} + \|\alpha - \alpha_\flat\|^2_{L^{2q}}) \\
&\leq C\|d^*_\alpha d_\alpha f\|_{L^q} + C''\delta_0\|f\|_{W^{2,q}},
\end{aligned}$$

where we used (28) and $\|F_\alpha\|_{L^q} < \delta_0$ in the second line. Taking δ_0 sufficiently small finishes the proof. □

Now we move on to study the action of d_α on 1-forms. First we establish a Hodge-decomposition result for connections with small curvature. The standard Hodge decomposition says

$$W^{k,q}(T^*\Sigma \otimes P(\mathfrak{g})) = H^1_{\alpha_\flat} \oplus \left(\operatorname{im} d_{\alpha_\flat}\right)^{k,q} \oplus \left(\operatorname{im} d^*_{\alpha_\flat}\right)^{k,q}, \tag{29}$$

for any flat connection $\alpha_\flat$. Here we are using the convention that superscripts k, q denote the $W^{k,q}$-Sobolev completion of the given space. Note also that $H^1_{\alpha_\flat}$ is finite dimensional, and so is equal to its own $W^{k,q}$-closure; moreover, the dimension of $H^1_{\alpha_\flat}$ is independent of the flat connection $\alpha_\flat$. The direct sum in (29) is L^2-orthogonal, even though the spaces need not be complete in the L^2-metric. We have a similar situation whenever α has small curvature, as the next lemma shows.

Lemma 3.5. *Assume that $P \to \Sigma$ satisfies the conditions of Lemma 3.4, and let $1 < q < \infty$ and $k \geq 0$. Then there are constants $\delta_0 > 0$ and $C > 0$ with the following significance. If $\alpha \in \mathcal{A}^{1,q}(P)$ has $\|F_\alpha\|_{L^q(\Sigma)} < \delta_0$, then*

$$H^1_\alpha := (\ker d_\alpha)^{k,q} \cap (\ker d^*_\alpha)^{k,q} \subseteq W^{k,q}(T^*\Sigma \otimes P(\mathfrak{g}))$$

has finite dimension equal to $\dim H^1_{\alpha_\flat}$*, for any flat connection* $\alpha_\flat$*. Furthermore, the space* H^1_α *equals the* L^2*-orthogonal complement of the image of* $d_\alpha \oplus d^*_\alpha$*:*

$$H^1_\alpha = \left((\operatorname{im} d_\alpha)^{k,q} \oplus (\operatorname{im} d^*_\alpha)^{k,q} \right)^\perp,$$

and so we have a direct sum decomposition

$$W^{k,q}\left(T^*\Sigma \otimes P(\mathfrak{g})\right) = H^1_\alpha \oplus \left((\operatorname{im} d_\alpha)^{k,q} \oplus (\operatorname{im} *d_\alpha)^{k,q} \right). \tag{30}$$

In particular, the L^2*-orthogonal projection*

$$\operatorname{proj}_\alpha : W^{k,q}(T^*\Sigma \otimes P(\mathfrak{g})) \longrightarrow H^1_\alpha$$

is well-defined.

Remark 3.6. *Elliptic regularity for* $d_\alpha \oplus d^*_\alpha$ *shows that the definition of* H^1_α *is independent of the choice of Sobolev constants* k, q*.*

Proof of Lemma 3.5. We first show that, when $\|F_\alpha\|_{L^q(\Sigma)}$ is sufficiently small, we have a direct sum decomposition

$$W^{k,q}\left(T^*\Sigma \otimes P(\mathfrak{g})\right) = H^1_\alpha \oplus (\operatorname{im} d_\alpha)^{k,q} \oplus (\operatorname{im} *d_\alpha)^{k,q}.$$

We prove this in the case $k = 0$; the case $k > 0$ is similar but slightly easier. By definition of H^1_α, it suffices to show that the images of d_α and $*d_\alpha$ intersect trivially. Towards this end, assume $d_\alpha f = *d_\alpha g$ for some 0-forms f, g of Sobolev class $L^q = W^{0,q}$. Acting by d_α and then $d_\alpha *$ gives

$$[F_\alpha, f] = d_\alpha * d_\alpha g, \qquad [F_\alpha, g] = -d_\alpha * d_\alpha f.$$

A priori, $d_\alpha * d_\alpha g$ and $d_\alpha * d_\alpha f$ are only of Sobolev class $W^{-2,q}$, however, the left-hand side of each of these equations is in L^r, where $1/r = 1/q + 1/p$. So elliptic regularity implies that f and g are each $W^{2,q}$. (This bootstrapping can be continued to show that f, g are smooth, but we will see in a minute that they are both zero.) By Lemma 3.4 and the embedding $W^{2,q} \hookrightarrow L^\infty$, it follows that, whenever $\|F_\alpha\|_{L^q}$ is sufficiently small, we have

$$\|f\|_{L^\infty} \leq C\|d_\alpha * d_\alpha f\|_{L^q} = C\| [F_\alpha, g] \|_{L^q} \leq 2C\|F_\alpha\|_{L^q}\|g\|_{L^\infty}.$$

Similarly, $\|g\|_{L^\infty} \leq 2C\|F_\alpha\|_{L^q}\|f\|_{L^\infty}$, and hence

$$\|f\|_{L^\infty} \leq 4C^2\|F_\alpha\|^2_{L^q}\|f\|_{L^\infty}.$$

If $\|F_\alpha\|^2_{L^q} < (2C)^{-2}$, then this can happen only if $f = g = 0$. This establishes the direct sum (30).

Now we prove that the dimension of H^1_α is finite and equals that of $H^1_{\alpha_\flat}$ for any flat connection $\alpha_\flat$. It is well-known that the operator

$$d_{\alpha_\flat} \oplus *d_{\alpha_\flat} : W^{k+1,q}(P(\mathfrak{g})) \oplus W^{k+1,q}(P(\mathfrak{g})) \longrightarrow W^{k,q}(T^*\Sigma \otimes P(\mathfrak{g}))$$

is elliptic, and hence Fredholm, whenever $\alpha_\flat$ is flat. The irreducibility condition implies that it has trivial kernel, and so has index given by $-\dim(H^1_{\alpha_\flat})$, which is a constant independent of $\alpha_\flat$. Then for any other connection α, the operator $d_\alpha \oplus *d_\alpha$ differs from $d_{\alpha_\flat} \oplus *d_{\alpha_\flat}$ by a compact operator, and so $d_\alpha \oplus *d_\alpha$ is Fredholm with the same index $-\dim(H^1_{\alpha_\flat})$ [26, Theorem A.1.5]. It follows from Lemma 3.4 that the (bounded) operator

$$d_\alpha \oplus *d_\alpha : W^{k+1,q}(P(\mathfrak{g})) \oplus W^{k+1,q}(P(\mathfrak{g})) \longrightarrow W^{k,q}(T^*\Sigma \otimes P(\mathfrak{g}))$$

is injective whenever $\|F_\alpha\|_{L^q(\Sigma)}$ is sufficiently small, and hence the cokernel has finite dimension $\dim(H^1_{\alpha_\flat})$. That is, $\dim(H^1_\alpha) = \dim(H^1_{\alpha_\flat})$, so this finishes the proof of Lemma 3.5. □

Next we show that the L^2-orthogonal projection to $H^1_\alpha = \ker d_\alpha \cap \ker d^*_\alpha$ depends smoothly on α in the L^q-topology.

Proposition 3.7. *Suppose that $P \to \Sigma$ and $\delta_0 > 0$ are as in Lemma 3.5, and let $1 < q < \infty$. Then the assignment $\alpha \mapsto \mathrm{proj}_\alpha$ is affine-linear and bounded*

$$\|\mathrm{proj}_\alpha - \mathrm{proj}_{\alpha'}\|_{\mathrm{op}} \leq C\|\alpha - \alpha'\|_{L^q(\Sigma)},$$

provided $\|F_\alpha\|_{L^q}, \|F_{\alpha'}\|_{L^q} < \delta_0$. Here $\|\cdot\|_{\mathrm{op}}$ is the strong operator norm on the space of linear maps $W^{0,q}(T^*\Sigma \otimes P(\mathfrak{g})) \to W^{0,q}(T^*\Sigma \otimes P(\mathfrak{g}))$.

Proof. We will see that the defining equations for proj_α are affine linear, and so the statement will follow from the implicit function theorem in the affine-linear setting.

First, we introduce the following shorthand:

$$W^{k,q}(\Omega^j) := W^{k,q}\left(\Lambda^j T^*\Sigma \otimes P(\mathfrak{g})\right), \quad L^q(\Omega^j) := W^{0,q}(\Omega^j).$$

Next, we note that for $\mu \in L^q(\Omega^1)$, the L^2-orthogonal projection $\mathrm{proj}_\alpha \mu$ is uniquely characterized by the following properties:

$$\underline{\text{Property A}}: \quad \exists (u,v) \in W^{1,q}(\Omega^0) \oplus W^{1,q}(\Omega^2),\ \mu - \mathrm{proj}_\alpha \mu = d_\alpha u + d^*_\alpha v,$$

$$\underline{\text{Property B}}: \quad \forall (a,b) \in W^{1,q^*}(\Omega^0) \oplus W^{1,q^*}(\Omega^2),\ \langle \mathrm{proj}_\alpha \mu,\ d_\alpha a + d^*_\alpha b\rangle = 0,$$

where q^* is the Sobolev dual to q: $1/q + 1/q^* = 1$. Here and below, we use $\langle \mu, \nu\rangle$ to denote the L^2-pairing on forms. This induces an identification of the dual space $(W^{k,q^*})^*$ with $W^{-k,q}$. Note that by Lemma 3.4 the operators d_α and d^*_α are injective on 0- and 2-forms, respectively, so any pair (u,v) satisfying Property A is unique.

Consider the map

$$\begin{aligned}\left(\mathcal{A}^{0,q} \times L^q(\Omega^1)\right) \times \left(L^q(\Omega^1) \times W^{1,q}(\Omega^0) \times W^{1,q}(\Omega^0)\right) \\ \longrightarrow W^{-1,q}(\Omega^0) \oplus W^{-1,q}(\Omega^2) \oplus L^q(\Omega^1)\end{aligned} \tag{31}$$

defined by

$$(\alpha, \mu; \nu, u, v) \longmapsto (d_\alpha^* \nu,\ d_\alpha \nu,\ \mu - \nu - d_\alpha u - d_\alpha^* v)\,.$$

The key point in this definition is that a tuple $(\alpha, \mu; \nu, u, v)$ maps to zero under (31) if and only if this tuple satisfies Properties A and B above.

Claim 1: The map (31) is bounded affine linear in the $\mathcal{A}^{0,q}$-variable, and bounded linear in the other 4 variables.

Claim 2: The linearization at $(\alpha, 0; 0, 0, 0)$ of (31) in the last 3-variables is a Banach space isomorphism, provided $\|\alpha - \alpha_\flat\|_{L^q}$ is sufficiently small for some flat connection $\alpha_\flat$.

Before proving the claims, we describe how they prove the lemma. Observe that $(\alpha, 0; 0, 0, 0)$ is clearly a zero of (31) for any α. Claim 1 implies that (31) is smooth, and so by Claim 2 we can use the implicit function theorem to show that, for each pair $(\alpha, \mu) \in \mathcal{A}^{0,q} \oplus L^q(\Omega^1)$, with $\|\alpha - \alpha_\flat\|_{L^q}$ sufficiently small, there is a unique $(\nu, u, v) \in L^q(\Omega^1) \oplus W^{1,q}(\Omega^0) \oplus W^{1,q}(\Omega^0)$ such that $(\alpha, \mu; \nu, u, v)$ is a zero of (31). (A priori this only holds for μ in a small neighborhood of the origin, but since (31) is linear in that variable, it extends to all μ.) It will then follow that $\nu = \mathrm{proj}_\alpha \mu$ depends smoothly on α in the L^q-metric. In fact, (31) is affine linear in α and linear in the other variables, so the uniqueness assertion of the implicit function theorem implies that proj_α depends affine-linearly on α, and so it follows that $\|\mathrm{proj}_\alpha - \mathrm{proj}_{\alpha_\flat}\|_{\mathrm{op}, L^q}$ is bounded by

$$\inf_{\|\mu\|_{L^q}=1} \left\| \left(\mathrm{proj}_\alpha - \mathrm{proj}_{\alpha_\flat}\right) \mu \right\|_{L^q} \leq C \inf_{\|\mu\|_{L^q}=1} \|\alpha - \alpha_\flat\|_{L^q} \|\mu\|_{L^q} = C \|\alpha - \alpha_\flat\|_{L^q}.$$

This proves the lemma for all α sufficiently L^q-close to $\mathcal{A}_{\mathrm{flat}}$. Now use (28) to conclude the result for all α with $\|F_\alpha\|_{L^q}$ sufficiently small.

Proof of Claim 1: It suffices to verify boundedness for each of the three (codomain) components separately. The first component is the map

$$\mathcal{A}^{0,q} \times L^q(\Omega^1) \longrightarrow W^{-1,q}(\Omega^0), \quad (\alpha, \nu) \longmapsto d_\alpha^* \nu \tag{32}$$

It is a standard consequence from the principle of uniform boundedness that a bilinear map is continuous if it is continuous in each variable separately. The same holds if the map is linear in one variable and affine-linear in the second, so it suffices to show that (32) is bounded in each of the two coordinates separately. Fix α and a flat connection $\alpha_\flat$. Then

$$\begin{aligned} \|d_\alpha \nu\|_{W^{-1,q}} &\leq \|d_{\alpha_\flat} \nu\|_{W^{-1,q}} + \| \left[\alpha - \alpha_\flat \wedge \nu\right] \|_{W^{-1,q}} \\ &\leq \|d_{\alpha_\flat} \nu\|_{W^{-1,q}} + 2 \|\alpha - \alpha_\flat\|_{L^q} \|\nu\|_{L^q} \\ &\leq C \left(1 + \|\alpha - \alpha_\flat\|_{L^q}\right) \|\nu\|_{L^q} \end{aligned}$$

which shows that the map is bounded in the variable ν, with α fixed. Next, fix ν and write

$$\|d_\alpha \nu - d_{\alpha_\flat}\nu\|_{W^{-1,q}} = \| \, [\alpha - \alpha_\flat \wedge \nu] \, \|_{W^{-1,q}} \leq 2\|\nu\|_{L^q}\|\alpha - \alpha_\flat\|_{L^q},$$

which shows it is bounded in the α-variable. This shows the first component of (31) is bounded. The other two components are similar.

Proof of Claim 2: The linearization of (31) at $(\alpha, 0; 0, 0, 0)$ in the last three variables is the map

$$\begin{array}{rcl} L^q(\Omega^1) \times W^{1,q}(\Omega^0) \times W^{1,q}(\Omega^0) & \longrightarrow & W^{-1,q}(\Omega^0) \oplus W^{-1,q}(\Omega^2) \oplus L^q(\Omega^1) \\ (\nu, u, v) & \longmapsto & (d_\alpha^*\nu, \; d_\alpha \nu, \; -\nu - d_\alpha u - d_\alpha^* v) \end{array}$$

By Claim 1, this is bounded linear, so by the open mapping theorem, it suffices to show that it is bijective. Suppose

$$(d_\alpha^*\nu, \; d_\alpha \nu, \; -\nu - d_\alpha u - d_\alpha^* v) = (0, 0, 0). \tag{33}$$

Then by Lemma 3.5, we can write ν uniquely as $\nu = \nu_H + d_\alpha a + d_\alpha^* b$ for $\nu_H \in H^1_\alpha = \ker d_\alpha \cap \ker d_\alpha^*$, and $(a, b) \in W^{1,q}(\Omega^0) \times W^{1,q}(\Omega^2)$, provided $\|F_\alpha\|_{L^q}$ is sufficiently small. This uniqueness, together with the first two components of (33), imply that $\nu = \nu_H$. The third component then reads

$$\nu_H = -d_\alpha u - d_\alpha^* v,$$

which is only possible if $\nu_H = d_\alpha u = d_\alpha^* v = 0$. By Lemma 3.4, this implies $(\nu, u, v) = (0, 0, 0)$, which proves injectivity.

To prove surjectivity, suppose the contrary. Then by the Hahn-Banach theorem, there are non-zero dual elements

$$(f, g, \eta) \in W^{1,q^*}(\Omega^0) \oplus W^{1,q^*}(\Omega^2) \oplus L^{q^*}(\Omega^1)$$

satisfying

$$0 = \langle f, d_\alpha^*\nu\rangle, \quad 0 = \langle g, d_\alpha \nu\rangle, \quad 0 = \langle \eta, \nu + d_\alpha u + d_\alpha^* v\rangle$$

for all (ν, u, v). The first two equations imply $\langle d_\alpha f, \nu\rangle = 0$ and $\langle d_\alpha^* g, \nu\rangle = 0$ for all ν. This implies $d_\alpha f = 0$ and $d^* g = 0$, and so $f = 0$ and $g = 0$ by Lemma 3.4. For the third equation, take $(u, v) = (0, 0)$ and we get $0 = \langle \eta, \nu\rangle$ for all ν. But this can only happen if $\eta = 0$, which is a contradiction to the tuple (f, g, η) being non-zero. □

We end this preparatory section by establishing the analogue of Lemma 3.4 for 1-forms.

Lemma 3.8. *Assume that $P \to \Sigma$ satisfies the conditions of Lemma 3.4, and let $1 < q < \infty$. Then there are constants $C > 0$ and $\delta_0 > 0$ such that*

$$\|\eta - \mathrm{proj}_\alpha \eta\|_{W^{1,q}(\Sigma)} \leq C \left(\|d_\alpha \eta\|_{L^q(\Sigma)} + \|d_\alpha * \eta\|_{L^q(\Sigma)} \right) \tag{34}$$

for all $\eta \in W^{1,q}(T^*\Sigma \otimes P(\mathfrak{g}))$ and all $\alpha \in \mathcal{A}^{1,q}(\Sigma)$ with $\|F_\alpha\|_{L^q(\Sigma)} < \delta_0$.

Proof. Since proj_α is the projection to the kernel of the operator $d_\alpha \oplus d_\alpha^*$, the estimate (34) follows immediately from the fact that this operator is Fredholm. That the constant can be taken to be independent of the connection α follows from (28), the identity $d_\alpha = d_{\alpha_\flat} + [\alpha - \alpha_\flat \wedge \cdot]$, and the compactness of the moduli space of flat connections. □

3.4. **Proof of Proposition 3.2.** We begin by proving (i). Let $\mathcal{G}_0^{2,q}(P)^{\mathbb{C}} \subseteq \mathcal{G}^{2,q}(P)^{\mathbb{C}}$ denote the identity component. This can be described as

$$\mathcal{G}_0^{2,q}(P)^{\mathbb{C}} = \left\{ \mu \exp(i\xi) \,\middle|\, \mu \in \mathcal{G}_0^{2,q}(P), \quad \xi \in W^{2,q}(P(\mathfrak{g})) \right\}.$$

Just as in the real case, this acts freely on the space of connections with sufficiently small curvature. Moreover, by Remark 3.3, the map NS_P is equivariant under a neighborhood of $\mathcal{G}_0^{2,q}(P)$ in $\mathcal{G}_0^{2,q}(P)^{\mathbb{C}}$. It follows that NS_P has a unique $\mathcal{G}_0^{2,q}(P)^{\mathbb{C}}$-equivariant extension to the flow-out

$$\mathcal{A}^{ss}(P) := \left(\mathcal{G}_0^{2,q}(P)^{\mathbb{C}}\right)^* \left\{ \alpha \in \mathcal{A}^{1,q}(P) \,\middle|\, \|F_\alpha\|_{L^q} < \epsilon_0 \right\}$$

of the domain of NS_P; the superscript stands for *semistable*. The group $\mathcal{G}_0^{2,q}(P)^{\mathbb{C}}$ acts freely on $\mathcal{A}^{ss}(P)$.

Consider the projection $\Pi^{\mathbb{C}} : \mathcal{A}^{ss}(P) \longrightarrow \mathcal{A}^{ss}(P)/\mathcal{G}_0^{2,q}(P)^{\mathbb{C}}$. Using NS_P, we have an identification $\mathcal{A}^{ss}(P)/\mathcal{G}_0^{2,q}(P)^{\mathbb{C}} \cong M(P)$, and hence a commutative diagram

$$\begin{array}{ccc} \mathcal{A}^{ss}(P) & \xrightarrow{\mathrm{NS}_P} & \mathcal{A}^{1,q}_{\mathrm{flat}}(P) \\ {\scriptstyle \Pi^{\mathbb{C}}}\downarrow & & \downarrow{\scriptstyle \Pi} \\ \mathcal{A}^{ss}(P)/\mathcal{G}_0^{2,q}(P)^{\mathbb{C}} & \xrightarrow{\cong} & \mathcal{A}^{1,q}_{\mathrm{flat}}(P)/\mathcal{G}_0^{2,q}(P) = M(P) \end{array}$$

The complex gauge group $\mathcal{G}(P)^{\mathbb{C}}$ acts on $\mathcal{C}(P)$, and hence $\mathcal{A}(P)$, in a way that preserves the complex structure, and this holds true in the Sobolev completions of these spaces. In particular, the infinitesimal action of $\mathcal{G}_0^{2,q}(P)^{\mathbb{C}}$ is complex linear. This implies that the linearization of $\Pi^{\mathbb{C}} : \mathcal{A}^{ss}(P) \to M(P)$ is complex-linear, but $\Pi^{\mathbb{C}} = \Pi \circ \mathrm{NS}_P$, so this shows $D_\alpha(\Pi \circ \mathrm{NS}_P)$ is complex-linear as well.

The inclusion

$$\mathrm{im}\, d_\alpha \oplus \mathrm{im}\, d_\alpha^* \subseteq \ker D_\alpha(\Pi \circ \mathrm{NS}_P)$$

now follows immediately from Remark 3.3. The reverse inclusion follows because the linear operator $D_\alpha(\Pi \circ \mathrm{NS}_P)$ maps surjectively onto $T_{\Pi \circ \mathrm{NS}_P(\alpha)} M(P)$, and the dimension of this tangent space equals the codimension of $\mathrm{im}\, d_\alpha \oplus \mathrm{im}\, d_\alpha^*$, by Lemma 3.5. This completes the proof of (i) in the statement of Proposition 3.2.

Now we prove (ii). Note that by Lemma 3.5 there is a decomposition

$$T_\alpha \mathcal{A}^{1,q}(P) = H^1_\alpha \oplus (\mathrm{im}\, d_\alpha \oplus \mathrm{im}\, d_\alpha^*),$$

whenever α has sufficiently small curvature. Moreover, the first summand is L^2-orthogonal. Denote by $\mathrm{proj}_\alpha : T_\alpha \mathcal{A}^{1,q}(P) \to H^1_\alpha$ the projection to the d_α-harmonic space, and note that this is continuous with respect to the L^q-norm on the domain and codomain. We claim that the operator

$$D_\alpha\left(\Pi \circ \mathrm{NS}_P\right) : T_\alpha \mathcal{A}^{1,q} \longrightarrow H_{\mathrm{NS}_P(\alpha)}$$

can be written as a composition

$$T_\alpha \mathcal{A}^{1,q} \longrightarrow H^1_\alpha \xrightarrow{M_\alpha} H_{\mathrm{NS}_P(\alpha)}$$

for some bounded linear map M_α, where the first map is proj_α. Indeed, we have

$$D_\alpha\left(\Pi \circ \mathrm{NS}_P\right)(\mu) = D_\alpha\left(\Pi \circ \mathrm{NS}_P\right)(\mathrm{proj}_\alpha \mu)$$

since the difference $\mu - \mathrm{proj}_\alpha \mu$ lies in the kernel of $D_\alpha(\Pi \circ \mathrm{NS}_P)$ by (i). So the claim follows by taking

$$M_\alpha := D_\alpha\left(\Pi \circ \mathrm{NS}_P\right)|_{H_\alpha}$$

to be the restriction.

Since M_α is a linear map between finite-dimensional spaces, it is bounded with respect to any norm. We take the L^q-norm on these harmonic spaces. Then $D_\alpha\left(\Pi \circ \mathrm{NS}_P\right)$ is the composition of two functions that are continuous with respect to the L^q-norm:

$$\begin{aligned} |D_\alpha\left(\Pi \circ \mathrm{NS}_P\right)\mu|_{M(P)} &= C\|D_\alpha\left(\Pi \circ \mathrm{NS}_P\right)\mu\|_{L^q(\Sigma)} \\ &= C\|M_\alpha \circ \mathrm{proj}_\alpha \mu\|_{L^q(\Sigma)} \quad \leq \quad C_\alpha \|\mu\|_{L^q(\Sigma)}. \end{aligned}$$

That this constant can be taken independent of α, for F_α sufficiently small, follows from (28), and the compactness of the moduli space of flat connections. Here one needs to use the fact that $D_\alpha(\Pi \circ \mathrm{NS}_P) = \mathrm{proj}_\alpha$ when α is a flat connection, and so this has norm 1 (which is clearly independent of α). This finishes the proof of (ii) in the statement of Proposition 3.2.

To prove (iii), note that each of the operators proj_α and $D_\alpha\left(\Pi \circ \mathrm{NS}_P\right)$ have kernel $\mathrm{im}(d_\alpha) \oplus \mathrm{im}(*d_\alpha)$. This gives

$$\begin{aligned} \|(\mathrm{proj}_\alpha - D_\alpha\left(\Pi \circ \mathrm{NS}_P\right))\eta\|_{L^q} &\leq C_1 \|(\mathrm{proj}_\alpha - D_\alpha\left(\Pi \circ \mathrm{NS}_P\right))\eta\|_{L^{2q}} \\ &= C_1 \| \left(\mathrm{proj}_\alpha - D_\alpha(\Pi \circ \mathrm{NS}_P)\right)\left(\mathrm{proj}_\alpha \eta\right) \|_{L^{2q}}. \end{aligned}$$

(We have converted to the L^{2q}-norm since $2 < 2q < \infty$, which will allow us to appeal to Theorem 3.1 (ii) shortly.) This last term is bounded by

$$C_1 \left\|(\mathrm{proj}_\alpha - D_\alpha\left(\Pi \circ \mathrm{NS}_P\right))\right\|_{\mathrm{op}, L^{2q}} \ \|\mathrm{proj}_\alpha \eta\|_{L^{2q}},$$

where $\|\cdot\|_{\mathrm{op}, L^{2q}}$ is the operator norm relative to the L^{2q}-topologies. On the finite-dimensional space H^1_α, the L^q- and L^{2q}-norms are equivalent: $\|\mathrm{proj}_\alpha \eta\|_{L^{2q}} \leq C_2 \|\mathrm{proj}_\alpha \eta\|_{L^q}$. This constant C_2 is independent of $\mathrm{proj}_\alpha \eta \in H_\alpha$, however it may depend on α. On the other hand, Proposition 3.7 tells

us that C_2 is independent of α provided $\|F_\alpha\|_{L^{2q}}$ is sufficiently small. So we have

$$\|\mathrm{proj}_\alpha \eta - D_\alpha \left(\Pi \circ \mathrm{NS}_P\right) \eta\|_{L^q} \leq f(\alpha) \|\mathrm{proj}_\alpha \eta\|_{L^q},$$

where we have set $f(\alpha) := C_1 C_2 \left\|(\mathrm{proj}_\alpha - D_\alpha \left(\Pi \circ \mathrm{NS}_P\right))\right\|_{\mathrm{op}, L^{2q}}$. By Theorem 3.1 (ii) and Proposition 3.7, the function $f(\alpha)$ depends continuously on α in the L^{2q}-topology. If $\alpha = \alpha_\flat$ is flat, then $D_\alpha \left(\Pi \circ \mathrm{NS}_P\right)$ equals the projection proj_α, and so $f(\alpha_\flat) = 0$. In particular, $f(\alpha) \to 0$ as α approaches $\mathcal{A}^{1,q}_{\mathrm{flat}}(P)$ in the L^{2q}-topology. The convergence statement relating $f(\alpha)$ and the curvature F_α now follows from (28). □

4. Proofs of the main results

In this section we prove main compactness results, Lemma 2.4 and Theorem 2.3. We refer freely to the notation from Section 2. In particular, we will write

$$A = a + p\, ds, \quad A|_{\mathbb{R}\times I\times \Sigma_\bullet} = \alpha + \phi\, ds + \psi\, dt$$

for the components of a connection A on $\mathbb{R} \times Y$, and

$$F_A = F_a + ds \wedge b_s, \quad F_A|_{\mathbb{R}\times I\times \Sigma_\bullet} = F_\alpha + ds \wedge \beta_s + dt \wedge \beta_t + ds \wedge dt\, \gamma,$$

for the components of the curvature of a connection A. These are related by the formulas

$$\begin{aligned} b_s &= \partial_s a - d_a p, \\ \beta_s = \partial_s \alpha - d_\alpha \phi, \quad \beta_t &= \partial_t \alpha - d_\alpha \psi, \quad \gamma = \partial_s \psi - \partial_t \phi - [\psi, \phi]. \end{aligned}$$

In the proofs of the main results, we will encounter ϵ-ASD connections that satisfies uniform estimates of the form

$$\sup_{\mathbb{R}\times I} \|\beta_s\|_{L^2(\Sigma_\bullet)} + \sup_{\mathbb{R}} \|b_s\|_{L^2(Y_\bullet)} \leq c_0 \tag{35}$$

for some fixed constant c_0. These will also have uniformly small slice-wise curvatures on $\Sigma_\bullet$ and $Y_\bullet$:

$$\sup_{\mathbb{R}\times I\times \Sigma_\bullet} |F_\alpha| + \sup_{\mathbb{R}\times Y_\bullet} |F_a| \leq \delta_0. \tag{36}$$

Here $\delta_0 > 0$ is a constant chosen so that if α and a satisfy (36), then there is some C for which

$$\|\rho\|_{L^p(\Sigma_\bullet)} \leq C \|d_\alpha \rho\|_{L^2(\Sigma_\bullet)}, \quad \|r\|_{L^q(Y_\bullet)} \leq C \|d_a r\|_{L^2(Y_\bullet)}$$

for all $1 \leq p < \infty, 1 \leq q \leq 6$, and all 0-forms ρ, r; see Lemma 3.4, and note that in dimension three Sobolev embedding $W^{1,2} \hookrightarrow L^q$ holds for $1 \leq q \leq 6$. The constants δ_0, C depend only on the bundle and the fixed metric. Note the similarities between (35-36), and hypotheses (i-iv) of Lemma 2.4. Intuitively, the conditions (36) assert that A is *almost* a representative of a map into the moduli space of flat connections with Lagrangian boundary

conditions. Condition (35) can be viewed as an analogue of a uniform bound on the energy density for such maps.

The proof of the Compactness Lemma 2.4 will proceed as follows; the details are carried out in Section 4.2. We will use the Narasimhan-Seshadri correspondence from Section 3 to push the α-component of A down to a map v from $\mathbb{R} \times I$ into the moduli space of flat connections; this uses the first estimate in (36). The properties of the Narasimhan-Seshadri correspondence will imply that v is holomorphic.

The holomorphic map v will typically not have Lagrangian boundary conditions. However, each restriction $v|_{\{s\}\times\partial I}$ is almost on the Lagrangians determined by $Y_\bullet$. To find these nearby points on the Lagrangians, we appeal to the Yang-Mills heat flow on the 3-manifold $Y_\bullet$ applied to the component a of A. This uses the second estimate in (36). The details of the heat flow are carried out in Section 4.1.

At this point, the conclusions (14) and (15) from the statement of Lemma 2.4 will follow from simple estimates on the Narasimhan-Seshadri correspondence and the Yang-Mills heat flow. This will establish certain pointwise convergence for the components of the ϵ-ASD connections. To prove Lemma 2.4, we will only need to establish (16) and (17), which posit certain pointwise convergence of the components of the *curvature*. The necessary estimates are provided by the following theorem.

Theorem 4.1. *(Elliptic Estimates Theorem) Fix a constant $c_0 > 0$, and let $\delta_0 > 0$ be as above. Then there are constants $\epsilon_0, C > 0$ so that*

$$\begin{aligned} &\|\nabla_s \beta_s\|_{L^2(K\times I\times \Sigma_\bullet)} + \|\nabla_t \beta_s\|_{L^2(K\times I\times \Sigma_\bullet)} \\ &\quad + \|\nabla_s^2 \beta_s\|_{L^2(K\times I\times \Sigma_\bullet)} + \|\nabla_t^2 \beta_s\|_{L^2(K\times I\times \Sigma_\bullet)} \\ &\quad\quad + \|\nabla_s b_s\|_{L^2(K\times Y_\bullet)} \\ &\leq C(1 + \mathrm{vol}(K) + E_\epsilon^{\mathrm{inst}}(A)) \end{aligned}$$

for all $0 < \epsilon < \epsilon_0$, all compact $K \subset \mathbb{R}$, and all ϵ-ASD connections A satisfying (35) and (36).

The above theorem will combine with the embeddings

$$W^{2,2}(K \times I) \hookrightarrow \mathcal{C}^0(K \times I), \quad W^{1,2}(K) \hookrightarrow \mathcal{C}^0(K)$$

to conclude (16) and (17) in the statement of the Compactness Lemma 2.4. We defer the proof of the Elliptic Estimates Theorem 4.1 until Section 4.4.

The proof of the Compactness Theorem 2.3 is essentially just a bubbling analysis. The idea is to rescale around any points for which the hypotheses of the Compactness Lemma 2.4 are not satisfied. Then we would like to show that at each point a bubble (as described in Section 2.5) appears. The difficulty is showing that each bubble has *positive* energy, which is a crucial ingredient in our compactness result (e.g., it is used to prove that

only finitely many bubbles can form). Positivity of energy will follow from the convergence statements (16) and (17) in the Compactness Lemma 2.4.

It will be convenient to use the ϵ-dependent norm

$$\|v\|^2_{L^2(U),\epsilon} := \int_U \langle v \wedge *_\epsilon v \rangle$$

where $U \subseteq \mathbb{R}\times Y$ is a measurable set, v is a form, and $*_\epsilon$ is the Hodge star on $\mathbb{R}\times Y$ induced from the metric ds^2+g_ϵ. We will also refer to the ϵ-dependent L^2-inner product $(\cdot,\cdot)_\epsilon$, and L^p-norms $\|\cdot\|_{L^p(U),\epsilon}$, which are defined in the obvious way. We drop the ϵ in the notation when we are working with the fixed metric $g = g_1$; that is,

$$\|\cdot\|_{L^p(U)} := \|\cdot\|_{L^p(U),1}.$$

(For example, all norms appearing in Theorem 4.1 are with respect to the fixed metric.) In particular, the scaling properties of the Hodge star imply that if v is a map from $\mathbb{R}$ to the space of k-forms on $Y_\bullet$, then

$$\|ds \wedge v\|^p_{L^p(\mathbb{R}\times Y_\bullet),\epsilon} = \|v\|^p_{L^p(\mathbb{R}\times Y_\bullet),\epsilon} = \epsilon^{3-pk}\|v\|^p_{L^p(\mathbb{R}\times Y_\bullet)}.$$

Similarly, if ν is a map from $\mathbb{R}\times I$ to the space of k-forms on $\Sigma_\bullet$, then

$$\|ds\wedge dt\wedge \nu\|^p_{L^p(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon} = \|\nu\|^p_{L^p(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon} = \epsilon^{2-pk}\|\nu\|^p_{L^p(\mathbb{R}\times I\times\Sigma_\bullet)}. \tag{37}$$

4.1. **The heat flow on cobordisms.** Suppose Q is principal G-bundle over a Riemannian 3-manifold Y. The *Yang-Mills heat flow* is the equation

$$\frac{d}{d\tau}a(\tau) = -d^*_{a(\tau)}F_{a(\tau)}, \quad a(0) = a, \tag{38}$$

where $\tau \mapsto a(\tau) \in \mathcal{A}(Q)$ is a path of connections, and $a \in \mathcal{A}(Q)$ is an initial condition. This is the negative gradient flow (relative to the L^2-metric) for the *Yang-Mills functional*

$$\mathcal{YM}_Q : \mathcal{A}(Q) \longrightarrow \mathbb{R}, \quad \mathcal{YM}_Q(a) = \frac{1}{2}\|F_a\|^2_{L^2(Y)}.$$

In his thesis [31], Råde proved the following existence and uniqueness result.

Theorem 4.2. *Suppose G is compact and Y is a closed, oriented manifold of dimension 3. Fix an initial condition $a \in \mathcal{A}^{1,2}(Q)$. Then (38) has a unique solution $\{\tau \mapsto a(\tau)\} \in C^0_{loc}\left([0,\infty), \mathcal{A}^{1,2}(Q)\right)$, with the further property that $F_{a(\cdot)} \in C^0_{loc}\left([0,\infty), L^2\right) \cap L^2_{loc}\left([0,\infty), W^{1,2}\right)$. Moreover, the limit $\lim_{\tau\to\infty} a(\tau)$ exists, is a critical point of the Yang-Mills functional, and varies continuously with the initial data a in the $W^{1,2}$-topology.*

Differentiating $\mathcal{YM}_Q(a(\tau))$ in τ and using (38) shows that $\mathcal{YM}_Q(a(\tau))$ decreases in τ. Moreover, it follows from Uhlenbeck's compactness theorem together with [31, Proposition 7.2] that the critical values of the Yang-Mills functional are discrete. Combining these two facts, we conclude that there

is some $\widetilde{\epsilon}_Q > 0$ such that if $\mathcal{YM}_Q(a) < \widetilde{\epsilon}_Q$, then the associated limiting connection $\lim_\tau a(\tau)$ is flat. The flow therefore defines a continuous gauge equivariant deformation retract

$$\text{Heat}_Q : \left\{ a \in \mathcal{A}^{1,2}(Q) \,|\, \mathcal{YM}_Q(a) < \widetilde{\epsilon}_Q \right\} \longrightarrow \mathcal{A}^{1,2}_{\text{flat}}(Q) \tag{39}$$

whenever Y is a closed 3-manifold.

In this section we prove a version of Råde's Theorem 4.2, but for bundles Q over 3-manifolds *with boundary*. The most natural boundary condition for our application is of Neumann type. This will allow us to use a reflection principle and thereby appeal directly to Råde's result for closed 3-manifolds. Before defining the specific boundary condition, we note that Råde's result holds with the $W^{1,2}$-topology. However, on 3-manifolds, not all $W^{1,2}$-sections are continuous. This makes the issue of boundary conditions rather subtle. One way to get around this is to observe that, in dimension 3, restricting $W^{1,2}$-functions to codimension-1 subspaces is in fact well-defined. We take an equivalent approach by considering the space $\mathcal{A}^{1,2}(Q, \partial Q)$, which we define to be the $W^{1,2}$-closure of the set of smooth connections $a \in \mathcal{A}(Q)$ that satisfy

$$\iota_{\partial_n} a|_U = 0 \tag{40}$$

on some neighborhood U of ∂Q (U may depend on a). Here $\partial_n \in \Gamma(TQ)$ is a fixed extension of the outward pointing unit normal of ∂Q; we may assume that the set U is always contained in the region in which ∂_n is non-zero. Use the normalized gradient flow of ∂_n to write $U = [0, \epsilon) \times \partial Q$, and let t denote the coordinate on $[0, \epsilon)$. Then in these coordinates we can write any connection as $a|_{\{t\} \times \partial Y} = \alpha(t) + \psi(t)\, dt$. Then (40) is equivalent to requiring $\psi(t) = 0$.

Set

$$\mathcal{A}^{1,2}_{\text{flat}}(Q, \partial Q) := \mathcal{A}^{1,2}(Q, \partial Q) \cap \mathcal{A}^{1,2}_{\text{flat}}(Q).$$

Both of the spaces $\mathcal{A}^{1,2}(Q, \partial Q)$ and $\mathcal{A}^{1,2}_{\text{flat}}(Q, \partial Q)$ admit the action of the subgroup $\mathcal{G}(Q, \partial Q) \subset \mathcal{G}(Q)$ consisting of gauge transformations that restrict to the identity in a neighborhood of ∂Q. (We are purposefully only working with the *smooth* gauge transformations here.) The main result of this section is the following.

Theorem 4.3. *Let G be a compact, connected Lie group, and $Q \to Y$ be a principal G-bundle over a compact, connected, oriented Riemannian 3-manifold Y with non-empty boundary.*

(i) *There is some $\epsilon_Q > 0$ and a continuous strong deformation retract*

$$\text{Heat}_Q : \left\{ a \in \mathcal{A}^{1,2}(Q, \partial Q) \,|\, \mathcal{YM}_Q(a) < \epsilon_Q \right\} \longrightarrow \mathcal{A}^{1,2}_{\text{flat}}(Q, \partial Q).$$

Furthermore, Heat_Q *intertwines the action of* $\mathcal{G}(Q, \partial Q)$.

(ii) *Fix a boundary component $\Sigma \subset \partial Y$. Then for every $\epsilon > 0$, there is some $\delta > 0$ such that if $a \in \mathcal{A}^{1,2}(Q, \partial Q)$ satisfies $\|F_a\|_{L^2(Y)} < \delta$, then*

$$\|(\mathrm{Heat}_Q(a) - a)\,|_\Sigma\|_{L^q(\Sigma)} < \epsilon,$$

for every $1 \leq q \leq 4$.

Remark 4.4. *N. Charalambous and L. Gross [5] have proven similar results for manifolds with boundary. They adopt a technique of Donaldson [6] that replaces (38) by a related parabolic equation, at the cost of breaking the gauge invariance. It seems their strategy is more convenient than Råde's approach when dealing with boundary value problems in general (Charalambous and Gross deal with Dirichlet and Marini boundary conditions, in addition to Neumann boundary conditions). As mentioned above, our proof relies on a simple doubling procedure for Neumann problems that allows us to appeal directly to Råde's Theorem 4.2.*

Proof of Theorem 4.3. Consider the *double* $Y^{(2)} := \overline{Y} \cup_{\partial Y} Y$, which is a closed 3-manifold. Denote by $\iota_Y : Y \hookrightarrow Y^{(2)}$ the inclusion of the second factor. We will identify Y with its image under ι_Y. There is a natural involution $\sigma : Y^{(2)} \to Y^{(2)}$ defined by switching the factors in the obvious way. Then $Y^{(2)}$ has a natural smooth structure making ι_Y smooth and σ a diffeomorphism (this is just the smooth structure obtained by choosing the same collar on each side of ∂Y). Clearly the map σ is orientation-reversing, satisfies $\sigma^2 = \mathrm{Id}$, and has fixed point set equal to ∂Y. Similarly, we can form $Q^{(2)} := \overline{Q} \cup_{\partial Q} Q$ and an involution $\widetilde{\sigma} : Q^{(2)} \to Q^{(2)}$. Then $Q^{(2)}$ is naturally a principal G-bundle over $Y^{(2)}$ and $\widetilde{\sigma}$ is a bundle map covering σ. Furthermore, $\widetilde{\sigma}$ commutes with the G-action on $Q^{(2)}$.

Though $\widetilde{\sigma}$ is not a gauge transformation (it does not cover the identity), it behaves as one in many ways. For example, since $\widetilde{\sigma}$ a bundle map, the space of connections $\mathcal{A}(Q^{(2)})$ is invariant under pullback by $\widetilde{\sigma}$. The action on covariant derivatives takes the form $d_{\widetilde{\sigma}^* a} = \sigma^* \circ d_a \circ \sigma^*$, where

$$\sigma^* : \Omega(Y^{(2)}, Q^{(2)}(\mathfrak{g})) \to \Omega(Y^{(2)}, Q^{(2)}(\mathfrak{g}))$$

is pullback by σ. The induced action on the tangent space $T_a\mathcal{A}(Q^{(2)}) = \Omega^1(Y^{(2)}, Q^{(2)}(\mathfrak{g}))$ is given by pullback by σ. Likewise, the curvature satisfies $F_{\widetilde{\sigma}^* a} = \sigma^* F_a$. In particular, the flow equation (38) on the double $Y^{(2)}$ is invariant under the action of $\widetilde{\sigma}$. We set

$$\epsilon_Q := \widetilde{\epsilon}_{Q^{(2)}}/2,$$

where $\widetilde{\epsilon}_{Q^{(2)}} > 0$ is as in (39).

Suppose $a \in \mathcal{A}^{1,2}(Q, \partial Q)$ has $\mathcal{YM}_Q(a) < \epsilon_Q$. Then a has a unique extension $a^{(2)}$ to the rest of $Q^{(2)}$, satisfying $\widetilde{\sigma}^* a^{(2)} = a^{(2)}$. We call $a^{(2)}$ the *double* of a.

Claim 1: The connection $a^{(2)}$ lies in the space $\mathcal{A}^{1,2}(Q^{(2)})$.

To see this, first suppose that a is smooth. Then the boundary condition on a implies that $a^{(2)}$ is continuous on all of $Q^{(2)}$ and smooth on the complement of $\partial Q \subset Q^{(2)}$. In particular, $a^{(2)}$ is of Sobolev class $W^{1,2}$ when a is smooth. More generally, note that every $a \in \mathcal{A}^{1,2}(Q, \partial Q)$ is a $W^{1,2}$-limit of smooth connections a_j whose normal component vanishes in a neighborhood of the boundary. By the linearity of the integral it is immediate that the doubles of the a_j converge to $a^{(2)}$ in $W^{1,2}$. This proves the claim.

Remark 4.5. *Note that, in general, the connection $a^{(2)}$ will not be smooth, even if a is. Indeed, the derivative of $a^{(2)}$ normal to $\Sigma \subset Y^{(2)}$ will have a jump discontinuity (unless this derivative vanishes). This phenomenon is typical in reflection principles.*

By assumption, we have

$$\mathcal{YM}_{Q^{(2)}}(a^{(2)}) < \widetilde{\epsilon}_{Q^{(2)}},$$

so by the discussion at the beginning of this section, there is a unique solution $a^{(2)}(\tau)$ to the flow equation (38) on the *closed* 3-manifold $Y^{(2)}$, with initial condition $a^{(2)}(0) = a^{(2)}$. Furthermore, the limit

$$\mathrm{Heat}_{Q^{(2)}}(a^{(2)}) := \lim_{\tau \to \infty} a^{(2)}(\tau)$$

exists and is flat. Since (38) is $\widetilde{\sigma}$-invariant, the uniqueness assertion guarantees that $\widetilde{\sigma}^* a^{(2)}(\tau) = a^{(2)}(\tau)$ for all τ. In particular,

$$\widetilde{\sigma}^* \mathrm{Heat}_{Q^{(2)}}(a^{(2)}) = \mathrm{Heat}_{Q^{(2)}}(a^{(2)}). \tag{41}$$

Define $\mathrm{Heat}_Q(a) := \mathrm{Heat}_{Q^{(2)}}(a^{(2)})\Big|_Q$. Then (41) shows that

$$\iota_{\partial_n} \mathrm{Heat}_Q(a)|_{\partial Y} = 0,$$

so Heat_Q does map into $\mathcal{A}^{1,2}_{\mathrm{flat}}(Q, \partial Q)$. Similarly, each gauge transformation $u \in \mathcal{G}(Q, \partial Q)$ has a unique extension to a $\widetilde{\sigma}$-invariant gauge transformation in $\mathcal{G}(Q^{(2)})$. In particular, $\mathrm{Heat}_Q(u^* a) = u^* \mathrm{Heat}_Q(a)$ follows from the $\mathcal{G}(Q^{(2)})$-equivariance of $\mathrm{Heat}_{Q^{(2)}}$. This finishes the proof of (i) in the statement of Theorem 4.3.

Now we prove (ii). By replacing Y with its double, we may assume that Y is closed and $\Sigma \subset Y$ is a closed two-sided surface. It therefore suffices to prove the following claim.

Claim 2: For every $\epsilon > 0$, there is some $\delta > 0$ such that if $a \in \mathcal{A}^{1,2}(Q)$ satisfies $\|F_a\|_{L^2(Y)} < \delta$, then $\|(\mathrm{Heat}_Q(a) - a)\,|_\Sigma\|_{L^q(\Sigma)} < \epsilon$, for every $1 \leq q \leq 4$.

For sake of contradiction, suppose there is some sequence $a_\nu \in \mathcal{A}^{1,2}(Q)$ with $\|F_{a_\nu}\|_{L^2} \to 0$, but

$$c_0 \leq \|(\mathrm{Heat}_Q(a_\nu) - a_\nu)\,|_\Sigma\|_{L^q(\Sigma)} \tag{42}$$

for some fixed $c_0 > 0$. By Uhlenbeck's weak compactness theorem, there is a sequence of gauge transformations $u_\nu \in \mathcal{G}^{2,2}$ such that $u_\nu^* a_\nu$ converges weakly in $W^{1,2}$ (hence strongly in L^r for $1 \leq r < 6$) to a limiting connection $a_\infty \in \mathcal{A}^{1,2}(Q)$, after possibly passing to a subsequence. Then a_∞ is necessarily flat. By redefining u_ν, if necessary, we may assume that each $u_\nu^* a_\nu$ is in Coulomb gauge with respect to a_∞, and still retain the fact that $u_\nu^* a_\nu$ converges to a_∞ strongly in L^r for $1 \leq r < 6$. Then

$$\begin{aligned} \|u_\nu^* a_\nu - a_\infty\|^2_{W^{1,2}} &= \|u_\nu^* a_\nu - a_\infty\|^2_{L^2} + \|d_{a_\infty}(u_\nu^* a_\nu - a_\infty)\|^2_{L^2} \\ &\leq C_1 \left(\|u_\nu^* a_\nu - a_\infty\|^2_{L^2} + \|F_{a_\nu}\|^2_{L^2} + \|u_\nu^* a_\nu - a_\infty\|^4_{L^4} \right) \end{aligned}$$

for some constant C_1. Observe that the right-hand side is going to zero, so a_ν is converging in $W^{1,2}$ to the space of flat connections, and so

$$\|a_\nu - (u_\nu^{-1})^* a_\infty\|_{W^{1,2}} \longrightarrow 0. \tag{43}$$

On the other hand, by the trace theorem [38, Theorem B.10], we have

$$\|\mathrm{Heat}_Q(a_\nu) - a_\nu|_\Sigma\|_{L^q(\Sigma)} \leq C_2 \|\mathrm{Heat}_Q(a_\nu) - a_\nu\|_{W^{1,2}(Y)} \tag{44}$$

for some C_2 depending only on Y and $1 \leq q \leq 4$. By (42), the left-hand side of (44) is bounded from below by c_0. Since Heat_Q is continuous in the $W^{1,2}$-topology, and restricts to the identity on the space of flat connections, there is some $\epsilon' > 0$ such that if a_ν is within ϵ' of the space of flat connections, then $C_2 \|\mathrm{Heat}_Q(a_\nu) - a_\nu\|_{W^{1,2}(Y)} \leq \frac{c_0}{2}$. By (43) a_ν is within ϵ' of $\mathcal{A}_{\mathrm{flat}}(Q)$ for ν large, and so we have a contradiction. □

The next lemma states that we can always put a connection $a \in \mathcal{A}(Q)$ in a gauge so that it is an element of $\mathcal{A}(Q, \partial Q)$. This is basically just a variation on the fact that connections can be put in temporal gauge, so we omit the proof. We state a version with an additional $\mathbb{R}$ parameter, since this is the context in which the lemma will be used.

Lemma 4.6. *Let $Q \to Y$ be as in Section 2.2. Then for every $A \in \mathcal{A}^{1,2}_{loc}(\mathbb{R} \times Q)$ there is an identity-component gauge transformation $U \in \mathcal{G}^{2,2}_{loc}(\mathbb{R} \times Q)$ with*

$$U^* A|_{\{s\} \times Y_{i(i+1)}} \in \mathcal{A}^{1,2}(Q_{i(i+1)}, \partial Q_{i(i+1)}), \quad \forall i \in \{0, \ldots, N-1\}, \forall s \in \mathbb{R}.$$

Furthermore, if A is smooth then $U^* A$ is smooth as well.

4.2. **Proof of the Compactness Lemma 2.4.** Let A_ν be as in the statement of Lemma 2.4, and view α_ν as a map from $\mathbb{R} \times I$ into the space of connections on the surface $\Sigma_\bullet$. Assumption (i) in the Compactness Lemma states that the curvatures of the α_ν are uniformly small, provided ν is sufficiently large. For such ν, it follows that the Narasimhan-Seshadri Theorem 3.1 applies to $\alpha_\nu(s,t)$ for all s, t. Consider the map NS_{P_j} constructed in Theorem 3.1 for the bundle $P_j \to \Sigma_j$, and let $\Pi : \mathcal{A}_{\mathrm{flat}}(P_j) \to M(P_j)$ be the

quotient map. Let M be the product of the $M(P_j)$ from Section 2.4, and define a map $v_\nu : \mathbb{R} \times I \to M$ by sending $(s,t) \in \mathbb{R} \times I$ to

$$\left(\Pi \circ \mathrm{NS}_{P_0}\left(\alpha_\nu(s,1-t)|_{\Sigma_0}\right), \ldots, \Pi \circ \mathrm{NS}_{P_{N-1}}\left(\alpha_\nu(s,t)|_{\Sigma_{N-1}}\right)\right); \tag{45}$$

the terms on Σ_j with j even have $1-t$, and those with j odd have t. Then Proposition 3.2 (i) implies that v_ν is holomorphic with respect to the complex structure J on M from Section 2.4.

We will denote by $|\cdot|_M$ the norm on the fibers of TM induced by J and the symplectic form. Then using Proposition 3.2, we have an (s,t)-pointwise estimate:

$$\begin{aligned} |\partial_s v_\nu|_M^2 &= \sum_j \|D_{\alpha_\nu}(\Pi \circ \mathrm{NS}_{P_j})(\partial_s \alpha_\nu)\|^2_{L^2(\Sigma_j)} \\ &= \sum_j \|D_{\alpha_\nu}(\Pi \circ \mathrm{NS}_{P_j})(\partial_s \alpha_\nu - d_{\alpha_\nu}\phi_\nu)\|^2_{L^2(\Sigma_j)} \\ &\leq C_0 \sum_j \|\partial_s \alpha_\nu - d_{\alpha_\nu}\phi_\nu\|^2_{L^2(\Sigma_j)}. \end{aligned}$$

Integrating these energy densities over S_0, this shows that the energy of v_ν is bounded by the energy of A_ν:

$$\int_{S_0} |\partial_s v_\nu|_M^2 \, ds \wedge dt \leq C_0 \sum_j \|\partial_s \alpha_\nu - d_{\alpha_\nu}\phi_\nu\|^2_{L^2(\Sigma_j),\epsilon_\nu} \leq C_0 \|F_{A_\nu}\|^2_{L^2(\mathbb{R}\times Y),\epsilon_\nu},$$

where we have converted to the ϵ_ν-dependent norm using (37). By assumption (v) in the statement of Lemma 2.4, this implies that the energies of the v_ν are uniformly bounded

$$\sup_\nu \int_{S_0} |\partial_s v_\nu|_M^2 \, ds \wedge dt < \infty.$$

Since $S_0 = \mathbb{R} \times I \backslash B$, where B is a finite set, it follows from the removal of singularities theorem for holomorphic maps [26, Theorem 4.1.2 (ii)] that each v_ν extends to a holomorphic map defined on all of the interior $\mathbb{R} \times (0,1)$.

Remark 4.7. *Note that since v_ν does not necessarily have Lagrangian boundary conditions, it may not extend over the bad points at the boundary $B \cap \mathbb{R} \times \{0,1\}$. For future reference, we denote the set of* good *points on the boundary by*

$$S_{\mathbb{R}} := \{s \in \mathbb{R} \,|\, (s,0), (s,1) \notin B\}.$$

Now consider the suprema of the energy densities. Assumption (iii) in the statement of Lemma 2.4 implies that these suprema are uniformly bounded on each compact set $K \subset S_0$:

$$\sup_\nu \sup_{(s,t)\in K} |\partial_s v_\nu|_M^2 \leq C_0 \sup_\nu \sup_{(s,t)\in K} \|\partial_s \alpha_\nu - d_{\alpha_\nu}\phi_\nu\|^2_{L^2(\Sigma_\bullet)} < \infty.$$

In particular, there is a subsequence, still denoted by $(v_\nu)_\nu$, that converges weakly in $\mathcal{C}^1$, and strongly in $\mathcal{C}^0$, on compact subsets of S_0, including the boundary. Let $v_\infty \in \mathcal{C}^1(S_0, M)$ denote the limiting holomorphic curve. As with the v_ν, the curve v_∞ extends to $\mathbb{R} \times (0,1)$ and is $\mathcal{C}^\infty$ in this region [26, Theorem B.4.1].

Remark 4.8. *(a) We can actually say quite a bit more. The uniform energy bound implies that, after possibly passing to a further subsequence, we have that the $v_\nu : \mathbb{R} \times I \to M$ converge to v_∞ in $\mathcal{C}^\infty$ on compact subsets of the* interior*, $S_0 \cap (\mathbb{R} \times (0,1))$ (see* [26, Theorem 4.1.1]*). In particular, this automatically proves (16) for $K \subset S_0 \cap (\mathbb{R} \times (0,1))$. However, for the proof of Theorem 2.3 we will need convergence for K intersecting the boundary ∂S_0; we address this in Claim 2, below. See also Remark 4.10.*

(b) Everything we have discussed carries over, with only minor changes in notation, to the case where the Morse function $f : Y \to S^1$ has no critical points. In this case, and in light of the above remark, the proof of Lemma 2.4 would be complete at this stage.

The claim below states that v_∞ actually does have Lagrangian boundary conditions. To state this claim, we recall the definition of $S_\mathbb{R}$ above, and of the Lagrangians $L_{(0)}, L_{(1)} \subset M$ from Section 2.4.

Claim 1: For $j \in \{0,1\}$ and for each $s \in S_\mathbb{R}$, the sequence $(v_\nu(s,j))_\nu$ converges (in the metric on M) to a point in $L_{(j)}$. In particular, since the complement of $S_\mathbb{R}$ in $\mathbb{R}$ is finite, the map v_∞ extends to a map defined on all of $\mathbb{R} \times I$ with Lagrangian boundary conditions: $v_\infty(\cdot, j) : \mathbb{R} \to L_{(j)}$.

To prove the claim, we first note that by applying suitable gauge transformations, we may assume each A_ν satisfies the conclusions of Lemma 4.6. Fix a compact $K \subset S_\mathbb{R}$. Consider assumption (ii) in Lemma 2.4, stating that the norms $\|F_{a_\nu(s)}\|_{L^\infty}$ decay to zero uniformly on K. This implies that Theorem 4.3 applies to each $a_\nu(s)$ for each $s \in K$ and for all ν sufficiently large (exactly how large will depend on K). For each i, let $\mathrm{Heat}_{Q_{i(i+1)}}$ be the map constructed in Theorem 4.3. It will be useful to keep track of the bundle in the projection maps $\Pi_{Q_{i(i+1)}} : \mathcal{A}_{\mathrm{flat}}(Q_{i(i+1)}) \to \mathcal{A}_{\mathrm{flat}}(Q_{i(i+1)})/\mathcal{G}_0(Q_{i(i+1)})$. By restriction to $\partial Y_\bullet$, the assignment sending $s \in K$ to

$$\begin{aligned}\big(&\Pi_{Q_{01}} \circ \mathrm{Heat}_{Q_{01}}\left(a_\nu(s)|_{Y_{01}}\right), \Pi_{Q_{23}} \circ \mathrm{Heat}_{Q_{23}}\left(a_\nu(s)|_{Y_{23}}\right), \\ &\ldots, \Pi_{Q_{(N-2)(N_1)}} \circ \mathrm{Heat}_{Q_{(N-2)(N-1)}}\left(a_\nu(s)|_{Y_{(N-2)(N-1)}}\right)\big)\end{aligned}$$

determines a map $\ell_{\nu,0} : K \to L_{(0)}$. Similarly, we obtain a map $\ell_{\nu,1} : K \to L_{(1)}$ by using $\Pi_{Q_{i(i+1)}} \circ \mathrm{Heat}_{Q_{i(i+1)}}(a_\nu(s)|_{Y_{i(i+1)}})$ with i odd. We will show

$$\sup_{s \in K} \mathrm{dist}_M\left(\ell_{\nu,j}(s), v_\nu(s,j)\right) \xrightarrow{\nu} 0. \tag{46}$$

Then Claim 1 follows by repeating this argument for a sequence of compact K that exhaust $S_\mathbb{R}$.

The proof of (46) is just a computation. Let dist_M denote the distance function on M coming from the metric induced by the L^2-inner product on the harmonic spaces $H^1_\alpha \cong T_{[\alpha]}M$. Then fixing $s \in K$, we find that $\mathrm{dist}_M\left(\ell_{\nu,j}(s),\, v_\nu(s,j)\right)^2$ is equal to

$$\begin{aligned}
&\sum_i \mathrm{dist}_{M(P_{i+j})}\Big(\Big\{\Pi_{Q_{i(i+1)}} \circ \mathrm{Heat}_{Q_{i(i+1)}}\left(a_\nu\left(s\right)|_{Y_{i(i+1)}}\right)\Big\}|_{\Sigma_{i+j}}, \\
&\qquad\qquad \Pi_{P_{i+j}} \circ \mathrm{NS}_{P_{i+j}}\Big(\alpha_\nu\left(s,j\right)|_{\Sigma_{i+j}}\Big)\Big)^2 \\
&= \sum_i \mathrm{dist}_{M(P_{i+j})}\Big(\Pi_{P_{i+j}}\Big\{\Big(\mathrm{Heat}_{Q_{i(i+1)}}a_\nu(s)|_{Y_{i(i+1)}}\Big)|_{\Sigma_{i+j}}\Big\}, \\
&\qquad\qquad \Pi_{P_{i+j}}\Big\{\mathrm{NS}_{P_{i+j}}\left(\alpha_\nu\left(s,j\right)|_{\Sigma_{i+j}}\right)\Big\}\Big)^2 \\
&\leq \sum_i \Big\|\Big(\mathrm{Heat}_{Q_{i(i+1)}}a_\nu(s)|_{Y_{i(i+1)}}\Big)|_{\Sigma_{i+j}} - \mathrm{NS}_{P_{i+j}}\left(\alpha_\nu\left(s,j\right)|_{\Sigma_{i+j}}\right)\Big\|^2_{L^2(\Sigma_{i+j})}
\end{aligned}$$

The equality holds because restricting a flat connection on $Q_{i(i+1)}$ to the boundary commutes with the harmonic projections $\Pi_{Q_{i(i+1)}}$ and $\Pi_{P_{i+j}}$; the inequality holds by the definition of the distance on the $M(P_i)$, and because Π_{i+j} has operator norm equal to one. Taking the supremum over $s \in K$ and using the triangle inequality, we can continue this to get that $\sup_s \mathrm{dist}_M\left(\ell_{\nu,j}(s),\, v_\nu(s,j)\right)^2$ is bounded by

$$\begin{aligned}
&\sup_s \sum_i \Big\{\Big\|\Big(\mathrm{Heat}_{Q_{i(i+1)}}a_\nu(s)|_{Y_{i(i+1)}}\Big)|_{\Sigma_{i+j}} - \Big(a_\nu(s)|_{Y_{i(i+1)}}\Big)|_{\Sigma_{i+j}}\Big\|^2_{L^2(\Sigma_{i+j})} \\
&\qquad\qquad + \Big\|\Big(a_\nu(s)|_{Y_{i(i+1)}}\Big)|_{\Sigma_{i+j}} - \mathrm{NS}_{P_{i+j}}\left(\alpha_\nu\left(s,j\right)|_{\Sigma_{i+j}}\right)\Big\|^2_{L^2(\Sigma_{i+j})}\Big\} \\
&= \sup_s \sum_i \Big\{\Big\|\Big(\mathrm{Heat}_{Q_{i(i+1)}}a_\nu(s)|_{Y_{i(i+1)}}\Big)|_{\Sigma_{i+j}} - \Big(a_\nu(s)|_{Y_{i(i+1)}}\Big)|_{\Sigma_{i+j}}\Big\|^2_{L^2(\Sigma_{i+j})} \\
&\qquad\qquad + \big\|\alpha_\nu\left(s,j\right)|_{\Sigma_{i+j}} - \mathrm{NS}_{P_{i+j}}\left(\alpha_\nu\left(s,j\right)|_{\Sigma_{i+j}}\right)\big\|^2_{L^2(\Sigma_{i+j})}\Big\}.
\end{aligned}$$

The equality holds because α_ν agrees with a_ν at the boundary of $Y_\bullet$. The second part of Theorem 4.3 shows that the first term in the summand goes to zero as $\nu \to \infty$, since F_{a_ν} converges to zero in L^∞ (uniformly in $s \in K$). Similarly, the second term in the summand goes to zero by Theorem 3.1 (iii). This verifies (46) and so proves Claim 1. In light of Theorem 4.3 (ii), this also proves (15) in the statement of Lemma 2.4.

As in Section 2.4, the holomorphic map $v_\infty : \mathbb{R} \times I \to M$ lifts to a holomorphic strip representative $A_\infty \in \mathcal{A}^{1,q}_{loc}(\mathbb{R} \times Q)$. To prove the convergence statement (14), we simply translate the convergence of the v_ν to v_∞ into a statement of the convergence of α_ν to α_∞. Note that, because M is finite-dimensional, we can choose any metric we want. At this point it is convenient to choose the metric on the tangent space induced from the $\mathcal{C}^0$-norm on the harmonic spaces. In particular, the $\mathcal{C}^0$-convergence in the

$\mathbb{R} \times I$-directions immediately implies that, for each $(s,t) \in S_0$, there are gauge transformations $\mu_\nu(s,t) \in \mathcal{G}_0^{2,q}(P)$ such that

$$\sup_{(s,t)\in K} \|\alpha_\infty(s,t) - \mu_\nu(s,t)^* \mathrm{NS}(\alpha_\nu(s,t))\|_{\mathcal{C}^0(\Sigma_\bullet)} \longrightarrow 0. \tag{47}$$

By perturbing the gauge transformations, we may suppose that each $\mu_\nu(s,t)$ is smooth in s,t. This gives that $\|\alpha_\infty - \mu_\nu^* \alpha_\nu\|_{\mathcal{C}^0(K\times\Sigma_\bullet)}$ is bounded by

$$\begin{aligned} &\sup_{(s,t)\in K} \left\{ \left\| \mu_\nu^*\alpha_\nu - \mu_\nu^* \mathrm{NS}(\alpha_\nu) \right\|_{\mathcal{C}^0(\Sigma_\bullet)} + \left\| \alpha_\infty - \mu_\nu^* \mathrm{NS}(\alpha_\nu) \right\|_{\mathcal{C}^0(\Sigma_\bullet)} \right\} \\ &\leq C \sup_{(s,t)\in K} \left\{ \left\| F_{\alpha_\nu} \right\|_{\mathcal{C}^0(\Sigma_\bullet)} + \left\| \alpha_\infty - \mu_\nu^* \mathrm{NS}(\alpha_\nu) \right\|_{\mathcal{C}^0(\Sigma_\bullet)} \right\} \end{aligned}$$

where the inequality follows from Theorem 3.1 (iii). This last term goes to zero by assumption (i) from Lemma 2.4 and (47). This proves (14).

It remains to prove (16) and (17). For this, consider the real-valued functions

$$e_\nu := \|\beta_{s,\nu}\|_{L^2(\Sigma_\bullet)} : S_0 \to \mathbb{R}, \quad f_\nu := \|b_{s,\nu}\|_{L^2(Y_\bullet)} : S_{\mathbb{R}} \to \mathbb{R}.$$

In light of the Sobolev embeddings $W^{2,2} \hookrightarrow \mathcal{C}^0$ and $W^{1,2} \hookrightarrow \mathcal{C}^0$ for compact sets in dimensions 2 and 1, respectively, the convergence statements (16) and (17) follow immediately from the next claim (after passing to a suitable subsequence).

Claim 2: There is a constant C so that

$$\sup_\nu \|e_\nu\|_{W^{2,2}(\mathbb{R}\times I)} + \|f_\nu\|_{W^{1,2}(\mathbb{R})} \leq C.$$

To prove the bound for e_ν, apply Kato's inequality

$$|d|V|| \leq |\nabla V|,$$

with $V = \beta_{s,\nu}$ and $\nabla = ds \otimes \nabla_s + dt \otimes \nabla_t$. Now the claim for e_ν follows immediately from Theorem 4.1 together with assumption (v) in the statement of Lemma 2.4 asserting a uniform energy bound on the A_ν. The bound for f_ν is similar, but use Kato's inequality with $V = b_{s,\nu}$ and $\nabla = ds \otimes \nabla_s$. This completes the proof of the Compactness Lemma 2.4. □

4.3. **Proof of the Compactness Theorem 2.3.** We begin by considering the cases where the hypotheses of the Compactness Lemma 2.4 are *not* satisfied. That is, we assume there is some compact $K \subset \mathbb{R}$ for which one of the following cases holds

Case 1: $\|F_{\alpha_\nu}\|_{L^\infty(K\times I\times\Sigma_\bullet)} + \|F_{a_\nu}\|_{L^\infty(K\times Y_\bullet)} \to \infty$;

Case 2: $\|F_{\alpha_\nu}\|_{L^\infty(K\times I\times\Sigma_\bullet)} + \|F_{a_\nu}\|_{L^\infty(K\times Y_\bullet)} \to \Delta > 0$;

Case 3: $\sup_{(s,t)\in K\times I} \|\beta_{s,\nu}(s,t)\|_{L^2(\Sigma_\bullet)} + \sup_{s\in K} \|b_{s,\nu}(s)\|_{L^2(Y_\bullet)} \to \infty$, and for all compact $K' \subset \mathbb{R}\times Y$ we have

$$\|F_{\alpha_\nu}\|_{L^\infty(K')} + \|F_{a_\nu}\|_{L^\infty(K')} \to 0.$$

To prove the theorem, we will show below that each of these cases leads to *energy quantization*. That is, we will show in each case that there is some *bubbling point* $(s,t) \in \mathbb{R}\times I$, and a set $T_{(s,t)} \subset \mathbb{R}\times Y$, such that there is some integer $m > 0$ with the property that, for every neighborhood U of $T_{(s,t)}$, the ϵ_ν-energies of the A_ν on U converge in ν to m/r. Here r is that appearing in PU(r). The set $T_{(s,t)}$ will either be a point in $\{(s,t)\}\times\Sigma_\bullet$, or the whole fiber $\{(s,t)\}\times\Sigma_\bullet$. Let B denote the set of bubbling points.

Assume, for the moment, that we have established that each of the above cases leads to energy quantization. We will show now how the Compactness Theorem follows from the Compactness Lemma 2.4. We have assumed $A_\nu \in A^{1,q}_{\epsilon_\nu}(a^-,a^+)$, so the ϵ_ν-ASD condition implies a uniform energy bound

$$E^{\mathrm{inst}}_{\epsilon_\nu}(A_\nu) = \mathcal{CS}(a^+) - \mathcal{CS}(a^-),$$

where $\mathcal{CS}$ is the Chern-Simons functional for Q. This bound combines with energy quantization to imply that the set B of bubbling points must be finite. Then the hypotheses of Lemma 2.4 hold on $S_0 := \mathbb{R}\times I\backslash B$, so a subsequence of the A_ν converges in the sense of the statement of Theorem 2.3 (with $J=1$ and $s^1_\nu = 0$) to a limiting holomorphic curve representative on $\mathbb{R}\times Q$. Finally, to obtain the equality in (13), we need to incorporate time translations, thereby producing the tuple $(A^1,\dots,A^J)$ of holomorphic curve representatives. We defer a complete discussion of the translations until after of our case analysis for energy quantization.

Turning now to the case analysis, we begin by introducing the following notation. Given $r > 0$ and a smooth manifold-with-boundary X, we set

$$X(r) := X \cup_{\partial X} ([0,r)\times\partial X)\,, \quad X^\infty := X\cup_{\partial X}([0,\infty)\times\partial X)\,. \tag{48}$$

There exist smooth structures on these spaces that are compatible in the sense that the inclusions

$$X(r) \subseteq X(r') \subseteq X^\infty \tag{49}$$

are smooth embeddings for $r \le r'$. If X has a metric g, then we will consider the metric on $X(r)$ and X^∞ that is given by g on X and $dt^2 + g|_{\partial X}$ on the end. In particular, the embeddings (49) become *metric* embeddings. This will be called the *fixed metric* on the given manifold, and we denote its various norms by $|\cdot|, \|\cdot\|_{L^p}$, etc. If X is equipped with a bundle $E\to X$ then we define bundles $E(r)\to X(r)$ and $E^\infty \to X^\infty$ in the obvious way. Note also that we have the following decomposition

$$\mathbb{R}\times X^\infty = (\mathbb{R}\times X)\cup(\mathbb{H}\times\partial X)\,. \tag{50}$$

Case 1. (Instantons on S^4) By passing to a subsequence, we may assume there is some j so that the L^∞-norm of each curvature is always achieved on Σ_{j+1} or $Y_{j(j+1)}$. That is, we may assume one of the following holds for all ν:

$$\|F_{\alpha_\nu}\|_{L^\infty(K\times I\times\Sigma_{j+1})} \geq \max\left\{\|F_{\alpha_\nu}\|_{L^\infty(K\times I\times\Sigma_\bullet)}, \|F_{a_\nu}\|_{L^\infty(K\times Y_\bullet)}\right\}, \text{or} \tag{51}$$

$$\|F_{a_\nu}\|_{L^\infty(K\times Y_{j(j+1)})} \geq \max\left\{\|F_{\alpha_\nu}\|_{L^\infty(K\times I\times\Sigma_\bullet)}, \|F_{a_\nu}\|_{L^\infty(K\times Y_\bullet)}\right\}. \tag{52}$$

If (51) holds, find a point $(s_\nu, t_\nu) \in K \times I$ so that $\|F_{\alpha_\nu(s_\nu,t_\nu)}\|_{L^\infty(\Sigma_{j+1})} = \|F_{\alpha_\nu}\|_{L^\infty(K\times I\times\Sigma_{j+1})}$. Similarly, if (52) holds, then find a point $s_\nu \in K$ with $\|F_{a_\nu(s_\nu)}\|_{L^\infty(Y_{j(j+1)})} = \|F_{a_\nu}\|_{L^\infty(K\times Y_{j(j+1)})}$. By passing to a subsequence, we may suppose the s_ν converge to some element of $K \subset \mathbb{R}$. Similarly, we may assume $t_\nu \to t_\infty \in I$ converges. Strictly speaking, we need to distinguish between whether t_∞ lies in the interior of I, or on the boundary. However, the analysis for when t_∞ lies in the boundary can be incorporated to the analysis for when (52) holds. Precisely, we find ourselves considering the following subcases:

Subcase 1: (51) holds and $t_\infty \notin \{0, 1\}$;

Subcase 2: (52) holds, or (51) holds and $t_\infty \in \{0, 1\}$.

Without loss of generality, we may suppose $j = 0$ and $t_\infty \in [0, 1)$.

In Subcase 1, define a rescaled connection $\widetilde{A}_\nu$ in terms of its components as follows:

$$\begin{aligned} \widetilde{\alpha}_\nu(s,t) &:= \alpha(\epsilon_\nu s + s_\nu, \epsilon_\nu t + t_\nu)|_{\Sigma_1} \\ \widetilde{\phi}_\nu(s,t) &:= \epsilon_\nu\phi(\epsilon_\nu s + s_\nu, \epsilon_\nu t + t_\nu)|_{\Sigma_1} \\ \widetilde{\psi}_\nu(s,t) &:= \epsilon_\nu\psi(\epsilon_\nu s + s_\nu, \epsilon_\nu t + t_\nu)|_{\Sigma_1}. \end{aligned} \tag{53}$$

We view these as connections and 0-forms defined on $D_{\epsilon_\nu^{-1}\eta} \times \Sigma_1 \subseteq \mathbb{C} \times \Sigma_1$, where $\eta = \frac{1}{2}\min\{t_\infty, 1 - t_\infty\}$, $D_r \subset \mathbb{C}$ is the ball of radius r centered at zero, and we assume ν is large enough so $t_\nu \leq \eta$.

In Subcase 2, define a connection $\widetilde{A}_\nu$ on a neighborhood of $\mathbb{R} \times Y_{01}$ as follows:

$$\widetilde{a}_\nu(s) := a_\nu(\epsilon_\nu s + s_\nu)|_{Y_{01}}, \quad \widetilde{p}_\nu(s) := \epsilon_\nu p_\nu(\epsilon_\nu s + s_\nu)|_{Y_{01}}$$

$$\begin{aligned} \widetilde{\alpha}_\nu(s,t) &:= \begin{cases} \alpha(\epsilon_\nu s + s_\nu, -\epsilon_\nu t + 1)|_{\Sigma_0} & \text{on } \Sigma_0 \\ \alpha(\epsilon_\nu s + s_\nu, \epsilon_\nu t)|_{\Sigma_1} & \text{on } \Sigma_1 \end{cases} \\ \widetilde{\phi}_\nu(s,t) &:= \begin{cases} \epsilon_\nu\phi(\epsilon_\nu s + s_\nu, -\epsilon_\nu t + 1)|_{\Sigma_0} & \text{on } \Sigma_0 \\ \epsilon_\nu\phi(\epsilon_\nu s + s_\nu, \epsilon_\nu t)|_{\Sigma_1} & \text{on } \Sigma_1 \end{cases} \\ \widetilde{\psi}_\nu(s,t) &:= \begin{cases} \epsilon_\nu\psi(\epsilon_\nu s + s_\nu, -\epsilon_\nu t + 1)|_{\Sigma_0} & \text{on } \Sigma_0 \\ \epsilon_\nu\psi(\epsilon_\nu s + s_\nu, \epsilon_\nu t)|_{\Sigma_1} & \text{on } \Sigma_1 \end{cases} \end{aligned} \tag{54}$$

We view this as a connection defined on $\mathbb{R} \times Y_{01}(\epsilon_\nu^{-1})$, where $Y_{01}(r)$ is constructed as in (48) using $X = Y_{01}$.

In both Subcases, the connections $\widetilde{A}_\nu$ are defined on increasing spaces that exhaust an ambient space ($\mathbb{C} \times \Sigma_1$ in Subcase 1 and $\mathbb{R} \times Y_{01}^\infty$ in Subcase 2). Moreover, the $\widetilde{A}_\nu$ are ASD with respect to the fixed metric on the ambient space.

Remark 4.9. *Note that we have changed perspective with our treatment of $\widetilde{A}_\nu$, relative to the ϵ_ν-ASD connections A_ν: The A_ν are all defined on the same underlying space, but with the ϵ_ν-dependence appearing in the metric (and hence in the ASD equation). On the other hand, the $\widetilde{A}_\nu$ are defined on increasing spaces, but are ASD relative to a fixed metric.*

The additivity of the integral implies that the $\widetilde{A}_\nu$ have uniformly bounded energy $\frac{1}{2}\|F_{\widetilde{A}_\nu}\|_{L^2}^2 \leq \mathcal{CS}(a^+) - \mathcal{CS}(a^-)$; here the norm should be taken on the domain on which the connection is defined. Furthermore, the energy densities are bounded from below:

$$\|F_{\widetilde{A}_\nu}\|_{L^\infty} \geq \|F_{\widetilde{\alpha}_\nu}\|_{L^\infty} + \|F_{\widetilde{a}_\nu}\|_{L^\infty} = \|F_{\alpha_\nu}\|_{L^\infty} + \|F_{a_\nu}\|_{L^\infty}. \tag{55}$$

In particular, the condition of Case 1 implies that $\|F_{\widetilde{A}_\nu}\|_{L^\infty} \to \infty$. Following the usual rescaling argument [35] [10, Section 9] (see also [26, Theorem 4.6.1] for the closely-related case of J-holomorphic curves), we can conformally rescale in a small neighborhood U of the blow-up point to obtain yet another sequence $\overline{A}_\nu$ of finite-energy instantons with $\|F_{\overline{A}_\nu}\|_{L^\infty} = 1$, and defined on increasing balls in $\mathbb{R}^4$. By Uhlenbeck's strong compactness theorem, there is a subsequence of the $\overline{A}_\nu$ that converges (modulo gauge and in $\mathcal{C}^\infty$ on compact sets) to a finite-energy, non-constant instanton $\overline{A}_\infty$ on $\mathbb{R}^4$. By Uhlenbeck's removable singularities theorem this extends to a non-constant instanton, also denoted by $\overline{A}_\infty$, on a $\mathrm{PU}(r)$-bundle $R_\infty \to S^4$. Since $\overline{A}_\infty$ is ASD and non-constant we have

$$0 < \frac{1}{2}\int_{S^4} |F_{\overline{A}_\infty}|^2 = -\frac{1}{2}\int_{S^4} \langle F_{\overline{A}_\infty} \wedge F_{\overline{A}_\infty} \rangle = \frac{1}{2r} q_4(R_\infty).$$

The characteristic class $t_2(R_\infty) \in H^2(S^4, \mathbb{Z}_r)$ is necessarily zero, so the bundle R_∞ is the reduction of an $\mathrm{SU}(r)$-bundle, and the quantity $q_4(R_\infty)/2r$ is an integer. Energy quantization for this case follows from a straightforward comparison of the energy of the rescaled connection $\overline{A}_\nu$ with that of A_ν.

Case 2. (Instantons on non-compact domains) This case is much the same as the previous, in that instantons near the blow-up point bubble off. Define $\widetilde{A}_\nu$ exactly as in Case 1 above. Everything up to and including (55) continues to hold. The condition of this case implies that $\liminf \|F_{\widetilde{A}_\nu}\|_{L^\infty}$ is bounded from below by $\Delta > 0$. After possibly passing to a subsequence, we may assume the quantities $\|F_{\widetilde{A}_\nu}\|_{L^\infty}$ converge to some $\Delta' \in [\Delta, \infty]$. If

$\Delta' = \infty$ then we are done by precisely the same analysis as in Case 1. So we may assume $0 < \Delta' < \infty$, in which case we can apply Uhlenbeck's strong compactness theorem directly to the sequence $\widetilde{A}_\nu$. We may therefore assume this sequence converges to a non-flat finite-energy instanton $\widetilde{A}_\infty$ on a bundle over one of the spaces $\mathbb{R} \times Y_{01}^\infty$ or $\mathbb{C} \times \Sigma_1$, depending on whether we are in Subcase 1 or 2. We show in [14] that the energy of any such instanton $\widetilde{A}_\infty$ is m/r for some positive integer m, so this finishes the analysis for Case 2.

Case 3. (Holomorphic spheres and disks in $M(P)$) For each ν, let c_ν be the supremum over $s \in K$ and $t \in I$ of the numbers

$$\|\beta_{s,\nu}(s,t)\|_{L^2(\Sigma_\bullet)}, \quad \|b_{s,\nu}(s)\|_{L^2(Y_\bullet)}.$$

The conditions of this case imply that $c_\nu \to \infty$. Find $j_\nu \in \{0, \dots, N-1\}$ and points $(s_\nu, t_\nu) \in K \times I$ for which

$$c_\nu = \|\beta_{s,\nu}(s_\nu, t_\nu)\|_{L^2(\Sigma_{j_\nu})}, \quad \text{or} \quad c_\nu = \|b_{s,\nu}(s_\nu)\|_{L^2(Y_{(j_\nu-1)j_\nu})}.$$

(Such points exist since A_ν has finite energy, so $\beta_{s,\nu}, b_{s,\nu}$ decay to zero at $\pm\infty$; alternatively, one could replace c_ν by $c_\nu/2$, without changing the argument below.) If $c_\nu = \|b_{s,\nu}(s_\nu)\|_{L^2(Y_{(j_\nu-1)j_\nu})}$, then we just declare $t_\nu = 0$. By passing to a subsequence, we can assume that $j_\nu = 1$ for all ν, and that the (s_ν, t_ν) converge to some $(s_\infty, t_\infty) \in K \times I$. The two relevant subcases to consider are as follows:

Subcase 1: $t_\infty \in (0,1)$

Subcase 2: $t_\infty \in \{0,1\}$

We may assume, without loss of generality, that if Subcase 2 holds then $t_\infty = 0$. Define rescaled connections $\hat{A}_\nu$ using (53) and (54), except replace every ϵ_ν by c_ν^{-1} (the subcases here correspond to those from Case 1).

Remark 4.10. *It turns out that Subcase 1 implies that a nontrivial holomorphic sphere in $M(P_1)$ bubbles off; non-triviality will follow from (16) from Lemma 2.4. Similarly, we will see that Subcase 2 implies that the bubble is a nontrivial holomorphic disk in $M(P_0) \times M(P_1)$ with Lagrangian boundary conditions in $L(Q_{01}) \subset M(P_0) \times M(P_1)$; non-triviality will follow from (16)* at the boundary*, as well as (17) from the same lemma. Since these latter conclusions of Lemma 2.4 were the technically demanding part of the lemma (see Remark 4.8), we will prove Subcase 2, leaving the (in many ways simpler) argument of Subcase 1 to the reader.*

We view the rescaled connections $\hat{A}_\nu$ as being defined on the increasing manifolds $\mathbb{R} \times Y_{01}(c_\nu)$. The components of $F_{\hat{A}_\nu}$ satisfy

$$\hat{\beta}_{s,\nu} + *_\Sigma \hat{\beta}_{t,\nu} = 0, \quad \hat{\gamma} = -\hat{\epsilon}_\nu^{-2} *_\Sigma F_{\hat{\alpha}_\nu}, \quad \hat{b}_{s,\nu} = -\hat{\epsilon}_\nu^{-1} *_Y F_{\hat{a}_\nu},$$

where $\hat{\epsilon}_\nu := c_\nu \epsilon_\nu$, and the Hodge stars are relative to the fixed metrics on $\Sigma_0 \sqcup \Sigma_1$ and Y_{01}. It may not be the case that the $\hat{\epsilon}_\nu$ are decaying to zero;

this is replaced by the assumption of Case 3 that the slice-wise curvatures converge to zero in L^∞:

$$\|F_{\hat{\alpha}_\nu}\|_{L^\infty} = \|F_{\alpha_\nu}\|_{L^\infty} \longrightarrow 0, \quad \|F_{\hat{a}_\nu}\|_{L^\infty} = \|F_{a_\nu}\|_{L^\infty} \longrightarrow 0.$$

Our choice of rescaling also gives

$$1 \leq \|\hat{\beta}_{s,\nu}(0,0)\|_{L^2(\Sigma_1)} + \|\hat{b}_{s,\nu}(0)\|_{L^2(Y_{01})} \leq 2. \tag{56}$$

By Lemma 2.4 it follows that, after possibly passing to a subsequence, there exists a sequence of gauge transformations $\mu_\nu : \mathbb{H} \to \in \mathcal{G}(P_0 \sqcup P_1)$, and a limiting connection $\hat{A}_\infty \in \mathcal{A}(\mathbb{R} \times Q_{01}^\infty)$ that is a holomorphic curve representative

$$\hat{\beta}_{s,\infty} + *\hat{\beta}_{t,\infty} = 0, \quad F_{\hat{\alpha}_\infty} = 0, \quad F_{\hat{a}_\infty} = 0,$$

and satisfies (16) and (17); here we are using (50) to view α_∞ as a map defined on $\mathbb{H}$. Let $\Pi_{P_j} : \mathcal{A}_{\text{flat}}(P_j) \to M(P_j)$ and $\Pi_{Q_{01}} : \mathcal{A}_{\text{flat}}(Q_{01}) \to L(Q_{01})$ be the projections to the moduli spaces. Then

$$v_\infty := (\Pi_{P_0}(\hat{\alpha}_\infty|_{\Sigma_0}), \Pi_{P_1}(\hat{\alpha}_\infty|_{\Sigma_1})) : \mathbb{H} \longrightarrow M(P_0) \times M(P_1)$$

is a holomorphic curve with Lagrangian boundary conditions $\mathbb{R} \to L(Q_{01}) \subset M(P_0) \times M(P_1)$ determined by $a_\infty : \mathbb{R} \to \mathcal{A}_{\text{flat}}(Q_{01})$. Furthermore, v_∞ has bounded energy

$$\begin{aligned} \int_{\mathbb{H}} |\partial_s v_\infty|^2 &= \int_{\mathbb{H}\times\Sigma_1 \sqcup \Sigma_2} |\hat{\beta}_{s,\infty}|^2 \leq \liminf_\nu \|\beta_{s,\nu}\|^2_{L^2(\mathbb{R}\times Y),\epsilon_\nu} \\ &\leq \liminf_\nu \tfrac{1}{2}\|F_{A_\nu}\|^2_{L^2(\mathbb{R}\times Y),\epsilon_\nu} = \left(\mathcal{CS}(a^+) - \mathcal{CS}(a^-)\right). \end{aligned}$$

In particular, the removal of singularities theorem [26, Theorem 4.1.2 (ii)] applies and so v_∞ extends to a holomorphic disk $v_\infty : \mathbb{D} \to M(P_1) \times M(P_2)$ with Lagrangian boundary conditions. Then (16) and (17) combine with (56) to give

$$2|\partial_s v_\infty(0,0)| \geq \liminf_\nu \|\hat{\beta}_{s,\nu}(0,0)\|_{L^2(\Sigma_1)} + \|\hat{b}_{s,\nu}(0)\|_{L^2(Y_{01})} \geq 1.$$

In particular, v_∞ is non-constant. Since v_∞ is a disk with boundary conditions in a simply-connected Lagrangian, it follows that v_∞ has an extension to a map of the form $\hat{v}_\infty : S^2 \to M(P_1) \times M(P_2)$ that agrees with v_∞ on one hemisphere $\mathbb{D} \subset S^2$ and lies in the Lagrangian in the other hemisphere. Then the energy of v_∞ is given by

$$\int_{\mathbb{D}} |\partial_s v_\infty|^2 = -\int_{\mathbb{D}} \omega(\partial_s v_\infty, \partial_t v_\infty) = -\int_{S^2} \omega(\partial_s \hat{v}_\infty, \partial_t \hat{v}_\infty) = (\hat{v}_\infty^* \omega)\left[S^2\right],$$

where the second equality holds since ω vanishes on the Lagrangian. This gives energy quantization. To obtain the appropriate constant for energy

quantization, we note that the symplectic manifold $M(P_0)\times M(P_1)$ is monotone with monotonicity constant $1/2r$ (see [39, Theorem 3.3.3] and the references therein). It follows that the energy is $c_1(\hat{v}_\infty^* TM)/2r$. The Chern number $c_1(\hat{v}_\infty^* TM)$ is even; see [8, Corollary 6.3]. This implies that the energy of v_∞ is m/r for some positive integer m. (Alternatively, one could lift v_∞ to a connection on $\mathbb{R}\times Y_{01}^\infty$ and then use the main result in [14].) This concludes our case analysis for energy quantization.

Finally, we address translations; we follow the strategy of [34]. The moduli space of flat connections on Q is canonically identified with the set of Lagrangian intersection points $L_{(0)}\cap L_{(1)}$, and the non-degeneracy assumption on the elements of $\mathcal{A}_{\text{flat}}(Q)$ implies that $L_{(0)}\cap L_{(1)}$ is a finite set in M; see [12, Section 4]. In particular, there is some $\epsilon_0>0$ so that $B_{\epsilon_0}(p)\cap B_{\epsilon_0}(p')=\emptyset$, for all distinct $p,p'\in L_{(0)}\cap L_{(1)}$. Use A_ν to define v_ν as in (45). By assumption, each A_ν converges at $\pm\infty$ to the flat connection $a^\pm$. Since the maps NS_{P_i} preserve flat connections, it follows that each v_ν converges at $\pm\infty$ to the Lagrangian intersection point $p^\pm\in L_{(0)}\cap L_{(1)}$ associated to $a^\pm$. Define

$$s_\nu^1 := \sup\left\{ s\in\mathbb{R} \,\middle|\, \operatorname{dist}_{M(P_\bullet)}\left(p^-,\ v_\nu(s,t)\right)\le\epsilon_0, \quad \text{for all } t\in I \right\}.$$

(We may assume $p^-\neq p^+$, otherwise all instantons are trivial and all holomorphic curves are constant. This assumption implies the set defining s_ν^1 is non-empty.) Then for each ν we have

$$\operatorname{dist}_{M(P_\bullet)}\left(p^-,\ \left(\tau_{s_\nu^1}^* v_\nu\right)(s,t)\right)\le\epsilon_0 \quad \forall t\in I, \forall s\le 0, \tag{57}$$

$$\operatorname{dist}_{M(P_\bullet)}\left(\left(\tau_{s_\nu^1}^* v_\nu\right)(0,t), p^+\right)=\epsilon_0 \quad \text{for some } t\in I \tag{58}$$

The case analysis above combines with the Compactness Lemma 2.4 to show that, after passing to a subsequence, the translates $\tau_{s_\nu^1}^* v_\nu$ converge on compact sets of bubbling set complement to a limiting holomorphic strip v^1. This holomorphic strip is asymptotic on the ends to some Lagrangian intersection points p^0 at $-\infty$ and p^1 at ∞. By (57) and the definition of ϵ_0, we must have $p^0=p^-$. On the other hand, (58) shows that $v^1(0,t)$ is *not* at a Lagrangian intersection point for some $t\in I$. This proves that v^1 has positive energy, and shows that the $\tau_{s_\nu^1}^* v_\nu$ become arbitrarily close to p^1.

Continue inductively with p^1 replacing p^0, etc. to obtain a sequence of limiting holomorphic strips v^j that limit to Lagrangian intersection points p^{j-1} and p^j. The theorem follows by lifting the v^j and p^j to representatives, and converting the convergence of the $\tau_{s_\nu^j}^* v_\nu$ to statements about the representatives, as we did in the proof of Lemma 2.4. The equality in (13) follows as in [28, Proposition 6.3.1], but is easier since we have assumed non-degeneracy of all flat connections. □

4.4. **Proof of the Elliptic Estimates Theorem 4.1.** We will prove Theorem 4.1 in four steps. The first is Proposition 4.11, and establishes a general

estimate that bounds first derivatives on the interval I and surface $\Sigma_\bullet$ in terms of the 3-dimensional operators d_a and d_a^* on Y. The second step is Proposition 4.12, and bounds the s-derivative $\nabla_s F_A$ for an ϵ-ASD connection A in terms of the energy $\|F_A\|_{L^2,\epsilon}$. This represents a version of the standard elliptic bootstrapping estimate for instantons, where here we keep track of how the constants depend on ϵ. We have restricted only to the s-derivative because this is the only direction in which no ϵ-scaling occurs; in Corollary 4.14 we bound various other first derivatives of the curvature. As a third step we establish a second order version of the first step; this is stated in Proposition 4.15. The last step is Proposition 4.16, and is a certain second order version of the second step, in which we bound the second derivatives $\nabla_s^2 F_A$ and $\nabla_t \nabla_s F_A$. Then Theorem 4.1 is an immediate corollary of these latter two propositions.

Proposition 4.11. *(General elliptic estimates; 1st order) Suppose A is an ϵ-ASD connection satisfying (35-36) for some $c_0 > 0$. Then there are constants $\epsilon_0, C > 0$ such that the following holds for all $0 < \epsilon < \epsilon_0$ and all ϵ-smooth 1-forms v on Y with $v|_{I\times\Sigma_\bullet} = \nu + \theta\, dt$:*

$$\begin{aligned} &\|d_\alpha \nu\|_{L^2(I\times\Sigma_\bullet),\epsilon} + \|d_\alpha^{*_\epsilon}\nu\|_{L^2(I\times\Sigma_\bullet),\epsilon} + \|\nabla_t \nu\|_{L^2(I\times\Sigma_\bullet),\epsilon} \\ &\quad + \|d_\alpha \theta\|_{L^2(I\times\Sigma_\bullet),\epsilon} + \|\nabla_t \theta\|_{L^2(I\times\Sigma_\bullet),\epsilon} \\ &\qquad \leq C\left(\|v\|_{L^2(Y),\epsilon} + \|d_a v\|_{L^2(Y),\epsilon} + \|d_a^{*_\epsilon} v\|_{L^2(Y),\epsilon}\right). \end{aligned} \tag{59}$$

The assumption that A is ϵ-ASD is only to simplify the exposition (for general A, statements (35-36) would need to be modified to control the remaining curvature components β_t and γ).

The proof will show that the bound (59) continues to hold with the same constants when $I \times \Sigma_\bullet$ is replaced by any measurable subset in the complement of the critical fibers of the Morse function $f : Y \to S^1$. We also note that the bound in (59) can be integrated over any interval in $\mathbb{R}$ to obtain analogous L^2-bounds for 4-manifolds.

Proof of Proposition 4.11. To illustrate the basic argument, we temporarily forget the fact that $f : Y \to S^1$ has critical points. We will still use $\Sigma_\bullet$ and $Y_\bullet$ as before, but we are simply assuming that $Y_\bullet$ is now a product cobordism. We are continuing to use the rescaled metric taking the form $dt^2 + \epsilon^2 g_\Sigma$ on $I \times \Sigma_\bullet$ and $\epsilon^2 g_Y$ on $Y_\bullet$. We will use dt to denote the 1-form that is parallel to df, but that is scaled so that $|dt|_\epsilon = 1$ everywhere on Y. Then dt is well-defined since we have assumed f has no critical points (so $df \neq 0$), and we have decompositions

$$v = \nu + \theta\, dt, \quad a = \alpha + \psi\, dt \tag{60}$$

that are defined globally on Y. This gives

$$d_a v = d_\alpha \nu + dt \wedge (\nabla_t \nu - d_\alpha \theta), \quad \text{and} \quad d_a^{*_\epsilon} v = d_\alpha^{*_\epsilon} \nu - \nabla_t \theta. \tag{61}$$

In this simplified setting we will prove

$$\begin{aligned} &\|d_\alpha \nu\|_{L^2(Y),\epsilon} + \|d_\alpha^{*_\epsilon} \nu\|_{L^2(Y),\epsilon} + \|\nabla_t \nu\|_{L^2(Y),\epsilon} + \|d_\alpha \theta\|_{L^2(Y),\epsilon} + \|\nabla_t \theta\|_{L^2(Y),\epsilon} \\ &\quad \leq C \left(\|v\|_{L^2(Y),\epsilon} + \|d_a v\|_{L^2(Y),\epsilon} + \|d_a^{*_\epsilon} v\|_{L^2(Y),\epsilon} \right) \end{aligned}$$

for a constant C that is independent of ϵ. We will then describe how to adjust the proof to accommodate the more general situation in which f has critical points. Here and below all norms, inner products and Hodge stars are over Y, unless otherwise specified. One exception is that the notation $d_\alpha^{*_\epsilon}$ is the adjoint taken with respect to the inner product on the surface $\Sigma_\bullet$, whereas $d_a^{*_\epsilon}$ is that on the 3-manifold; the rule being that we view d_α as an operator on the surface and d_a as an operator on the 3-manifold.

Take the L^2-norm square of each term in (61), and then add to get

$$\begin{aligned} \|d_a v\|_\epsilon^2 + \|d_a^{*_\epsilon} v\|_\epsilon^2 &= \|d_\alpha \nu\|_\epsilon^2 + \|\nabla_t \nu - d_\alpha \theta\|_\epsilon^2 + \|\nabla_t \theta - d_\alpha^{*_\epsilon} \nu\|_\epsilon^2 \\ &= \|d_\alpha \nu\|_\epsilon^2 + \|d_\alpha^{*_\epsilon} \nu\|_\epsilon^2 + \|\nabla_t \nu\|_\epsilon^2 \\ &\quad + \|d_\alpha \theta\|_\epsilon^2 + \|\nabla_t \theta\|_\epsilon^2 \\ &\quad - 2 \left(\nabla_t \theta, d_\alpha^{*_\epsilon} \nu \right)_\epsilon - 2 \left(\nabla_t \nu, d_\alpha \theta \right)_\epsilon . \end{aligned} \tag{62}$$

It suffices to bound these last two terms. We integrate by parts in the second of these to get

$$\begin{aligned} -2 \left(\nabla_t \theta, d_\alpha^{*_\epsilon} \nu \right)_\epsilon - 2 \left(\nabla_t \nu, d_\alpha \theta \right)_\epsilon &= -2 \left(\nabla_t \theta, d_\alpha^{*_\epsilon} \nu \right)_\epsilon + 2 \left(\nu, \nabla_t d_\alpha \theta \right)_\epsilon \\ &= -2 \left(\nabla_t \theta, d_\alpha^{*_\epsilon} \nu \right)_\epsilon + 2 \left(\nu, d_\alpha \nabla_t \theta \right)_\epsilon \\ &\quad + 2 \left(\nu, [\beta_t, \theta] \right)_\epsilon \\ &= 2 \left(\nu, [\beta_t, \theta] \right)_\epsilon , \end{aligned} \tag{63}$$

where we canceled the first two terms in the last step after a second integration by parts. To control this, write

$$\left(\nu, [\beta_t, \theta] \right)_\epsilon = \int_{Y_\bullet} \langle \nu \wedge [*_\epsilon \beta_t, \theta] \rangle + \int_{I \times \Sigma_\bullet} \langle \nu \wedge [*_\epsilon \beta_t, \theta] \rangle. \tag{64}$$

We begin by estimating the second term on the right in (64). For this we note that pointwise on $I \times \Sigma_\bullet$ we have $*_\epsilon \beta_t = *\beta_t$, since β_t is a 1-form and the ϵ-scaling is only in the $\Sigma_\bullet$-direction. This shows that $\left| \int_{I \times \Sigma_\bullet} \langle \nu \wedge [*_\epsilon \beta_t, \theta] \rangle \right|$ is bounded by

$$\begin{aligned} &\left(\sup_I \|\beta_t\|_{L^2(\Sigma_\bullet)} \right) \|\nu\|_{L^2(I, L^4(\Sigma_\bullet))} \|\theta\|_{L^2(I, L^4(\Sigma_\bullet))} \\ &\quad \leq c_0 C_1 \left(\|\nu\|_{L^2(I \times \Sigma_\bullet)} + \|d_\alpha \nu\|_{L^2(I \times \Sigma_\bullet)} + \|d_\alpha^* \nu\|_{L^2(I \times \Sigma_\bullet)} \right) \left(\|d_\alpha \theta\|_{L^2(I \times \Sigma_\bullet)} \right) \end{aligned}$$

where we have used the standard elliptic estimates for the operator $d_\alpha \oplus d_\alpha^*$ (with respect to the fixed metric). Converting back to the ϵ-dependent norms, we can continue this as follows:

$$\begin{aligned} &= c_0 C_1 \left(\|\nu\|_{L^2(I\times\Sigma_\bullet),\epsilon} + \epsilon \|d_\alpha \nu\|_{L^2(I\times\Sigma_\bullet),\epsilon} + \epsilon \|d_\alpha^{*_\epsilon} \nu\|_{L^2(I\times\Sigma_\bullet),\epsilon} \right) \\ &\quad \times \left(\|d_\alpha \theta\|_{L^2(I\times\Sigma_\bullet),\epsilon} \right) \\ &\leq C_2 \|\nu\|^2_{L^2(I\times\Sigma_\bullet),\epsilon} \\ &\quad + \tfrac{1}{2} \left(\|d_\alpha \nu\|^2_{L^2(I\times\Sigma_\bullet),\epsilon} + \|d_\alpha^{*_\epsilon} \nu\|^2_{L^2(I\times\Sigma_\bullet),\epsilon} + \|d_\alpha \theta\|^2_{L^2(I\times\Sigma_\bullet),\epsilon} \right), \end{aligned}$$

provided ϵ is sufficiently small. This finishes the bound on the $I \times \Sigma_\bullet$ part of (64), since the derivative terms appearing here can be absorbed by the analogous terms in (62).

For the first term on the right-hand side of (64) ($Y_\bullet$ part), we use the fact that β_t is the dt-component of F_a and we have a bound of the form $\|F_a\|_{L^2(Y_\bullet)} \leq \epsilon c_0$. This gives

$$\begin{aligned} \| *_\epsilon \beta_t \|_{L^2(Y_\bullet)} &= \epsilon \|\beta_t\|_{L^2(Y_\bullet)} = |dt| \|\beta_t\|_{L^2(Y_\bullet)} \\ = \|dt \wedge \beta_t\|_{L^2(Y_\bullet)} &\leq \|F_a\|_{L^2(Y_\bullet)} \leq \epsilon c_0, \end{aligned}$$

where in the first equality we used $*_\epsilon \beta_t = \epsilon * \beta_t$, and in the second we used $\epsilon = \epsilon |dt|_\epsilon = |dt|$, which holds due to the ϵ-scaling on $Y_\bullet$. At this point the computation is similar to the one above. We begin by writing

$$\left| \int_{Y_\bullet} \langle \nu \wedge [*_\epsilon \beta_t, \theta] \rangle \right| \leq \| *_\epsilon \beta_t \|_{L^2(Y_\bullet)} \|\nu\|_{L^4(Y_\bullet)} \|\theta\|_{L^4(Y_\bullet)}.$$

As before we want to use the elliptic estimates for the operator $d_a \oplus d_a^*$, with respect to the fixed metric. We recall that the appropriate ϵ-independent 't'-derivative on $Y_\bullet$ is actually $\epsilon \nabla_t$, since ∇_t is the object defined with respect to the ϵ-dependent metric. This allows us to bound $|\int_{Y_\bullet} \langle \nu \wedge [*_\epsilon \beta_t, \theta] \rangle|$ by

$$\begin{aligned} &\epsilon c_0 C_3 \left(\|\nu\|_{L^2(Y_\bullet)} + \|\epsilon \nabla_t \nu\|_{L^2(Y_\bullet)} + \|d_\alpha \nu\|_{L^2(Y_\bullet)} + \|d_\alpha^* \nu\|_{L^2(Y_\bullet)} \right) \\ &\quad \times \left(\|\epsilon \nabla_t \theta\|_{L^2(Y_\bullet)} + \|d_\alpha \theta\|_{L^2(Y_\bullet)} \right). \end{aligned}$$

Now we continue this by converting back to the ϵ-dependent metric:

$$\begin{aligned} &\leq \epsilon c_0 C_3 \Big(\epsilon^{-1/2} \|\nu\|_{L^2(Y_\bullet),\epsilon} + \epsilon^{-1/2} \|\epsilon \nabla_t \nu\|_{L^2(Y_\bullet),\epsilon} + \epsilon^{1/2} \|d_a \nu\|_{L^2(Y_\bullet),\epsilon} \\ &\quad + \epsilon^{1/2} \|d_a^{*_\epsilon} \nu\|_{L^2(Y_\bullet),\epsilon} \Big) \cdot \left(\epsilon^{-1/2} \|\epsilon \nabla_t \theta\|_{L^2(Y_\bullet),\epsilon} + \epsilon^{1/2} \|d_a \theta\|_{L^2(Y_\bullet),\epsilon} \right) \\ &= c_0 C_3 \Big(\|\nu\|_{L^2(Y_\bullet),\epsilon} + \epsilon \|\nabla_t \nu\|_{L^2(Y_\bullet),\epsilon} + \epsilon \|d_a \nu\|_{L^2(Y_\bullet),\epsilon} \\ &\quad + \epsilon \|d_a^{*_\epsilon} \nu\|_{L^2(Y_\bullet),\epsilon} \Big) \cdot \left(\epsilon \|\nabla_t \theta\|_{L^2(Y_\bullet),\epsilon} + \epsilon \|d_a \theta\|_{L^2(Y_\bullet),\epsilon} \right) \\ &\leq c_0 C_4 \|\nu\|^2_{L^2(Y_\bullet),\epsilon} + \epsilon^2 \|\nabla_t \nu\|^2_{L^2(Y_\bullet),\epsilon} + \epsilon^2 \|d_a \nu\|^2_{L^2(Y_\bullet),\epsilon} \\ &\quad + \epsilon^2 \|d_a^{*_\epsilon} \nu\|^2_{L^2(Y_\bullet),\epsilon} + \epsilon^2 \|\nabla_t \theta\|^2_{L^2(Y_\bullet),\epsilon} + \epsilon^2 \|d_a \theta\|^2_{L^2(Y_\bullet),\epsilon}. \end{aligned}$$

When ϵ^2 is small, the derivative terms appearing here can be absorbed into the analogous terms in (62), so this completes the proof in the simplified situation with no critical points.

Now we describe how to adjust the above argument to accommodate the case where f has N critical points $\{p_i\}$. We recall that N is even and each $Y_{i(i+1)}$ contains a unique critical point p_i with index 1 or 2. Note that the 1-form $dt = df/|df|_\epsilon$ is well-defined on the complement $Y \setminus \{p_i\}_i$ of the critical points. Consequently, the component functions appearing in (60) are also defined in this region. Write $c_i = f(p_i)$ for the critical value associated to p_i. Fix $r > 0$ small, and set

$$Y_r := f^{-1}\left(S^1 \setminus \cup_i B_r(c_i)\right), \tag{65}$$

where $B_r(c_i) \subset S^1$ is the closed interval of radius r around c_i. We extend this to $r = 0$ by declaring Y_0 to be the complement in Y of the critical fibers $\cup_i f^{-1}(c_i)$. Note that Y_0 has full measure in Y.

We will repeat the calculations above with Y replaced by Y_r. Due to the integration by parts, this will result in some additional boundary terms, but we will see that these cancel as r goes to zero. Explicitly, note that the computation of (62) holds with all norms interpreted as being over Y_r instead of Y. The new feature occurs in (63) where the integration by parts in ∇_t gives the aforementioned boundary terms:

$$\begin{aligned} &-2\left(\nabla_t\theta, d_\alpha^{*_\epsilon}\nu\right)_{L^2(Y_r),\epsilon} - 2\left(\nabla_t\nu, d_\alpha\theta\right)_{L^2(Y_r),\epsilon} \\ &= 2\left(\nu, [\beta_t, \theta]\right)_{L^2(Y_r),\epsilon} - 2\int_{\partial Y_r} (\nu, d_\alpha\theta)_\epsilon. \end{aligned}$$

Our above estimates for the first term in the second line remains valid, so we therefore have

$$\begin{aligned} &\|d_\alpha\nu\|^2_{L^2(Y_r),\epsilon} + \|d_\alpha^{*_\epsilon}\nu\|^2_{L^2(Y_r),\epsilon} + \|\nabla_t\nu\|^2_{L^2(Y_r),\epsilon} \\ &+\|d_\alpha\theta\|^2_{L^2(Y_r),\epsilon} + \|\nabla_t\theta\|^2_{L^2(Y_r),\epsilon} - 2\int_{\partial Y_r} (\nu, d_\alpha\theta)_\epsilon \\ \leq\ & C\left(\|v\|^2_{L^2(Y_r),\epsilon} + \|d_a v\|^2_{L^2(Y_r),\epsilon} + \|d_a^{*_\epsilon} v\|^2_{L^2(Y_r),\epsilon}\right) \\ \leq\ & C\left(\|v\|^2_{L^2(Y),\epsilon} + \|d_a v\|^2_{L^2(Y),\epsilon} + \|d_a^{*_\epsilon} v\|^2_{L^2(Y),\epsilon}\right). \end{aligned}$$

Since the right-hand side is independent of r, we will be done if we can show

$$\lim_{r\to 0^+} \int_{\partial Y_r} (\nu, d_\alpha\theta)_\epsilon = 0.$$

This inner product is independent of ϵ by the conformal scaling properties of 1-forms on surfaces. The manifold Y_r is a deformation retract of the cobordism $I \times \Sigma_\bullet$, and so can be viewed as a cobordism between surfaces S_r^- and S_r^+, with $S_r^\pm \cong \Sigma_\bullet$. See Figure 5.

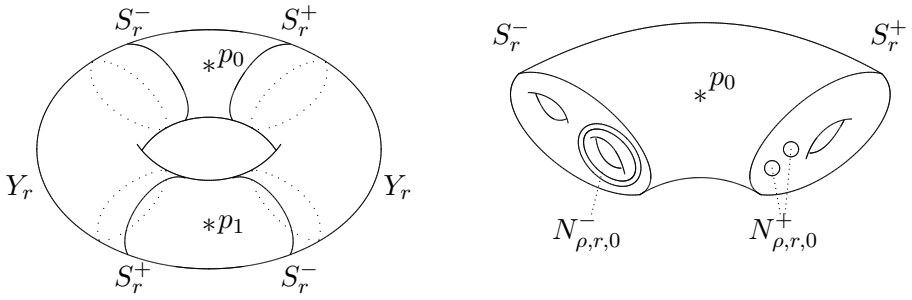

FIGURE 5. *The manifolds Y_r, S_r^- and S_r^+ each have N connected components, where N is the number of critical points of the Morse function $f : Y \to S^1$. The case $N = 2$ is illustrated on the left above. The surfaces $S_r^\pm$ are indicated by the solid lines, and these bound the cobordism Y_r. The dotted lines indicate the seams (where the cobordisms $Y_\bullet$ and $I \times \Sigma_\bullet$ meet). The figure on the right is a larger illustration of the region around the critical point p_0. Taking the standard orientation of the circle, this is a critical point of index 1. The annulus labeled $N^-_{\rho,r,0}$ is the portion of the neighborhood $N^-_{\rho,r}$ that lies in the component of S_r^- closest to p_0. Similarly, the pair of disks labeled $N^+_{\rho,r,0}$ is the portion of $N^+_{\rho,r}$ in the component of S_r^+ closest to p_0. Though this is not illustrated, the intersection of S_r^- with the stable manifold of p_0 is a loop lying in the middle of the annulus $N^-_{\rho,r,0}$. Similarly, the unstable manifold of p_0 determines a pair of points in S_r^+, and these are the centers of the disks $N^+_{\rho,r,0}$.*

In terms of the notation $S_r^\pm$, to prove the proposition it suffices to show that

$$\left(\int_{S_r^+} (\nu, d_\alpha \theta) - \int_{S_r^-} (\nu, d_\alpha \theta) \right) \longrightarrow 0 \tag{66}$$

as r decreases to zero. Since S_r^- and S_r^+ are both converging to the same fiber $f^{-1}(\cup_i c_i)$ as r approaches 0, verifying the limit (66) would be straightforward if, for example, we knew the integrands were L^∞. Unfortunately, in general these integrands are not L^∞. The issue arises from the fact that the coordinate decomposition $v = \nu + \theta\, dt$ is only valid *away* from the critical points of f. Though this implies ν, θ are L^∞ on all of Y (e.g., $|\nu| \leq |v|$), the derivative $d_\alpha \theta$ is typically not L^∞. However, we will see that it is L^1, and this will be enough to verify (66).

To carry this out, we begin by isolating the problem. For $\rho > 0$, let $N^\pm_{\rho,r} \subset S_r^\pm$ denote the set of points in $S_r^\pm$ that are within a distance of ρ to a gradient trajectory of the restriction $f| : Y_r \to S^1$. Then $N^+_{\rho,r}$ is a collection of $N/2$ annuli and $N/2$ pairs of disks; a pair of disks arise from

the unstable (resp. stable) manifold of an index 1 (resp. 2) critical point, and an annulus from the stable (resp. unstable) manifold of an index 1 (resp. 2) critical point. The same holds for $N^-_{\rho,r}$. The neighborhoods $N^{\pm}_{\rho,r}$ contain the regions in which $d_\alpha\theta$ is poorly behaved. The next claim is the key estimate in controlling $d_\alpha\theta$ in this region.

Claim: There is a constant C so that

$$\int_{N^{\pm}_{\rho,r}} |(\nu, d_\alpha\theta)| \leq C(\rho + r)$$

for all $\rho > 0$, $r > 0$ sufficiently small. The constant depends on f, v, and the fixed metric.

Before proving the claim, we will show how it used to finish the proof of (66), and therefore the proposition. Let $S_0 := f^{-1}(\cup_i c_i)$ denote the critical fibers of f. Then the normalized gradient flow of $\mp f$ provides embeddings

$$\varphi^{\pm}_r : S_0 \backslash \{p_j\}_j \to S^{\pm}_r.$$

The image of $\varphi^{\pm}_r$ is the complement in $S^{\pm}_r$ of the stable and unstable manifolds of $f|_{Y\backslash Y_r}$. The family $\{\varphi^+_r\}_r$ varies continuously in r and approaches the inclusion

$$\varphi_0 : S_0 \backslash \{p_j\} \hookrightarrow S_0$$

as r approaches 0; the same holds for $\{\varphi^-_r\}_r$. Each function $(\nu, d_\alpha\theta)|_{S^{\pm}_r}$ is well-defined and smooth on the complement of the critical points p_j. These observations imply that the family of functions

$$(\varphi^+_r)^* \left((\nu, d_\alpha\theta) \,|_{S^+_r} \right) - (\varphi^-_r)^* \left((\nu, d_\alpha\theta) \,|_{S^-_r} \right) : S_0 \backslash \{p_j\} \longrightarrow \mathbb{R}$$

is continuous in r and converges pointwise, as r approaches zero, to the zero function on $S_0 \backslash \{p_j\}$. On any compact $K \subset S_0 \backslash \{p_j\}$ this convergence is therefore uniform, and so we have

$$\begin{aligned} &\int_{\varphi^+_r(K)} (\nu, d_\alpha\theta) - \int_{\varphi^-_r(K)} (\nu, d_\alpha\theta) \\ &\quad = \int_K (\varphi^+_r)^* \left((\nu, d_\alpha\theta) \,|_{S^+_r} \right) - (\varphi^-_r)^* \left((\nu, d_\alpha\theta) \,|_{S^-_r} \right) \xrightarrow{r} 0. \end{aligned} \tag{67}$$

Combining this observation with the claim, we obtain the desired convergence in (66): Fix $\delta > 0$. By the claim we can ensure

$$\int_{N^{\pm}_{\rho,r}} |(\nu, d_\alpha\theta)| \leq \delta/3$$

for all $\rho, r > 0$ sufficiently small, which we assume is the case; we will refine the choice of r momentarily. Notice that the functions $\varphi^{\pm}_r$ only drastically change the metric near the stable/unstable manifold. In particular, there is

some compact $K \subset S_0 \backslash \{p_j\}$ so that $S_r^{\pm} \backslash \varphi_r^{\pm}(K) \subset N_{\rho,r}$ for all $r > 0$. By (67), we have

$$\left| \int_{\varphi_r^+(K)} (\nu, d_\alpha \theta) - \int_{\varphi_r^-(K)} (\nu, d_\alpha \theta) \right| \leq \delta/3,$$

for any sufficiently small $r > 0$. Putting this all together gives

$$\begin{aligned} \left| \int_{S_r^+} (\nu, d_\alpha \theta) - \int_{S_r^-} (\nu, d_\alpha \theta) \right| &\leq \int_{S_r^+ \backslash \varphi_r^+(K)} |(\nu, d_\alpha \theta)| + \int_{S_r^- \backslash \varphi_r^-(K)} |(\nu, d_\alpha \theta)| \\ &\quad + \left| \int_{\varphi_r^+(K)} (\nu, d_\alpha \theta) - \int_{\varphi_r^+(K)} (\nu, d_\alpha \theta) \right| \\ &\leq \int_{N_{\rho,r}^+} |(\nu, d_\alpha \theta)| + \int_{N_{\rho,r}^-} |(\nu, d_\alpha \theta)| + \delta/3 \\ &\leq \delta \end{aligned}$$

Since this holds for all r sufficiently small, this completes the proof of (66).

To finish the proof of the proposition it therefore suffices to prove the claim. Fix a critical point p_i and denote by $N_{\rho,r,i}^{\pm}$ the component of $N_{\rho,r}^{\pm}$ that is closest to p_i; see Figure 5. With this notation, we then have

$$\int_{N_{\rho,r}^{\pm}} |(\nu, d_\alpha \theta)| \leq \|v\|_{L^\infty(Y)} \sum_i \|d_\alpha \theta\|_{L^1(N_{\rho,r,i}^{\pm})}.$$

Since $\|v\|_{L^\infty(Y)}$ is independent of ρ, r, it suffices to bound each $\|d_\alpha \theta\|_{L^1(N_{\rho,r,i}^{\pm})}$. We will do this by expressing $d_\alpha \theta$ in terms of smooth coordinates on Y, and then estimating to show that the coefficients in these coordinates are L^1. Fix a critical point p_i. Without loss of generality, we may assume p_i has index 1 with respect to the positive orientation of the circle. Then the Morse lemma allows us to identify a neighborhood of p_i with the set

$$\{(x_1, x_2, x_3, \zeta) \in U \times B_r(0) \subset \mathbb{R}^3 \times \mathbb{R} \mid -x_1^2 + x_2^2 + x_3^2 = \zeta\} \tag{68}$$

for some open neighborhood U of the origin in $\mathbb{R}^3$. To simplify the discussion we assume the metric on Y near p_i agrees with the metric on (68) induced from the standard metric on $\mathbb{R}^3 \times \mathbb{R}$; the more general situation can be reduced to this by noting that (i) two different metrics on Y induce equivalent L^1-norms on $N_{\rho,r,i}^{\pm}$, and (ii) the constants defining this equivalence can be chosen to depend only on the metrics (that is, the constants do not depend ρ, r since the neighborhoods $N_{\rho,r,i}^{\pm}$ are all contained in a compact region).

In terms of the Morse coordinates (68), we have the following:

- p_i is identified with $(0, 0, 0, 0)$,
- the function f is the projection $(x_1, x_2, x_3, \zeta) \mapsto \zeta + f(p_i)$,
- the unstable manifold is the set of points $(x_1, 0, 0, -x_1^2)$, and

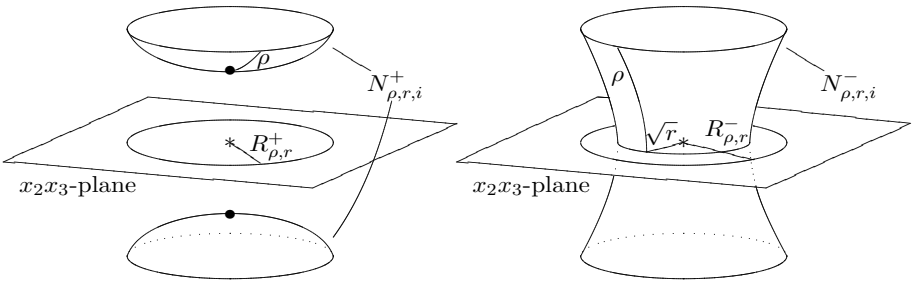

FIGURE 6. *The figure on the left illustrates the pair of disks $N^+_{\rho,r,i}$ with center $\bullet$ and radius ρ. The vertical (x_1-)axis is the unstable manifold of p_i, and the intersection of this with $N^+_{\rho,r,i}$ consists of the two centers $\bullet$. The projection of $N^+_{\rho,r,i}$ to the x_2x_3-plane is illustrated as the circle of radius $R^+_{\rho,r}$ with center $*$, the origin. The figure on the right illustrates the annulus $N^-_{\rho,r,i}$ and its projection to the x_2x_3-plane. This projection is another annulus with inner radius $\sqrt{r}$ and outer radius $R^-_{\rho,r}$. The stable manifold of p_i is the x_2x_3-plane and ρ is the distance in $N^+_{\rho,r}$ from the stable manifold to the boundary of $N^+_{\rho,r}$.*

- the stable manifold is the set of points $(0, x_2, x_3, x_2^2 + x_3^2)$.

By choosing U appropriately, we can arrange so that $N^\pm_{\rho,r,i}$ is identified, under these coordinates, with the fiber $\zeta = \mp r$. That is, $N^+_{\rho,r,i}$ is the set of points $(x_1, x_2, x_3, -r)$ in (68) that are within a distance of ρ to the unstable manifold, and $N^-_{\rho,r,i}$ is the set of points (x_1, x_2, x_3, r) in (68) that are within a distance of ρ to the stable manifold. See Figure 6.

The restriction of the principal bundle Q to the coordinate patch (68) is trivializable, and we let d denote the trivial connection coming from a fixed trivialization. Then $d_\alpha\theta$ and $d\theta$ differ by the L^∞ form $[\alpha, \theta]$, so to prove the claim it suffices to bound the L^1-norm of $d\theta$. The coordinates x_1, x_2, x_3 are smooth functions near (and at) the critical points of f. In particular, writing

$$v = w_1 dx_1 + w_2 dx_2 + w_3 dx_3$$

it follows that the w_i are smooth because v is smooth. We also have

$$dt = \frac{1}{|df|} df = \frac{2}{|df|} \left(x_1 dx_1 - x_2 dx_2 - x_3 dx_3\right)$$

and so comparing with the coordinates $v = \nu + \theta\, dt$, we find $\theta = \frac{1}{2x_1}|df| w_1$. This gives

$$d\theta \;=\; \frac{dx_2}{2x_1}\left((\partial_2|df|)w_1 + |df|\partial_2 w_1\right) + \frac{dx_3}{2x_1}\left((\partial_3|df|)w_1 + |df|\partial_3 w_1\right).$$

Taking the norm, we obtain

$$|d\theta| \leq \frac{C}{|x_1|} = \frac{C}{\sqrt{-\zeta + x_2^2 + x_3^2}}, \tag{69}$$

where C depends only on v and f.

First we analyze the integral of $|d\theta|$ over $N^+_{\rho,r,i}$. Recall $N^+_{\rho,r,i}$ is a pair of disks corresponding to the fiber $\zeta = -r$. Projecting $N^+_{\rho,r,i}$ to the x_2x_3-plane is 2-1 with image $\left\{(x_2, x_3) \,|\, x_2^2 + x_3^2 \leq (R^+_{\rho,r})^2\right\}$, a disk with some radius $R^+_{\rho,r} > 0$. It is easy to check that $R^+_{\rho,r} \leq \rho$. Now integrating (69) over both disks in $N^+_{\rho,r,i}$ we get

$$\|d\theta\|_{L^1(N^+_{\rho,r,i})} \leq 4\pi C \left(\sqrt{r + (R^+_{\rho,r})^2} - \sqrt{r}\right) \leq 4\pi C \left(\sqrt{r + \rho^2} - \sqrt{r}\right),$$

which is the desired estimate for the region $N^+_{\rho,r,i}$.

Now we move on to $N^-_{\rho,r,i}$, which is an annulus corresponding to the fiber $\zeta = r$. Projecting this annulus to the x_2x_3-plane is a map that is 2-1 off of the intersection of this annulus with this plane. The image is the annulus $\left\{(x_2, x_3) \,|\, r \leq x_2^2 + x_3^2 \leq (R^-_{\rho,r})^2\right\}$ for some radius $\sqrt{r} < R^-_{\rho,r} \leq \rho + \sqrt{r}$. The claim then follows by integrating (69) over $N^-_{\rho,r,i}$:

$$\|d\theta\|_{L^1(N^-_{\rho,r,i})} \leq 4\pi C\sqrt{-r + (R^+_{\rho,r})^2} \leq 4\pi C\sqrt{\rho^2 + 2\sqrt{r}\rho}.$$

□

Now we begin to estimate the derivatives of F_A for an ϵ-ASD connection.

Proposition 4.12. *(Instanton bootstrapping estimate; 1st order) Fix $c_0 > 0$, open $\Omega \subset \mathbb{R}$, and compact $K \subset \Omega$. Then there are $\epsilon_0, C > 0$ so that*

$$\|\nabla_s F_A\|_{L^2(K\times Y),\epsilon} \leq C(1 + \mathrm{vol}(K) + \|F_A\|_{L^2(\Omega\times Y),\epsilon})$$

for all $0 < \epsilon < \epsilon_0$ and all ϵ-ASD connections A satisfying (35) and (36).

Remark 4.13. *The proof will show that the constants ϵ_0, C can be chosen to depend on K, Ω only through the value $1/\mathrm{dist}(\partial K, \partial\Omega)$ (actually, polynomially in this quantity). In particular, if $\Omega = \mathbb{R}$, then they can be taken to be independent of K. Similarly, by taking $K = [n, n+1]$, $\Omega = (n-1/2, n+3/2)$, and then summing over n we get*

$$\|\nabla_s F_A\|_{L^2(\mathbb{R}\times Y),\epsilon} \leq 2C(2 + \|F_A\|_{L^2(\mathbb{R}\times Y),\epsilon}).$$

Before proving Proposition 4.12, we note that combining Propositions 4.11 and 4.12 we obtain uniform bounds for various other first derivatives of the curvature.

Corollary 4.14. *Under the assumptions of Proposition 4.12, the following are bounded by $C(1 + \|F_A\|_{L^2(\mathbb{R}\times Y),\epsilon})$:*

$$\begin{gathered}\|d_\alpha \beta_s\|_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon} + \|d^{*_\epsilon}_\alpha \beta_s\|_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon} + \|\nabla_t \beta_s\|_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon} \\ \|d_\alpha \gamma\|_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon} + \|d^{*_\epsilon}_\alpha \gamma\|_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon} + \|\nabla_t \gamma\|_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon} \\ \|\beta_s\|_{L^4(\mathbb{R}\times I\times\Sigma_\bullet)} + \|\,[\beta_s,\gamma]\,\|_{L^2(\mathbb{R}\times I\times\Sigma_\bullet)}.\end{gathered} \tag{70}$$

The constant C depends only on c_0. If $K \subseteq \mathbb{R}$ is any compact set, then

$$\|b_s\|_{L^4(K\times Y_\bullet)} \leq C(1 + \mathrm{vol}(K) + \|F_A\|_{L^2(\mathbb{R}\times Y),\epsilon}).$$

Note that the first six norms in (70) are ϵ-dependent, while the remaining norms in Corollary 4.14 are relative to the fixed metric.

Proof of Corollary 4.14. For the items in the first and second rows of (70), apply Proposition 4.12 with $v = b_s$, and use the identities

$$d_a b_s = \nabla_s F_a, \quad d^{*_\epsilon}_a b_s = 0,$$

together with $\nabla_s F_A = \nabla_s F_a + ds \wedge \nabla_s F_a$.

We can estimate the items in the third row by writing

$$\|\beta_s\|_{L^4(\mathbb{R}\times I\times\Sigma_\bullet)} + \|\,[\beta_s,\gamma]\,\|_{L^2(\mathbb{R}\times I\times\Sigma_\bullet)} \quad\leq\quad c_0\left(\|\beta_s\|_{L^2(L^4)} + \|\gamma\|_{L^2(L^4)}\right),$$

where we have set $L^2(L^4) := L^2(\mathbb{R}\times I, L^4(\Sigma_\bullet))$ and used (35). Using the embedding $W^{1,2} \hookrightarrow L^4$ on $\Sigma_\bullet$, we can bound this by

$$\begin{aligned} & c_0 C\left(\|\beta_s\|_{L^2} + \|d_\alpha\beta_s\|_{L^2} + \|d^*_\alpha\beta_s\|_{L^2} + \|d_\alpha\gamma\|_{L^2}\right) \\ = \quad & c_0 C\left(\|\beta_s\|_{L^2,\epsilon} + \epsilon\|d_\alpha\beta_s\|_{L^2,\epsilon} + \epsilon\|d^{*_\epsilon}_\alpha\beta_s\|_{L^2,\epsilon} + \|d_\alpha\gamma\|_{L^2,\epsilon}\right)\end{aligned}$$

where $L^2 := L^2(\mathbb{R}\times I\times\Sigma_\bullet)$. We have already bounded these in the first and second row, so the result for the third row of (70) follows.

It remains to bound $\|b_s\|_{L^4(K\times Y_\bullet)}$. For this, write

$$\|b_s\|_{L^4(K\times Y_\bullet)} \leq c_0\|b_s\|_{L^2(L^4)},$$

where now $L^2(L^4) := L^2(K, L^4(Y_\bullet))$ and we have used (35) again. This is bounded by

$$\begin{aligned} & c_0 C\left(\|b_s\|_{L^2(K\times Y_\bullet)} + \|d_a b_s\|_{L^2(K\times Y_\bullet)}\right) \\ \leq\ & c_0 C\left(\|b_s\|_{L^2(K\times Y_\bullet)} + \epsilon^{1/2}\|\nabla_s F_A\|_{L^2(\mathbb{R}\times Y),\epsilon}\right).\end{aligned}$$

Using (35) one last time, we have $\|b_s\|_{L^2(K\times Y)} \leq \mathrm{vol}(K)c_0$ is bounded. Finally, use Proposition 4.12 to control $\epsilon^{1/2}\|\nabla_s F_A\|_{L^2,\epsilon}$. □

Proof of Proposition 4.12. In what follows, all unspecified integrals and Hodge stars are over $\mathbb{R}\times Y$; for example, $L^2 = L^2(\mathbb{R}\times Y)$. Observe that

$$\nabla_s F_A = \partial_s F_A + [p, F_A] = d_A\partial_s A - d_A(d_A p) = d_A\left(\partial_s A - d_A p\right).$$

and

$$\partial_s A - d_A p = \partial_s a - d_a p + (\partial_s p - \nabla_s p)\ ds = \partial_s a - d_a p = b_s.$$

In particular, we get

$$\nabla_s F_A = d_A b_s,$$

where d_A is the derivative on the *4-manifold* $\mathbb{R} \times Y$, and in this formula we are viewing b_s as a 1-form on this 4-manifold (as opposed to a path of 1-forms on a 3-manifold).

Let $h : \mathbb{R} \to [0, 1]$ be a compactly supported bump function for $K \subset \Omega$. Then there is a constant C_h so that $|\partial_s h| \leq C_h$; the dependence on K, Ω, h of all constants will only be through the values C_h and $\mathrm{vol}(\mathrm{supp}(h))$.

We have

$$\|h \nabla_s F_A\|^2_{L^2,\epsilon} = \int h^2 \langle \nabla_s F_A \wedge *_\epsilon \nabla_s F_A \rangle = \int h^2 \langle d_A b_s \wedge *_\epsilon \nabla_s F_A \rangle.$$

By Stokes' theorem, we obtain

$$\begin{aligned} \|h \nabla_s F_A\|^2_{L^2,\epsilon} &= \int h^2 \langle d_A b_s \wedge *_\epsilon \nabla_s F_A \rangle \\ &= -\int 2h\ \partial_s h\ ds \wedge \langle b_s \wedge *_\epsilon \nabla_s F_A \rangle + \int h^2 \langle b_s \wedge d_A \nabla_s *_\epsilon F_A \rangle \\ &= -\int 2h\ \partial_s h\ ds \wedge \langle b_s \wedge *_\epsilon \nabla_s F_A \rangle - \int h^2 \langle b_s \wedge [b_s \wedge *_\epsilon F_A] \rangle, \end{aligned}$$

where, in the last step, we used $\nabla_s d_A = d_A \nabla_s + [b_s, \cdot]$ and the ϵ-ASD condition. Next, use the inequality

$$2ab \leq \delta^{-1} a^2 + \delta b^2, \quad \delta > 0 \tag{71}$$

with $\delta = 5$ to get

$$\|h \nabla_s F_A\|^2_{L^2,\epsilon} \leq 5C_h \|b_s\|^2_{L^2,\epsilon} + \tfrac{1}{5} \|h \nabla_s F_A\|^2_{L^2,\epsilon} + \int h^2 \langle b_s \wedge [b_s \wedge *_\epsilon F_A] \rangle.$$

Subtract the term $\frac{1}{5} \|h \nabla_s F_A\|^2_{L^2,\epsilon}$ from both sides to get

$$\frac{4}{5} \|h \nabla_s F_A\|^2_{L^2,\epsilon} \leq 5C_h \|b_s\|^2_{L^2,\epsilon} + \int h^2 \langle b_s \wedge [b_s \wedge *_\epsilon F_A] \rangle. \tag{72}$$

It suffices to bound the second term on the right. For this, we note that in terms of the Hodge star $*^Y_\epsilon$ on Y we have

$$*_\epsilon F_A = ds \wedge *^Y_\epsilon F_a + *^Y_\epsilon b_s,$$

so this second term is just

(73)
$$\int h^2 ds \wedge \langle b_s \wedge \left[b_s \wedge *_\epsilon^Y F_a\right]\rangle = \int_{\mathbb{R}\times Y_\bullet} h^2 ds \wedge \langle b_s \wedge \left[b_s \wedge *_\epsilon^Y F_a\right]\rangle + \int_{\mathbb{R}\times I\times\Sigma_\bullet} h^2 ds \wedge \langle b_s \wedge \left[b_s \wedge *_\epsilon^Y F_a\right]\rangle.$$

We will be done if we can satisfactorily estimate (73); we begin by estimating the first integral on the right. Note that on $Y_\bullet$ we have $*_\epsilon^Y F_a = \epsilon^{-1} *^Y F_a$, so by (35) the integral $\int_{\mathbb{R}\times Y_\bullet} h^2 ds \wedge \langle b_s \wedge \left[b_s \wedge *_\epsilon^Y F_a\right]\rangle$ is controlled by

$$c_0 \|hb_s\|^2_{L^2(\mathbb{R},L^4(Y_\bullet))} \leq c_0 C_1 \left(\|hb_s\|^2_{L^2(\mathbb{R}\times Y_\bullet)} + \|hd_a b_s\|^2_{L^2(\mathbb{R}\times Y_\bullet)}\right),$$

where we have used the Sobolev embedding $W^{1,2}(Y_\bullet) \hookrightarrow L^4(Y_\bullet)$ and the ϵ-ASD condition $d_a^* b_s = \epsilon^2 d_a^{*_\epsilon} b_s = 0$ on $Y_\bullet$. Using (35) again, we can bound $\|hb_s\|^2_{L^2(\mathbb{R}\times Y_\bullet)}$ by c_0^2 times the volume of the support of h. To control $\|hd_a b_s\|^2_{L^2(\mathbb{R}\times Y_\bullet)}$, we convert back to the ϵ-dependent norm to write

$$\|hd_a b_s\|^2_{L^2(\mathbb{R}\times Y_\bullet)} = \epsilon \|h\nabla_s F_a\|^2_{L^2(\mathbb{R}\times Y_\bullet),\epsilon} \leq \epsilon \|h\nabla_s F_A\|^2_{L^2,\epsilon}.$$

Taking $\epsilon < 1/5$, this can be absorbed by the left-hand side of (72).

It remains to estimate the second integral in (73). Expanding b_s and F_a into components on $I \times \Sigma_\bullet$, this becomes

$$\int_{\mathbb{R}\times I\times\Sigma_\bullet} h^2 ds\wedge dt\wedge\langle\beta_s\wedge\left[\beta_s \wedge *_\epsilon^\Sigma F_\alpha\right]\rangle + 2\int_{\mathbb{R}\times I\times\Sigma_\bullet} h^2 ds\wedge dt\wedge\langle\gamma\wedge\left[\beta_s \wedge *_\epsilon^\Sigma \beta_t\right]\rangle.$$

Using (35), this is controlled by

$$c_0 C_2 \|h\beta_s\|_{L^2(\mathbb{R}\times I,L^4(\Sigma_\bullet))} \left(\|h\epsilon^{-2} F_\alpha\|_{L^2(\mathbb{R}\times I,L^4(\Sigma_\bullet))} + \|h\gamma\|_{L^2(\mathbb{R}\times I,L^4(\Sigma_\bullet))}\right)$$

$$\leq C_3 \left(\|h\beta_s\|_{L^2} + \|hd_\alpha\beta_s\|_{L^2} + \|hd_\alpha^*\beta_s\|_{L^2}\right)\left(\|h\epsilon^{-2} d_\alpha^* F_\alpha\|_{L^2} + \|hd_\alpha\gamma\|_{L^2}\right),$$

where the L^2-norms are over $\mathbb{R}\times I\times\Sigma_\bullet$. Using (71) and converting back to the ϵ-dependent norms, we can bound this by

(74)
$$C_4\delta^{-1}\left(\|h\beta_s\|^2_{L^2,\epsilon} + \epsilon^2\|hd_\alpha\beta_s\|^2_{L^2,\epsilon} + \epsilon^2\|hd_\alpha^{*_\epsilon}\beta_s\|^2_{L^2,\epsilon}\right) + \delta\|hd_\alpha^{*_\epsilon}F_\alpha\|^2_{L^2,\epsilon} + \delta\|hd_\alpha\gamma\|^2_{L^2,\epsilon}.$$

By Proposition 4.11 applied to $v = b_s$ and $v = *_\epsilon F_a$, the last two terms are bounded by

$$\delta C\left(\|F_A\|_{L^2,\epsilon} + \|h\nabla_s F_A\|^2_{L^2,\epsilon}\right).$$

By taking δ so that $\delta C < 1/5$, the derivative term can be absorbed into the left-hand side of (72); as usual, the non-derivative term is fine. Having fixed δ, we focus now on the remaining derivative terms $C_4\delta^{-1}\epsilon^2(\|hd_\alpha\beta_s\|^2_{L^2,\epsilon} + \|hd_\alpha^{*_\epsilon}\beta_s\|^2_{L^2,\epsilon})$ in (74). By Proposition 4.11, these terms are controlled by

$$CC_4\delta^{-1}\epsilon^2\left(\|F_A\|_{L^2,\epsilon}+\|h\nabla_s F_A\|^2_{L^2,\epsilon}\right).$$

Now take ϵ small enough so $CC_4\delta^{-1}\epsilon^2<1/5$. Then the $\nabla_s F_A$ term can be subtracted and absorbed into the left-hand side of (72). □

The following is a second order version of Proposition 4.11. To simplify the discussion we state it for the special case $v=b_s$ and $v=*_\epsilon F_a$. The point is that we can bound various second order derivatives by the norms of $\nabla_s^2 F_A, \nabla_t\nabla_s F_A$, plus lower order terms. We have singled out the derivatives $\nabla_s^2 F_A, \nabla_t\nabla_s F_A$ because these derivatives scale favorably in ϵ; these terms will then be estimated in Proposition 4.16, below.

Proposition 4.15. *(General elliptic estimates; 2nd order) Fix $c_0>0$, open $\Omega\subset\mathbb{R}$, and compact $K\subset\Omega$. There are $\epsilon_0, C>0$ so that the following holds for all $0<\epsilon<\epsilon_0$ and all ϵ-ASD connections A satisfying (35) and (36):*

$$\begin{aligned}&\|\nabla_s d_\alpha\beta_s\|^2_{L^2(K\times I\times\Sigma_\bullet),\epsilon}+\|\nabla_s d^{*_\epsilon}_\alpha\beta_s\|^2_{L^2(K\times I\times\Sigma_\bullet),\epsilon}\\ &\quad+\|\nabla_t d_\alpha\beta_s\|^2_{L^2(K\times I\times\Sigma_\bullet),\epsilon}+\|\nabla_t d^{*_\epsilon}_\alpha\beta_s\|^2_{L^2(K\times I\times\Sigma_\bullet),\epsilon}\\ &\quad+\|\nabla_s\nabla_t\beta_s\|^2_{L^2(K\times I\times\Sigma_\bullet),\epsilon}+\|\nabla_t\nabla_t\beta_s\|^2_{L^2(K\times I\times\Sigma_\bullet),\epsilon}\\ &\quad+\|\nabla_s d_\alpha\gamma\|^2_{L^2(K\times I\times\Sigma_\bullet),\epsilon}+\|\nabla_t d_\alpha\gamma\|^2_{L^2(K\times I\times\Sigma_\bullet),\epsilon}\\ &\quad+\|\nabla_s\nabla_t\gamma\|^2_{L^2(K\times I\times\Sigma_\bullet),\epsilon}+\|\nabla_t^2\gamma\|^2_{L^2(K\times I\times\Sigma_\bullet),\epsilon}\end{aligned}\tag{75}$$

$$\begin{aligned}&\le C\Big(1+\mathrm{vol}(K)+\|F_A\|_{L^2(\Omega\times Y),\epsilon}\\ &\qquad+\|\nabla_s^2 F_A\|_{L^2(\Omega\times Y),\epsilon}+\sup_{r>0}\|\nabla_t\nabla_s F_A\|_{L^2(\Omega\times Y_r),\epsilon}\Big),\end{aligned}$$

where Y_r is as in (65). The same result holds with β_t (resp. $*_\epsilon F_\alpha$) in place of β_s (resp. γ) on the left.

As with Proposition 4.11, the proof will show that the bound (59) continues to hold with $I\times\Sigma_\bullet$ replaced by any subset of Y_r. Also, the constants only depend on the choice of K,Ω through the distance from ∂K to $\partial\Omega$; see Remark 4.13.

Proof of Proposition 4.15. The only reason we restrict to compact K (rather than all of $\mathbb{R}$) is so we can appeal to Proposition 4.12. To simplify notation, we will ignore this and work with $K=\Omega=\mathbb{R}$, with the understanding that the 'true' computation would involve a bump function with compact support in Ω; the extension to this 'true' case is no more complicated than the situation appearing in Proposition 4.12.

With this understood, we will bound the left-hand side of (75) by a constant times

$$1+\|F_A\|_{L^2(\mathbb{R}\times Y),\epsilon}+\|\nabla_s d_a b_s\|^2_{L^2(\mathbb{R}\times Y),\epsilon}+\sup_{r>0}\|\nabla_t d_a b_s\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon}.$$

Let $r>0$ be small. Then on Y_r we have

$$d_a b_s=d_\alpha\beta_s+dt\wedge(\nabla_t\beta_s-d_\alpha\gamma),\quad 0=d_a^{*_\epsilon}b_s=d_\alpha^{*_\epsilon}\beta_s-\nabla_t\gamma.$$

Apply ∇_s and then ∇_t to both equations, take the norm square and add everything to get that $\|\nabla_s d_a b_s\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} + \|\nabla_t d_a b_s\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon}$ is greater than or equal to

(76)
$$\begin{aligned}
& \|\nabla_s d_\alpha \beta_s\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} + \|\nabla_s d^{*_\epsilon}_\alpha \beta_s\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} + \|\nabla_s \nabla_t \beta_s\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} \\
+ \quad & \|\nabla_t d_\alpha \beta_s\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} + \|\nabla_t d^{*_\epsilon}_\alpha \beta_s\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} + \|\nabla_t \nabla_t \beta_s\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} \\
+ \quad & \|\nabla_s d_\alpha \gamma\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} + \|\nabla_t d_\alpha \gamma\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} \\
+ \quad & \|\nabla_s \nabla_t \gamma\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} + \|\nabla_t^2 \gamma\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} \\
- \quad & 2(\nabla_s \nabla_t \beta_s, \nabla_s d_\alpha \gamma)_{L^2(\mathbb{R}\times Y_r),\epsilon} - 2(\nabla_s d^{*_\epsilon}_\alpha \beta_s, \nabla_s \nabla_t \gamma)_{L^2(\mathbb{R}\times Y_r),\epsilon} \\
- \quad & 2(\nabla_t^2 \beta_s, \nabla_t d_\alpha \gamma)_{L^2(\mathbb{R}\times Y_r),\epsilon} - 2(\nabla_t d^{*_\epsilon}_\alpha \beta_s, \nabla_t^2 \gamma)_{L^2(\mathbb{R}\times Y_r),\epsilon}.
\end{aligned}$$

It suffices to show that we can control the four cross terms at the end. We will work this out explicitly for the terms

(77)
$$- 2(\nabla_s \nabla_t \beta_s, \nabla_s d_\alpha \gamma)_{L^2(\mathbb{R}\times Y_r),\epsilon} - 2(\nabla_s d^{*_\epsilon}_\alpha \beta_s, \nabla_s \nabla_t \gamma)_{L^2(\mathbb{R}\times Y_r),\epsilon};$$

the analysis for the remaining terms (i.e., the last two terms in (76)) is similar. As in the proof of Proposition 4.11, the idea is to integrate by parts. These will then cancel, up to some boundary terms coming from ∂Y_r plus lower-order terms coming from the commutation relations

$$\nabla_s \nabla_t - \nabla_t \nabla_s = \gamma, \quad \nabla_s d_\alpha - d_\alpha \nabla_s = \beta_s, \quad \nabla_t d_\alpha - d_\alpha \nabla_t = \beta_t.$$

Explicitly, we integrate by parts in ∇_t and then in d_α to get that (77) is equal to a linear combination of the boundary term

(78)
$$\int_{\mathbb{R}\times\partial Y_r} (\nabla_s \beta_s, \nabla_s d_\alpha \gamma)$$

together with the following lower order cross terms

(79)
$$([\gamma, \beta_s], \nabla_s d_\alpha \gamma)_{L^2(\mathbb{R}\times Y_r),\epsilon}$$

(80)
$$(\nabla_s \beta_s, \nabla_s [\beta_t, \gamma])_{L^2(\mathbb{R}\times Y_r),\epsilon}$$

(81)
$$(\nabla_s \beta_s, [\gamma, d_\alpha \gamma])_{L^2(\mathbb{R}\times Y_r),\epsilon}$$

(82)
$$(\nabla_s \beta_s, [\beta_s, \nabla_t \gamma])_{L^2(\mathbb{R}\times Y_r),\epsilon}$$

(83)
$$(*^{\Sigma}_\epsilon [\beta_s \wedge *^{\Sigma}_\epsilon \beta_s], \nabla_s \nabla_t \gamma)_{L^2(\mathbb{R}\times Y_r),\epsilon}.$$

The star that appears in (83) is the Hodge star on surfaces. As in the proof of Proposition 4.11, the boundary term (78) goes to zero as r decreases to zero. It therefore suffices to show that the lower order terms (79-83) are suitably bounded with constants independent of r and ϵ.

- (79): This is bounded by

$$\delta \|\nabla_s d_\alpha \gamma\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} + \delta^{-1} C \| [\gamma, \beta_s] \|^2_{L^2(\mathbb{R}\times Y_r),\epsilon}.$$

The first of these terms is good for δ small, since it can be absorbed by analogous terms. To bound the second term, first notice that the portion of the integral over $I \times \Sigma_\bullet$ is controlled by Corollary 4.14. The portion over the complementary region $Y_r \cap Y_\bullet$ can be bounded by the following similar argument: By the scaling properties of the Hodge star on 3-manifolds, we have

$$\| [\gamma, \beta_s] \|^2_{L^2(\mathbb{R}\times Y_r \cap Y_\bullet),\epsilon} = \epsilon \| [\gamma, \beta_s] \|^2_{L^2(\mathbb{R}\times Y_r \cap Y_\bullet)} \leq \epsilon c_0 \|\gamma\|^2_{L^2(\mathbb{R}, L^4(Y_r \cap Y_\bullet))},$$

where we used that β_s is a component of b_s, together with (35). The embedding $W^{1,2} \hookrightarrow L^4$ on $Y^r \cap Y_\bullet$, together with the fact that a is irreducible, implies that there is a bound of the form

$$\begin{aligned} \|\gamma\|^2_{L^2(\mathbb{R}, L^4(Y_r\cap Y_\bullet))} &\leq C\|d_a\gamma\|^2_{L^2(\mathbb{R}\times Y_r\cap Y_\bullet)} \\ &= C\left(\epsilon^2\|\nabla_t\gamma\|^2_{L^2(\mathbb{R}\times Y_r\cap Y_\bullet)} + \|d_\alpha\gamma\|^2_{L^2(\mathbb{R}\times Y_r\cap Y_\bullet)}\right) \\ &= \epsilon^{-1}C\left(\|\nabla_t\gamma\|^2_{L^2(\mathbb{R}\times Y_r\cap Y_\bullet),\epsilon} + \|d_\alpha\gamma\|^2_{L^2(\mathbb{R}\times Y_r\cap Y_\bullet),\epsilon}\right), \end{aligned}$$

where in the second line we used $|dt| = \epsilon$ on the $\nabla_t\gamma$-term. Apply Proposition 4.11 with $v = b_s$ to bound $\nabla_t\gamma$ and $d_\alpha\gamma$ in terms of $d_a b_s$ and $d_a^{*_\epsilon} b_s = 0$. This gives

$$\| [\gamma, \beta_s] \|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} \leq c_0 C' \left(\|b_s\|^2_{L^2(\mathbb{R}\times Y)} + \|d_a b_s\|^2_{L^2(\mathbb{R}\times Y)}\right),$$

which, by Proposition 4.12, is bounded by a constant times $\|F_A\|^2_{L^2(\mathbb{R}\times Y),\epsilon}$.

- (80): Integrate by parts in ∇_s. Then (80) is bounded by

$$\delta\|\nabla_s^2\beta_s\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon} + \delta^{-1}\| [\beta_t, \gamma] \|_{L^2(\mathbb{R}\times Y_r),\epsilon}.$$

The first of these can be absorbed for small δ. The second is controlled as in (79); here note that β_t is a component of F_a, and in particular satisfies

$$\sup_{\mathbb{R}} \|\beta_t\|_{L^2(Y_r\cap Y_\bullet)} \leq \epsilon^{-1} \sup_{\mathbb{R}} \|F_a\|_{L^2(Y_\bullet)} \leq c_0.$$

- (81): Integrate by parts in ∇_s to get that (81) is equal to

$$-(\beta_s, [\nabla_s\gamma, d_\alpha\gamma])_{L^2(\mathbb{R}\times Y_r),\epsilon} - (\beta_s, [\gamma, \nabla_s d_\alpha\gamma])_{L^2(\mathbb{R}\times Y_r),\epsilon}. \tag{84}$$

The second term is exactly (79). The first term in (84) can be controlled as follows: The portion of the integral over $I \times \Sigma_\bullet$ is bounded by

$$c_0(\delta\|d_\alpha\nabla_s\gamma\|^2_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon} + \delta^{-1}\epsilon^2\|d_\alpha^{*_\epsilon}d_\alpha\gamma\|^2_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon}).$$

For the first of these terms, use $d_\alpha\nabla_s\gamma = \nabla_s d_\alpha\gamma + [\beta_s, \gamma]$, then Corollary 4.14 and take δ small to absorb the second order terms. For the second term, use the identity $d_\alpha\gamma = -\nabla_s\beta_t + \nabla_t\beta_s$ to produce terms of the form

$\nabla_s d_\alpha^{*_\epsilon}\beta_t, \nabla_t d_\alpha^{*_\epsilon}\beta_s$ and lower order terms. Having chosen δ, now pick ϵ small enough to absorb the second order terms.

The portion of the integral over $Y_r \cap Y_\bullet$ is bounded similarly. Here one should use

$$\sup_{\mathbb{R}} \|\beta_s\|^2_{L^2(Y_\bullet),\epsilon} = \epsilon \sup_{\mathbb{R}} \|\beta_s\|^2_{L^2(Y_\bullet)} \leq c_0\epsilon$$

as well as the embedding $W^{1,2} \hookrightarrow L^4$ on the 3-manifold $Y_r \cap Y_\bullet$.

- (82): This is similar to the first term in (84).
- (83): Integrate by parts in ∇_s to get (82). □

Proposition 4.16. *(Instanton bootstrapping estimates; 2nd order) Fix $c_0 > 0$, open $\Omega \subset \mathbb{R}$, and compact $K \subset \Omega$. Then there are $\epsilon_0, C > 0$ so that*

$$\|\nabla_s^2 F_A\|_{L^2(K\times Y),\epsilon} + \sup_{r>0} \|\nabla_t \nabla_s F_A\|_{L^2(K\times Y_r),\epsilon} \leq C(1 + \|F_A\|_{L^2(\Omega\times Y),\epsilon})$$

for all $0 < \epsilon < \epsilon_0$ and all ϵ-ASD connections A satisfying (35) and (36).

As usual, the constants only depend on the choice of K, Ω through the distance from ∂K to $\partial\Omega$.

Proof of Proposition 4.16. Fix a compactly supported bump function h for $K \subset \Omega$. All integrals, inner products, norms, etc. are over $\mathbb{R} \times Y$ unless otherwise specified. We begin with $\nabla_s^2 F_A$, but in the end we need to compute this simultaneously with $\nabla_t \nabla_s F_A$. The proof is in many ways quite similar to that of Proposition 4.12, so we will be brief, putting most emphasis on the new features. Use integration by parts and the ϵ-ASD relation $d_A^* F_A = 0$ to get that the quantity $\|h\nabla_s^2 F_A\|^2_{L^2,\epsilon}$ is given by a linear combination of the following terms

$$\int h^2 ds \wedge \langle [\nabla_s b_s \wedge \nabla_s b_s] \wedge b_s \rangle \tag{85}$$

$$\int h(\partial_s h) ds \wedge \langle \nabla_s b_s \wedge *_\epsilon^Y \nabla_s^2 b_s \rangle, \tag{86}$$

where $*_\epsilon^Y$ is the ϵ-dependent Hodge star on Y. We will estimate these in the bullets below (we separate (85) into two integrals, one over $\mathbb{R} \times Y_\bullet$ and one over $\mathbb{R} \times I \times \Sigma_\bullet$).

- (85) on $\mathbb{R} \times Y_\bullet$: Use (35) to control this by

$$c_0 \|h\nabla_s b_s\|^2_{L^2(\mathbb{R}, L^4(Y_\bullet))}.$$

Note that by the ϵ-ASD relation on $\mathbb{R} \times Y_\bullet$, the term $\nabla_s b_s = \epsilon^{-1} d_a^* F_a$ is coexact. Combining this observation with the embedding $W^{1,2} \hookrightarrow L^4$ for $Y_\bullet$, we can bound this by a constant times

$$\|h d_a \nabla_s b_s\|^2_{L^2(\mathbb{R}\times Y_\bullet)} \leq \|h\nabla_s d_a b_s\|^2_{L^2(\mathbb{R}\times Y_\bullet)} + \|h b_s\|^2_{L^4(\mathbb{R}\times Y_\bullet)},$$

where we used $[\nabla_s, d_a] = [b_s, \cdot]$ in the inequality. Corollary 4.14 provides a uniform bound for the L^4-norm. For the derivative term, convert to the ϵ-dependent norm to get

$$\|h\nabla_s d_a b_s\|^2_{L^2(\mathbb{R}\times Y_\bullet)} = \epsilon\|h\nabla_s d_a b_s\|^2_{L^2(\mathbb{R}\times Y_\bullet),\epsilon} \leq \epsilon\|h\nabla_s^2 F_A\|^2_{L^2,\epsilon}.$$

This can absorbed for ϵ small.

• (85) on $\mathbb{R}\times I\times\Sigma_\bullet$: In coordinates we have $\nabla_s b_s = \nabla_s\beta_s + dt\wedge\nabla_s\gamma$, so (85) becomes a linear combination of the two terms

$$\begin{aligned}&\int_{\mathbb{R}\times I\times\Sigma_\bullet} h^2 ds\wedge dt\wedge\langle[\nabla_s\beta_s\wedge\nabla_s\gamma]\,,\beta_s\rangle,\\ &\int_{\mathbb{R}\times I\times\Sigma_\bullet} h^2 ds\wedge dt\wedge\langle[\nabla_s\beta_s\wedge\nabla_s\beta_s]\,,\gamma\rangle.\end{aligned}\tag{87}$$

For the first term in (87), we use (35) on β_s, together with the embedding $W^{1,2}\hookrightarrow L^4$ for $\Sigma_\bullet$ to control this by a constant times

$$\begin{aligned}&\delta^{-1}\left(\|h\nabla_s\beta_s\|^2_{L^2(\mathbb{R}\times I\times\Sigma_\bullet)} + \|hd_\alpha\nabla_s\beta_s\|_{L^2(\mathbb{R}\times I\times\Sigma_\bullet)} + \|hd_\alpha^*\nabla_s\beta_s\|_{L^2(\mathbb{R}\times I\times\Sigma_\bullet)}\right)\\ &\quad+\delta\|hd_\alpha\nabla_s\gamma\|^2_{L^2(\mathbb{R}\times I\times\Sigma_\bullet)}\end{aligned}$$

for some $\delta>0$ that we will determine momentarily. The norm of $h\nabla_s\beta_s$ is controlled by Proposition 4.12. For the remaining terms we commute ∇_s and d_α and then convert to the ϵ-dependent norms to get that these remaining terms are bounded by

$$\begin{aligned}&\delta^{-1}\left(\epsilon\|h\nabla_s d_\alpha\beta_s\|_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon} + \|h\nabla_s d_\alpha^{*_\epsilon}\beta_s\|_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon}\right)\\ &\quad+\delta\|h\nabla_s d_\alpha\gamma\|^2_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon}\end{aligned}\tag{88}$$

plus terms of the form $\|h\,[\beta_s\wedge\beta_s]\,\|^2_{L^2}$ and $\|h\,[\beta_s,\gamma]\,\|^2_{L^2}$, which are uniformly bounded by Corollary 4.14. By Proposition 4.15, the quantity (88) is controlled by

$$C_1(\delta^{-1}\epsilon+\delta)\left(1+\|hF_A\|^2_{L^2,\epsilon}+\|h\nabla_s^2F_A\|^2_{L^2,\epsilon}+\|h\nabla_t\nabla_sF_A\|^2_{L^2,\epsilon}\right).$$

By taking δ small, and then ϵ smaller, we can absorb these second derivative terms (recall we should really be estimating $\nabla_s^2F_A$ and $\nabla_t\nabla_sF_A$ simultaneously).

To bound the second term in (87), integrate by parts in ∇_s to get that this second term equals a linear combination of

$$\int_{\mathbb{R}\times I\times\Sigma_\bullet} h^2 ds\wedge dt\wedge\langle\big[\nabla_s\beta_s\wedge\beta_s\big],\nabla_s\gamma\rangle,\qquad \int_{\mathbb{R}\times I\times\Sigma_\bullet} h^2 ds\wedge dt\wedge\langle\big[\nabla_s^2\beta_s\wedge\beta_s\big],\gamma\rangle,$$

$$\int_{\mathbb{R}\times I\times\Sigma_\bullet} h(\partial_s h) ds\wedge dt\wedge\langle\big[\nabla_s\beta_s\wedge\beta_s\big],\gamma\rangle.$$

The first of these is exactly the first term in (87) and was already bounded. The second of these is controlled by

$$\delta\|h\nabla_s^2\beta_s\|^2_{L^2(\mathbb{R}\times I\times\Sigma_\bullet)} + \delta^{-1}\|h\left[\beta_s,\gamma\right]\|^2_{L^2(\mathbb{R}\times I\times\Sigma_\bullet)}$$

which is fine for δ small by Corollary 4.14. The last of these terms is bounded by a constant times

$$\|h\nabla_s\beta_s\|^2_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon} + \|hd_\alpha\gamma\|^2_{L^2(\mathbb{R}\times I\times\Sigma_\bullet),\epsilon},$$

which is controlled by Proposition 4.12.

- (86) on $\mathbb{R}\times Y$: This is bounded by

$$\delta\|h\nabla_s^2 b_s\|^2_{L^2,\epsilon} + \delta^{-1}\|\nabla_s b_s\|^2_{L^2(\mathrm{supp(h)}\times Y),\epsilon}.$$

The first of these can be absorbed for small δ; the second is bounded by Proposition 4.12 since h has compact support.

This completes the argument for $\nabla_s^2 F_A$, so we move on to the quantity $\nabla_t\nabla_s F_A$. Integration by parts as above shows that $\|h\nabla_t\nabla_s F_A\|^2_{L^2(\mathbb{R}\times Y_r),\epsilon}$ is equal to a linear combination of terms of the form

$$\begin{gathered}
\int_{\mathbb{R}\times Y_r} h^2 ds\wedge dt\wedge\langle\left[\nabla_t\nabla_s\beta_s\wedge\beta_s\right],\gamma\rangle\\
\int_{\mathbb{R}\times Y_r} h^2 ds\wedge dt\wedge\langle\left[\nabla_s\beta_s\wedge\beta_t\right],\nabla_t\gamma\rangle\\
\int_{\mathbb{R}\times Y_r} h^2 ds\wedge dt\wedge\langle\left[\nabla_s F_\alpha,\gamma\right],\nabla_t\gamma\rangle\\
\int_{\mathbb{R}\times Y_r} h(\partial_s h) ds\wedge\langle\nabla_t b_s\wedge\nabla_t\nabla_s F_a\rangle\\
\int_{\mathbb{R}\times Y_r} h(\partial_s h) ds\wedge dt\wedge\langle\left[\beta_s\wedge\beta_t\right],\nabla_t\gamma\rangle
\end{gathered}$$

together with the terms

$$\begin{gathered}
\int_{\mathbb{R}\times\partial Y_r} h^2 ds\wedge\langle\nabla_t\nabla_s\beta_s\wedge\nabla_t\beta_s\rangle,\\
\int_{\mathbb{R}\times\partial Y_r} h^2 ds\wedge\langle\left[\nabla_s\beta_t\wedge\beta_s\right],\gamma\rangle,\quad \int_{\mathbb{R}\times\partial Y_r} h^2 ds\wedge\langle\left[\nabla_t\beta_s\wedge\beta_s\right],\gamma\rangle,\\
\int_{\mathbb{R}\times\partial Y_r} h^2 ds\wedge\langle\left[\beta_s\wedge\beta_s\right],\nabla_t\gamma\rangle,\quad \int_{\mathbb{R}\times\partial Y_r} h^2 ds\wedge\langle\left[\beta_s\wedge\beta_t\right],\nabla_s\gamma\rangle
\end{gathered}$$

coming from integration in the t-direction. The first group can be bounded as we did in the case of $\nabla_s^2 F_A$ (recall $|dt| = \epsilon$ on $Y_r\cap Y_\bullet$), and the second group goes to zero as r decreases to 0, just as with Proposition 4.11.

□

References

[1] M. Atiyah. New invariants of three and four dimensional manifolds. *The Math. Heritage of Hermann Weyl, Proc. Sympos. Pure Math.* 48, 285-299, 1988.

[2] M. Atiyah, R. Bott. The Yang-Mills equations over Riemann surfaces. *Phil. Trans. Roy. Soc. London Ser. A,* 308:523-615, 1982.

[3] S. Akbulut, J. McCarthy. *Casson's invariant for oriented homology 3-spheres, an exposition.* Math. Notes, No. 36, Princeton Univ. Press, Princeton, 1990.

[4] P.J Braam, S.K. Donaldson. Floer's work on instanton homology, knots and surgery. *The Floer Memorial Volume*, Ed. Hofer et al. Birkhäuser 195-256, 1995.

[5] N. Charalambous, L. Gross. The Yang-Mills heat semigroup on three-manifolds with boundary. 2010. arXiv:1004.1639.

[6] S. Donaldson. Anti-self dual Yang-Mills connections over complex algebraic surfaces and stable vector bundles. *Proc. London Math. Soc.* (3) 50, no. 1, 1-26, 1985.

[7] S. Donaldson, P. Kronheimer. *The Geometry of Four-Manifolds.* Oxford Mathematical Monographs, Oxford: Clarendon Press, 1997.

[8] S. Dostoglou, D. Salamon. Cauchy-Riemann operators, self-duality and the spectral flow. *First European Congress of Mathematics, Vol.* I (Paris, 1992), 511-545, Progr. Math., 119, Birkhäuser, Basel, 1994.

[9] S. Dostoglou, D. Salamon. Instanton homology and symplectic fixed points. *Symplectic geometry*, 57-93, *London Math. Soc. Lecture Note Ser.*, 192, Cambridge Univ. Press, Cambridge, 1993.

[10] S. Dostoglou, D. Salamon. Self-dual instantons and holomorphic curves. *Ann. of Math.* (2) 139 (1994), no. 3, 581-640.

[11] S. Dostoglou, D. Salamon. Corrigendum: Self-dual instantons and holomorphic curves. *Ann. of Math.* 165 (2007), 665-673.

[12] D. Duncan. Higher-rank instanton cohomology and the quilted Atiyah-Floer conjecture. 2015. arXiv:1311.5609.

[13] D. Duncan. On the components of the gauge group for $\mathrm{PU}(r)$-bundles. 2013. arXiv:1311.5611.

[14] D. Duncan. The Chern-Simons invariants for doubles of compression bodies. In preparation.

[15] A. Floer. An instanton-invariant for 3-manifolds. *Comm. Math. Phys.*, 118(2):215-240, 1988.

[16] A. Floer. Instanton homology and Dehn surgery. *Floer Memorial Volume, Progress in Mathematics, Vol.* 133, Birkhäuser Verlag, 1995. ISBN 3-7643-5044-X, Basel; ISBN 0-8176-5044-8, Boston.

[17] A. Floer. Morse theory for Lagrangian intersections. *J. Differential Geom.* 28 (1988), no. 3, 513-547.

[18] K. Fukaya. Anti-self-dual equation on 4-manifolds with degenerate metric. *Geom. Funct. Anal.* 8, 466-528, 1998.

[19] D. Gay, R. Kirby. Indefinite Morse 2-functions: broken fibrations and generalizations. 2011. arXiv:1102.0750.

[20] S. Grundel. Moment maps and diffeomorphisms. *Diploma Thesis*, ETH Zürich, 2005.

[21] V. Guillemin S. Sternberg. Geometric quantization and multiplicities of group representations. *Invent. Math.*, 67:515-538, 1982.

[22] F. Kirwan. *Cohomology of quotients in symplectic and algebraic geometry*. Mathematical Notes Series, Vol. 31, Princeton University Press, 1984. ISBN 0691083703, 9780691083704.
[23] R. Lee, W. Li. Floer homology for Lagrangian intersections and instantons. 1995. arXiv:math/9506221v1.
[24] M. Lipyanskiy. Gromov-Uhlenbeck compactness. 2014. arXiv:1409.1129.
[25] C. Manolescu, C. Woodward. Floer homology on the extended moduli space. 2010. arXiv:0811.0805v3.
[26] D. McDuff, D. Salamon. *J-holomorphic curves and symplectic topology*. American Mathematical Society Colloquium Publications, Vol. 52. American Mathematical Society, Providence, RI, 2004.
[27] J. Milnor. *Lectures on the h-Cobordism Theorem. Notes by L. Siebenmann and J. Sondow*. Princeton University Press, Princeton, N.J., 1965.
[28] J. Morgan, T. Mrowka, D. Ruberman. *The L^2-moduli space and a vanishing theorem for Donaldson polynomial invariants*, Monographs in Geometry and Topology, II. International Press, Cambridge, MA, 1994.
[29] M. Narasimhan, C. Seshadri. Stable and unitary vector bundles on a compact Riemann surface. *Annals of Mathematics. Second Series* 82 (3): 540-567, 1965.
[30] T. Nishinou, Convergence of adiabatic family of anti-self-dual connections on products of Riemann surfaces. *Journal of Mathematical Physics*, 51, 022306 (2010); doi: 10.1063/1.3318164
[31] J. Råde. On the Yang-Mills heat equation in two and three dimensions. *J. Reine Angew.* 431, 123-163 (1992).
[32] A. Ramanathan. Stable Principal bundles on a compact Riemann surface. *Math. Ann.* 213, 29-152 (1975).
[33] D. Salamon, K. Wehrheim. Instanton Floer homology with Lagrangian boundary conditions. 2007. arXiv:math/0607318.
[34] D. Salamon. Morse theory, the Conley index and Floer homology. *Bull. London Math. Soc.* (1990) 22 (2): 113-140. doi: 10.1112/blms/22.2.113
[35] C. Taubes. Self-dual Yang-Mills connections on non-self-dual 4-manifolds. *J. Di. Geom.* 17 (1982), 139-170.
[36] F. Warner. *Foundations of differentiable manifolds and Lie groups*. Scott-Foreman, Glen view, Illinois, 1971.
[37] K. Wehrheim. Outline of compactness proofs for Atiyah-Floer degenerations in the absence of reducibles, private communication, and talks at Columbia, UW Madison, ETH Zurich. 2004-6.
[38] K. Wehrheim. *Uhlenbeck compactness*. EMS Series of Lectures in Mathematics, 2004.
[39] K. Wehrheim, C. Woodward. Floer field theory for coprime rank and degree. Preprint.
[40] L. M. Woodward. The classification of principal PU_n-bundles over a 4-complex. *J. London Math. Soc.* (2) 25 (1982), no. 3, 513-524.