ELSEVIER

Journal of Geometry and Physics 36 (2000) 324–384

JOURNAL OF GEOMETRY AND PHYSICS

Simple type and the boundary of moduli space

David Groisser [a,*,1], Lorenzo Sadun [b,2]

[a] *Department of Mathematics, University of Florida, Gainesville, FL 32611-8105, USA*
[b] *Department of Mathematics, University of Texas, Austin, TX 78712, USA*

Received 13 April 2000

Abstract

We measure, in two distinct ways, the extent to which the boundary region of moduli space contributes to the "simple type" condition of Donaldson theory. Using the natural geometric representative of $\mu(\mathrm{pt})$ defined in [L. Sadun, Commun. Math. Phys. 178 (1996) 107–113], the boundary region of moduli space contributes $\frac{6}{64}$ of the homology required for simple type, regardless of the topology or geometry of the underlying 4-manifold. The simple type condition thus reduces to the interior of the $(k+1)$th ASD moduli space, intersected with two representatives of (4 times) the point class, being homologous to 58 copies of the kth moduli space. This is peculiar, since the only known embeddings of the kth moduli space into the $(k+1)$th involve Taubes gluing, and the images of such embeddings lie entirely in the boundary region.

When using the natural de Rham representatives of $\mu(\mathrm{pt})$ considered by Witten [Commun. Math. Phys. 117 (1988) 353], the boundary region contributes $\frac{1}{8}$ of what is needed for simple type, again regardless of the topology or geometry of the underlying 4-manifold. The difference between this and the geometric representative answer is not contradictory, as the contribution of a fixed region to the Donaldson invariants is geometric, not topological.

MSC: 57R57; 58D27; 53C07; 58G99
Subj. Class.: Quantum field theory, Differential geometry
Keywords: Simple type; Donaldson theory; μ-Map; Yang–Mills

1. Introduction

This paper is a study in the geometry and topology of anti-self-dual Yang–Mills moduli spaces. Although moduli spaces were studied extensively for their own sake in the 1970s and early 1980s, in the late 1980s and early 1990s such studies were primarily a means to an end.

* Corresponding author.
E-mail addresses: groisser@math.ufl.edu (D. Groisser), sadun@math.utexas.edu (L. Sadun).

[1] Research supported in part by National Science Foundation Grant DMS-9307648.
[2] Research supported in part by National Science Foundation Grant DMS-9626698, an NSF Mathematical Sciences Postdoctoral Fellowship and Texas ARP Grant 003658-037.

0393-0440/00/$ – see front matter
PII: S0393-0440(00)00030-9

Moduli spaces were studied to compute Donaldson invariants, and Donaldson invariants were computed for their applications in classifying smooth 4-manifolds. Seiberg–Witten theory has, of course, made that last road obsolete. It is believed that the Seiberg–Witten invariants determine the Donaldson invariants, and the former are far easier to handle.

However, Seiberg–Witten theory has opened up new uses for Donaldson theory. From Seiberg–Witten theory, we now have a much better understanding of Donaldson invariants. Instead of using moduli spaces as a tool for computing Donaldson invariants, we can now use Donaldson invariants as a tool for understanding moduli spaces. This paper is an exercise along those lines.

A basic problem in four-dimensional gauge theory is to understand the "simple type" condition. In Donaldson theory, a manifold is said to have simple type if its Donaldson invariants satisfy a certain recursion relation ([12]; see (1.2) below). In Seiberg–Witten theory, a manifold has simple type if it has no Seiberg–Witten classes of nonzero index. The two notions of simple type are believed to be equivalent so that theorems proved about one form of simple type should yield information about the other.

In this paper we work with the Donaldson theory sense of simple type, examining what simple type implies about the geometry of anti-self-dual moduli spaces. In two ways — with intersection theory and with de Rham theory using natural (and therefore *nongeneric*) geometric representatives in both cases — we measure the extent to which the boundary region of moduli space contributes to the simple type recursion relation. Our results imply that the anti-self-dual moduli spaces associated to any manifold of simple type have a very surprising interior geometric structure. Widely satisfied sufficient conditions are known for a manifold to be of simple type [12], and it is conjectured that indeed *all* 4-manifolds with $b_+ > 1$ are of simple type. (This conjecture is known to be false for $b_+ = 1$, $\mathbf{C}P^2$ is a counterexample; see [7,11].) Hence our results apply to a great many manifolds.

Simple type says that the $(k+1)$st moduli space $\mathcal{M}_{k+1}$, intersected with certain varieties, has the homology of a certain multiple of the kth moduli space $\mathcal{M}_k$. Our intersection theory approach is based on the construction in [13] of a geometric representative of μ of a point (see below). Using this representative we show that the portion of (a small perturbation of) $\mathcal{M}_{k+1}$ near the boundary contributes $\frac{6}{64}$ of the homology required for simple type, regardless of the topology or geometry of the underlying 4-manifold. (For a quick, heuristic derivation of this $\frac{6}{64}$, see [14].) Simple type thus reduces to a statement relating $\mathcal{M}_k$ to nontrivial structure in the *interior* of $\mathcal{M}_{k+1}$ (unless our small perturbation of $\mathcal{M}_{k+1}$ is drastically unfaithful topologically, which seems highly unlikely). This is surprising, since the only known relations between $\mathcal{M}_k$ and $\mathcal{M}_{k+1}$ involve Taubes patching, and relate $\mathcal{M}_k$ to the boundary of $\mathcal{M}_{k+1}$.

Our second approach is to use differential form representatives of the images of the μ-map. One then takes the wedge product of these forms and integrates over $\mathcal{M}_{k+1}$. If we restrict the domain of integration to a neighborhood of the boundary of $\mathcal{M}_{k+1}$, we can reinterpret the simple type condition in terms of the integral of a certain 8-form over a submanifold that represents the space of "bubble parameters" in the neighborhood of a background connection in $\mathcal{M}_k$. We show that, again independent of the topology and geometry of the base manifold, this integral has precisely $\frac{1}{8}$ the value of what one would

naively expect if the relation between our representatives and simple type were captured purely by a neighborhood of the boundary. Thus again simple type becomes a statement about the nontrivial structure of the interior of moduli space.

It is curious but no contradiction that the two approaches yield the different numerical answers $\frac{6}{64}$ and $\frac{1}{8}$. While the Donaldson polynomial is topological, hence independent of the choice of geometric or de Rham representatives, the contribution of each region of moduli space is geometric, and need not be the same for two different representatives. Indeed, the de Rham and geometric representative calculations not only disagree on the contribution of the boundary region, but also disagree on how close to the boundary the essential contributions are. In terms of the small parameter L described below, the geometric representative picks up contributions from bubbles of size $\mathrm{O}(L^2)$, while the bulk of the support of the de Rham representative is on bubbles of size $\mathrm{O}(L)$.

Since homological statements are by their nature nonlocal, one might arrange for the boundary-neighborhood contribution to intersection numbers to be anything one likes by choosing appropriate representatives of μ of a point. Indeed, Donaldson invariants are usually defined using generic representatives of the μ classes (cf. [4, Section 9.2]), which force the intersections to stay away from the boundary of moduli space. By contrast, our representatives are nongeneric but geometrically natural, depending only on the choice of a point in the base manifold N — not on any other details of N, choice of representatives of other classes in $H_*(N)$, or other data. The intersections are all compact, so the total intersection number is the same in both approaches, but in our approach the locations of the intersections as well as their number gives geometric information about the structure of moduli spaces. Similar considerations apply to the de Rham theory calculations; we will comment on these more specifically below.

To state our results more precisely, we must review the definition of the Donaldson invariants, and of simple type. Let N be an oriented 4-manifold, let $G = SU(2)$ or $SO(3)$, and let $\mathcal{B}_k^*$ be the space of irreducible connections (up to gauge equivalence) on P_k, the principal G-bundle of instanton number k over N. Let $\mathcal{M}_k \subset \mathcal{B}_k$ be the space of irreducible connections on P_k with anti-self-dual curvature, modulo gauge transformations. We will frequently omit the index k.

Donaldson [1,2] defined a map $\mu : H_i(N, \mathbf{Q}) \to H^{4-i}(\mathcal{B}_k^*, \mathbf{Q})$, $i = 0, 1, 2, 3$, whose image freely generates the rational cohomology of $\mathcal{B}_k^*$. Donaldson invariants are then defined by pairing the fundamental class of $\mathcal{M}_k$ with products of μ of the homology classes of N, where k is chosen so that the dimensions match. Formally, for elements $[\Sigma_1], \ldots, [\Sigma_n] \in H_*(N)$, we write

$$D([\Sigma_1] \cdots [\Sigma_n]) = \mu([\Sigma_1]) \smile \cdots \smile \mu([\Sigma_n])[\mathcal{M}_k]. \tag{1.1}$$

Now let x be the point class in $H_0(N)$, and let ω be any formal product of classes in $H_*(N)$. The simple type condition is that, for all ω,

$$D(x^2\omega) = 4D(\omega). \tag{1.2}$$

Of course, the "fundamental class of $\mathcal{M}_k$" is usually not well defined, as $\mathcal{M}_k$ is typically not compact. The usual way to make sense of (1.1) and (1.2) is with geometric rep-

resentatives. One finds finite-codimension varieties V_Σ in $\mathcal{B}_k^*$ that are Poincaré dual to $\mu([\Sigma])$; we say simply that V_Σ *represents* $\mu([\Sigma])$. One then counts points, with sign, in $V_{\Sigma_1} \cap \cdots \cap V_{\Sigma_n} \cap \mathcal{M}_k$. To make a topological invariant one must show that the number of intersection points is independent of auxiliary data, such as the metric and the choice of representatives. This requires careful analysis of the bubbling phenomena that make $\mathcal{M}_k$ noncompact.

To compute the left-hand side of (1.2) we need a variety that represents μ of the point class x. In general $\mu(x)$ is not an integral class in $H^4(\mathcal{B}^*)$, so strictly speaking it has no geometric representative. However, $-4\mu(x)$ is an integral class, hence is Poincaré dual to a cycle V_x in $\mathcal{B}_k^*$. Let us suppose that to each $p \in N$ we can, by some natural procedure, associate a cycle V_p' homologous to V_x. Then the simple type condition can be rewritten as

$$\#(\mathcal{M}_{k+1} \cap V_p' \cap V_q' \cap V_\omega) = 64\#(\mathcal{M}_k \cap V_\omega), \tag{1.3}$$

where p and q are any two points in N, ω is an arbitrary formal product of homology cycles of N, and V_ω is a geometric representative of $\mu(\omega)$. Of course, to *compute* the intersection number by point-counting one may have to perturb V_p', V_q', and V_ω to achieve transversality, but the intersection number is well defined as long as the intersection is compact.

More formally, one can write (1.3) as

$$[\mathcal{M}_{k+1} \cap V_p' \cap V_q'] = 64[\mathcal{M}_k]. \tag{1.4}$$

Strictly speaking, the left-hand side is an element of $H_*(\mathcal{B}_{k+1})$, while the right-hand side is in $H_*(\mathcal{B}_k)$. However, $\mathcal{B}_k$ and $\mathcal{B}_{k+1}$ are homotopy equivalent spaces, and their homology classes can be identified. Essentially, then, (1.4) says that $\mathcal{M}_{k+1} \cap V_p' \cap V_q'$ consists homologically of 64 copies of $\mathcal{M}_k$.

In view of (1.3), a natural geometric question is whether (and if so, how) one can associate to each point in $\mathcal{M}_k \cap V_\omega$ a *natural and specific* set of 64 points in $\mathcal{M}_{k+1} \cap V_p' \cap V_q' \cap V_\omega$. Using *generic* representatives of the point class (as is usually done when the goal is to define and prove relations among Donaldson invariants) there is no hope, as there is no known way to associate general points of $\mathcal{M}_{k+1}$ with points of $\mathcal{M}_k$. One can, however, hope that a geometrically natural choice of representatives pushes the points of $\mathcal{M}_{k+1} \cap V_p' \cap V_q' \cap V_\omega$ towards the $\mathcal{M}_k \times N$ stratum of the Uhlenbeck boundary of $\mathcal{M}_{k+1}$, where we can simply project onto the $\mathcal{M}_k$ factor.

Towards this end, for $p \in N$ let

$$\nu_p = \{[A] \in \mathcal{B}_{k+1}^* \mid F_A^- \text{ is reducible at } p\}. \tag{1.5}$$

Here $F_A^- = \frac{1}{2}(F_A - *F_A)$ is the anti-self-dual part of the curvature F_A, and by "reducible at p" we mean that the components $F_{ij}^-(p)$ are all collinear as elements of the Lie algebra of G. In [13] it was shown that ν_p, despite being noncompact in general, is a geometric representative of $-4\mu([p])$ — to our knowledge, the only such representative that has been canonically defined. Given ω, the intersection of ν_p with a generic representive V_ω of $\mu(\omega)$ is compact, so that (perturbing if necessary) the intersection numbers in (1.3), with V_p', V_q' replaced by ν_p, ν_q, are well defined.

The first question studied in this paper, then, is this: *Suppose p and q are extremely close points in N, separated by a distance 2L. How many of the points on the left-hand side of* (1.3), *with* V_p', V_q' *replaced by* ν_p, ν_q, *lie near the boundary of* $\mathcal{M}_{k+1}$? The answer is quite simple, but surprising.

Theorem 1.1. *Let* (N, g) *be a compact oriented Riemannian* 4*-manifold of arbitrary topology and geometry and let* $4k \geq 3b_+ + 5$. *Fix a coordinate patch on N, and let p and q be the points with coordinates* $(\pm L, 0, 0, 0)$. *Fix* $\omega \in \mathrm{Sym}^*(H_*N)$, $K > 0$, *and* $\alpha \in (0, 2)$. *Let* $\tilde{\mathcal{M}}^0_{k+1}$ *be the portion of the* (*perturbed*) *moduli space* $\tilde{\mathcal{M}}_{k+1}$ *consisting of a background of charge k and a charge-one bubble of size* $\lambda < KL^\alpha$. *For generic choices of geometric representatives* V_ω *of* $\mu(\omega)$, *and for all sufficiently small L, the intersection number of* $\tilde{\mathcal{M}}^0_{k+1}$ *with* $V_\omega \cap \nu_p \cap \nu_q$ *is* $6D(\omega)$.

The perturbed moduli space $\tilde{\mathcal{M}}_{k+1}$ is constructed, and the genericity conditions specified in Sections 3 and 4. This theorem is restated, more precisely, as Theorem 4.1. In this theorem, and throughout Sections 2–4, we assume that k is in the indicated "stable range" to avoid contributions from lower strata of the compactified moduli space.

To understand why Theorem 1.1 is surprising, observe that it is essentially a statement about the intersection $\tilde{\mathcal{M}}^0_{k+1} \cap V_\omega \cap \nu_p \cap \nu_q$ in the limit as $p \to q$. (This is more or less equivalent to a "neck stretch", separating the points p and q from the rest of the manifold.) There are three obvious guesses for what might happen in this limit. The first guess is that, for each point in $V(\omega) \cap \mathcal{M}_k$, 64 points of $V(\omega) \cap V(x_1) \cap V(x_2) \cap \mathcal{M}_{k+1}$ get pushed to the boundary, converging to points in $\mathcal{M}_k \times \{q\} \subset \mathcal{M}_k \times N$ as $p \to q$. (This was our initial hope.) A second guess, voiced by most of the experts with whom we discussed the project, is that *none* of the points get pushed to the boundary. A third possibility is that the behavior depends on the details of the manifold N, the metric g, and the homology polynomial ω.

Theorem 1.1 shows that all three guesses are wrong: for the representatives ν_p, the number of points pushed to the boundary is independent of the manifold, the metric, and ω, but the number is always 6, not 0 or 64. Moreover, the proof of Theorem 1.1 shows that the intersection points are pushed to the boundary in a highly regular way. For each point $A_0 \in V(\omega) \cap \mathcal{M}_k$, all the six points converge to $(A_0, q) \in \mathcal{M}_k \times N$.

(While there are other potential ways to push points of $V(\omega) \cap V(p) \cap V(q) \cap \mathcal{M}_{k+1}$ toward the boundary, such as artificially placing cycles defining ω in other than general position, or by letting p and/or q approach these cycles, these methods differ from ours in that they involve choices that are specific to the manifold (N, g) and the cycles defining ω.)

It should be noted that our intersection-theoretic results do not distinguish between $b_+ = 1$ and $b_+ > 1$. Our calculation is essentially local, involving only the curvature of the background connection A_0 at q, so it is not surprising that the "6" in our formula for the boundary-neighborhood intersection number is independent of b_+. What is surprising, at least to the authors, is that this number is not 64. As manifolds with $b_+ > 1$ appear to have simple type, the authors had expected manifolds with $b_+ > 1$ to have a boundary contribution of $64D(\omega)$ and zero interior contribution, while manifolds with $b_+ = 1$ would have a boundary contribution of $64D(\omega)$ plus a mysterious interior contribution that our methods

could not probe. But Theorem 1.1 yields a boundary contribution of $6D(\omega)$ *regardless of simple type*. Simple type thus reduces to a statement that, for p and q sufficiently close, the *interior* of $\mathcal{M}_{k+1} \cap \nu_p \cap \nu_q$ is homologous to 58 copies of $\mathcal{M}_k$. This is striking, since in general very little is known about the interior of $\mathcal{M}_{k+1}$. As noted earlier, the only known embeddings of $\mathcal{M}_k$ into $\mathcal{M}_{k+1}$ involve Taubes patching, and have an image near the boundary of $\mathcal{M}_{k+1}$. Theorem 1.1 implies that for any manifold of simple type, the intersection number has a peculiar interior contribution of $58D(\omega)$, while $\mathbf{C}P^2$, which does not have simple type, has an interior contribution of something other than $58D(\omega)$.

On the level of differential forms, the de Rham-theoretic version of the μ-map is represented by a map

$$\mu_{\mathrm{d}} : \Omega^i(N) \to \Omega^i(\mathcal{B}^*_{k+1}), \quad i = 0, \ldots, 4; \tag{1.6}$$

the argument of μ_{d} is a form representing the Poincaré dual of the argument of μ. In particular, $\mu(x)$ is represented by a 4-form $\mu_{\mathrm{d}}(\omega) \in H^4(\mathcal{B}^*_{k+1})$ for any $\omega \in \Omega^4(N)$ with $\int_M \omega = 1$.

One can write down an explicit formula for such a representative $\mu_{\mathrm{d}}(\omega)$ by appealing to Chern–Weil theory on the canonical $SO(3)$-bundle $\mathcal{P} \to \mathcal{B}^*_{k+1} \times M$ (see Section 5). Furthermore given $p \in M$, if we replace ω by δ_p, a delta-form supported at a point p, then the resulting form on $\mathcal{B}^*_{k+1}$ is still de Rham cohomologous to a form obtained using smooth ω (although there is an important difference that we will discuss later). Let us write $\mu_{\mathrm{d}}(p) := \mu_{\mathrm{d}}(\delta_p)$. Note that for smooth $\omega \in \Omega^4(N)$ we have

$$\mu_{\mathrm{d}}(\omega)|_A = \int_N \mu_{\mathrm{d}}(p)|_A \omega(p). \tag{1.7}$$

It is generally believed that the integrals of wedge products of the forms $\mu_{\mathrm{d}}(\cdot)$ over moduli space compute the Donaldson invariants (the "de Rham-theoretic conjecture"). This conjecture has been largely unapproachable because of formidable analytic problems posed by these differential forms: they are nonlocal in p, involving covariant Green operators, and have noncompact support in A.

However, even absent a proof of the de Rham-theoretic conjecture, these differential forms have been of interest to physicists, appearing, for example, as correlation functions of massless Fermion fields in Witten's $N = 2$ supersymmetric topologic quantum field theory (TQFT) approach to Donaldson theory [17]. The poor localization of $\mu_{\mathrm{d}}(\cdot)$ is reflective of the supersymmetry that Seiberg and Witten [15,18] used to relate Donaldson invariants to solutions to the Seiberg–Witten equations. Short-distance properties of the Seiberg–Witten TQFT are said to be related to long-distance properties of the Donaldson TQFT, e.g., the nonlocality of differential forms. The extreme fruitfulness of this approach argues for more rigorous analysis of these forms. While our original motivation for considering these differential forms was primarily to study their relation to simple type, we hope that the material in Sections 5–10 will provide the foundation for a rigorous understanding of these forms. A proof of the de Rham-theoretic conjecture, for example, would necessarily entail showing that the relevant differential forms are integrable over the whole moduli space. This boils down to integrability over the ends. Our analysis in Sections 5–10 is a step in

this direction: essentially we show integrability near one stratum of the boundary (modulo certain technical assumptions). With more work along the same lines, we suspect that one could both eliminate the technical assumptions and show integrability over neighborhoods of all the boundary strata. As is already evident from the work in our paper, such a proof would require an enormous amount of additional technical work tangential to our primary purpose. However, someone wishing to prove the de Rham-theoretic conjecture could take this paper as a starting point.

A necessary condition for de Rham-theoretic conjecture to be true is that for manifolds of simple type, the integrals of products of μ_{d}-images obey exactly the same calculus that one would expect from simple type. To examine this expected calculus in greater detail, let Z be an eight-dimensional cycle in $\mathcal{B}^*_{k+1}$. Since the cohomology class of $\mu_{\mathrm{d}}(p)$ is independent of p, for any points $p, q \in M$ we have $\int_Z \mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(p) = \int_Z \mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q)$, and moreover this integral depends only on the homology class of Z.

Pretend for a moment that the moduli spaces $\mathcal{M}_{k+1}$ and $\mathcal{M}_k$ are, respectively, the total space and base space of a compact, connected, oriented fiber bundle $\pi\colon \mathcal{M}_{k+1} \to \mathcal{M}_k$; the fibers would then be mutually homologous compact submanifolds $Z \subset \mathcal{M}_{k+1}$. For any form $\phi \in \Omega^{\mathrm{top}}(\mathcal{M}_k)$, we would have a product formula

$$\int_{\mathcal{M}_{k+1}} \mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q) \wedge \pi^*\phi = \left(\int_Z \mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q)\right)\left(\int_{\mathcal{M}_k} \phi\right) \tag{1.8}$$

(assuming compatible orientations), so the simple-type condition (1.2) would be equivalent to

$$\int_Z \mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q) = 4. \tag{1.9}$$

In reality the moduli spaces are not compact and there is no such global fibration. However, from the current understanding of the forms $\mu_{\mathrm{d}}(\cdot)$ one might speculate that the relevant integrals are supported in a region near the ideal boundary of $\mathcal{M}_{k+1}$, in some sense of "near" to be determined later — i.e., that this choice of de Rham representative pushes cohomological information out towards the boundary. Of course a random de Rham representative of a cohomology class can be supported wherever it likes, but $\mu_{\mathrm{d}}(\cdot)$ is not random, and there is evidence that its properties near the boundary of moduli space do indeed capture a lot of cohomological information. For example, consider the five-dimensional moduli spaces of 1-instantons over simply connected manifolds with $b_+ = 0$. In such cases the inverse of a collar map gives embeddings $\tau_\lambda : N \to \mathcal{M}_1 \subset \mathcal{B}^*$ for λ sufficiently small (the image of τ_λ consisting of instantons of scale λ), and one has Donaldson's theorem that the composition $\tau_\lambda^* \circ \mu : H_2(N, \mathbf{Z}) \to H^2(N, \mathbf{Z})$ is precisely Poincaré duality [4, Corollary 5.3.3]. The corresponding assertion in de Rham cohomology would be that $\tau_\lambda^* \circ \mu_{\mathrm{d}} : \Omega^2(N) \to \Omega^2(N)$ induces the identity on cohomology. But in fact in this context one can show that $\lim_{\lambda\to 0}\tau_\lambda^* \circ \mu_{\mathrm{d}}$ is already the identity map *on the level of forms* in all degrees [10].

There is no reason a priori to expect all such information to be lost when one moves from $b_+ = 0$ to $b_+ > 1$, the realm of Donaldson invariants. Therefore consider that portion $\mathcal{M}'_{k+1,\lambda_0}$ of $\mathcal{M}_{k+1,\lambda_0}$ near the highest-dimensional boundary stratum $\mathcal{M}_k \times N$. There is

indeed a fibration $\mathcal{M}'_{k+1,\lambda_0} \to \mathcal{M}'_k$ whose fibers can be identified with subsets of an eight-dimensional space of framed ASD connections on $\mathbf{R}^4$. (Here $\mathcal{M}'_k$ denotes the space of nonconcentrated irreducible k-instantons, and $\mathcal{M}'_{k+1,\lambda_0}$ the space of $(k+1)$-instantons with only a single "bubble", of charge 1, and scale less than some small number λ_0.) The typical fiber $Z = Z_{\lambda_0}$ is itself a bundle over $(0, \lambda_0) \times N$ for some small λ_0, whose fiber over $(\lambda, p) \in (0, \lambda_0) \times N$ is the space of "gluing parameters" $\mathrm{Hom}_{SO(3)}(\Lambda^2_+ T^*N, Ad\, P_k) \cong SO(3)$ (see [4, p. 324]). Since $\mathcal{M}'_{k+1,\lambda_0}$ is such a large portion of the end of $\mathcal{M}_{k+1}$, one might then expect that an approximate version of (1.9) holds under the assumption of simple type.

What we show below (Theorem 1.2) is that (1.9) fails in a very precise way: independent of the topology and geometry of N, if λ_0 is small enough and $\mathrm{dist}(p, q)$ is small compared to λ_0 (but nonzero), then

$$\int_{Z_{\lambda_0}} \mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q) \approx \frac{1}{2} \tag{1.10}$$

under certain technical but intuitively reasonable assumptions about the fiber Z_{λ_0}. Taking a limit as $q \to p$ and then as $\lambda_0 \to 0$, the integral above approaches an integral over the space of framed instantons on $\mathbf{R}^4$, and this latter integral has the precise value $\frac{1}{2}$. Despite the technical assumptions, Theorem 1.2 gives a picture of the *best* one can hope for by integrating the μ_{d} products over the top stratum of the ends of moduli space.

At this stage the reader may wonder why we do not simply take $p = q$ in (1.10). The reason is that for purposes of integration, the $\mu_{\mathrm{d}}(p)$ turn out to be more singular than the representatives $\mu_{\mathrm{d}}(\omega)$ for smooth ω. Were we to set $p = q$ in (1.10), $\frac{1}{2}$ would be replaced by 0. This discontinuity can be modeled by the following two-dimensional example. Let H be the upper half-plane $\{(x, \lambda) \in \mathbf{R}^2 | \lambda > 0\}$ and for each $L \in \mathbf{R}$ let $\theta_L : H \to (0, \pi)$ be the usual polar-coordinate angle as measured from $(L, 0)$ (so $\mathrm{d}\theta_L = ((x - L)\,\mathrm{d}\lambda - \lambda\,\mathrm{d}x)/((x - L)^2 + \lambda^2)$). As forms on H, the $\mathrm{d}\theta_L$ are all cohomologous (in fact cohomologous to zero). However, $\int_H \mathrm{d}\theta_0 \wedge \mathrm{d}\theta_0 = 0$, while for $L > 0$ we have $\int_H \mathrm{d}\theta_0 \wedge \mathrm{d}\theta_L = \frac{1}{2}\pi^2$. Essentially, $\mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q)$ behaves like a quaternionic version of this example.

The technical assumptions on the fiber Z are enumerated as (Z1)–(Z5) in Section 7. The first three of these assumptions are known to be satisfied by the fiber constructed in [4], but we have not determined whether the latter two are satisfied. These two are assumptions on the tangent space to $T_{[A]}Z$, where $[A] \in Z$, and we prove that they are satisfied by a subspace of $T_{[A]}\mathcal{M}$ (the "approximate tangent space") that we argue is close to $T_{[A]}Z$. Because this step is only a plausibility argument, (1.10) implies one of two things: either $\mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q)$ has most of its support in the interior of moduli space (or near higher codimension boundary strata), or the intuitive picture of the fiber Z is significantly wrong. Either way, the conclusion is surprising.

Our second main theorem is then the following:

Theorem 1.2. *Let N be a compact oriented Riemannian 4-manifold of arbitrary topology and geometry and let $k \geq 1$. Assume that a typical fiber Z_{λ_0} of the fibration $\mathcal{M}'_{k+1,\lambda_0} \to \mathcal{M}'_k$*

satisfies (Z1)–(Z5) *of Section* 7. *Then for any* $p, q \in N$, *the form* $\mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q)$ *is integrable over* Z_{λ_0} *for* λ_0 *sufficiently small, and*

$$\lim_{\lambda_0 \to 0} \left(\lim_{q \to p} \int_{Z_{\lambda_0}} \mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q) \right) = \frac{1}{2}, \tag{1.11}$$

while

$$\lim_{\lambda_0 \to 0} \int_{Z_{\lambda_0}} \mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(p) = 0. \tag{1.12}$$

The convergence in (1.11) *and* (1.12) *is uniform in p, q. Hence if* $\tilde{\delta}_{p,L}$ *denotes a smooth* 4*-form on N of total integral* 1 *supported in a ball of radius L about p, then* (*using* (1.7))

$$\lim_{\lambda_0 \to 0} \left(\lim_{L \to 0} \int_{Z_{\lambda_0}} \mu_{\mathrm{d}}(\tilde{\delta}_{p,L}) \wedge \mu_{\mathrm{d}}(\tilde{\delta}_{p,L}) \right) = \frac{1}{2}. \tag{1.13}$$

By uniform convergence in (1.11) we mean that for all $\epsilon > 0$ there exist $\lambda_1, \delta(\cdot) > 0$ such that if $0 < \lambda_0 < \lambda_1$ and $0 < \mathrm{dist}(p, q) < \delta(\lambda_0)$ then the integral in (1.11) differs from $\frac{1}{2}$ by less than ϵ.

It is not necessary to take the limits in (1.11) completely independently as long as $q \to p$ much faster than $\lambda_0 \to 0$. If, for example, we require that $\mathrm{dist}(p, q) = \mathrm{const.}\, \lambda_0^{1+\alpha}$ for some $\alpha > 0$, and then take a limit as $\lambda_0 \to 0$, we again get $\frac{1}{2}$.

Note that if we held p and q fixed rather than taking $\lim_{q \to p}$ in (1.11), the limit as $\lambda_0 \to 0$ would necessarily be zero (since $\mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q)$ is integrable). It turns out that for $q \neq p$ the integrand in (1.11) is supported in a region in which λ is of the order $\mathrm{dist}(p, q)$. Thus if we wish to extend $\mu_{\mathrm{d}}(p)$ and $\mu_{\mathrm{d}}(\tilde{\delta}_{p,L})$ to forms on the Uhlenbeck compactification of $\mathcal{M}$, with $\lim_{L \to 0} \mu_{\mathrm{d}}(\tilde{\delta}_{p,L}) = \mu_{\mathrm{d}}(p)$ in a distributional sense, then $\mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(p)$ should be viewed as the sum of a delta-form supported on the boundary of moduli space and a smooth form supported away from the boundary.

Theorem 1.2 does not require k to be in the "stable range" unlike Theorem 1.1. However (assuming the de Rham-theoretic conjecture), Theorem 1.2 is most interesting for k in the stable range, since only then can the Donaldson invariant $D([\Sigma_1] \cdots [\Sigma_n])$ be expressed as a topologically invariant integral $\int_{\mathcal{M}} \mu_{\mathrm{d}}(\omega_1) \wedge \cdots \wedge \mu_{\mathrm{d}}(\omega_n)$.

Additionally, note that like our intersection-theoretic calculation, a posteriori our differential-forms result is insensitive to b_+. This was not obvious a priori since the differential forms $\mu_{\mathrm{d}}(\cdot)$ are nonlocal. Only after the rather detailed analysis in Sections 9 and 10 will it be clear that integrals in Theorem 1.2 are independent of b_+. Furthermore, even given the insensitivity to b_+, a priori the limit in Theorem 1.2 could have been 4 rather than $\frac{1}{2}$; i.e. the boundary behavior of the forms $\mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q)$ could have given simple-type calculus for *all* manifolds. The conclusion then would have been just that for manifolds not of simple type, the interior behavior of these forms is more complicated than it is for manifolds of simple type. While this hoped-for conclusion is false, it is false in a very precise and universal way that is interesting in itself. Modulo the technical hypotheses, Theorem 1.2 shows that the restriction of $\mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q)$ to a neighborhood of the boundary exhibits

delta-form behavior, concentrating with universal amplitude on the boundary as $p \to q$ — but exactly $\frac{1}{8}$ the amplitude the authors had anticipated.

The rest of this paper is organized into two main parts, with Sections 2–4 devoted to proving Theorem 1.1 and Sections 5–10 devoted to proving Theorem 1.2. The strategy of proof, and the division of the paper, is as follows:

Let A be a connection obtained by gluing a small bubble onto a background connection A_0. It turns out that the curvature of A is well approximated by the sum of the curvature F_0 of A_0 and the curvature F_{std} of a standard $k = 1$ instanton, viewed in the correct gauge. We are thus led to the following model problem: *Given a connection* $[A_0] \in \mathcal{M}_k$ *and two closely spaced points p and q, for how many triples* (x, λ, g) *is the sum of the curvature* F_0 *of* A_0 *and the curvature* F_{std} *of a standard instanton, centered at x with size* λ *and gluing angle g, reducible at both p and q*? In Section 2 we solve this model problem and show that, for generic A_0, the answer is 6.

In Section 3 we construct a family of approximately ASD connections based on an explicit gluing formula. This is a perturbation, which we denote by $\tilde{\mathcal{M}}_{k+1}$, of the boundary region of $\mathcal{M}_{k+1}$. We check explicitly that in this family the curvature is well approximated by $F_0 + F_{\text{std}}$. By linearly interpolating between $F_0 + F_{\text{std}}$ and the actual curvatures of connections in $\tilde{\mathcal{M}}_{k+1}$, we show that corresponding to each generic $A_0 \in \mathcal{M}_k$ there are exactly six points in $\tilde{\mathcal{M}}_{k+1} \cap \nu_p \cap \nu_q$ with λ sufficiently small.

In Section 4 we apply these results to show that if we consider only the boundary region of the (perturbed) moduli space, we obtain (1.4) with 6 on the right-hand side rather than 64, thereby completing the proof of Theorem 1.1.

Ideally, one would then like to interpolate from $\tilde{\mathcal{M}}_{k+1}$ to $\mathcal{M}_{k+1}$. This is quite difficult as ν_p and ν_q are defined by pointwise conditions on the curvature. We know of no pointwise estimates relating the curvature of an almost-ASD connection to that of a nearby ASD connection. In order to make use of the integral estimates available in the literature, one would have to replace ν_p and ν_q by geometric representatives defined by integral conditions. While certainly possible, this is beyond the scope of this paper.

We prove Theorem 1.2 by exhibiting $\mu_{\text{d}}(p)$ as a purely local piece $\mu_{\text{loc}}(p)$ plus a nonlocal remainder. The local piece dominates in (1.11): as $q \to p$ the integral of $\mu_{\text{loc}}(p) \wedge \mu_{\text{loc}}(q)$ approaches a calculable integral on $\mathbf{R}^8$, with value $\frac{1}{2}$, independent of λ_0. (However, $\mu_{\text{loc}}(p) \wedge \mu_{\text{loc}}(p) \equiv 0$.) We establish (1.11) and (1.12) by showing that the integral of the remainder terms in $\mu(p) \wedge \mu(q)$ approaches zero as $\lambda_0 \to 0$, independent of p and q. Thus taking a limit as $q \to p$ is relevant only to the purely local part of $\mu_{\text{d}}(p) \wedge \mu_{\text{d}}(q)$ (and taking a limit as $\lambda_0 \to 0$ is relevant only to the nonlocal part); the delta-form behavior of $\mu_{\text{d}}(p) \wedge \mu_{\text{d}}(p)$ is due solely to $\mu_{\text{loc}}(p) \wedge \mu_{\text{loc}}(p)$. The uniformity assertion in Theorem 1.2 follows from the proofs of (1.11) and (1.12), and the final assertion (1.13) then follows from (1.7).

In Section 5 we begin our work on Theorem 1.2 by constructing the de Rham representatives $\mu_{\text{d}}(p)$. The splitup $\mu_{\text{d}}(p) = \mu_{\text{loc}}(p) +$ remainder is based on the "approximate tangent space" mentioned above. This approximation is built by lifting the action of certain vector fields on N to $\mathcal{B}^*$. In Section 6 we discuss this lifted action (the "canonical flow"), use it to define the approximate tangent spaces $\mathcal{H}_A$, and discuss how close the $\mathcal{H}_A$ are

to being tangent to $\mathcal{M}$. We then exhibit the relation between a fiber constructed from the canonical flow (whose tangent space is essentially the projection to $T_{[A]}\mathcal{M}$ of approximate tangent space above) and the fiber constructed in [4]. This digression is needed to motivate the technical assumptions (Z1)–(Z5) given and discussed in Section 7. In Section 8 we return to the main track, defining $\mu_{\rm loc}(p)$ and computing the limiting integral of $\mu_{\rm loc}(p) \wedge \mu_{\rm loc}(q)$. Sections 9 and 10 are devoted to a study of the remainder terms $\mu_{\rm d}(p) \wedge \mu_{\rm d}(q) - \mu_{\rm loc}(p) \wedge \mu_{\rm loc}(q)$. In Section 9 we state the main technical theorem that yields the pointwise norm of these terms (Proposition 9.2), and use this theorem to establish that the integral of the remainder terms tends to zero as $\lambda_0 \to 0$. Finally in Section 10, we prove Proposition 9.2. It is this section that contains the core of the analysis underpinning the validity of all the earlier calculations. The estimates in Section 10 require a weighted Sobolev inequality, proven in Appendix A, that the authors have not seen elsewhere.

2. The model intersection theory calculation

In this section we begin to compare the boundary region of $\mathcal{M}_{k+1} \cap \nu_p \cap \nu_q$ with $\mathcal{M}_k$ by looking at a model problem. Pick a small neighborhood $\tilde{U}$ of our manifold N and give it a flat metric with corresponding Euclidean coordinates. Let U be the corresponding ball in $\mathbf{R}^4$. We will denote points either by four real coordinates $(x^0, \ldots, x^3)$ or by a single quaternionic coordinate $x^0 + ix^1 + jx^2 + kx^3$. Let p and q be the points $(\pm L, 0, 0, 0)$. Let A_0 be an ASD connection on N expressed in a smooth gauge on $\tilde{U}$.

An important notational tool is the identification of ASD curvatures with 3×3 real matrices. Let F_0 be the pullback to U of the curvature F_{A_0} of an ASD connection on $\tilde{U}$. Relative to the standard oriented basis of $\Lambda^2_- T^*\mathbf{R}^4$ ($\omega_1 = \mathrm{d}x^0\,\mathrm{d}x^1 - \mathrm{d}x^2\,\mathrm{d}x^3$, $\omega_2 = \mathrm{d}x^0\,\mathrm{d}x^2 - \mathrm{d}x^3\,\mathrm{d}x^1$, $\omega_3 = \mathrm{d}x^0\,\mathrm{d}x^3 - \mathrm{d}x^1\,\mathrm{d}x^2$), F_0 has at each point three Lie-algebra-valued components, and so can be viewed as a triple of 3-vectors. We package this triple of vectors into a 3×3 real matrix, which we denote by $Mat(F_0)$. More precisely, the first, second and third columns of $Mat(F_0)$ are half the ω_1, ω_2 and ω_3 components of F_0, while the first, second and third entries of each column refer to the three directions in $\mathfrak{su}(2)$, the Lie algebra of $SU(2)$. A_0 is reducible at a point if and only if $Mat(F_0)$ has rank 1 (or 0) there.

Of course, this construction is dependent on gauge and a choice of basis for TN. A gauge transformation is a change of basis in $\mathfrak{su}(2)$, and thus changes $Mat(F_0)$ by left-multiplication by an element of $SO(3)$. An orthogonal change of basis in TN changes $Mat(F_0)$ by right-multiplication by an element of $SO(3)$. Thus the singular values of $Mat(F_0)$, and in particular the rank of $Mat(F_0)$, are gauge- and basis-independent. We shall frequently be thinking of curvatures as 3×3 matrices in this way. When the context is clear, we will omit the explicit function "Mat".

Now consider a standard $k = 1$ instanton on $\mathbf{R}^4$ of scale size λ and center y, viewed in a radial gauge that is singular at y and regular at ∞. There are many such gauges parametrized by a gluing angle $m \in SO(3)$. For fixed (y, λ, m), let $F_{\rm std}$ be the curvature of this connection restricted to U.

Let A be an ASD connection obtained by gluing in a bubble with data (y, λ, m) to the background A_0. In Section 3 we shall see that F_A, in an appropriate gauge, is approximately equal to $F_0 + F_{\text{std}}$. This reduces our main question to the following model problem:

When L is small, for what values of (y, λ, m)*, with* λ *small, is* $F_0 + F_{\text{std}}$ *reducible at both p and q?*

Of course, to obtain sensible answers, we must define what we mean by λ being "small". Pick constants $K > 0$ and $\alpha \in (0, 2)$. We say λ is small (or that the corresponding bubble is small) if $\lambda < KL^\alpha$. The set of gluing data for small bubbles near p and q is $B = U \times (0, KL^\alpha) \times SO(3)$. Let $\tilde{\nu}_p$ (resp. $\tilde{\nu}_q$) be the set of points $(\lambda, y, m) \in B$ such that $F_0(p) + F_{\text{std}}(p)$ (resp. $F_0(q) + F_{\text{std}}(q)$) is reducible. We must count the intersection points of $\tilde{\nu}_p$ and $\tilde{\nu}_q$.

Recall that the singular values $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq 0$ of a 3×3 real matrix M are the square roots of the eigenvalues of $M^{\mathrm{T}}M$. For M generic, these are distinct and positive. The nongeneric cases are as follows: Matrices in a codimension-1 set have $\sigma_3 = 0$. Matrices in a codimension-2 set either have $\sigma_1 = \sigma_2$ or $\sigma_2 = \sigma_3$. Matrices in a codimension-4 set have $\sigma_2 = \sigma_3 = 0$; these matrices have rank 1 or 0. Matrices in a codimension-5 set have $\sigma_1 = \sigma_2 = \sigma_3$; these are all scalar multiples of $SO(3)$ matrices. Only the zero matrix (codimension-9) has $\sigma_1 = \sigma_2 = \sigma_3 = 0$.

Theorem 2.1. *Fix* $K > 0$, $\alpha \in (0, 2)$, *and a background connection* A_0. *If the singular values of* $Mat(F_0(0))$ *are all distinct, then for all sufficiently small L,* $\tilde{\nu}_p$ *and* $\tilde{\nu}_q$ *intersect at exactly six points. These six intersections are all transverse, and the local intersection number is* $+1$ *at each point.*

Remark. We shall see that, under the assumptions of the theorem, the intersection points all have $\lambda = \mathrm{O}(L^2)$. However, when two of the singular values of $Mat(F_0(0))$ are the same, then there are only four intersection points with $\lambda = \mathrm{O}(L^2)$. In that case there are generically four additional intersection points with $\lambda = \mathrm{O}(L)$. The intersection number of $\tilde{\nu}_p$ and $\tilde{\nu}_q$ then 4 if $\alpha > 1$ and 8 if $\alpha < 1$.

Before beginning the proof of Theorem 2.1 we need some basic facts about $k = 1$ instantons on $\mathbf{R}^4 = \mathbf{H}$, we need to fix some conventions, and we need a linear algebra lemma. Think of $SU(2)$ as the unit quaternions with $\mathfrak{su}(2)$ as the imaginary quaternions. The connection form of a standard instanton of scale size 1, centered at the origin, is $A_{\text{std}_0} = \mathrm{Im}(\bar{x}\,\mathrm{d}x/(1 + |x|^2))$. The curvature of this connection is

$$F_{\text{std}_0} = \frac{\mathrm{d}\bar{x}\,\mathrm{d}x}{(1 + |x|^2)^2} = \frac{2i\omega_1 + 2j\omega_2 + 2k\omega_3}{(1 + |x|^2)^2}. \tag{2.1}$$

Note that the matrix $Mat(F_{\text{std}_0})$ is $1/(1 + |x|^2)^2$ times the identity matrix.

That is in the usual regular gauge, in which $A \sim \phi^{-1}\,\mathrm{d}\phi$ as $|x| \to \infty$, where $\phi(x) = x/|x|$. We do a gauge transformation by ϕ^{-1} to get a radial gauge in which $A = \mathrm{O}(|x|^{-3})$ as $|x| \to \infty$ (and in which A is singular at the origin). We then do a further gauge transformation by a constant g_0 to get the most general radial gauge with this property. Let F_{std} be the curvature form in this gauge. We have $F_{\text{std}} = g_0^{-1}\phi F_{\text{std}_0}\phi^{-1}g_0$. In terms of matrices,

$Mat(F_{\mathrm{std}}) = \rho(g_0^{-1})\rho(\phi)Mat(F_{\mathrm{std}_0})$, where ρ is the standard double covering map from $SU(2)$ to $SO(3)$; the three columns of $\rho(\phi)$ are $\phi i\phi^{-1}$, $\phi j\phi^{-1}$, and $\phi k\phi^{-1}$. The matrix $\rho(g_0)$ is our gluing angle m.

Now suppose that we have a $k = 1$ instanton, centered at a point y, with scale size λ. The curvature matrix, expressed in the exterior radial gauge of gluing angle m, is

$$Mat(F_{\mathrm{std}}(x)) = \frac{\lambda^2}{(\lambda^2 + |x-y|^2)^2} m^{-1}\rho\left(\frac{x-y}{|x-y|}\right). \tag{2.2}$$

Note that the matrix $Mat(F_{\mathrm{std}}(x))$ is a positive multiple of an $SO(3)$ matrix. The multiple is determined by λ and $|x-y|$, while the $SO(3)$ matrix is determined by m and $(x-y)/|x-y|$. (We henceforth will not explicitly distinguish between a curvature and its matrix.)

Our problem is thus one of adding positive multiples of $SO(3)$ matrices to $F_0(p)$ and $F_0(q)$ to make them reducible. The following lemma is essential.

Lemma 2.2. *Let P be a 3×3 real matrix with singular values $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq 0$. If these singular values are all distinct, then there are exactly two pairs $(s, M) \in (0,\infty)\times SO(3)$ for which $P + sM$ has rank 1 (and no pairs (s, M) for which $P + sM = 0$). In both cases $s = \sigma_2(P)$. If exactly two of the singular values of P are the same and nonzero, then the two solutions (s, M) coalesce to a double root.*

Proof. Let $W = -(P + sM)$. Adding sM to P to make it reducible is the same as decomposing $-P$ as $sM + W$ with W reducible. We therefore count the ways to decompose a matrix $-P$ into the sum of a positive multiple of an $SO(3)$ matrix and a rank 1 matrix. First we show that the desired decompositions can occur *only* with $s = \sigma_2$ by assuming a decomposition $-P = sM + W$ and computing $\sigma_2(P)$. Multiplying P on the left and right by $SO(3)$ matrices does not change the singular values, but does allow us to set $M = I$ and put W into the form

$$W = \begin{pmatrix} a & b & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{2.3}$$

Then

$$P^{\mathrm{T}}P = \begin{pmatrix} (s+a)^2 & (s+a)b & 0 \\ (s+a)b & s^2+b^2 & 0 \\ 0 & 0 & s^2 \end{pmatrix}. \tag{2.4}$$

One of the eigenvalues of $P^{\mathrm{T}}P$ is obviously s^2 with eigenvector $(0,0,1)$. Restricting to the upper left 2×2 block, we subtract $s^2 I$ and get a matrix whose determinant, $-s^2b^2$, is nonpositive. Thus at most one eigenvalue of $P^{\mathrm{T}}P$ is greater than s^2 and at most one eigenvalue is less than s^2. Since s^2 is the middle eigenvalue of $P^{\mathrm{T}}P$, $\sigma_2(P) = s$.

Next we show that $P + sM$ can have rank 1, with $s = \sigma_2(P)$, in two ways. By multiplying on the left and right by $SO(3)$ matrices, we can take P diagonal with entries $P_{11} \geq P_{22} \geq$

$|P_{33}|$. Next we look for orthogonal matrices of the form

$$M_\theta = \begin{pmatrix} -\cos\theta & 0 & \sin\theta \\ 0 & -1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix}. \tag{2.5}$$

We then have

$$P + sM_\theta = P + P_{22}M_\theta = \begin{pmatrix} P_{11} - P_{22}\cos\theta & 0 & P_{22}\sin\theta \\ 0 & 0 & 0 \\ P_{22}\sin\theta & 0 & P_{33} + P_{22}\cos\theta \end{pmatrix}. \tag{2.6}$$

This matrix has an obvious null vector $(0, 1, 0)$. $P + sM_\theta$ has rank 1 (or 0) if and only if there is a second null vector. To see if there is a second null vector, we restrict $P + sM_\theta$ to the 1–3 plane and take its determinant, which equals $-P_{22}^2 + P_{11}P_{33} + (P_{11} - P_{33})P_{22}\cos\theta$. This is a periodic function of θ with a single maximum of $(P_{11} - P_{22})(P_{22} + P_{33})$ at $\theta = 0$ and a single minimum of $-(P_{11} + P_{22})(P_{22} - P_{33})$ at $\theta = \pi$. If $P_{11} > P_{22} > |P_{33}|$, the maximum and minimum values have opposite signs, so the function must cross zero exactly twice at the points $\theta = \pm\cos^{-1}([P_{22}^2 - P_{11}P_{33}]/(P_{11} - P_{33})P_{22})$. If $P_{11} = P_{22}$ or $P_{33} = -P_{22}$, then the maximum value becomes zero, while if $P_{22} = P_{33}$, then the minimum becomes zero. In these cases we have a double root at $\theta = 0$ or π. Finally, if $P_{11} = P_{22} = P_{33}$, then the function is identically zero and we have an infinite number of roots. This corresponds to the original P being a positive multiple of an $SO(3)$ matrix.

Finally, we show that these are the only possible decompositions with $s = P_{22}$. Suppose that M is an $SO(3)$ matrix with $P + sM$ having rank 1. Then every 2×2 block of $P + sM$ has determinant 0, and in particular the upper left 2×2 block has a null vector v. However, P_{11} and P_{22} are both at least s, so $|Pv| \geq s$. The upper left corner of sM has operator norm at most s, so $|sMv| \leq s$. Thus we must have $|Pv| = s|Mv| = s = P_{22}$. If $P_{11} > P_{22}$, this means $v = (0, 1, 0)$, so $Mv = (0, -1, 0)$, so M must take the form (2.5). The case $P_{11} = P_{22}$ must be checked separately, but leads only to the solution $M = diag(-1, -1, 1)$. □

The form of the explicit solutions found above also demonstrates the continuous dependence of M on P. Expressed invariantly, M is a rotation by π about an axis. This axis is orthogonal to the second principal axis of $P^{\mathrm{T}}P$, and makes an angle $\theta/2 = (\pm\frac{1}{2})\cos^{-1}([\sigma_2^2 \pm \sigma_1\sigma_3]/[(\sigma_1 \pm \sigma_3)\sigma_2])$ with the third principal axis of $P^{\mathrm{T}}P$, where the $\pm$ is determined by the sign of the determinant of P. A small change in P can only change θ by an amount of order $|\delta P|/\min(\sigma_1 - \sigma_2, \sigma_2 - \sigma_3)$, and, by first order perturbation theory (integrated to get rigorous bounds), can only change the principal axes of $P^{\mathrm{T}}P$ by a similar amount. Thus if δP is a small perturbation of P, the norm of the corresponding δM is bounded by a constant times $|\delta P|/\min(\sigma_1 - \sigma_2, \sigma_2 - \sigma_3)$.

Not surprisingly, this stability breaks down when we approach the double root. If two of the singular values are equal, then a small perturbation may change M by as much as $O(\sqrt{|\delta P|})$.

Proof of Theorem 2.1. Let s_p be the second singular value of $F_0(p)$, and let $M_p \in SO(3)$ be a matrix such that $F_0(p) + s_pM_p$ is reducible (with similar definitions for s_q and M_q).

Let s_0 be the second singular value of $F_0(0)$. Note that $s_0 > 0$ since the three singular values of $F_0(0)$ were assumed distinct. Since s_p and s_q are within $O(L)$ of s_0, we can bound s_p and s_q away from zero.

We shall count the ways to simultaneously make $F_{\text{std}}(p) = s_p M_p$ and $F_{\text{std}}(q) = s_q M_q$. The condition for the standard curvature F_{std} to have magnitude s_p at p is

$$\frac{\lambda^2}{(|y-p|^2+\lambda^2)^2} = s_p, \tag{2.7}$$

or equivalently

$$\lambda^2 + |y-p|^2 = \lambda/\sqrt{s_p}. \tag{2.8}$$

As long as $|y-p| < 1/2\sqrt{s_p}$ there are two solutions to (2.8), while for $|y-p| > 1/2\sqrt{s_p}$ there are none. When $|y-p| < 1/2\sqrt{s_p}$, one solution has $\lambda > 1/2\sqrt{s_p}$, which is greater than KL^α for L small. The other solution qualifies as small if $|y-p|$ is small enough and, for $|y-p| \ll 1/\sqrt{s_p}$, is approximately $\lambda = |y-p|^2\sqrt{s_p}$. As a set in $\mathbf{R}^5 = (N, \lambda)$ space, the solutions to (2.8) are a 4-sphere. Projected onto N, they form (two copies of) a 4-disk. In either case, only a small subset of solutions qualifies as "small".

The interesting question, of course, is how many times we can solve the equations for p and q simultaneously. We begin with Eq. (2.8) and the corresponding equation for q. The intersection of two 4-spheres in $\mathbf{R}^5$ is a 3-sphere. Projected onto N, we get a three-dimensional ellipsoid, possibly degenerating to two disks. As before, only a small patch of the ellipsoid (or alternative part of one of the two disks) gives a small enough value of λ. It is this region that we consider.

Recall that p and q are at $\pm L$, where we are using quaternionic coordinates. For L small, $s_q = s_p + O(L)$. Let s_m be such that $2/\sqrt{s_m} = 1/\sqrt{s_p} + 1/\sqrt{s_q}$. Let $\Delta = (1/\sqrt{s_p} - 1/\sqrt{s_q})/L$. As $L \to 0$, $s_m = s_0 + O(L^2)$, while Δ approaches $-(\mathrm{d}s_p/\mathrm{d}L)|_{L=0}/s_0^{3/2}$. Let y_0 and y_I be the real and imaginary parts of y. Adding and subtracting (2.8) to the corresponding equation for q we obtain

$$-4y_0 = \lambda\Delta, \qquad \lambda^2 + L^2 + y_0^2 + |y_I|^2 = \lambda/\sqrt{s_m}. \tag{2.9}$$

Plugging the first equation into the second, we get

$$\lambda^2\left(1 + \frac{\Delta^2}{16}\right) - \frac{\lambda}{\sqrt{s_m}} + L^2 + |y_I|^2 = 0. \tag{2.10}$$

This equation shows that λ, and thus y_0, may be viewed as functions of y_I. As long as $L^2+|y_I|^2 \ll 1/\sqrt{s_m}$ there are two solutions to (2.10), one of which has $\lambda \approx (L^2+|y_I|^2)\sqrt{s_m}$, the other of which has $\lambda \approx ((1+\Delta^2/16)\sqrt{s_m})^{-1}$. The first solution has $\lambda < KL^\alpha$ if and only if $|y_I|$ is small enough, while the second solution always has $\lambda > KL^\alpha$. Let $R_{K,\alpha}$ be the largest number such that $|y_I| < R_{K,\alpha}$ implies $\lambda \le KL^\alpha$. Henceforth we consider only "admissible" y, i.e. those with $|y_I| < R_{K,\alpha}$. For L chosen small enough, as we assume henceforth it is, $R_{K,\alpha}^2 \sim KL^\alpha/\sqrt{s_m} - L^2 \sim KL^\alpha/\sqrt{s_m}$, since $\alpha < 2$. Note that

$$y_0 = -\tfrac{1}{4}\lambda\Delta \approx -\tfrac{1}{4}(L^2 + |y_I|^2)\sqrt{s_m}\Delta. \tag{2.11}$$

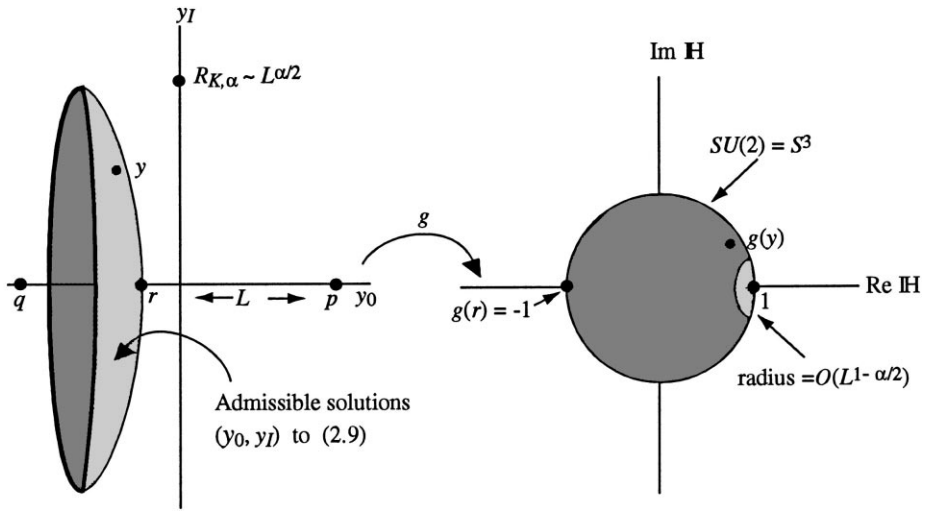

Fig. 1. Diagram for proof of Theorem 2.1.

Hence for admissible y, we have $|y_\mathrm{I}| < \text{const.}\, L^{\alpha/2}$ and $|y_0| < \text{const.}\, L^{\alpha}$. Let $r = (y_0(0), 0)$ be the unique admissible point where the ellipsoid of solutions (y_0, y_I) to (2.9) hits the real axis. Since $|r| = \mathrm{O}(L^2)$, r lies on the line segment $\overline{pq}$, and the ellipsoid has curvature $\mathrm{O}(1)$ at r (see Fig. 1).

We still have to get the $SO(3)$ matrices right. This means simultaneously solving the equations $m^{-1}\rho((y-p)/|y-p|) = M_p$ and $m^{-1}\rho((y-q)/|y-q|) = M_q$ for m. If a solution exists, it is obviously unique. A solution exists if and only if $\rho((y-p)/|y-p|)^{-1}\rho((y-q)/|y-q|) = M_p^{-1}M_q$. Let $g(y) = (\bar{y}-\bar{p})(y-q)/|(y-p)(y-q)|$. We must count the points on our 3-disk (of small solutions to (2.9) and (2.10)) for which the $SO(3)$-valued function $\rho(g(y))$ equals $M_p^{-1}M_q$. Note that

$$g(y) = -I + 2\frac{y_\mathrm{I}}{L}(1 + \mathrm{O}((|y_0|/L)^2)) + \mathrm{O}((|y_\mathrm{I}|/L)^2) \quad \text{for } |y_\mathrm{I}| \ll L, \tag{2.12}$$

while

$$g(y) = I + 2\frac{Ly_\mathrm{I}}{|y_\mathrm{I}|^2}(1 + \mathrm{O}((|y_0|/|y_\mathrm{I}|)^2)) + \mathrm{O}((L/|y_\mathrm{I}|)^2) \quad \text{for } |y_\mathrm{I}| \gg L. \tag{2.13}$$

In view of (2.11), we can replace $\mathrm{O}((|y_0|/L)^2)$ in (2.12) and $\mathrm{O}((|y_0|/|y_\mathrm{I}|)^2)$ in (2.13) by $\mathrm{O}(L^2)$ and $\mathrm{O}(L^{2\alpha})$, respectively.

Observe that $L/R_{K,\alpha}$ is $\mathrm{O}(L^{1-\alpha/2})$ and hence goes to zero as $L \to 0$. On the disk of admissible y_I, the map g covers all of $SU(2)$ except for a ball of radius $cL^{1-\alpha/2}$ around the identity for some constant c. Since ρ is a 2–1 map, $\rho(g(y))$ hits all of $SO(3)$ twice, except for a ball of radius $2cL^{1-\alpha/2} + \mathrm{O}(L^{2-\alpha})$ around the identity, which is only hit once. The number of solutions to our problem depends on whether, for small L, $M_p^{-1}M_q$ is in this ball or not.

If the singular values of $F_0(0)$ are distinct, then by Lemma 2.2, there are two distinct matrices $M_{1,2}(0)$ for which $F_0(0) + \sigma_2(0)M$ has rank 1. By the comment after the proof of Lemma 2.2, the two matrices for p and q satisfy $M_{1,2}(p,q) = M_{1,2}(0) + \mathrm{O}(L)$. As $L \to 0$, $M_1(p)^{-1}M_2(q)$ and $M_2(p)^{-1}M_1(q)$ are bounded away from the identity, but $M_1(p)^{-1}M_1(q)$ and $M_2(p)^{-1}M_2(q)$ are within $\mathrm{O}(L)$ (and hence within $2cL^{1-\alpha/2} + \mathrm{O}(L^{2-\alpha})$) of the identity. Thus we have two configurations in (y, λ, m) space that give $s_pM_1(p)$ at p and $s_qM_2(q)$ at q, two that give $s_pM_2(p)$ at p and $s_qM_1(q)$ at q, one that gives $s_pM_1(p)$ at p and $s_qM_1(q)$ at q and one that gives $s_pM_2(p)$ at p and $s_qM_2(q)$ at q. A total of six solutions in all.

On a codimension-2 set of background data, the background curvature at the origin has two equal singular values, so $M_1(0) = M_2(0)$ and $M_{1,2}(p,q) = M_1(0) + \mathrm{O}(L^{1/2})$. In that case all four possibilities have $M_p^{-1}M_q = 1 + \mathrm{O}(L^{1/2})$. If $\alpha > 1$, this is within $2cL^{1-\alpha/2}$ of the identity for small enough L, and so each possibility yields one solution. If $\alpha < 1$ and the $\mathrm{O}(L^{1/2})$ term in the expansion of $M_{1,2}(p,q)$ in powers of L is nonzero, then each possibility yields two solutions.

Finally we consider the orientations of our solutions. It is not immediately clear that all solutions have the same orientation, but in fact they do. The problem of matching amplitudes is the same in all cases. The problem of matching gluing angles reduces to the intersection of two 3-cycles in a 3-disk $\times$ $SO(3)$ (i.e., all possible pairs (y_{I}, m)), and is easily seen to be transverse. The intersection numbers are continuous functions of M_p and M_q as long as a solution continues to exist. Sending M_p around a noncontractible loop in $SO(3)$ interchanges the two solutions associated to a given pair (M_p, M_q), which shows that the two solutions for any given (M_p, M_q) have the same orientation. Also by continuity, this orientation is the same for all pairs (M_p, M_q).

All that remains is to compute this orientation in one case. Let $s_p = s_q = 1$, $M_p = M_q = I$, and look near the solution with $y = 0$ and $m = I$. The varieties $\tilde{\nu}_p$ and $\tilde{\nu}_q$ are just the zero sets of $F_{\mathrm{std}}(p) - I$ and $\tilde{F}_{\mathrm{std}}(q) - I$, which we view as functions of (y, λ, m). Taking derivatives, we find that changes in (y, λ, m) give the following first order changes in $F_{\mathrm{std}}(p)$ and $F_{\mathrm{std}}(q)$:

1. Increasing λ increases the magnitude of both $\tilde{F}_{\mathrm{std}}(p)$ and $F_{\mathrm{std}}(q)$ without changing either direction.
2. Increasing y_0 increases the magnitude of $F_{\mathrm{std}}(p)$ and decreases that of $F_{\mathrm{std}}(q)$, while keeping the directions fixed.
3. Increasing y_1 (resp. y_2, y_3) rotates $\tilde{F}_{\mathrm{std}}(p)$ in the direction defined by the Lie algebra element $-i$ (resp. $-j$, $-k$), and rotates $F_{\mathrm{std}}(p)$ an equal amount in the direction $+i$ (resp. $+j$, $+k$).
4. Rotating m in any direction rotates both $F_{\mathrm{std}}(p)$ and $F_{\mathrm{std}}(q)$ in the opposite direction.

>From this we deduce that the Jacobian $|\mathrm{d}(F_{\mathrm{std}}(p), F_{\mathrm{std}}(q))/\mathrm{d}(y, \lambda, m)|$ is positive, and that the local intersection number of $\tilde{\nu}_p$ and $\tilde{\nu}_q$ is $+1$ in this case. Thus the local intersection number of $\tilde{\nu}_p$ and $\tilde{\nu}_q$ is $+1$ in all cases. □

Having proven Theorem 2.1, we consider the question of stability. How much do our intersection points move around if we change M_p or M_q or s_p or s_q slightly? Since $F_0 + F_{\mathrm{std}}$

is only an approximation to the true curvature of a connection in $\mathcal{M}_{k+1}$, our results must be stable in order to be meaningful.

Let χ be the map that takes (y, λ, m) to $(F_{\text{std}}(p), F_{\text{std}}(q))$. Near our solutions, $\mathrm{d}\chi$ is never close to singular. By changing λ and one component of y we can adjust $|F_{\text{std}}(p)|$ and $|F_{\text{std}}(q)|$ independently, while by adjusting m and the remaining three components of y we can adjust the directions of $F_{\text{std}}(p)$ and $F_{\text{std}}(q)$ independently. It is not difficult to estimate the matrix elements of $(\mathrm{d}\chi)^{-1}$. Some are $\mathrm{O}(1)$, some are $\mathrm{O}(L)$, and some are $\mathrm{O}(L^2)$. If we know the required F_{std} (p or q) to within ϵ, we know m to within $\mathrm{O}(\epsilon)$, y to within $\mathrm{O}(\epsilon L)$, and λ to within $\mathrm{O}(\epsilon L^2)$. In short, small errors in the input data result in only small changes of the locations of our intersection points in (y, λ, m) space.

Finally, we consider a perturbation of our model problem that is more directly applicable in the sequel. Let $\tilde{F}_0(x)$ be the curvature of the background connection in the standard radial gauge about the gluing point y. (That is, use the original connection A_0 to trivialize the fiber over y, and then use parallel transport radially outwards from y to trivialize the bundle over U.) We wish to count the number of ways to make $\tilde{F}_0 + F_{\text{std}}$ reducible at both p and q.

Theorem 2.3. *Under the assumptions of Theorem* 2.1, *the number of ways to make* $\tilde{F}_0 + F_{\text{std}}$ *reducible at p and q* (*counted with sign*) *is the same as the number of ways to make* $F_0 + F_{\text{std}}$ *reducible at p and q* (*counted with sign*), *namely* $+6$.

Proof. We first put our background connection into a radial gauge with respect to the origin. This is a fixed gauge, and Theorem 2.1 applies. Since F_0 and $\tilde{F}_0$ are related by a gauge transformation, the singular values of F_0 and $\tilde{F}_0$ are the same. Thus we must solve (2.8) and the corresponding equation for q, exactly as in Theorem 2.1, with the same values of s_p and s_q. We then solve $m^{-1}\rho((y-p)/|y-p|) = M_p$ and $m^{-1}\rho((y-q)/|y-q|) = M_q$ for m as before. The only difference in our analysis is that M_p and M_q are now functions of y. We compute the extent to which they depend on y.

Let A be a connection in radial gauge with respect to y, and let A' be the same connection in radial gauge with respect to y'. The gauge transformation that relates these, evaluated at the point p, is the holonomy around a triangle from p to y to y' to p, and so its difference from the identity is bounded by the sup norm of $|F_A|$ times the area of the triangle (see Fig. 2).

In our case, A is the background connection, so $|F_A|$ is fixed and bounded, and y and y' are restricted to lie on the ellipsoid of solutions to (2.9) and (2.10) with $|y_{\text{I}}|$ and $|y'_{\text{I}}|$ both less than $R_{K,\alpha}$. Note that the area of a triangle is bounded by half the product of the length of *any* two of its legs. Because the curvature of the ellipsoid of solutions $y = (y_0, y_{\text{I}})$ is $\mathrm{O}(1)$ at admissible points, $|y_0 - y'_0|$ is bounded by a constant times $|y_{\text{I}} - y'_{\text{I}}|$. As a result,

$$|M_p(y) - M_p(y')| \leq \text{const.} \times \sqrt{L^2 + |y|^2}\,|y_{\text{I}} - y'_{\text{I}}|, \tag{2.14}$$

while

$$|M_p(y) - M_p(0)| \leq \text{const.} \times L|y_{\text{I}}| \tag{2.15}$$

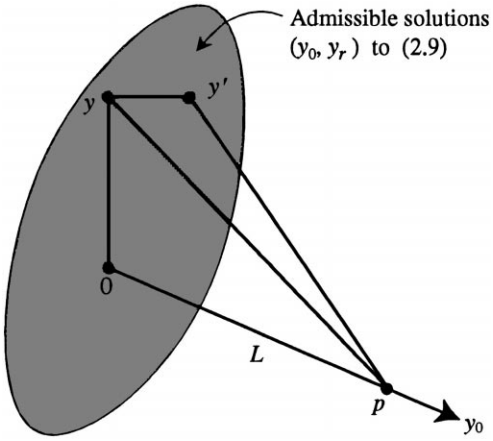

Fig. 2. Diagram for proof of Theorem 2.3.

with similar estimates for M_q. The second result is an estimate on M_p itself, while the first leads to a bound on the derivative of M with respect to y_I. By (2.11), $L^2 + |y|^2 \leq \text{const.}(L^2 + |y_\mathrm{I}|^2)$, so we obtain

$$\left|\frac{\partial M_p}{\partial y_\mathrm{I}}\right| \leq \text{const.} \times \sqrt{L^2 + |y_\mathrm{I}|^2}. \tag{2.16}$$

As before, we look for solutions to $\rho(g(y)) = M_p^{-1} M_q$, where now the right-hand side depends on y. We break the disk of radius $R_{K,\alpha} = \mathrm{O}(L^{\alpha/2})$ into two pieces, an inner disk D_1 and an annulus D_2. The radii will be chosen such that on D_1 the estimate (2.14) is strong enough to allow implicit function theorem arguments to apply. Here the solutions to $\rho(g(y)) = M_p^{-1}(y) M_q(y)$ are but small perturbations of the solutions to $\rho(g(y)) = M_p^{-1}(0) M_q(0)$. On D_2 the estimate (2.15) will be strong enough to show that there are no solutions to $\rho(g(y)) = M_p^{-1}(y) M_q(y)$. Taken together, this will prove the theorem.

On the disk D_1, the implicit function theorem will apply as long as the smallest singular value of $\partial(\rho \circ g)/\partial y_\mathrm{I}$ is at least twice the largest singular value of $\partial(M_p^{-1} M_q)/\partial y_\mathrm{I}$, which by (2.16) is bounded above by a multiple of $(L^2 + |y_\mathrm{I}|^2)^{1/2}$. Computing the derivative of $\rho \circ g$ is an easy geometrical calculation, and one finds that all singular values are bounded below by a constant times $L/(L^2 + |y_\mathrm{I}|^2)$. Comparing $L/(L^2 + |y_\mathrm{I}|^2)$ to $(L^2 + |y_\mathrm{I}|^2)^{1/2}$, we see that the implicit function theorem applies whenever $|y_\mathrm{I}|$ is smaller than a constant times $L^{1/3}$, and in particular whenever $|y_\mathrm{I}| < L^{1/2}$ (and L is sufficiently small). We take the radius of D_1 (and the inner radius of D_2) to be $L^{1/2}$.

Now consider $y_\mathrm{I} \in D_2$. If $\alpha > 1$, then D_2 is empty, so we assume $\alpha \leq 1$. By (2.13), $|I - \rho(g(y))| = 2L/|y_\mathrm{I}|(1 + \mathrm{O}(L^\alpha)) + \mathrm{O}(L^2/|y_\mathrm{I}|^2)$. Since $c_1 L^{1/2} < |y_\mathrm{I}| < c_2 L^{\alpha/2}$, $c_3 L^{1-\alpha/2} < |I - \rho(g(y))| < c_4 L^{1/2}$. Now recall that $M_p^{-1}(0) M_q(0)$ is either bounded away from the identity or is within $\mathrm{O}(L)$ of the identity (e.g. $M_1^{-1}(p) M_2(q)$ is bounded

away from the identity, while $M_1^{-1}(p)M_1(q)$ is within $O(L)$ of the identity). By (2.15), $M_p^{-1}(y)M_q(y)$ is also either bounded away from the identity or within $O(L)$ of the identity on D_2. Thus $|I - M_p^{-1}(y)M_q(y)|$ can never be between $c_3L^{1-\alpha/2}$ and $c_4L^{1/2}$, so there are no solutions to $\rho(g(y)) = M_p^{-1}(y)M_q(y)$ on D_2. □

3. The perturbed moduli space

In this section we show that the model problem of Section 2 correctly describes the intersection of ν_p, ν_q, and a perturbation (denoted by $\tilde{\mathcal{M}}_{k+1}$) of the boundary region of $\mathcal{M}_{k+1}$. $\tilde{\mathcal{M}}_{k+1}$ is parametrized by quadruples (A_0, y, λ, m), where $A_0 \in \mathcal{M}_k$ is a background connection, and the glued-in bubble has size λ, center y and gluing angle m. We construct $\tilde{\mathcal{M}}_{k+1}$ by an explicit gluing formula and show that, in the relevant region, the curvature of a connection in $\tilde{\mathcal{M}}_{k+1}$ is well approximated by the sum of the background curvature F_0 and the curvature F_{std} of a standard instanton of size λ, center y and gluing angle m. Our model problem was essentially to make this sum reducible at p and q. By interpolating between this sum and the actual curvature of a connection in $\tilde{\mathcal{M}}_{k+1}$, we show that the results of Section 2 carry over almost word for word.

As before, we pick a background connection $A_0 \in \mathcal{M}_k$ and constants $K > 0$ and $\alpha \in (0, 2)$. Let the neighborhood $\tilde{U}$ in N, and the corresponding neighborhood U of the origin in $\mathbf{R}^4$, be as in Section 2. We now allow bubbles to be glued in anywhere (not just in $\tilde{U}$), so the set B of gluing data is a $(0, KL^\alpha) \times SO(3)$ bundle over N, with local coordinates $(y, \lambda, m) \in N \times (0, KL^\alpha) \times SO(3)$. When the center of the bubble is in $\tilde{U}$, we identify the center point in N with the corresponding coordinate in U, and call both points y. For each $(y, \lambda, m) \in B$, let F be the curvature of the connection $(A_0, y, \lambda, m) \in \tilde{\mathcal{M}}_{k+1}$. The variety ν_p (resp. ν_q), restricted to the fiber of $\tilde{\mathcal{M}}_{k+1}$ over A_0, is the set of points $(\lambda, y, m) \in B$ such that $F^-(p)$ (resp. $F^-(q)$) is reducible. We must count the intersection points of ν_p and ν_q. In this section we prove the following theorem.

Theorem 3.1. *Fix $K > 0$, $\alpha \in (0, 2)$, and $A_0 \in \mathcal{M}_k$. If the singular values of $F_{A_0}(0)$ are all distinct, then for all sufficiently small L, the intersection number of ν_p, ν_q, and the fiber of $\tilde{\mathcal{M}}_{k+1}$ over A_0 is $+6$.*

We begin by constructing the space $\tilde{\mathcal{M}}_{k+1}$. For now, assume we are gluing a bubble of size λ in $\tilde{U}$ with the center point at the origin. There are three natural length scales determined by the background connection A_0. The first is the length scale $|F_{A_0}(0)|^{-1/2}$ of the background curvature at the origin. The second is the length scale $|F_{A_0}(0)|/|\nabla^A F_{A_0}(0)|$ at which this curvature varies. Let R_3 be the smaller of these two length scales. Finally, let s_0 be the second singular value of $F_{A_0}(0)$. It is easy to see that $s_0 < 1/R_3^2$, but there is no simple lower bound for s_0 (although, by assumption, s_0 is always positive). As we have seen, s_0 measures how far $F_{A_0}(0)$ is from being reducible.

Now pick additional length scales R_1 and R_2, which can depend on λ, R_3 and s_0 such that $R_1^2 < 10^{-6}\lambda R_3$ and $R_2^2 > 10^6\lambda/\sqrt{s_0}$. When $\lambda \ll R_3$, which is the only case we

will consider, we want $\lambda \ll R_1 \ll R_2 \ll R_3$. The points of interest x will all have $R_1 < |x| < R_2$. The number 10^6 is of course arbitrary. It is just chosen large enough that we can safely ignore small numerical factors.

Let $\beta(r)$ be a smooth monotonic function that equals zero for $r < \frac{1}{2}$ and equals 1 for $r > 2$, and such that $\beta' \leq 1$. We define cut-off functions $\beta_1(x) = \beta(|x|/R_1)$ and $\beta_2(x) = 1 - \beta(|x|/R_2)$.

Let A_0 be the background connection expressed in a smooth fixed radial gauge with respect to the origin. Let A_{std} be the connection of a standard instanton of size λ expressed in a radial gauge that is *singular* at the origin and regular at ∞. (This gauge is not unique; it depends on a gluing angle m. See the discussion before expression (2.2).) Note that $|A_{\text{std}}| \sim \lambda^2/r^3$ for $r \gg \lambda$, while $|A_0| \sim r|F_{A_0}| = r/R_3^2$ for $r \ll R_3$.

Our point $(A_0, 0, \lambda, m) \in \tilde{\mathcal{M}}_{k+1}$ is defined by the connection form

$$A' = \beta_1 A_0 + \beta_2 A_{\text{std}}. \tag{3.1}$$

We compute

$$\begin{aligned} F = F_{A'} = \mathrm{d}A' + A' \wedge A' = {} & \beta_1 F_{A_0} + \beta_2 F_{A_{\text{std}}} + (\beta_1^2 - \beta_1) A_0 \wedge A_0 \\ & + (\beta_2^2 - \beta_2) A_{\text{std}} \wedge A_{\text{std}} + \mathrm{d}\beta_1 \wedge A_0 + \mathrm{d}\beta_2 \wedge A_{\text{std}} \\ & + \beta_1\beta_2 (A_{\text{std}} \wedge A_0 + A_0 \wedge A_{\text{std}}), \end{aligned} \tag{3.2}$$

and the interpolating 2-form

$$F_t = t(F_{A_0} + F_{\text{std}}) + (1 - t)F, \tag{3.3}$$

where $0 \leq t \leq 1$.

If the bubble is to be glued in at a point y, rather than at the origin, we must adjust the formulas as follows. First suppose $y \in \tilde{U}$. Take A_0 as the connection of the background in radial gauge with respect to y (not with respect to 0). The quantities s_0, R_1, R_2, and R_3 are computed from the curvature $F_{A_0}(y)$, not $F_{A_0}(0)$. The connection A_{std} is in a singular radial gauge with respect to y (not with respect to 0). The cut-off functions are $\beta_1(x) = \beta(|x-y|/R_1)$ and $\beta_2(x) = 1-\beta(|x-y|/R_2)$. With these modifications, we still have $A' = \beta_1 A_0 + \beta_2 A_{\text{std}}$, and formulas (3.2) and (3.3) still apply. For $y \notin U$, just apply the same formulas, using geodesic normal coordinates around y. In this case the "standard instanton" A_{std} is no longer exactly anti-self-dual, but becomes anti-self-dual in the $\lambda \to 0$ limit. The gluing angle m depends on a local trivialization, but the set of gluing angles is invariant. This defines the space $\tilde{\mathcal{M}}_{k+1}$ for all y.

In Section 2 we distinguished notationally between radial gauge with respect to 0 and radial gauge with respect to y, calling the background curvature F_0 in the first case and $\tilde{F}_0$ in the second case. Theorem 2.1 discussed making $F_0 + F_{\text{std}}$ reducible at p and q, while Theorem 2.3 discussed making $\tilde{F}_0 + F_{\text{std}}$ reducible at p and q. In this section the background connection is *always* in radial gauge with respect to the gluing point y. With only one case to consider, we always write F_0, never $\tilde{F}_0$.

Note that we do not use the gluing formula found in standard works such as [4]. Traditionally, one takes $A'' = (1 - \beta_2)A_0 + (1 - \beta_1)A_{\text{std}}$, so that the resulting connection is

exactly flat in the annulus with radii $2R_1$ and $\frac{1}{2}R_2$ around y. This makes identifying the bundles on which A_0 and A_{std} live conceptually easier. However, such a procedure makes for a perturbed moduli space on which ν_p and ν_q intersect nontransversely, since $F^-_{A''}$ is reducible, indeed zero, on the entire annulus $2R_1 < r < \frac{1}{2}R_2$. This makes the intersection number effectively impossible to compute.

Instead, we allow the supports of $\beta_1 A_0$ and $\beta_2 A_{\text{std}}$ to overlap as in Taubes' work such as [16]. This allows us to observe the interaction between the background connection and the glued-in instanton. In the Donaldson and Kronheimer [4] method, the interaction only occurs when we go from our explicit approximate ASD connection to the true ASD connection (something we have relatively little analytic control over). In our method, the interaction is seen at the level of the approximate connection A' which we can calculate. Moreover, $F^+_{A'}$ is much smaller than $F^+_{A''}$ (in the L^2 norm), so our method should give a closer approximation to the properties of the true moduli space.

Let $\nu_{t,p}$ (resp. $\nu_{t,q}$) be the set of gluing data (y, λ, m) with $\lambda < KL^\alpha$ for which $F_t^-(p)$ (resp. $F_t^-(q)$) is reducible. If y is not in $\tilde{U}$, then for small enough λ, the connection form near p is exactly A_0. By assumption, F_0 is not reducible at the origin. For small enough L, therefore F_0 is not reducible at p, and $F_t(p) = F_0(p)$ is not reducible. We may therefore assume, without loss of generality, that our gluing point y is always in $\tilde{U}$. Indeed by picking L small enough, we may assume that y is in an arbitrarily small neighborhood of the origin, and therefore that $F_0(y)$ is arbitrarily close to $F_0(0)$. Thus we may take the length scales R_1, R_2, and R_3 to be independent of y (although R_1 and R_2 may depend on λ).

We consider five possibilities:

1. $|p - y| \leq \frac{1}{2}R_1$ (p is in the "interior zone", where $\beta_1 = 0$ and $\beta_2 = 1$),
2. $\frac{1}{2}R_1 < |p - y| < 2R_1$ (p is in the interior "shoulder"),
3. $2R_1 \leq |p - y| \leq \frac{1}{2}R_2$ (p is in the "plateau", where $\beta_1 = \beta_2 = 1$),
4. $\frac{1}{2}R_2 < |p - y| < 2R_2$ (p is in the exterior "shoulder"), and
5. $|p - y| \geq 2R_2$ (p is in the "exterior zone", where $\beta_1 = 1$ and $\beta_2 = 0$).

As in Section 2, we will be identifying curvatures with 3×3 real matrices. The phrase "the second singular value of F", for example, is shorthand for "the second singular value of $Mat(F^-)$".

The problem of Theorem 2.3 was to find $\nu_{1,p}$, $\nu_{1,q}$ and count their intersection points. In that problem condition 3 always applied with $|p - y|^2 \approx \lambda/\sqrt{s_p}$. We will show that $F_t^-(p)$ being reducible with $\lambda < KL^\alpha$ also implies condition 3, and that $\nu_{t,p}$ is a small perturbation of $\nu_{1,p}$. We establish condition 3 by showing that the other conditions lead to contradictions.

We begin by considering condition 1. Where $|x - y| \leq \frac{1}{2}R_1$, $A' = A_{\text{std}}$ has an ASD curvature, so $F_t^- = F_t = tF_{A_0} + F_{\text{std}}$. For F_t to be reducible at p, we need $|F_{\text{std}}(p)| = ts_p$. That is, $\lambda^2 + |p - y|^2 = \lambda/\sqrt{ts_p}$. This quadratic equation has two solutions, one with $\lambda \approx |p - y|^2\sqrt{ts_p}$, the other with $\lambda \approx 1/\sqrt{ts_p}$, but both of them are consistent with condition 1. Since $|p-y| < R_3$, s_p is close to s_0. Since $|p-y|^2 < R_1^2 < 10^{-6}\lambda R_3$, while $\sqrt{ts_0} \leq 1/R_3$, one cannot have $\lambda \approx |p - y|^2\sqrt{ts_p}$. The second solution has $\lambda \approx 1/\sqrt{ts_p} > R_3$, which contradicts $\lambda \ll R_3$. Thus condition 1 is impossible.

If p is in the interior shoulder, we have additional terms to consider:

$$F_t = F_{\text{std}} + (t + (1-t)\beta_1)F_{A_0} + (1-t)[(\beta_1^2 - \beta_1)A_0 \wedge A_0 + \mathrm{d}\beta_1 \wedge A_0 + \beta_1(A_{\text{std}} \wedge A_0 + A_0 \wedge A_{\text{std}})]. \tag{3.4}$$

The ASD part of the terms after F_{std} can be bounded in norm by $1/R_3^2 + 4R_1/R_3^4 + 4/R_3^2 + \lambda^2/R_1^2R_3^2 < 100/R_3^2$, and so the second singular value of F_t is within $100/R_3^2$ of the second singular value of F_{std}. For F_t^- to be reducible, $|F_{\text{std}}|$ can be at most $100/R_3^2$. Thus we need $\lambda/(\lambda^2 + |p-y|^2) < 10/R_3$, which in turn means that either $\lambda > \frac{1}{100}R_3$ or $\lambda < 100R_1^2/R_3$. The first is not allowed as λ is assumed small. The second contradicts the definition of R_1. So condition 2 is also impossible.

If p is in the exterior zone, we have $F_t = F_{A_0} + (1-t)F_{\text{std}}$, so we need $\lambda/(\lambda^2 + |p-y|^2) = \sqrt{s_p/(1-t)}$, or equivalently, $\lambda = (\lambda^2 + |p-y|^2)\sqrt{s_p/(1-t)}$. But $|p-y|^2 > 2R_2^2 > 10^6\lambda/\sqrt{s_0}$, so $\sqrt{s_p/(1-t)}$ always exceeds $\lambda/(\lambda^2 + |p-y|^2)$. So again we have a contradiction.

If p is in the exterior shoulder, we have

$$F_t = (t + (1-t)\beta_2)F_{\text{std}} + F_{A_0} + (1-t)[(\beta_2^2 - \beta_2)A_{\text{std}} \wedge A_{\text{std}} + \mathrm{d}\beta_2 \wedge A_{\text{std}} + \beta_2(A_{\text{std}} \wedge A_0 + A_0 \wedge A_{\text{std}})]. \tag{3.5}$$

The ASD parts of the terms other than F_{A_0} have total magnitude bounded by $\lambda^2/R_2^4 + \lambda^4/R_2^6 + \lambda^2/R_2^4 + \lambda^2/R_2^2R_3^2 < 10\lambda^2/R_2^4 < 10^{-11}s_0 < 10^{-10}s_p$. But F_{A_0} is a distance greater than $\frac{1}{2}s_p$ from the nearest reducible matrix, so $F_t^-(p)$ cannot be reducible.

Thus for all points in $\nu_{t,p}$, condition 3 applies, and here the analysis is relatively simple. The cut-off functions are both 1, so $F(p) = F_{A_0} + F_{\text{std}} + (1-t)(A_{\text{std}} \wedge A_0 + A_0 \wedge A_{\text{std}})$. This last term has magnitude bounded by $\lambda^2/R_1^2R_3^2$, and changes only slightly as (y, λ, m) are varied. It can thus be treated as a perturbation of F_{A_0}. We perturb $\nu_{1,p}$ to $\nu_{t,p}$ iteratively (as in the standard proof of the inverse function theorem): Given a point in $\nu_{1,p}$, compute $(1-t)(a \wedge A_0 + A_0 \wedge a)$, use that to adjust (y, λ, m), compute the change in $(1-t)(A_{\text{std}} \wedge A_0 + A_0 \wedge a)$, adjust (y, λ, m), and so on. The iteration converges geometrically. Similarly, a point in $\nu_{t,p}$ can be perturbed to a point in $\nu_{1,p}$. Of course, the same analysis applies to $\nu_{t,q}$.

Now we consider the number of intersection points of $\nu_{t,p}$ and $\nu_{t,q}$ as a function of t. The only way the intersection number can change is if intersection points appeared or disappeared at the ends of $\nu_{t,p}$ or $\nu_{t,q}$. However, we have shown that such intersection points can only occur when both p and q are in the plateau. In the proof of Theorem 2.3, we saw that, for $\lambda \gg L^2$ but $\lambda \ll 1$ (e.g., $\lambda \sim KL^\alpha$), the points of $\nu_{1,p}$ are bounded away from $\nu_{1,q}$. Since condition 3 applies, for $\lambda \sim KL^\alpha$, $\nu_{t,p}$ and $\nu_{t,q}$ are close to $\nu_{1,p}$ and $\nu_{1,q}$, respectively, and so are bounded away from each other. Thus intersection points between $\nu_{t,p}$ and $\nu_{t,q}$ may not appear from or disappear to the boundary. Thus $\#(\nu_{0,p} \cap \nu_{0,q}) = \#(\nu_{1,p} \cap \nu_{1,q})$. By Theorem 2.3, the latter number is $+6$, regardless of A_0.

4. Computing the Donaldson invariants

In Sections 2 and 3 we saw that, for a fixed generic background connection, there are six ways to glue in a small bubble near p and q so as to make the curvature reducible at p and q. In this section we demonstrate that this is sufficient information to compute the contribution of the boundary region of $\tilde{\mathcal{M}}_{k+1}$ to the simple type condition. For generic choices of representatives (of the classes other than $\mu(p)$ and $\mu(q)$), and for generic choice of the location of the origin of our coordinate system, the boundary region contributes $\frac{6}{64}$ of what is needed for simple type.

We continue the notation of Sections 2 and 3. $\tilde{\mathcal{M}}_{k+1}$ is the perturbed moduli space and $\tilde{U}$ is a fixed ball in N with a Euclidean metric, which we identify with a neighborhood, U, of the origin in $\mathbf{R}^4$. For fixed K, α, L, let $\tilde{\mathcal{M}}^0_{k+1}$ be the subset of $\tilde{\mathcal{M}}_{k+1}$ with $\lambda < KL^\alpha$. Let ω be a formal product of cycles $[\Sigma_1], \ldots, [\Sigma_n] \in H_*(X)$ such that $\deg(\mu(\omega)) = \dim(\mathcal{M}_k)$, so that the Donaldson invariant $D(\omega)$ is computed on the kth moduli space $\mathcal{M}_k$.

We assume that the classes $\{[\Sigma_i]\}$ are represented by smooth submanifolds $\{\Sigma_i\}$ in general position. In particular, a subset of the $\{\Sigma_i\}$ can intersect only if their codimensions add up to 4 or less. Pick tubular neighborhoods $\{\tilde{\Sigma}_i\}$ of $\{\Sigma_i\}$ small enough to have the same property: a subset of the $\{\tilde{\Sigma}_i\}$ can intersect only if the codimensions of the corresponding Σ_i's add up to 4 or less. Similarly, we assume that the $\tilde{\Sigma}_i$'s do not intersect our fixed ball $\tilde{U}$. Choose a geometric representative V_i of each $\mu([\Sigma_i])$ that depends only on the connection restricted to $\tilde{\Sigma}_i$. This may be done for the one-, two-, and three-dimensional cycles as in [4], and for the zero-dimensional Σ's as in [4] or [13]. (This allows us to identify the geometric representative of $\mu([\Sigma])$ on $\mathcal{B}_k$ with the geometric representative of $\mu([\Sigma])$ on $\mathcal{B}_{k+1}$. In each case it is the set of connections whose restriction to $\tilde{\Sigma}$ satisfies a certain condition.) Note that the codimension of V_i in $\mathcal{B}$ is the codimension of Σ_i in N. Let $V_\omega = \cap_i V_i$. V_ω is a geometric representative of $\mu(\omega)$. Generically, V_ω will intersect $\mathcal{M}_k$ at a finite number of points (this number, counted with sign, is the Donaldson invariant $D(\omega)$), and each of these points will exhibit generic behavior. In particular, for each such point A_0, we can assume that $Mat(F_{A_0}(0))$ has three distinct singular values.

Theorem 4.1. *Fix $\tilde{U}$, ω, $K > 0$, and $\alpha \in (0, 2)$. For generic choices of V_ω as described above and for all sufficiently small L, the intersection number of $\tilde{\mathcal{M}}^0_{k+1}$ with $V_\omega \cap \nu_p \cap \nu_q$ is $6D(\omega)$.*

Proof. We need to show that the only way for the boundary region of $\tilde{\mathcal{M}}_{k+1}$ to intersect $V_\omega \cap \nu_p \cap \nu_q$ is if a bubble is pinching off near p and q, while the background connection in $\mathcal{M}_k$ is contributing to $D(\omega)$. We then must demonstrate that, under these circumstances, the problem reduces to the counting problems studied in Sections 2 and 3.

Suppose we have a small bubble centered at a point y that is not in $\tilde{U}$. The point y can lie in at most four of the $\tilde{\Sigma}_i$'s, with the corresponding Σ_i's having total codimension 4 or less. Recall that we are using the explicit formula (3.1), and that outside a neighborhood

of y, the new connection is *identical* to the background connection. For small λ, therefore, the bubble inserted at y has no effect on the connection restricted to the remaining $\tilde{\Sigma}$'s (which we index by j). Therefore for a connection $(A_0, \lambda, y, m) \in \tilde{\mathcal{M}}_{k+1}$ to lie in $\cap_i V_i$, the background connection $A_0 \in \mathcal{M}_k$ must lie in $\cap_j V_j$. However, $\mathcal{M}_k$ has dimension 8 less than $\tilde{\mathcal{M}}_{k+1}$, while $\cap_j V_j$ has dimension at most 4 more than $\cap_i V_i$. Since the dimension of $\mathcal{M}_k$ is less than the codimension of $\cap_j V_j$, $\cap_j V_j \cap \mathcal{M}_k$ is generically empty.

Next we consider the case where a small bubble is centered in $\tilde{U}$. Then $\{\Sigma_j\}$ is equal to all the cycles Σ except at the two points p and q. For small λ, on each of the $\tilde{\Sigma}_j$'s the connection form is equal to the background connection A_0, which must therefore be in $\cap_j V_j \cap \mathcal{M}_k$. However, now the dimension of $\mathcal{M}_k$ and the codimension of $\cap_j V_j$ match. $\cap_j V_j \cap \mathcal{M}_k$ is, by our genericity assumption, a discrete set of points, whose number (counted with sign) is the Donaldson invariant $D(\omega)$. For each of these points, the singular values of $Mat(F_{A_0}(0))$ are distinct.

By Theorem 3.1, for each such background A_0 and for L small enough, there are exactly six values of (λ, y, m) such that $(A_0, \lambda, y, m) \in \tilde{\mathcal{M}}^0_{k+1}$ has reducible curvature at p and q. Furthermore, the intersection numbers for the local problem are all $+1$. Now the orientation of $\tilde{\mathcal{M}}^0_{k+1}$ is the same as that of $\mathcal{M}_k \times U \times (0, KL^\alpha) \times SO(3)$ [3] [Section 3].

Thus the contribution of points (A_0, λ, y, m) to $D([p] \cdot [q] \cdot \omega)$, for fixed A_0, is exactly six times the contribution of A_0 to $D(\omega)$. Summing over the finite set $\{A_0\}$, we get that the contribution of $\tilde{\mathcal{M}}_{k+1}$ to $D([p] \cdot [q] \cdot \omega)$ is $6D(\omega)$. □

5. Differential forms and the μ-map: introduction

Theorem 1.1, restated precisely as Theorem 4.1, is one of the two major results of this paper. It quantifies the contribution of the boundary region of moduli space to the geometric representative computation of the Donaldson invariants that appear in the simple type recursion relation. The remainder of the paper is a proof of Theorem 1.2, which quantifies the contribution of the boundary region to a differential form calculation of the same Donaldson invariants.

In this section we construct a de Rham-theoretic version of Donaldson's μ-map using Chern–Weil theory. Recall that there is a canonical $SO(3)$-bundle $\mathcal{P} \to \mathcal{B}^* \times N$, and that the μ-map is defined by slant product with $-\frac{1}{4}p_1(\mathcal{P})$. Using the L^2 metric one can produce a natural connection on $\mathcal{P}$ with curvature $\mathcal{F}$; see [4, Sections 5.1 and 5.2]. By the Chern–Weil formula one has

$$-\tfrac{1}{4}p_1(\mathcal{P}) = \frac{1}{8\pi^2}\mathrm{tr}(\mathcal{F} \wedge \mathcal{F}) \in \Omega^4(\mathcal{B}^* \times N), \tag{5.1}$$

where the trace comes from the two-dimensional representation of $\mathfrak{so}(3) \cong \mathfrak{su}(2)$. Let us write tangent vectors to $\mathcal{B}^* \times N$ as pairs (α, X) with $\alpha \in T\mathcal{B}^*$ and $X \in TN$, and identify $T_A\mathcal{B}^*$ with $\ker((d^A)^*) \subset \Omega^1(Ad\,P)$. Further, for $\alpha, \beta \in \Omega^1(Ad\,P)$, define $\{\alpha, \beta\} = -\sum_{i=0}^4 [\alpha_i, \beta_i] \in \Omega^0(Ad\,P)$, where the local $Ad\,P$-valued functions α_i, β_i are the components of α, β relative to a local orthonormal basis of T^*N. If A is irreducible, then

$\mathcal{F}((\alpha,0),(\beta,0)) = -2G_0^A\{\alpha,\beta\}$, where G_0^A is the inverse of the covariant Laplacian on $\Omega^0(Ad\,P)$, and hence

$$\begin{aligned}\mu_{\mathrm{d}}(\omega)(\alpha,\beta,\gamma,\rho)|_A &= \int_N \left(\iota_{(\rho,0)}\iota_{(\gamma,0)}\iota_{(\beta,0)}\iota_{(\alpha,0)}\frac{1}{8\pi^2}\mathrm{tr}(\mathcal{F}\wedge\mathcal{F})\right)\omega \\ &= \frac{1}{\pi^2}\int_N \mathrm{tr}(G_0^A\{\alpha,\beta\}G_0^A\{\gamma,\rho\} + G_0^A\{\alpha,\gamma\}G_0^A\{\rho,\beta\} \\ &\quad + G_0^A\{\alpha,\rho\}G_0^A\{\beta,\gamma\})\omega. \end{aligned} \tag{5.2}$$

For our application it is crucial to get the combinatoric factors in (5.2) correct.

If we replace ω by δ_p, a delta-form supported at a point p, the resulting form on $\mathcal{B}^*$ is still de Rham cohomologous to a form obtained using smooth ω. Henceforth we write $\mu_{\mathrm{d}}(p) := \mu_{\mathrm{d}}(\delta_p)$. For any $p \in N$, a 4-form representing $\mu_{\mathrm{d}}(p)$ is given by

$$\begin{aligned}\mu_{\mathrm{d}}(\delta_p)(\alpha,\beta,\gamma,\rho) = \frac{1}{\pi^2}&\mathrm{tr}(G_0^A\{\alpha,\beta\}G_0^A\{\gamma,\rho\} \\ &+ G_0^A\{\alpha,\gamma\}G_0^A\{\rho,\beta\} + G_0^A\{\alpha,\rho\}G_0^A\{\beta,\gamma\})\Big|_p . \end{aligned} \tag{5.3}$$

To make use of (5.3) we need some concrete formulas — with calculable leading terms and small remainders — for $G_0^A\{\alpha,\beta\}$. We can obtain such formulas when A is a concentrated instanton with a "charge-one bubble" and α, β come from infinitesimal changes in the bubble parameters (center, scale, and gluing angle). Tangent vectors of this type span an "approximate tangent space" on which very strong estimates are possible. This space, its relation to the action of the quaternionic affine group on $\mathbf{R}^4$, and its relation to the gluing construction in [4] are central to the proof of Theorem 1.2. In the next section, we define the approximate tangent space precisely and study these relations in detail.

6. Group actions and the approximate tangent space

Let $\mathbf{H}$ denote the quaternions and $\mathbf{H}^*$ the nonzero quaternions. The eight-dimensional approximate tangent space we define later is obtained by an "almost-action" of $\mathbf{H}^*\times\mathbf{H} \cong \mathbf{R}_+\times SU(2)\times\mathbf{R}^4$ on $\mathcal{B}$ induced by an almost-action on P (what "almost-action" means is explained below). Essentially, we lift from N to P cut-off versions of translations, dilations, and "self-dual rotations" in a gauge-invariant way.

To make this more precise, let X be a vector field on N, and A a connection on P. The pair (X, A) defines a flow on P obtained by lifting X A-horizontally to P. We thereby obtain from X the "canonical flow of X on $\mathcal{A}$" with associated vector field $A \mapsto \tilde{X} := \iota_X F_A \in \Omega^1(Ad\,P) \cong T_A\mathcal{A}$ (see [8, Proposition 4.3]). The canonical flow is invariant under the gauge group, hence it descends to $\mathcal{B}$. Moreover, any two lifts to P of a diffeomorphism of N differ by a gauge transformation, and hence given an action on N by any connected Lie group G on N, the canonical flow integrates to a well-defined action of G on $\mathcal{B}$, though in general not on $\mathcal{A}$. Of interest to us later will be the comparison of the canonical lift to that obtained by lifting X horizontally with respect to a reference connection A_0. In that case

the difference between the two tangent vectors in $T_A\mathcal{A}$ induced by the two flows is $d^A u$, where $u = \iota_X(A - A_0)$.

Now let G be a Lie group acting from the left on N. Suppose that for each $g \in G$, the action Φ_g of g on N lifts to a bundle map $\tilde{\Phi}_g : P \to P$; if G is connected, we can obtain such lifts by using the canonical flow. (We do not require the lifts to piece together to a G-action.) For later purposes we will need to calculate the differential of the induced action of G on $\mathcal{B}$ at any $g \in G$. This is not difficult, but it is easy to confuse the roles of g and g^{-1} in this calculation, and this mistake would be fatal for our application.

For each connection $A \in \mathcal{A}$, let $\Theta_A \in \Omega^1(P, \mathfrak{su}(2))$ denote the connection form of A. Given a lift $\tilde{\phi}_g$ as above, define $g \cdot A$ to be the connection with connection form $(\tilde{\Phi}_g^{-1})^* \Theta_A$. If the lifts piece together into an action of G on P (necessarily a left action), then $(g, A) \to g \cdot A$ defines a left action of G on $\mathcal{A}$. $\tilde{\Phi}_{g_1 g_2}$ and $\tilde{\Phi}_{g_1} \circ \tilde{\Phi}_{g_2}$ are gauge-equivalent, since both are lifts of $\Phi_{g_1 g_2}$, so an element-wise liftable G-action on N *always* induces a G-action on $\mathcal{B}$, whether or not it induces one on P.

Now fix $[A_0] \in \mathcal{B}$ and define $\overline{\rho} : G \to \mathcal{B}$ by $\overline{\rho(g)} = [g \cdot A_0]$. This is well defined and is independent of the choice of lifts. On a small enough neighborhood U of any $g \in G$, we can always choose the $\tilde{\Phi}_h$ to vary smoothly with h, so that on U the map $\overline{\rho}$ factors through a smooth map $\rho : G \to \mathcal{A}$ defined by $\rho(g) = g \cdot A_0$. Let $v = (\mathrm{d}/\mathrm{d}t) g_t|_{t=0} \in T_g G$ and write $v = R_{g*} w$, where $w \in T_e G = \mathfrak{g}$. Then

$$\rho_{*_g} v = \frac{\mathrm{d}}{\mathrm{d}t}((\exp(tw)g) \cdot A_0)\bigg|_{t=0}. \tag{6.1}$$

But $\tilde{\Phi}_{\exp(tw)g} = \gamma(t) \circ \tilde{\Phi}_{\exp(tw)} \circ \tilde{\Phi}_g$ for some gauge transformation $\gamma(t)$ varying smoothly in t, and hence $(\exp(tw)g) \cdot A_0 = ((\exp(tw)) \cdot g \cdot A_0) \cdot \gamma(t))$. Thus

$$\begin{aligned} \rho_{*_g} v &= \frac{\mathrm{d}}{\mathrm{d}t}(\exp(tw) \cdot g \cdot A_0)\bigg|_{t=0} \bmod \mathrm{Im}(d_{g \cdot A_0}) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}((\tilde{\Phi}^{-1}_{\exp(tw)})^* \omega_{g \cdot A_0})\bigg|_{t=0} \bmod \mathrm{Im}(d_{g \cdot A_0}). \end{aligned} \tag{6.2}$$

Let $\hat{w} \in \Gamma(TN)$ and $\hat{w}_P \in \Gamma(TP)$ be the vector fields on N and P induced by the infinitesimal action of w. Then

$$\frac{\mathrm{d}}{\mathrm{d}t}((\tilde{\Phi}^{-1}_{\exp(tw)})^* \omega_{g \cdot A_0})\bigg|_{t=0} = -\mathcal{L}_{\hat{w}_P} \omega_{g \cdot A_0} = -\iota_{\hat{w}} F_{g \cdot A_0} \bmod \mathrm{Im}(d_{g \cdot A_0}) \tag{6.3}$$

(if $\tilde{\Phi}$ is defined by the canonical flow, then "$\bmod \mathrm{Im}(d_{g \cdot A_0})$" can be erased in this line). Note that v directly defines a vector field on N by $\hat{v}|_{\Phi_g(x)} = (\mathrm{d}/\mathrm{d}t)(\Phi_{g_t}(x))|_{t=0}$. Since we can take $g_t = \exp(tw)g$, it immediately follows that $\hat{v}|_{\Phi_g(x)} = \hat{w}|_{\Phi_g(x)}$ for all $x \in N$, so the vector fields $\hat{v}$ and $\hat{w}$ are the same. Hence

$$\rho_{*_g} v = -\iota_{\hat{v}} F_{g \cdot A_0} \bmod \mathrm{Im}(d_{g \cdot A_0}). \tag{6.4}$$

Thus if we identify $T_{[A]}\mathcal{B}$ with $\ker((d^A)^*) \subset \Omega_1(Ad\, P)$, then

$$\overline{\rho}_{*_g} X = -\pi'_A \iota_{\hat{X}} F_{g \cdot A_0}, \tag{6.5}$$

where $\pi'_A : \Omega^1(Ad\,P) \to \ker((d^A)^*)$ is the L^2-orthogonal projection. (Here and below, for notational convenience we do not distinguish between a tangent vector to $\mathcal{B}$ of $\mathcal{M}$ at $[A]$, literally a gauge-invariant section of a certain vector bundle over the gauge-orbit through A, and the representative of this section at A.)

We would like to apply these ideas to the situation of a local action of $\mathbf{H}^* \times \mathbf{H}$ on a neighborhood of a point in N. Given a concentrated connection A, with scale $\lambda = \lambda(A)$ and center point p_A, fix a positively oriented normal coordinate system centered at p_A.[3] Near p_A it makes sense to speak of the translation, dilation, and rotation vector fields. These are determined invariantly by data $(\mathbf{b}, a, \alpha) \in T_{p_A}M \oplus \mathbf{R} \oplus \Lambda^2(TN)$ by setting

$$\hat{X}_{(\mathbf{b},a,\alpha)} = b^j \frac{\partial}{\partial x^j} + (\sqrt{2}\lambda^{-1}) \left(a x^i \frac{\partial}{\partial x^i} + \alpha_{ij} x^i \frac{\partial}{\partial x^j} \right), \tag{6.6}$$

where $\{x^i\}$ are normal coordinates centered at p_A and b^j, α_{ij} are the associated components of $\mathbf{b}$, α. We include the normalization factor $(\sqrt{2}\lambda)^{-1}$ to arrange $\|\iota_{\hat{X}} F_A\|_2 \approx$ const. (independent of λ); see Proposition 6.4. We call $\alpha_{ij} x^i (\partial/\partial x^j)$ a self-dual/anti-self-dual rotation vector field if $\alpha_{ij}\, \mathrm{d}x^i \wedge \mathrm{d}x^j$ is an SD/ASD 2-form at p_A.

Since $\hat{X}$ are only defined locally, we extend them to N by cutting them off outside a small ball. For this purpose with $\hat{X}$ as above, we define $X = \beta \hat{X}$, where β is a cut-off of scale

$$\epsilon = 4 n_0 \lambda^{1/2}. \tag{6.7}$$

Here n_0 is a constant taken large enough to ensure that β can be used in the gluing constructions of Donaldson and Kronheimer [4], but for all of our other applications we can ignore n_0. For convenience we take $\beta = \beta_{\mathrm{std}}(r_A/\epsilon)$, where r_A is the distance to p_A and β_{std} is a cut-off function with support in $[0, 2]$, identically 1 on $[0, 1]$. (These cut-offs, which will be used for the rest of this paper, are different from the ones in Section 3.)

We define

$$\mathfrak{h}_A = \{ X_{(\mathbf{b},a,\alpha)} = \beta \hat{X}_{(\mathbf{b},a,\alpha)} \in \Gamma(TN) \,|\, (\mathbf{b}, a, \alpha) \in T_{p_A}M \oplus \mathbf{R} \oplus \Lambda^2_+(TN) \}. \tag{6.8}$$

It is worthwhile to observe that the definition of (A)SD rotation vector fields is necessarily a local definition, since globally a nontrivial exact 2-form cannot be SD or ASD on an orientable compact manifold. In fact on S^4, rotations that are SD at one pole are ASD at the other. This is most easily seen by using stereographic projection to identify $S^4 - \{\infty\}$ with $\mathbf{R}^4$, then with $\mathbf{H}$. Left-multiplication by unit quaternions induces SD rotations near 0, while right-multiplication induces ASD rotations near 0. But coordinates near ∞ on S^4 are related to those near 0 by quaternionic inversion (the orientation-preserving map $x \mapsto x^{-1}$), which interchanges the roles of left- and right-multiplication.

When A is an ASD, we make the following definition.

[3] The precise definitions of λ and p_A are not important here. There are several definitions in the literature leading to some arbitrariness in the definition of "near", "bubble", etc. In all instances in which the differences among these definitions have been carefully analyzed, it has been found that these differences do not affect the estimates we need in any material way (cf. [9, Section 5]). We will simply assume in this paper that the same is true here, and will freely quote results proved using different definitions as if they had been proved using consistent definitions of scale and center.

Definition 6.1. The *approximate tangent space* $\mathcal{H}_A$ at A to the moduli space is the space

$$\{\tilde{X}_A := \iota_X F_A | X \in \mathfrak{h}_A\}. \tag{6.9}$$

We usually write simply $\tilde{X}$ and leave the A-dependence implicit.

To justify this terminology, we consider the action induced by such X on an ASD connection. Since the X's are nearly conformal vector fields, one expects the induced flow to map an ASD connection to a nearly ASD connection. Proposition 6.4 shows that this is the case, and more — but first we need a definition and a lemma.

Definition 6.2. Given κ, ν, $\lambda_0 > 0$, let $\mathcal{M}^{\kappa,\nu}_{k+1,\lambda_0} \subset \mathcal{M}'_{k+1,\lambda_0}$ denote the subset of instantons $[A]$ obeying the conditions

1. the first eigenvalues of the Laplacians $(d^A)^* d^A$ on 0-forms, $d^A_+ (d^A_+)^*$ on SD 2-forms are both greater than ν, and
2. for all $p \in N$,

$$|F_A(p)| \leq \frac{C\lambda^2}{\lambda^2 + r_A(p)^2)^2 + \kappa}, \tag{6.10}$$

where F_A is the curvature of A, $r_A(p) = \mathrm{dist}(p_A, p)$, and λ and p_A are the scale and center point of A, respectively.

The pointwise bound (6.10) essentially says that $|F_A|$ is bounded by the curvature of a standard instanton plus a contribution κ from a background connection. At small distances from p_A, the latter term is negligible, but far from p_A, the background term dominates.

When dealing with estimates for the approximate tangent space, one must decide at what scale ϵ to cut off the vector fields in $\mathfrak{h}_A$. If one takes ϵ to be too small, the derivatives of the cut-off function become inconveniently large, while if one takes ϵ to be too large, the contribution from the background connection swamps the contribution from the concentrated curvature. If we require that ϵ scale as a power of λ, we get the optimal balance between these undesirable features only if $\epsilon \sim \lambda^{1/2}$. Earlier we chose $\epsilon = 4n_0\lambda^{1/2}$, and we now take n_0 to be large enough so that in the gluing construction of Donaldson and Kronheimer [4] one is assured of landing in the domain of Taubes' contracting mapping argument. Once n_0 has been so chosen, it (like κ) is simply another ignorable constant for the computations we need in this section. In particular, note that on the ball $B(p_A, \epsilon)$ or radius ϵ centered at p_A, we have $\lambda^2/(\lambda^2 + r_A^2)^2 \geq (\lambda_0 + 64n_0^2)^{-2} \geq \text{const.}\,\kappa$. Hence, with a new constant $c = c(\kappa)$,

$$|F_A| \leq c\lambda^2/(\lambda^2 + r_A^2)^2 \quad \text{on supp}(\beta). \tag{6.11}$$

This enormously simplifies our computations.

The next lemma shows that we can always arrange the fiber Z to lie in some $\mathcal{M}^{\kappa,\nu}_{k+1,\lambda_0}$.

Lemma 6.3. *Given small λ_0, let $T : \mathcal{M}'_{k+1,\lambda_0} \to \mathcal{M}_k$ be the projection sending a concentrated connection to a "background" connection. Let $[A_0] \in \mathcal{M}'_k$ and assume that the first eigenvalues of the Laplacians $d^*_{A_0} d_{A_0}$, $d^+_{A_0}(d^+_{A_0})^*$ on Ad P-valued 0-forms and SD*

2-forms, respectively, are positive. Then there exists $\lambda_0 > 0$, $\nu > 0$, $\kappa > 0$, $C > 0$, *and a neighborhood* $\mathcal{U}$ *of* $[A_0]$ *in* $\mathcal{M}_{k+1}$ *such that* $T^{-1}(\mathcal{U}) \subset \mathcal{M}^{\kappa,\nu}_{k+1,\lambda_0}$.

Proof. That condition 1 in the definition of $\mathcal{M}^{\kappa,\nu}_{k+1,\lambda_0}$ can be satisfied follows from the proof of Lemma 7.1.24 in [4]; that condition 2 can be satisfied follows from modifying several ideas in [9, Definition 4.1, Lemma 4.3d, and Proposition 4.4]. □

Henceforth we will always assume that instantons $[A]$ lie in a fixed $\mathcal{M}^{\kappa,\nu}$. For such connections we have the following.

Proposition 6.4. *Fix* κ, ν. *Let* $\pi_A : \Omega^1(Ad\,P) \to H^1_A := \ker((d^A)^*) \cap \ker(d^A_+)$ (*naturally identified with the tangent space* $T_{[A]}\mathcal{M}$) *be the* L^2*-orthogonal projection. For all sufficiently small, positive* δ, *there exist* c, $\epsilon_1(\lambda)$ *such that if* $A \in \mathcal{M}^{\kappa,\nu}_{k+1,\lambda_0}$, *then*

$$|\|\tilde{X}_{(\mathbf{b},a,\alpha)}\|_2^2 - 8\pi^2(|\mathbf{b}|^2 + |a|^2 + |\alpha|^2)| \le \epsilon_1(\lambda)(|\mathbf{b}|^2 + |a|^2 + |\alpha|^2), \tag{6.12}$$

where $\epsilon_1(\lambda) \to 0$ *as* $\lambda \to 0$, *and*

$$\|\tilde{X} - \pi_A\tilde{X}\|_2 \le c\lambda^{\delta}(|\mathbf{b}|\lambda + (|a| + |\alpha|)\lambda^{1/2}). \tag{6.13}$$

Proof. The proof of the first statement is similar to that of Ref. [8, Proposition 3.6]; we omit the details. We prove the second statement later as Proposition 10.8(b). □

Thus by taking λ small enough, we can ensure that $\pi_A : \mathcal{H}_A \to H^1_A$ is injective. Let $O_0 \subset \mathfrak{h}_A$ be an open neighborhood of zero. For $X \in \mathfrak{h}_A$, let A^X denote the connection that results from acting on A by the canonical flow of X for unit time, and let $O_A = \{A^X \mid X \in O_0\}$. Proposition 6.4 has two implications once we take O_0 small enough. First, O_A lies in a neighborhood of the ASD connections on which Taubes' contracting-mapping argument lets us "project" the image of A to an ASD connection. Second, by the implicit function theorem, the image of O_A in $\mathcal{M}_{k+1}$ is an eight-dimensional submanifold of $\mathcal{M}_{k+1}$.

The quantity $\tilde{X} - \pi_A\tilde{X}$ will be central to the definition of the remainder terms in $\mu_{\mathrm{loc}}(p)$ and to the analysis in Section 10. We define

$$\xi_X = \tilde{X} - \pi_A\tilde{X} = d^A G^A_0 (d^A)^*\tilde{X} + (d^A_+)^* G^A_+ d^A_+ \tilde{X}. \tag{6.14}$$

Here G^A_0 and G^A_+ are the inverses of the Laplacians $(d^A)^*d^A$ and $d^A(d^A_+)^*$ on $\Omega^0(Ad\,P)$ and $\Omega^2_+(Ad\,P)$, respectively.

We make three observations here. First, if A is ASD, $\|(d^A_+)^* G^A_+ d^A_+ \tilde{X}\|_2/\|\tilde{X}\|_2$ is small for *any* rotation vector field, not just SD ones. This is to be expected since any rotation vector field is an approximate isometry and hence should approximately preserve anti-self-duality. However, $\|d^A G^A_0 (d^A)^*\tilde{X}\|_2/\|\tilde{X}\|_2$ is small only for the rotation vector fields of duality opposite to that of the connection. Second, to deduce from this smallness that the parameter space injects (locally) into $\mathcal{B}$, one must know that the first eigenvalue of the Laplacians on 0-forms does not tend to zero as λ tends to zero as it will if the "background" connection is flat (or merely reducible). Indeed on $\mathcal{M}_1(S^4)$, all rotation vector fields (not cut off), lifted

as above, preserve the standard instanton. (On $\mathcal{M}_1(\mathbf{R}^4)$, if one writes instantons in the usual gauge and instead lifts rotations using the flat connection, then ASD rotations preserve the standard ASD instantons centered at the origin, while SD vector fields induce the effect of a gauge transformation.) Third, because of the cut-off β, $\mathfrak{h}_A$ is not a Lie subalgebra of $\Gamma(TN)$, although in some sense it is close to being one. Thus, while intuitively $\mathfrak{h}_A$ is associated with the Lie algebra of an eight-dimensional group of translations, dilations, and rotations, O_A is not quite the orbit of an eight-dimensional local Lie group, hence the term "almost-action".

We will return to this point at the end of this section, but first we wish to relate O_A to the gluing construction in [4]. The fibration of a region in $\mathcal{M}'_{k+1,\lambda_0}$ over $\mathcal{M}'_k$ is usually viewed in terms of center point, scale, and gluing parameter. We claim that on an infinitesimal level, these are essentially the eight parameters used to define the approximate tangent space. Indeed [9] [Section 5] it was shown that lifts using the translation and dilation vector fields do correspond to infinitesimal changes in center point and scale up to an error that is essentially $O(\lambda)$. (Ref. [9] dealt only with $\mathcal{M}_1$, but under a suitable definition of "concentrated", the same argument works more generally.) It remains to identify our action of $SO(3)$ (the SD rotations) with the "gluing parameters" of the construction in [4, Section 7.2]. As both constructions are noncanonical we content ourselves with a somewhat heuristic correspondence.

Instantons in the subspace $\mathcal{M}^{\kappa,\nu}_{k+1,\lambda_0}$ have a single "charge-one bubble" and are otherwise not concentrated. For any such ASD reference connection $A_0 = A$, there exists a gauge over the ball $B(p_A, K\lambda)$ such that after pulling back to $B(0, K\lambda) \subset \mathbf{R}^4$ by a positively oriented normal coordinate system $\{x^i\}$, the connection form is close to A^{std}_λ, the standard instanton on $\mathbf{R}^4$ of scale λ and center the origin (see [4, Section 8.2.1]). Here $K > 1$ is any fixed number and "close" means that after dilating by λ, the two connections are C^2-close on $B(0, K) \subset \mathbf{R}^4$; the undilated connections satisfy $|\nabla^j(A - A^{\text{std}}_\lambda)| \leq \epsilon_1 \lambda^{-1-j}$, $0 \leq j \leq 2$, where by taking λ_0 small enough we can take ϵ_1 as small as we please. After a choice of normal coordinate system, the identification $\mathbf{R}^4 \cong \mathbf{H}$, and an identification of $SU(2)$ with the unit quaternions, the connection form for A^{std}_λ on our ball is

$$\omega_0 = \frac{\operatorname{Im}(\bar{x}\, \mathrm{d}x)}{\lambda^2 + |x|^2}. \tag{6.15}$$

For integers $j = 1, \ldots, 10$, define the sets $U_j = B(p_A, jn_0\lambda^{1/2})$ and $V_j = N - U_j$. Also let U_∞ denote the annulus $U_{10} \cap V_1$ and let Ω denote the smaller annulus $U_9 \cap V_2$. We choose gauges s_0, s_∞ (local sections of P) over U_{10}, U_∞, respectively, such that the transition function between U_{10} and U_∞ is $g_{0\infty}(x) := \overline{x}/|x|$ (i.e. $s_\infty = s_0 g_{0\infty}$); furthermore we take s_0 to be the radial gauge for A with respect to p_A with which (6.15) is written. The connection form for A^{std}_λ with respect to s_∞ on U_∞ is then

$$\omega_\infty = \frac{\lambda^2 \operatorname{Im}(x\, \overline{\mathrm{d}x})}{|x|^2(\lambda^2 + |x|^2)}. \tag{6.16}$$

To make contact with the construction in [4], we will pretend that on the ball U_8, our connection A is exactly standard (so that the connection form relative to s_0 on U_8 is (6.15)). Let $\tilde{\beta}$ be a function that is identically 1 on $N - U_\infty$ and identically 0 on Ω with $|\nabla\tilde{\beta}| \leq$

$cn_0^{-1}\lambda^{-1/2}$. (Note that the "interior" part of the support of $\tilde{\beta}$ occurs where $\beta \equiv 1$.) On U_∞, let ω'_∞ denote the connection form $\tilde{\beta}\omega'$. We then define a new connection A' on P by declaring the connection form for A' in the gauge s_∞ over U_∞ to be ω'_∞, and declaring $A' = A$ on $N - U_\infty$. We think of $A'|_{V_8}$ as a cut-off "background connection". In fact, we can define a bundle P_k of Pontryagin index k by replacing the transition function $g_{0\infty}$ by the identity; $A'|_{V_8}$ extends to a connection on P_k that is flat near p_A. Note that over Ω the connection A' is flat; its connection form there, relative to s_0, is $g_{0\infty}\,\mathrm{d}g_{0\infty}^{-1} = \mathrm{Im}(\overline{x}\,\mathrm{d}x)/|x|^2$.

Our choice of normal coordinates and identification $\mathbf{R}^4 \cong \mathbf{H}$ induces a Lie algebra isomorphism $\theta : \Lambda^2_+ T^*_{p_A} M \to \mathfrak{su}(2) = \mathrm{Im}(\mathbf{H})$, mapping the standard basis of $\Lambda^2_+ T_0^*\mathbf{R}^4$ to $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$. (Alternatively, θ^{-1} is given by mapping $v \in \mathfrak{su}(2)$ first to the vector field induced by quaternionic left-multiplication on $\mathbf{H}$, then to the 2-form obtained by lowering an index using the metric.) Let $v = \theta(\alpha)$ and assume $|v|$ is not too large. We consider the canonical flow of $X = \beta\alpha_{ij}x^i(\partial/\partial x^j)$ acting on the cut-off connection A'. After integrating the flow for time 1, the action on the base is $x \mapsto \phi(x) = h_1(r)x$, where $r = |x|$ and $h_1(r) = \exp(\beta v)$. Let A'_v be the connection determined by this integrated canonical flow.

An alternative flow, the "s_0-flat" flow, is obtained by lifting X to $P|_{U_8}$ using the flat connection determined by s_0, and extending this flow to the complement of U_8 by the identity (since X is supported in U_8). If we integrate the s_0-flat flow for time 1, the resulting connection form on Ω with respect to s_0 is

$$\omega'_v = \frac{\mathrm{Im}(\overline{h_1(r)x}\,\mathrm{d}(h_1(r)x))}{|x|^2} = g_{0\infty}h_1^{-1}\,\mathrm{d}h_1 g_{0\infty}^{-1} + g_{0\infty}\,\mathrm{d}g_{0\infty}^{-1}. \tag{6.17}$$

By our earlier comments, the connections resulting from the canonical flow and the s_0-flat flow are gauge equivalent (and in fact are equal outside U_8). Thus A'_v is gauge-equivalent to a connection A''_v equaling A' on $N - U_\infty$, and whose connection form in Ω (in the gauge s_0) is (6.17).

We claim that the connection A''_v is the one constructed in [4, p. 296]. The latter essentially begins with the connection A' (thought of as a cut-off connection on $P_k|_{V_1}$ glued to a connection on a $k = 1$-bundle $P_1|_{U_{10}}$) and modifies it on U_∞ as follows. Let $h_1(r)$ be as above, let $h_2(r) = \exp(-(1-\beta)v)$ and consider the two gauge transformations $\tilde{h}_1$, $\tilde{h}_2$ over U_∞ given by $\tilde{h}_i(s_\infty(x)) = s_\infty(x)h_i(r)$. The gauge transformation $\tilde{h}_1$ does not extend to all of P (unless $\exp v = \pm 1$), but it does extend to the bundle P_k defined earlier, and for $r \le 4n_0\lambda^{1/2}$ changes the trivialization s_∞ (extended to P_k) by the constant $\exp v$. Similarly the gauge transformation $\tilde{h}_2$ extends to $P|_{U_{10}}$, changing the trivialization s_∞ for $r \ge 8n_0\lambda^{1/2}$ by $\exp(-v)$. Because $h_1^{-1}\,\mathrm{d}h_1 = h_2^{-1}\,\mathrm{d}h_2$, the two gauge transformations have the same effect on the flat connection $A'|_\Omega$. Therefore we can define a new connection A_v^{DK} by

$$A_v^{\mathrm{DK}} = \begin{cases} \tilde{h}_2(A') \text{ on } U_9, \\ \tilde{h}_1(A') \text{ on } V_2. \end{cases} \tag{6.18}$$

The connection form for A_v^{DK} with respect to s_∞ on Ω is $h_1^{-1}\,\mathrm{d}h_1 = h_2^{-1}\,\mathrm{d}h_2$, so with respect to s_0 the connection form is precisely (6.17). Thus A''_v and A_v^{DK} coincide on Ω.

Since $\tilde{h}_1 \equiv 1$ on V_9 we have $A_v^{\mathrm{DK}} = A'$ on this region, and since $X \equiv 0$ on V_9 we have $A_v'' = A'$ there as well. Thus $A_v'' = A^{\mathrm{DK}}$ on V_2. It remains to consider only U_2. On this ball, a computation shows that the connection form for A' relative to s_0 is

$$\omega_0' = \frac{\mathrm{Im}(\overline{x}\,\mathrm{d}x)}{|x|^2}\left(1 - \tilde{\beta}(r)\frac{\lambda^2}{\lambda^2 + |x|^2}\right). \tag{6.19}$$

On U_2, we have $h_2 \equiv 1$, so the connection form for A_v^{DK} remains ω_0'. But the connection A' is also preserved by the "s_0-flat" flow of X; replacing x by $h_1(r)x$ in (6.19) does not change ω_0, since h_1 is constant on U_2. Therefore $A_v'' = A_v^{\mathrm{DK}}$ over all of N.

Now let A_v be the connection obtained from applying the canonical flow of X for time 1 to A (rather than to A'). The preceding shows that up to gauge equivalence, when $|v|$ is not too large, the only differences between A_v and A_v^{DK} arise from the facts that (i) A is only approximately standard on a small ball $B(p_A, K\lambda)$ rather than exactly standard on the larger ball $B(p_A, 10n_0\lambda^{1/2})$, and (ii) we do not cut off A_v before applying the flow.

It should also be noted that since the subspace $\mathfrak{h}_A^{\mathrm{rot}}$ corresponding to the SD rotation vector fields is not closed under Lie bracket, if we let $\mathfrak{h}_A^{\mathrm{rot}}$ act on A by the canonical flow for time 1, we should not expect to get a closed "orbit". But the construction in [4] shows that the space of gluing parameters is a copy of $SO(3)$.

To address this discrepancy, first note that if N were $\mathbf{R}^4$ we could dispense with the cut-offs in the definition of $\mathfrak{h}_A$. The vector fields would be globally defined, and would generate a Lie algebra exponentiating to the group of motions of $\mathbf{R}^4$,

$$\{x \mapsto ax + b \,|\, (a, b) \in \mathbf{H}^* \times \mathbf{H}\}. \tag{6.20}$$

The stabilizer of the origin would be $\mathbf{H}^* \times \{0\}$, and if the initial connection A were standard, the set of connections generated by letting $SU(2) \subset \mathbf{H}^*$ act via the canonical flow would give two copies of the space obtained by the construction in [4], as $(a, 0)$ and $(-a, 0)$ yield the same connection. (Alternatively, if $v \neq 0$ is small enough, the connections $A' = A^{\mathrm{DK}}$ and $A_v'' = A_v^{\mathrm{DK}}$ are gauge-inequivalent, because the gauge transformation $\tilde{h}_1$ defined earlier — which always extends to $P|_{N-\{p_A\}}$ — extends to P if and only if $\exp v = \pm 1$.) From our earlier discussion the action of (a, b) on the standard instanton is given by pulling back the connection form by the *inverse* of (a, b), which results in a connection of scale $|a|$ and center b (cf. (8.23)). The unit quaternion $a/|a|$ corresponds to gluing angle doubly parametrized.

Intuitively then, we have the following picture. Fix a reference connection $A = A_0 \in \mathcal{M}_{k+1,\lambda_0}^{\kappa,\nu}$. Let $B \subset \mathfrak{su}(2)$ be the ball centered at the origin that is carried diffeomorphically to $SU(2) - \{-1\}$ by the exponential map, and let $B' \subset \mathfrak{h}_A^{\mathrm{rot}}$ be the corresponding set of SD rotation vector fields. If we let the canonical flow of elements in B' act for time 1 on $[A_0]$, we obtain a space that (for purposes of integrating reasonably behaved differential forms) approximates two copies of the fiber Z_{DK}. This correspondence becomes sharper as $\lambda_0 \to 0$: as we take the limit and rescale the (local) metric and normal coordinates correspondingly, the failure of $\mathfrak{h}_A$ to close under Lie bracket disappears on any ball of fixed rescaled size. Furthermore, because the rescaled metric becomes flat, the limiting space of vector fields $\mathfrak{h}_A$ is the same whether we center the rotations and dilations at p_A or at p. Thus the limiting action of $\mathbf{H}^* \times \mathbf{H}$ above appears to generate an immersed manifold that we can

treat "homologically" as two copies of Z_{DK}. This discussion motivates the assumptions we make on Z in the next section.

7. The fiber Z

For our purpose we need only consider one fiber $Z = Z_{\lambda_0}$ of the projection $\mathcal{M}'_{k+1,\lambda_0} \to \mathcal{M}'_k$; we do not need to construct the whole fibration. We will assume that Z has the following five properties. The first three are known to be satisfied by Z_{DK}, so the key assumptions are really the last two, which require the tangent spaces of Z and Z_{DK} to be close in various norms. The assumptions can almost certainly be weakened from those below at the cost of considerably more technical work.

(Z1) Z fibers over N via the projection $Z \to N$ sending a concentrated connection to its center. Given $U \subset N$ we let $Z|_U$ denote the inverse image of U under the projection. We assume that N can be covered by a finite number of normal coordinate charts U_i (which we may take to be geodesic balls) such that for each i there is a two-to-one fiber-preserving covering map $\overline{\rho}_i : \mathbf{H}^*_{\lambda_0} \times U_i \to Z|_{U_i}$ having additional properties listed below. Here $\mathbf{H}^*_{\lambda_0} = \{a \in \mathbf{H}^* \mid |a| < \lambda_0\} \cong (0, \lambda_0) \times SU(2)$, where the isomorphism is $a \mapsto (|a|, a/|a|)$. (Note that the center point and scale maps are defined globally on Z; it is only for the purpose of handling gluing parameters that we need to chop up N.)

In general a normal coordinate system $\{x^j\}$ on U_i determines an identification between U_i and a ball in $\mathbf{H}$ centered at the origin, and hence a local action of $\mathbf{H}^*_{\lambda_0} \times \mathbf{H}$ on U_i given by $((a, b), x) \mapsto ax + b$. We assume that on each U_i there is a positively oriented normal coordinate system $\{x^j\}$ on U_i such that $\overline{\rho}_i$ is approximately given by the induced canonical flow of this $\mathbf{H}^*_{\lambda_0} \times \mathbf{H}$-action, based at the standard instanton on $\mathbf{R}^4$, in the sense that (Z2)–(Z5) below are true. From [1, Section 3], the orientation induced on the fiber Z by the standard orientation of $\mathbf{H} \times \mathbf{H}$ as a complex vector space is then compatible with the standard orientations of $\mathcal{M}_{k+1}$ and $\mathcal{M}_k$ (i.e. the orientation of $\mathcal{M}'_{k+1,\lambda_0}$ is the product of the orientation of Z and the pullback of the orientation of $\mathcal{M}_k$). These are the orientations used in (1.8).

(Z2) We assume that for each i, the scale and center point of $A = \overline{\rho}_i(a, b)$ are $\lambda(A) = |a|$ and $p_A = b$ (in quaternionic normal coordinates), respectively.

(Z3) Given i, let $[A_{a,b}] = \overline{\rho}_i(a, b)$ and let $F_{a,b} = F_{A_{a,b}}$. We assume that for any $K > 0$, on the ball $B(p_A, K\lambda(A))$ we have

$$\left| |F_{a,b}| - \frac{\sqrt{48}|a|^2}{(|a|^2 + |x - b|^2)^2} \right| \leq \epsilon_1(\lambda)\lambda^{-2}, \tag{7.1}$$

where $\epsilon_1(\lambda) \to 0$ as $\lambda \to 0$.

(Z4) Let B be the component of $\exp^{-1}(SU(2) - \{-1\})$ containing 0. A tangent vector $v \in T_{(a,b)}(\mathbf{H}^*_{\lambda_0} \times B)$ gives rise to a vector field on a neighborhood of $b \in B$ that determines an element $X_v \in \mathfrak{h}_A$. Writing $\tilde{X}_v = \iota_{X_v} F_A$, we require that

$$\|\overline{\rho}_* v + \tilde{X}_v\|_{L^2} \leq \epsilon_2(\lambda)|v|, \tag{7.2}$$

where $\epsilon_2(\lambda) \to 0$ as $\lambda \to 0$. Observe that because of (6.12), we can alternatively write (7.2) as $\|\overline{\rho}_* v - (-\pi_A \tilde{X}_v)\|_{L^2} \leq \epsilon_2(\lambda)|v|$; cf. (6.5).

(Z5) Letting $\xi'_v = \overline{\rho}_* v + \tilde{X}_v$ and $\xi_v = \tilde{X}_v - \pi^A \tilde{X}_v$, we further require that ξ'_v satisfy the same weighted L^4 bounds as ξ_v given in Proposition 9.1 (Eqs. (10.74) and (10.75)), and the pointwise bound (10.72).

If not for (Z4) and (Z5), we would not need to assume (Z1)–(Z3). By itself, (Z1) follows from the description of the ends of moduli space in [4, Sections 7.2 and 8.2]; we simply take the local diffeomorphism $(0, \lambda_0) \times SO(3) \times U_i \cong Z|_{U_i}$, and pre-compose with the covering map $SU(2) \to SO(3)$. Similarly, (Z2) and (Z3) follow from [4, Section 8.2.1].

What is not clear is whether the construction in [4] yields a fiber whose tangent space at $[A]$ is sufficiently close to $\pi_A(\mathcal{H}_A)$ in the norms required for our analysis. If the subspace $\mathfrak{h}_A \subset \Gamma(TN)$ were a Lie subalgebra, then by (6.5) the canonical flow would generate a fiber whose tangent space at $[A]$ would be precisely $\pi_A(\mathcal{H}_A)$. However, $\mathfrak{h}_A$ is not closed under Lie bracket and the canonical flow of vector fields in $\mathfrak{h}_{A_0}$ acting on a single reference connection $[A_0]$ has no chance of generating an orbit that reasonably approximates *all* of Z_{DK}; the cut-off in the translation vector fields prevents the canonical flow from moving the center point very far from p_{A_0}, whereas all points in N can occur as center in Z_{DK}. But the estimates relevant to proving Theorem 1.2 are much less sensitive to changing the definition of of translations than to changing the definition of rotations and dilations, so it seems plausible that by a patching argument altering the definitions of only the translation vector fields in any significant way, we can splice together canonical flows based at connections with nearby center points. Presumably by splicing enough flows together we can obtain a fiber that is C^1-close globally to Z_{DK} and C^1-close locally to the orbit of some canonical flow. Even if the splicing construction fails, there are two reasons why, for purposes of integration, we may not need to define a true fiber (such as Z_{DK}) in a topological sense. First, when we integrate $\mu(p) \wedge \mu(q)$ over an orbit of the canonical flow, only connections with center point near p and q contribute significantly to the integral. It is likely that the same holds for an integral over Z_{DK}, so that it suffices to approximate only a region of Z_{DK} consisting of connections with center point in a fixed small ball. Second, although the canonical flow of the subspace $\mathfrak{h}^{\mathrm{rot}}_{A_0}$ acting on $[A_0]$ does not generate a *closed* manifold, it does generate an immersed copy of $SU(2) - \{-1\}$ lying in a small neighborhood of an $SO(3)$-orbit in Z_{DK}, and which geometrically wraps twice around this orbit. A careful analysis may show that there is a homotopy from the immersed punctured $SU(2)$ to a punctured double cover of the $SO(3)$ in Z_{DK}, small enough in all relevant norms that there is only a negligible difference between integrating over Z_{DK} and over the orbit of the canonical flow.

Thus the idea behind (Z1)–(Z5) is basically that there is fiber that interpolates between Z_{DK} and the not-quite-fiber generated by splicing together canonical flows. The hypotheses (Z4) and (Z5) amount to assuming that in this interpolated fiber, the bounds on ξ' are as good as they would be if the tangent space to the fiber were the one determined by the canonical flow. We need such an assumption because when we pull $\mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q)$ back to Z, we need to insert true tangent vectors to Z into (5.3); the $\pi_A \tilde{X}$'s in the expansion (8.1) below should be replaced by $\overline{\rho}_* v$'s — which has the effect of replacing each ξ in (8.6) with ξ'.

There is other evidence making the simultaneous satisfaction of at least (Z1)–(Z4) very plausible. On $\mathbf{R}^4$, if we remove the cut-offs in the definition of $\mathfrak{h}_A$ and define Z from the canonical flow acting on the standard instanton, then the spaces $T_{[A]}Z$ and $\mathcal{H}_A$ coincide. In the case of 1-instantons over simply connected definite manifolds (where the background connection is flat and there are no gluing parameters, so Z is five-dimensional), (Z4) was shown in [9] to be true with $\epsilon_2(\lambda) \leq c\lambda^{1-\delta}$ for small $\delta > 0$; in [6] this was strengthened to $\lambda^{1+\delta}$.

The technical hypothesis (Z5) is more ad hoc, and stronger than necessary, but is not without basis. In the setting of the five-dimensional moduli spaces mentioned above, certain estimates of this paper and Ref. [6] can be combined to show that $\|r_p^{-\delta}\xi'\|_4 \lesssim \lambda^{-1+\delta}(b \cdot \lambda + a \cdot \lambda^{1/2})$ and $\|r_A\xi'\|_4 \leq b \cdot \lambda^{\delta}(b \cdot \lambda + a \cdot \lambda^{1/2})$, much stronger than the L^p bounds assumed in (Z5). (Here r_p denotes distance to an arbitrary point $p \in N$.)

An important implication of (Z4) and (6.12) is the following. Let $\mathrm{d}vol_Z^{L^2}$ be the volume form on Z induced by the L^2 metric on $\mathcal{B}$. (Hypothesis (Z1) determines an orientation on Z, so there is no sign ambiguity here.) Let $a \in \mathbf{R}^4$ denote the quaternionic variable in $\mathbf{H}^*_{\lambda_0}$. Then

$$\overline{\rho}_i^*(\mathrm{d}vol_Z^{L^2}) \approx \text{const.} \times \mathrm{d}^4a \wedge \mathrm{d}vol_N = \text{const.} \times \lambda^3\,\mathrm{d}\lambda \wedge \mathrm{d}vol_{S^3} \wedge \mathrm{d}vol_N, \tag{7.3}$$

where the approximation becomes exact as $\lambda \to 0$ (and the constant is of course nonzero). Our chief use of (7.3) will be to help estimate the integrals of the nonlocal terms in $\mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q)$. For this purpose, we do not actually need "$\approx$" in (7.3); "$\leq$" would suffice. Thus hypothesis (Z4) can be weakened.

8. Localizing $\mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q)$

From now on we assume there is a fiber Z with properties (Z1)–(Z5). To motivate the leading-term calculation in (1.10), suppose for the moment that for $[A] \in Z$ the tangent space $T_{[A]}Z$ is precisely, rather than approximately, the space $\pi_A(\mathcal{H}_A) \subset T_{[A]}\mathcal{M}$. Then if we pull $\mu_{\mathrm{d}}(p)$ back to Z, we need only apply (5.3) to arguments of the form $\pi_A\tilde{X}_A$ with $X \in \mathfrak{h}_A$. Recalling the definition of ξ_X in (6.14), we then have

$$G_0^A\{\pi^A\tilde{X}, \pi^A\tilde{Y}\} = G_0^A(\{\tilde{X}, \tilde{Y}\} + Rem_1(X, Y; A)), \tag{8.1}$$

where

$$Rem_1(X, Y) = \{\tilde{X}, \xi_Y\} - \{\tilde{Y}, \xi_X\} + \{\xi_X, \xi_Y\}. \tag{8.2}$$

(We omit writing most of the A-dependence in these formulas explicitly.) Here $\{\cdot, \cdot\}$ is a universal, local, anti-symmetric bilinear pairing that takes two $Ad\,P$-valued 1-forms and produces an $Ad\,P$-valued 0-form. Note that $Rem_1(X, Y)$ is anti-symmetric in X and Y.

In [[5], Proposition 2.1], it was shown how to expand several of the expressions appearing in (8.1) and (8.2) as a leading-order local term plus a nonlocal remainder smaller in appropriate norms. In particular, for any vector fields X, Y on N, we have

$$G_0^A\{\tilde{X}, \tilde{Y}\} = -\tfrac{1}{2}F(X, Y) + G_0^A(R''(X, Y)), \tag{8.3}$$

where $F = F_A$ and where

$$2R''(X,Y) = \mathcal{R}(F)(X,Y) + F(\Delta X, Y) + F(X, \Delta Y) - 2(\nabla_i^A F)(\nabla_i X, Y) \\ -2(\nabla_i^A F)(X, \nabla_i Y) - 2F(\nabla_i X, \nabla_i Y). \tag{8.4}$$

Here $\mathcal{R}$ is an endomorphism proportional to the Riemann tensor whose precise form does not concern us. As a consequence of (8.3), $R''(X,Y)$ is anti-symmetric in X and Y. The precise way in which the derivatives of F and the derivatives of X and Y are hooked together in (8.4) is critical for certain estimates (Lemma 10.4).

Applying (8.3) to the first term in (8.1), we find

$$G_0^A\{\pi\tilde{X}, \pi\tilde{Y}\} = -\tfrac{1}{2}(F(X,Y) - Rem_2(X,Y)), \tag{8.5}$$

where

$$\begin{aligned} \tfrac{1}{2}Rem_2(X,Y) &= G_0^A(R''(X,Y) + Rem_1(X,Y)) \\ &= G_0^A(R''(X,Y) + \{\tilde{X}, \xi_Y\} - \{\tilde{Y}, \xi_X\} + \{\xi_X, \xi_Y\}) \\ &:= G_0^A(Rem_2'(X,Y)). \end{aligned} \tag{8.6}$$

Inserting all this into (5.3) we find

$$\begin{aligned} \mu_{\mathrm{d}}(p)(\pi\tilde{X}, \pi\tilde{Y}, \pi\tilde{V}, \pi\tilde{W}) = \frac{1}{4\pi^2}\mathrm{tr}((&F(X,Y)F(V,W) + F(X,V)F(W,Y) \\ &+F(X,W)F(Y,V))|_p) + Rem_3(X,Y,V,W)|_p \\ &+Rem_4(X,Y,V,W)|_p, \end{aligned} \tag{8.7}$$

where

$$Rem_3 = \text{const.} \times \mathrm{tr}(F \wedge Rem_2), \qquad Rem_4 = \text{const.} \times \mathrm{tr}(Rem_2 \wedge Rem_2). \tag{8.8}$$

(In (8.8) we regard F and Rem_2 as $\Gamma(Ad\,P)$-valued 2-forms on the space of vector fields.) The first term in (8.7) is just $(8\pi^2)^{-1}\,\mathrm{tr}(F \wedge F)(X,Y,V,W)|_p$, which, since F is an ASD, can be rewritten as $(8\pi^2)^{-1}|F|^2\,\mathrm{d}vol(X,Y,V,W)|_p$. Thus if we define

$$\mu_{\mathrm{loc}}(p)(\pi\tilde{X}, \pi\tilde{Y}, \pi\tilde{V}, \pi\tilde{W}) = \frac{1}{8\pi^2}|F(p)|^2\,\mathrm{d}vol(X,Y,V,W)|_p, \tag{8.9}$$

then (8.7) simplifies to

$$\begin{aligned} \mu_{\mathrm{d}}(p)(\pi\tilde{X}, \pi\tilde{Y}, \pi\tilde{V}, \pi\tilde{W}) = \mu_{\mathrm{loc}}(p)(&\pi\tilde{X}, \pi\tilde{Y}, \pi\tilde{V}, \pi\tilde{W}) + Rem_3(X,Y,V,W)|_p \\ &+Rem_4(X,Y,V,W)|_p. \end{aligned} \tag{8.10}$$

Of greatest concern to us will be the local part $\mu_{\mathrm{loc}}(p)$ of this expression. Note that $\mu_{\mathrm{loc}}(p) \wedge \mu_{\mathrm{loc}}(p) = 0$ since $\mathrm{d}vol_p \wedge \mathrm{d}vol_p = 0$. However, we will see that $\lim_{q\to p}\int_Z \mu_{\mathrm{loc}}(p) \wedge \mu_{\mathrm{loc}}(q) \neq 0$. In this integral it turns out that instantons of scale $\approx \mathrm{dist}(p,q)$ give the main contribution to the integral. Thus the pullback of $\mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(p)$ to Z can be thought of loosely as a δ-form concentrated on instantons of scale zero.

To integrate $\mu_{\rm d}(p)\wedge\mu_{\rm d}(q)$ we must still worry about the nonlocal remainder terms Rem_i as well as the fact that the tangent space $T_{[A]}Z$ is not precisely $\pi_A\mathcal{H}_A$. We will see later that as $\lambda_0\to 0$, the contributions to the integral of $\mu_{\rm d}(p)\wedge\mu_{\rm d}(q)$ over $Z=Z_{\lambda_0}$ from both of these corrections tend to zero. What we wish to compute now is

$$\lim_{q\to p}\int_{Z_{\lambda_0}}\mu_{\rm loc}(p)\wedge\mu_{\rm loc}(q), \tag{8.11}$$

where p and λ_0 are fixed.

For given p, q, as we integrate $\mu_{\rm loc}(p)\wedge\mu_{\rm loc}(q)$ over Z, the center point p_A of $[A]$ in Z moves around, affecting the support of the vector fields X, Y, v, w in (8.9). Thus for $\mu_{\rm loc}(p)\wedge\mu_{\rm loc}(q)(\pi\tilde{X}_1,\ldots,\pi\tilde{X}_8)$ to be nonzero, p_A must lie in $B(p,8n_0\lambda^{1/2})\cap B(q,8n_0\lambda^{1/2})$. In particular we can restrict p_A to a small normal-coordinate ball U centered at p (which we can take to be one of the U_i in (Z1)) without affecting $\int_Z\mu_{\rm loc}(p)\wedge\mu_{\rm loc}(q)$. Since we are interested in the limit as $q\to p$, we may also assume $q\in U$.

Let $2L=\mathrm{dist}(p,q)$; we will later send L to zero. Define $Z_1\subset Z$ to be the set of instantons in Z obeying the two criteria

$$L^{0.1}\geq\lambda^{1/2}\geq L, \tag{8.12}$$

$$p_A\in B(p,n_0\lambda^{1/2}). \tag{8.13}$$

Note that if $[A]\in Z_1$ then $p_A\in B(q,(n_0+1)\lambda^{1/2})$, so that the cut-off β in the definition of the vector fields X_i equals 1 at both p and q. We will see later that the contribution to (8.11) from the complement of Z_1 is negligible.

Let $\{x^i_{\rm old}\}$ denote normal coordinates on U. We change coordinates by setting $x_{\rm new}=L^{-1}x_{\rm old}$ and replace the metric $g_{\rm old}$ on U by $g_{\rm new}=L^{-2}g_{\rm old}$. Because of the conformal invariance of $|F|^2\,\mathrm{d}vol$, $\mu_{\rm loc}(p)\wedge\mu_{\rm loc}(q)$ is unaffected by this change. However, since $\lambda=\lambda_{\rm old}$ represented a distance in the old coordinate system, we now have a rescaled upper cut-off for $\lambda_{\rm new}=L^{-1}\lambda_{\rm old}$ on Z, namely $\lambda_{0,\rm new}=\lambda_0/L$. Measuring all distances in the new metric, the defining conditions for Z_1 become

$$L^{-0.8}\geq\lambda_{\rm new}\geq L, \tag{8.14}$$

$$p_A\in B(p,NL^{-1/2}\lambda_{\rm new}^{1/2}). \tag{8.15}$$

As $L\to 0$, several things happen. For $A\in Z$, $|F_A|$ becomes approximately standard on any fixed ball $B(p,K)$; $g_{\rm new}$ approaches the flat metric $\sum(\mathrm{d}x^i_{\rm new})^2$; and (in the rescaled metric and coordinates) Z_1 becomes an $SO(3)$-bundle over monotonically increasing regions of center-scale space that exhaust $(0,\infty)\times(\mathbf{R}^4-\{0\})$ as $L\to 0$. Because of (Z1), we can identify Z_1 with ever-increasing subsets G_L of $G:=(\mathbf{H}^*\times\mathbf{H})/\mathbf{Z}_2$. Letting $\mu'_{\rm loc}$ denote the pullback of $\mu_{\rm loc}$ to $\mathbf{H}^*\times\mathbf{H}$, we therefore have

$$\lim_{L\to 0}\int_{Z_1}\mu_{\rm loc}(p)\wedge\mu_{\rm loc}(q)=\frac{1}{2}\lim_{L\to 0}\int_{G_L}\mu'_{\rm loc}(p)\wedge\mu'_{\rm loc}(q), \tag{8.16}$$

provided this integral converges.

Let $\bar{\rho}$ be as in (Z1)–(Z5). Write elements of G as pairs (a, b), and write $A_{(a,b)} = \rho(a, b)$, $F_{(a,b)} = F_{A_{(a,b)}}$ as in (Z3). If we define $\mu'_{\mathrm{loc}} = \bar{\rho}^* \mu_{\mathrm{loc}} \in \Omega^4(G)$, then

$$\int_{G_L} \mu'_{\mathrm{loc}}(p) \wedge \mu'_{\mathrm{loc}}(q) = \int_{G_L} \mu'_{\mathrm{loc}}(p) \wedge \mu'_{\mathrm{loc}}(q) \left(\frac{\partial}{\partial a^1}, \ldots, \frac{\partial}{\partial b^4} \right) \mathrm{d}a^1 \wedge \cdots \wedge \mathrm{d}a^4 \wedge \mathrm{d}b^1 \wedge \cdots \wedge \mathrm{d}b^4. \tag{8.17}$$

To compute this we need to know $\bar{\rho}_{*(a,b)}(\partial/\partial a^i)$, $\bar{\rho}_{*(a,b)}(\partial/\partial b^i)$. At each (a, b) define X_i and Y_i to be the vector fields on $\mathbf{R}^4$ induced by $\partial/\partial a^i$ and $\partial/\partial b^i$, respectively. Temporarily writing $b^i = a^{i+4}$ and $Y_i = X_{i+4}$, from (Z4) there is an 8×8 matrix $C = Id + \mathrm{O}(\epsilon_2(\lambda_{\mathrm{old}}))$ for which we have

$$\bar{\rho}_{*_{(a,b)}} C_i^j \frac{\partial}{\partial a^j} = -\pi_{A_{(a,b)}} \iota_{X_i} F_{(a,b)}. \tag{8.18}$$

Hence from (8.10), if not for the correction matrix C, we would have

$$\mu'_{\mathrm{loc}}(p) \left(\frac{\partial}{\partial a^1}, \ldots, \frac{\partial}{\partial a^4} \right) = \mu_{\mathrm{loc}}(p)(\pi_{A_{(a,b)}} \tilde{X}_1^A, \ldots, \pi_{A_{(a,b)}} \tilde{X}_4^A) = (8\pi^2)^{-2} |F_A(p)|^2 \, \mathrm{d}vol_p(X_1, \ldots, X_4) \tag{8.19}$$

with a similar formula if we replace any of the $\partial/\partial a^i$'s by a $\partial/\partial b^i$.

Let us ignore, for now, (i) the $\mathrm{O}(\epsilon_2(\lambda_{\mathrm{old}})) = \mathrm{O}(\epsilon_2(L\lambda_{\mathrm{new}}))$ difference between the matrix C and the identity, and (ii) the $\mathrm{O}(|x_{\mathrm{old}}|^2) = \mathrm{O}(L^2|x_{\mathrm{new}}|^2)$ difference between the true metric on the rescaled ball and the flat metric; we will make the corrections later. Since the Euclidean volume form is $\mathrm{d}vol = \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^4$, we will write $\mathrm{d}vol_p = \mathrm{d}^4 x_p$, $\mathrm{d}vol_q = \mathrm{d}^4 x_q$ below. Hence

$$\mu'_{\mathrm{loc}}(p) \wedge \mu'_{\mathrm{loc}}(q) \left(\frac{\partial}{\partial a^1}, \ldots, \frac{\partial}{\partial b^4} \right) = (8\pi^2)^{-2} |F_A(p)|^2 |F_A(q)|^2 \, \mathrm{d}^4 x_p \wedge \mathrm{d}^4 x_q(X_1, \ldots, Y_4). \tag{8.20}$$

So far we have treated $\mathrm{d}^4 x_p \wedge \mathrm{d}^4 x_q$ as an 8-form whose arguments are vector fields, but we may as well consider it as an 8-form on the eight-dimensional space $T_pM \oplus T_qN$. Using the canonical isomorphisms $T_p\mathbf{R}^4 \cong T_q\mathbf{R}^4 \cong \mathbf{R}^4$ and our further identification of $\mathbf{R}^4$ with $\mathbf{H}$, we can write each X_i, Y_j in the form $(v, w) \in \mathbf{H} \oplus \mathbf{H}$. In the coordinate system $\{x^i_{\mathrm{new}}\}$, the origin represents p, and we may assume that q lies on the real axis with coordinate $2 \in \mathbf{H}$. Let $\tau_1 = 1 \in \mathbf{H}$ and let $\{\tau_i\}_2^4$ be the quaternions $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$. Then $X_i(x) = \tau_i a^{-1}(x - b)$ and $Y_i(x) = \tau_i$. So the corresponding elements in $T_p\mathbf{R}^4 \oplus T_q\mathbf{R}^4 \cong \mathbf{H} \oplus \mathbf{H}$ are $X'_i = (-\tau_i a^{-1} b, \tau_i a^{-1}(2 - b))$ and $Y'_i = (\tau_i, \tau_i)$. Modulo the span of the Y'_i, we have $X'_i = (0, 2\tau_i a^{-1}) := X''_i$, so $\mathrm{d}^4 x_p \wedge \mathrm{d}^4 x_q(X'_1, \ldots, X'_4, Y'_1, \ldots, Y'_4) = \mathrm{d}^4 x_p \wedge \mathrm{d}^4 x_q(X''_1, \ldots, X''_4, Y'_1, \ldots, Y'_4)$. Since $\mathrm{d}^4 x_p(X''_i, *, *, *) = 0$, it follows that $\mathrm{d}^4 x_p \wedge \mathrm{d}^4 x_q(X''_1, \ldots, X''_4, Y'_1, \ldots, Y'_4) = \mathrm{d}^4 x_p(Y'_1, \ldots, Y'_4) \, \mathrm{d}^4 x_q(X''_1, \ldots, X''_4)$. But $\mathrm{d}^4 x_p(Y'_1, \ldots, Y'_4) = 1$ and $\mathrm{d}^4 x_q(X_1, \ldots, X_4) = 2^4 |a|^{-4}$. Hence

$$\mu'_{\mathrm{loc}}(p) \wedge \mu'_{\mathrm{loc}}(q)|_{(a,b)} \left(\frac{\partial}{\partial a^1}, \ldots, \frac{\partial}{\partial b^4} \right) = (8\pi^2)^{-2} |F_{(a,b)}(0)|^2 |F_{(a,b)}(2)|^2 2^4 |a|^{-4}, \tag{8.21}$$

where $F_{(a,b)}$ is the curvature of the instanton obtained from the action of (a, b) on a reference connection in our fiber. Therefore

$$\mu'_{\rm loc}(p) \wedge \mu'_{\rm loc}(q)|_{(a,b)} = 2^4(8\pi^2)^{-2}|F_{(a,b)}(0)|^2|F_{(a,b)}(2)|^2|a|^{-4}\,{\rm d}^4a \wedge {\rm d}^4b. \quad (8.22)$$

Because λ_0 is small, there is a reference connection in our fiber that looks approximately standard on a ball of any fixed large radius with the approximation getting better as $\lambda_0 \to 0$ (the rescaling by L only improves this approximation). Our next approximation is to ignore the difference between the true reference connection A_0 and the standard instanton; we will deal with the error later. The connections in the limiting Z_1 are then the orbit of the standard instanton A_1 under the action of G. Hence

$$\begin{aligned} |F_{(a,b)}|^2(x) &= |\tilde{\Phi}^{-1}_{(a,b)}F_{A_1}|^2(x) = \left|\frac{{\rm d}\overline{(a^{-1}(x-b))} \wedge {\rm d}(a^{-1}(x-b))}{(1+|a^{-1}(x-b)|^2)^2}\right|^2 \\ &= \frac{48|a|^4}{(|a|^2+|x-b|^2)^4}. \end{aligned} \quad (8.23)$$

Thus

$$\begin{aligned} I_p &:= \lim_{L\to 0}\int_{G_L} \mu'_{\rm loc}(p) \wedge \mu'_{\rm loc}(q) \\ &= 36\pi^{-4}\int_{\boldsymbol{H}^*\times\boldsymbol{H}} \frac{2^4|a|^4\,{\rm d}^4a \wedge {\rm d}^4b}{(|a|^2+|b|^2)^4(|a|^2+|2-b|^2)^4}, \end{aligned} \quad (8.24)$$

and provided the error terms we have so far ignored are truly ignorable,

$$\lim_{L\to 0}\int_Z \mu_{\rm loc}(p) \wedge \mu_{\rm loc}(q) = \frac{1}{2}I_p \quad (8.25)$$

(see (8.16)).

Lemma 8.1. $I_p = 1$.

Proof. First introduce spherical coordinates in a-space (with radial variable we call λ) and cylindrical coordinates in b-space (with radial variable r). The integrals over the 3-sphere in a-space and the 2-sphere in the imaginary subspace of b-space are trivial, contributing factors $2\pi^2$ and 4π, respectively. Thus

$$\begin{aligned} I_p &= 36\pi^{-4} \\ &\quad \times \cdot 8\pi^3 \cdot 2^4 \int_{\lambda=0}^{\infty}\int_{r=0}^{\infty}\int_{z=-\infty}^{\infty} \frac{\lambda^7 r^2}{(\lambda^2+r^2+z^2)^4(\lambda^2+r^2+(z-2)^2)^4}\,{\rm d}z\,{\rm d}r\,{\rm d}\lambda. \end{aligned} \quad (8.26)$$

Introducing polar coordinates in the λ–r quarter-plane, the angular integration reduces to an integral over two real variables. Using the Residue Theorem to integrate over z leaves us with a one-dimensional integral that can be computed in closed form, yielding $I_p = 1$. □

In the local calculation we ignored errors from four sources: (i) the contribution from the complement of Z_1; (ii) the difference between the flat metric and the true metric on the

rescaled ball; (iii) the difference between $\bar{\rho}_* v$ and $-\pi_A \tilde{X}_v$ (i.e. the difference between the matrix C and the identity); and (iv) the difference between $|F|_{a,b}$ and the standard instanton of scale $|a|$ and center b.

Let us first deal with (i). Since the vector fields X_i we feed into $\mu_{\rm loc}$ are cut-off at distances $\geq 2n_0\lambda^{1/2}$ from p_A, the integrand $\mu_{\rm loc}(p) \wedge \mu_{\rm loc}(q)(X_1, \ldots, X_8)$ vanishes for p_A outside the ball $B(p, 2n_0\lambda^{1/2})$. For purposes of integration we therefore need only that portion of Z lying over a ball of fixed small radius centered at p. Because of (6.11), the integrand over such a region is bounded by a constant times the integrand we used in our previous calculation, cut off in certain regions. Since the integrand in (8.24) is integrable over all of $\mathbf{H}^* \times \mathbf{H}$, given any exhaustion $W_1 \subset W_2 \subset \cdots$ of $\mathbf{H} \times \mathbf{H}$, the integral over the complement of W_n goes to zero as $n \to \infty$. As the sets G_L provide such an exhaustion, the integral of $\mu_{\rm loc}(p) \wedge \mu_{\rm loc}(q)$ over the complement of Z_1 tends to zero.

Next we turn to the errors (ii)–(iv) listed above. In place of the set Z_1 considered in the derivation above, for $K > 0$ consider the sets $Z_{K,L}$ defined by $\{L^{0.1} \geq \lambda^{1/2} \geq L, p_A \in B(p, K\lambda)\}$. After rescaling by L as before these conditions become $\{\lambda_{\rm new} \geq L, p_A \in B(p, K\lambda_{\rm new})\}$. This time as $L \to 0$, $Z_{K,L}$ exhausts $Z_{K,0} := \{p_A \in B(p, K\lambda_{\rm new})\}$ with $\lambda_{\rm new}$ unrestricted. But convergence of the integral I_p implies that given $\epsilon_3 > 0$, we can choose k large enough and L small enough that the integral of the integrand in (8.24) over the complement of $Z_{K,L}$ is less than ϵ_3. On the interior set $Z_{K,L}$, hypotheses (Z3) and (Z4) imply that given $\epsilon_4 > 0$, by taking L sufficiently small we can arrange for the ratio of the true $\mu'_{\rm loc}(p) \wedge \mu'_{\rm loc}(q)$ to be within a multiple $(1+\epsilon_4)^{\pm 1}$ of the integrand in (8.24) over all of $Z_{K,L}$. (Error (ii) gives an $O(\lambda_{\rm old}) \leq O(L^{0.2})$ contribution to ϵ_4; error (iii) a contribution $\epsilon_2(\lambda_{\rm old}) \leq \epsilon_2(L^{0.2})$ through the matrix C. As for error (iv), in the rescaled metric and coordinates, (Z3) implies

$$\left| |F_{a,b}|_{g_{\rm new}} - \frac{\sqrt{48}|a|^2}{(|a|^2 + |x-b|^2)^2} \right| \leq \epsilon_1(\lambda_{\rm old})\lambda_{\rm new}^{-2}, \tag{8.27}$$

so that this error gives a contribution $\epsilon_1(\lambda_{\rm old}) \leq \epsilon_1(L^{0.2})$ to ϵ_4.) Hence we can arrange for the integral of $\int_{Z_{K,L}} \mu'_{\rm loc}(p) \wedge \mu'_{\rm loc}(q)$ to overlie within $\frac{1}{2}\epsilon_3$ of the integral over $Z_{K,L}$ of the integrand in (8.24). It follows that by choosing L small enough, the errors introduced by our approximations can be made arbitrarily small.

We have now proven the following.

Proposition 8.2. *For any $p \in N$,*

$$\lim_{q \to p} \int_{Z_{\lambda_0}} \mu_{\rm loc}(p) \wedge \mu_{\rm loc}(q) = \frac{1}{2}. \tag{8.28}$$

9. The nonlocal terms in $\mu_{\rm d}(p) \wedge \mu_{\rm d}(q)$

From (8.8)–(8.10), $\mu_{\rm d}(p) \wedge \mu_{\rm d}(q)$ can be expanded as $\mu_{\rm loc}(p) \wedge \mu_{\rm loc}(q)$ plus a remainder. Our next task is to show that, as $\lambda_0 \to 0$, the contribution of this remainder to $\int_Z \mu_{\rm d}(p) \wedge$

$\mu_{\mathrm{d}}(q)$ tends to zero. This will follow from the next proposition, whose proof occupies the remainder of this paper.

Proposition 9.1. *Let Ω be the restriction to Z_{λ_0} of $\Omega_{\mathcal{M}} := \mu_{\mathrm{d}}(p) \wedge \mu_{\mathrm{d}}(q) - \mu_{\mathrm{loc}}(p) \wedge \mu_{\mathrm{loc}}(q) \in \Omega^8(\mathcal{M})$. Assuming* (Z1)–(Z5)*, there exists $\delta > 0$ such that*

$$\int_{Z_{\lambda_0}} \Omega \leq \text{const.}\, \lambda_0^{\delta}, \tag{9.1}$$

where the constant is independent of p and q.

Observe that Propositions 8.2 and 9.1 together prove Theorem 1.2.

Proving Proposition 9.1 requires some bounds on $Rem_2(X, Y)$ for $X, Y \in \mathfrak{h}_A$. Before starting to derive these, we need some notational simplification. Below we will be computing many things that are multilinear in data of the form $(\mathbf{b}, a, \alpha) \in T_{p_A}N \oplus \mathbf{R} \oplus \Lambda^2_+ T_{p_A}N$. Given a single vector field X constructed from such data, we can denote the defining data of (6.6) by $(\mathbf{b}_X, a_X, \alpha_X)$. This notation becomes cumbersome especially when computing objects that involve more than a single vector field. However, because $|X_{(\mathbf{b},a,\alpha)}| \leq c(|\mathbf{b}| + (|a|+|\alpha|)\lambda^{-1}r_A)$, the a and α data always enter our bounds with precisely the same weight, so for shorthand we will generally lump the a and α terms together, and simply call them a. Furthermore, for simplicity we will often omit the subscripts $X, Y, \ldots$ in the defining data $(\mathbf{b}_X, a_X, \alpha_X), (\mathbf{b}_Y, a_Y, \alpha_Y), \ldots$; the dependence on $X, Y, \ldots$ can be reconstructed from the context. For example, if we write

$$|\text{something bilinear in } X, Y \in \mathfrak{h}_A| \leq c_1 b^2 + c_2 ba + c_3 a^2, \tag{9.2}$$

then on the RHS the notation has the following meaning:

$$\begin{aligned} b^2 &= |\mathbf{b}_X||\mathbf{b}_Y|, \qquad ba = (|\mathbf{b}_X|(|a_Y| + |\alpha_Y|) + |\mathbf{b}_Y|(|a_X| + |\alpha_X|)), \\ a^2 &= (|a_X| + |\alpha_X|)(|a_Y| + |\alpha_Y|). \end{aligned} \tag{9.3}$$

If the bilinear quantity is anti-symmetric in X, Y (as in Proposition 9.2), then the estimate factors through the wedge product, in which case we can take

$$\begin{aligned} b^2 &= |\mathbf{b}_X \wedge \mathbf{b}_Y|, \qquad ba = (|\mathbf{b}_X|(|a_Y| + |\alpha_Y|) + |\mathbf{b}_Y|(|a_X| + |\alpha_X|)), \\ a^2 &= (|a_X||\alpha_Y| + |\alpha_X||a_Y| + |\alpha_X \wedge \alpha_Y|). \end{aligned} \tag{9.4}$$

Finally, the notation $x \lesssim y$ means $x \leq cy$ for a constant c that is uniform in all relevant parameters.

With this notation in mind, we have the following proposition.

Proposition 9.2. (a) *There exists $\delta > 0$ such that*

$$\|Rem_2(X, Y)\|_\infty \lesssim \lambda^{-1+\delta}(b^2 + ba \cdot \lambda^{-1/2} + a^2 \cdot \lambda^{-1/2}). \tag{9.5}$$

Furthermore, there exists $c_1 > 0$ such that, for $r_A \geq c_1\lambda^{1/2}$, we have the pointwise decay

$$|Rem_2(X, Y)| \lesssim r_A^{-1}(b^2 + ba \cdot \lambda^{-1/2} + a^2 \cdot \lambda^{-1/2}). \tag{9.6}$$

(b) Let v, X_v, ξ_v, ξ_v' be as in hypothesis (Z5). *If we alter the definition of $Rem_2(X_v, X_w)$ by replacing ξ_v with ξ_v', then the bounds above still apply.*

We will prove Proposition 9.2 (actually a slightly stronger version) in Section 10. Let us assume it for now and move onto its application, the proof of Proposition 9.1. The decay estimate (9.6) is crucial in this proof; the global bound (9.5) does not suffice.

Proof of Proposition 9.1. By hypothesis (Z1), $\int_Z \Omega \leq \sum_i \rho_i^* \Omega$. Since $\Omega \in \Omega^8(Z)$ we can write $\Omega = f \, \mathrm{d}vol_Z^{L^2}$, where the function f can be computed at $[A] \in Z$ from any (positively) oriented L^2-orthonormal basis $\eta_1, \ldots, \eta_8$ of $T_{[A]}Z$ by

$$f([A]) = \Omega(\eta_1, \ldots, \eta_8). \tag{9.7}$$

Similarly, we define $f'([A]) = \Omega_{\mathcal{M}}(\eta_1', \ldots, \eta_8')$, where the $\{\eta_i'\}$ are an orthonormal basis for $\pi_A \mathcal{H}_A$, and set $\Omega' = f' \, \mathrm{d}vol_Z^{L^2} \in \Omega^8(Z)$.

We will first show that $\int_Z \Omega' \leq c\lambda_0^\delta$ (where δ is as in Proposition 9.2), and then deduce that the same is true for $\int_Z \Omega$.

We proceed to estimate Ω'. By Proposition 6.4, an approximately orthonormal basis of $\pi_A \mathcal{H}_A$, up to a scale factor $(8\pi^2)^{1/2}$, is given by $\{\eta_n' = \pi_A \tilde{X}_{(\mathbf{b}_n, a_n, \alpha_n)} := \pi \tilde{X}_n\}_1^8$ as $(\mathbf{b}_n, a_n, \alpha_n)$ run over an orthonormal basis of $T_{p_A}N \oplus \mathbf{R} \oplus \Lambda^2_+(T_{p_A}N)$. Applying Gram–Schmidt to $\{\pi \tilde{X}_n\}$, it follows that $f' \leq \text{const.} \times \Omega(\pi \tilde{X}_1, \ldots, \pi \tilde{X}_8) \, \mathrm{d}vol_Z^{L^2}$. Hence from (7.3),

$$\rho_i^*(\Omega') \leq c\Omega(\pi \tilde{X}_1, \ldots, \pi \tilde{X}_8)\lambda^3 \, \mathrm{d}\lambda \wedge \mathrm{d}vol_{S^3} \wedge \mathrm{d}vol_N. \tag{9.8}$$

Symbolically we can write $\Omega' = \sum_{i=0}^3 \Omega_i'$ as a sum of terms of the form $F^i \wedge Rem_2^{4-i}$, $0 \leq i \leq 3$. We estimate the integrals of Ω_i' one case at a time. Only the completely nonlocal term Ω_0' requires the pointwise decay estimate (9.6); for the remaining terms the uniform bound (9.5) suffices. Bounding $\int \Omega_3'$ requires some care but we shall see that the integrals of Ω_1' and Ω_2' can be estimated heavy-handedly.

Case 1: Terms of the form Rem_2^4. Let $Z_2 \subset Z$ denote the subset of connections for which both $\mathrm{dist}(p, p_A)$ and $\mathrm{dist}(q, p_A)$ are $\geq c_1\lambda^{1/2}$, where c_1 is as in Proposition 9.2, and let $Z_1 = Z - Z_2$. The sets Z_1, Z_2 are the inverse images of sets W_1, $W_2 \subset (0, \lambda_0) \times N$ under the map sending a connection to its scale and center. If for each $\lambda \in (0, \lambda_0)$ we define $W_{1,\lambda} := \{y \in N \,|\, (\lambda, y) \in W_1\}$, then $W_{1,\lambda}$ is contained in the union of a ball of radius $\lesssim \lambda^{1/2}$ centered at p and a similar ball centered at q, so $\mathrm{Vol}(W_{1,\lambda}) \lesssim \lambda^2$.

For the orthonormal set $\{(\mathbf{b}_n, a_n, \alpha_n)\}$ we may choose four elements of the type $(\mathbf{b}, 0, 0)$ and four of the type $(0, *, *)$, all normalized to unit length. Then from (9.6) on Z_1, we have

$$\begin{aligned} |Rem_2^4(X_1, \ldots, X_n)| &\lesssim \lambda^{-4+4\delta} \cdot \{\text{coefficient of } b^4a^4 \text{ in } (b^2 + ba\lambda^{-1/2} + a^2\lambda^{-1/2})^4\} \\ &\lesssim \lambda^{-6+4\delta}. \end{aligned} \tag{9.9}$$

Hence from (9.8),

$$\int_{\rho_i^{-1}(Z_1)} \rho_i^* \Omega_0' \lesssim \int_{W_1} \lambda^{-6+4\delta}\lambda^3 \, \mathrm{d}\lambda \, \mathrm{d}vol_N \lesssim \int_0^{\lambda_0} (\lambda^{-3+4\delta} \, vol(W_{1,\lambda})) \, \mathrm{d}\lambda \lesssim \lambda_0^{4\delta}; \tag{9.10}$$

the integral over the gluing-parameter space S^3 gives a constant factor.

Similarly on Z_2, $|Rem_2^4(\tilde{X}_1, \ldots, \tilde{X}_n)| \lesssim \lambda^{-2} r_A(p)^{-2} r_A(q)^{-2}$; the two distances $r_A(p)$, $r_A(q)$ enter this way because in the Rem_2^4 term in $\mu(p) \wedge \mu(q)$, two of the Rem_2's are evaluated at p and two at q (see (8.10)). Since $r_A(p)^{-2} r_A(q)^{-2} \leq r_A(p)^{-4} + r_A(q)^{-4}$ and on W_2 both $r_A(p)$ and $r_A(q)$ are $\geq c\lambda^{1/2}$, we have

$$\int_{\rho_i^{-1}(Z_2)} \rho_i^* \Omega_0' \lesssim \int_0^{\lambda_0} \lambda^3 \, d\lambda \left(\int_{c\lambda^{1/2}}^{\mathrm{diam}(N)} \lambda^{-2} r^{-4} r^3 \, dr \right) \lesssim \int_0^{\lambda_0} \lambda |\log \lambda| \, d\lambda \lesssim \lambda_0^{1.99}. \tag{9.11}$$

Combining this with the integral over Z_1 and summing over i,

$$\int_Z \Omega_0' \lesssim \lambda_0^{4\delta}. \tag{9.12}$$

Case 2: Terms of the form $F \wedge Rem_2^3$. In this and the remaining cases, $F(X_i, X_j)$ is computed either at p or at q, and since the X_i are cut-off outside a ball of radius $\sim \lambda^{1/2}$ centered at p_A, for $i \geq 1$ terms of the form $F^i \wedge Rem_2^{4-i}(X_1, \ldots, X_8)|_{p,q}$ vanish unless (λ, p_A) lies in the set Z_1 defined in Case 1. All points p_A in the remaining computations can thus be assumed to lie in one of our sets U_j, and $\int_{\rho_j^{-1}(Z)} \rho^* \Omega_i' = \int_Z \Omega_i'$.

Note that all vector fields $X, Y \in \mathfrak{h}_A$ satisfy $|X|, |Y| \leq \beta(b + a\lambda^{-2} r_A) \leq b + a\lambda^{-1/2}$, and hence $|F(X, Y)| \leq |F|(b + a\lambda^{-1/2})^2$. Using the uniform bound (9.5) to estimate the three Rem_2 terms, we obtain the pointwise bound

$$\begin{aligned} |F \wedge Rem_2^3(X_1, \ldots, X_8)| &\lesssim |F| \cdot \{\text{coefficient of } b^4 a^4 \text{ in} \\ &(b + a\lambda^{-1/2})^2 \lambda^{-3+3\delta} (b^2 + ba\lambda^{-1/2} + a^2 \lambda^{-1/2})^3\} \lesssim |F| \lambda^{-5+3\delta}, \end{aligned} \tag{9.13}$$

where F is evaluated at either p or q. Because of the cut-off in X_i we may assume that p_A is a distance $\lesssim c\lambda^{1/2}$ from whichever of these points at which we evaluate. Hence using (6.11),

$$\int_Z \Omega_1' \lesssim \int_0^{\lambda_0} \lambda^3 \, d\lambda \left(\int_0^{c\lambda^{1/2}} \frac{\lambda^{-5+3\delta} \lambda^2}{(\lambda^2 + r^2)^2} r^3 \, dr \right) \lesssim \int_0^{\lambda_0} \lambda^{3\delta} |\log \lambda| \, d\lambda \lesssim \lambda_0^{1+2\delta}. \tag{9.14}$$

Case 3: Terms of the form $F^2 \wedge Rem_2^2$. Here there are two subcases, depending on where the points at which F and Rem_2 are evaluated; we can have terms of type $F(p)F(p)Rem_2(q)$ $Rem_2(q)$ or of type $F(p)F(q)Rem_2(p)Rem_2(q)$. In each subcase we bound the Rem_2 terms using (9.5). At whichever point $F(X_i, X_j)$ is evaluated, we can again assume $r_A \lesssim \lambda^{1/2}$, so that $|X_i| \lesssim b + a\lambda^{-1/2}$. Letting p', p'' denote either of p, q, we then have

$$\begin{aligned} |F^2 \wedge Rem_2^2(X_1, \ldots, X_8)| &\lesssim |F|(p')|F|(p'') \cdot \{\text{coefficient of } b^4 a^4 \text{ in} \\ &(b + a\lambda^{-1/2})^4 \lambda^{-2+2\delta} (b^2 + ba\lambda^{-1/2} + a^2 \lambda^{-1/2})^2\} \\ &\lesssim (|F|^2(p') + |F|^2(p'')) \lambda^{-4+2\delta}. \end{aligned} \tag{9.15}$$

Hence the integral of the different types of $F^2 Rem_2^2$ terms can all be bounded by the integral of $|F|^2(p)\lambda^{-4+2\delta}$:

$$\int_Z \Omega_2' \lesssim \int_0^{\lambda_0} \lambda^3 \,\mathrm{d}\lambda \left(\int_0^{c\lambda^{1/2}} \frac{\lambda^{-4+2\delta}\lambda^4}{(\lambda^2+r^2)^4} r^3 \,\mathrm{d}r \right) \int_0^{\lambda_0} \lambda^{-1+2\delta} |\log\lambda| \,\mathrm{d}\lambda \lesssim \lambda_0^\delta. \tag{9.16}$$

Case 4: Terms of the form $F^3 \wedge Rem_2$. In the previous two cases we were rather wasteful in bounding $|X_i|$ pointwise; this time we must be more economical.

Since p and q enter the problem symmetrically it suffices to deal with terms of the form $F(p)F(p)F(q)Rem_2(q)$. Temporarily write $r_p = r_A(p)$, $r_q = r_A(q)$, $F_p = |F|(p)$, $F_q = |F|(q)$. Note that for our term to be nonzero, both r_p and r_q must be $\leq c\lambda^{1/2}$. Using this fact several times we find

$$\begin{aligned} |F^3 \wedge Rem_2(X_1,\ldots,X_8)| &\lesssim F_p^2 F_q \cdot \lambda^{-1+\delta} \cdot \{\text{coefficient of } b^4 a^4 \text{in } (b + a\lambda^{-1} r_p)^4 \\ &\quad (b + a\lambda^{-1} r_q)^2 (b^2 + ba\lambda^{-1/2} + a^2\lambda^{-1/2})\} \lesssim \lambda^{-4+\delta}(F_p^2 F_q r_p^2 + F_p^2 F_q r_q^2) \\ &\lesssim \lambda^{-4+\delta}(F_p^4 r_p^4 + F_q^2 + \lambda^2 F_p^3 + \lambda^{-4} F_q^3 r_q^6). \end{aligned} \tag{9.17}$$

We can now replace p by q and integrate over the region $\{(\lambda, p_A) \,|\, 0 < \lambda \leq \lambda_0, 0 \leq r_A(p) \leq c\lambda^{1/2}\}$ as in Cases 2 and 3. For each of the four terms $\lambda^i |F|^j r^k$ in parentheses in (9.17), one finds $\int_0^{c\lambda^{1/2}} \lambda^i (\lambda^2/(\lambda^2+r^2)^2)^j r^k r^3 \,\mathrm{d}r \leq$ const., so

$$\int_Z \Omega_3' \lesssim \int_0^{\lambda_0} \lambda^3 \lambda^{-4+\delta} \,\mathrm{d}\lambda \lesssim \lambda_0^\delta. \tag{9.18}$$

Combining the four cases, this proves that $\int_Z \Omega' \leq c\lambda_0^\delta$ (assuming Proposition 9.2).

Now define $Rem_2^{\text{true}}(X, Y)$ to be the right-hand side of (8.6), but with ξ_X, ξ_Y replaced by the objects ξ_X', ξ_Y' of (Z5). The form Ω is obtained from Ω' by replacing each occurrence of Rem_2 with Rem_2^{true}. Hence $\Omega - \Omega'$ can be expressed as a sum of terms of the form $F^i (Rem_2^{\text{true}} - Rem_2)^j Rem_2^k$ for appropriate i, j, k. By part (b) of Proposition 9.2, the bounds on $|(Rem_2^{\text{true}}(X, Y) - Rem_2(X, Y)|$ are of precisely the same form as in part (a), so the same argument as above shows that $\int_Z (\Omega - \Omega') \leq c\lambda_0^\delta$, establishing (9.1). □

10. The proof of Proposition 9.2

The proof of Proposition 9.2 is long, so we outline the strategy. To obtain (9.5), we need a pointwise bound on $G_0^A(Rem_2'(X, Y))$ (see (8.6)). If there were a four-dimensional Sobolev embedding $L_2^2 \hookrightarrow C^0$, then modulo extra terms arising from Weitzenböck identities that occur when comparing objects of the form $\|\nabla^A \nabla^A \phi\|_2$ to objects of the form $\Delta^A \phi$, we could get a C^0 bound on $Rem_2(X, Y)$ from an L^2 bound on $Rem_2'(X, Y)$. (This, in turn, would require some L^p and/or pointwise bounds on ξ.)

Of course there is no embedding $L_2^2 \hookrightarrow C^0$, but since the failure is borderline, *any* stronger Sobolev-type norm should give an embedding into C^0. The most efficient Sobolev

inequality for our purpose is the following one. This inequality is not surprising, but may not be widely known, so we prove it in the appendix (Corollary A.2).

Lemma 10.1 (Sobolev embedding lemma). *Let E be a vector bundle over a compact four-dimensional manifold N. For $p \in N$,* let r_p *denote the distance to p. Then for any $\delta > 0$, there exists a constant $c(\delta)$ such that for all connections ∇ on E, all $\phi \in \Gamma(E)$, and all $p \in N$,*

$$|\phi(p)| \leq c(\delta)(\|\phi\|_2 + \|r_p^{-\delta}\nabla\nabla\phi\|_2). \tag{10.1}$$

Hence

$$\|\phi\|_\infty \leq c(\delta) \sup_{p\in N}(\|\phi\|_2 + \|r_p^{-\delta}\nabla\nabla\phi\|_2). \tag{10.2}$$

We will use this lemma to get pointwise bounds on $\phi = G_0^A(Rem_2'(X, Y))$. Hence we will need to estimate $\|G_0^A\omega\|_2$ and $\|r_p^{-\delta}\nabla^A\nabla^A G_0^A\omega\|_2$ for $\omega = Rem_2'(X, Y)$. For general ω, Proposition 10.2 estimates these in terms of weighted L^2 norms of ω, providing bounds whose only dependence on the connection is explicitly through the center point and scale. (This type of uniformity in the connection is the hard part of all our elliptic estimates. Uniformity is important since to estimate an integral over a family of connections, we cannot use any bounds that depend on the connection in an uncontrolled way.) Proposition 10.2 also provides similar estimates of objects ξ of the form appearing in (6.14), which we need for reasons discussed below.

The pointwise estimates of $G_0^A\omega$ in terms of weighted L^2 norms of general ω will be summarized (and generalized) as part of Proposition 10.2, specifically the first half of (10.16). To apply these general estimates to $\omega = Rem_2'(X, Y)$ we still need to bound the weighted L^2 norms of $Rem_2'(X, Y)$. To understand what this entails, write $Rem_2'(X, Y) = Rem_{2,\mathrm{loc}}' + Rem_{2,\mathrm{semiloc}}' + Rem_{2,\mathrm{nonloc}}'$, where

$$\begin{aligned} &Rem_{2,\mathrm{loc}}' = R''(X, Y), \qquad Rem_{2,\mathrm{semiloc}}' = \{\tilde{X}, \xi_Y\} - \{\tilde{Y}, \xi_X\}, \\ &Rem_{2,\mathrm{nonloc}}' = \{\xi_X, \xi_Y\} \end{aligned} \tag{10.3}$$

(see (8.4)). Because of the cut-offs in X and Y, the expressions $Rem_{2,\mathrm{loc}}'$ and $Rem_{2,\mathrm{semiloc}}'$ are supported in $B(p_A, 2\epsilon)$, but $Rem_{2,\mathrm{nonloc}}'$ is not. Thus among the estimates we need are weighted L^2 bounds on $R''(X, Y)$. By Lemma 10.4 below, pointwise we find

$$|R''(X, Y)| \leq c|\hat{X}||\hat{Y}|(\beta + \epsilon^{-2}\chi)(|F| + r_A|\nabla^A F|) \tag{10.4}$$

(recall that $\hat{X}$ is the object that the cut-off β multiplies in the definition of X). Here χ is the characteristic function of the annulus $\epsilon \leq r_A \leq 2\epsilon$. Thus to apply the estimate (10.16) of Proposition 10.2 to obtain bounds on $G_0^A(Rem_2'(X, Y))$, we need to estimate certain expressions of the form $\|\beta r_p^{-\delta} r_A^m F\|_2$, and similar expressions with F replaced by $\nabla^A F$ and/or with β replaced by χ. This will be accomplished in Lemma 10.5, where we will list all the purely local estimates we need.

Weighted L^2-norm bounds on $Rem'_{2,\mathrm{semiloc}}$ and $Rem'_{2,\mathrm{nonloc}}$ can be obtained from weighted L^4-norm bounds on $\tilde{X}$ and ξ. The first of these is another purely local estimate. The second will be achieved in Proposition 10.8, where we will use the basic elliptic tools in Proposition 10.2 to turn the problem into a local estimate again.

Till now we have made no mention of the role the point p plays in affecting the weighted norms. If we compute these weighted norms as above and take the supremum over $p \in N$ as in (10.2), we obtain only the sup-norm bound (9.5) for $Rem_2(X, Y)$. To prove Proposition 9.1 we additionally need the pointwise decay bound (9.6). Since the local quantities we bound are supported near the center point p_A of A, decay is only an issue for the nonlocal quantities, but these are built out of Green operators applied to quantities supported near p_A. Thus one expects that as the distance between p and p_A increases, the bounds on our nonlocal quantities should decrease. This turns out to be true (at least for $\mathrm{dist}(p, p_A) \geq \text{const.} \times \lambda^{1/2}$); we simply have to work harder, establishing some general pointwise bounds in Proposition 10.3. Our basic estimates in Proposition 10.2 are most useful for p close to p_A; to get the bounds that lead to (9.6), in which p is farther from p_A, we will apply Proposition 10.3.

To establish (9.6) we again break up $Rem'_2(X, Y)$ into its local, semi-local, and nonlocal pieces as in (10.3). In the cases of $Rem'_{2,\mathrm{loc}}$ and $Rem'_{2,\mathrm{semiloc}}$, Proposition 10.3 again reduces our work to weighted L^p bounds of purely local quantities. For $Rem'_{2,\mathrm{nonloc}}$, however, Proposition 10.3 leaves us with bounding an expression of the form $\|r_A^{1+\delta'}\{\xi_X, \xi_Y\}\|_2$, and the obvious approach — Hölder's inequality and the weighted L^4 bounds already obtained — does not give us a strong enough bound for an adequate decay rate in (9.6). We will circumvent this by obtaining a pointwise decay estimate for ξ, which in turn gives us a satisfactory decay rate for Rem_2. (In fact, with the pointwise estimate on ξ in hand, it turns out that the contribution of $Rem'_{2,\mathrm{nonloc}}$ to Rem_2 is much smaller than the bounds we obtain from the other two terms.)

With this discussion behind us, our procedure is clear. First we will fill our elliptic toolbox by proving Propositions 10.2 and 10.3. To apply these we need to compute weighted L^p norms of various quantities appearing in Rem'_2, which is our next step. The final step is then a matter of bookkeeping, applying the general elliptic tools to bound G_0^A of $Rem'_{2,\mathrm{loc}}$, $Rem'_{2,\mathrm{semiloc}}$, and $Rem'_{2,\mathrm{nonloc}}$.

To avoid writing similar hypotheses over and over, and for notational simplicity, for the rest of this section we impose the following.

Blanket hypotheses and notation. A always denotes a connection with $[A] \in \mathcal{M}^{\kappa,\nu}_{k+1,\lambda_0}$ (see Definition 6.1). Every Proposition, Lemma, etc., has an implicit hypothesis "for λ_0 sufficiently small and for all $[A] \in \mathcal{M}'_{k+1,\lambda_0} \cap \mathcal{M}^{\kappa,\nu}$". Constants c are uniform in A and in all other relevant parameters not explicitly shown (though some would depend on κ and ν, if these were not fixed); e.g. $c(\delta)$ depends only on δ. Constants are continually updated, and when a hitherto unnamed c appears, there is an implicit "for some constant c". The notation "$x \lesssim y$" means $x \leq cy$. F always denotes the curvature of the connection A, and $\nabla = \nabla^A$ denotes the full covariant derivative on $\Gamma(Ad\,P \otimes \Lambda^* T^* N)$ (the tensor product connection determined by A and the Levi-Civita connection). Given any vector field X on N, we write $\tilde{X} = \iota_X F$ (thus there is an A-dependence we suppress). We write p_A for the center point of

A and λ for $\lambda(A)$. For any $p \in N$, we let r_p denote the distance to p, and write r_A for r_{p_A}. When a point p appears in a hypothesis, the letter d always means $r_A(p) = \mathrm{dist}(p, p_A)$. The scale $\epsilon = \text{const.}\,\lambda^{1/2}$ and cut-off $\beta = \beta_{\mathrm{std}}(r_A/\epsilon)$ are always as in (6.8), and χ denotes the characteristic function of the annulus $\{\epsilon \le r_A \le 2\epsilon\}$ containing the support of $\mathrm{d}\beta$. We also define the operators $\mathcal{D} = \mathcal{D}^A : \Omega^1(Ad\,P) \to \Omega^0(Ad\,P) \oplus \Omega^2_+(Ad\,P)$ by

$$\mathcal{D}^A\eta = ((d^A)^*\eta, \sqrt{2}d^A_+\eta); \tag{10.5}$$

thus $\ker(\mathcal{D}^A) = H^1_A$, the harmonic space in the middle of the elliptic complex

$$0 \to \Omega^0(Ad\,P) \overset{d^A}{\to} \Omega^1(Ad\,P) \overset{\sqrt{2}d^A_+}{\to} \Omega^2_+(Ad\,P) \to 0. \tag{10.6}$$

Define $\Delta^A_0, \Delta^A_1, \Delta^A_+$ to be the Laplacians on 0-forms, 1-forms, and SD 2-forms, respectively, constructed from this complex, and let G^A_0, G^A_+ be the inverses of Δ^A_0, Δ^A_+. Also define $\Delta^A_\oplus$, $G^A_\oplus$ on $\Omega^0(Ad\,P) \oplus \Omega^2_+(Ad\,P)$ by $\Delta^A_\oplus = \Delta^A_0 \oplus \Delta^A_+$, $G^A_\oplus = G^A_0 \oplus G^A_+$. Note that

$$(\mathcal{D}^A)^*(\phi_0, \phi_+) = d^A\phi_0 + \sqrt{2}(d^A_+)^*\phi_+, \tag{10.7}$$

so that the quantity $\xi_X = \xi^A_X$ of (6.14) can be written as

$$\xi_X = (\mathcal{D}^A)^* G^A_\oplus \mathcal{D}^A \tilde{X}. \tag{10.8}$$

Finally, observe that

$$\mathcal{D}^A(\mathcal{D}^A)^* = \Delta^A_\oplus, \qquad (\mathcal{D}^A)^*\mathcal{D}^A = \Delta^A_1. \tag{10.9}$$

Now we can finally begin proving Proposition 9.2. In the following proposition, what drives the estimates are two facts: (i) $\Delta^A_\oplus$ is uniformly bounded below, and (ii) in the Weitzenböck identity for $\Delta^A_\oplus$, only Riemannian curvature terms appear; F does not enter.

Proposition 10.2. *For $\delta_0 > 0$ sufficiently small and any $\delta, \delta', \delta''$ (possibly zero) of absolute value less than δ_0, such that for any $p \in N$ and any $\omega \in \Omega^0(Ad\,P) \oplus \Omega^2_+(Ad\,P)$:*

$$\|G^A_\oplus\omega\|_2 \lesssim \|r_p^{1+\delta'}\omega\|_2, \tag{10.10}$$

$$\|r_p^{-\delta}\nabla^A\nabla^A G^A_\oplus\omega\|_2 \lesssim \|r_p^{-\delta}\omega\|_2 + \lambda^{\delta'-1}\|r_A^{1-\delta-\delta'}\omega\|_2. \tag{10.11}$$

Furthermore if $\xi = (\mathcal{D}^A)^* G^A_\oplus\omega \in \Omega^1(Ad\,P)$ (cf. (10.8)), *then*

$$\|\xi\|_2 \lesssim \|r_A^{1+\delta}\omega\|_2, \tag{10.12}$$

$$\|r_A^{1-\delta}\xi\|_4 + \|r_A^{1-\delta}\nabla^A\xi\|_2 \lesssim \|r_A^{1-\delta}\omega\|_2, \tag{10.13}$$

$$\|r_p^{-1-\delta}\xi\|_2 + \|r_p^{-\delta}\xi\|_4 + \|r_p^{-\delta}\nabla^A\xi\|_2 \lesssim \|r_p^{-\delta}\omega\|_2 + \lambda^{\delta'-1}\|r_A^{1-\delta-\delta'}\omega\|_2, \tag{10.14}$$

$$\begin{aligned}&\|r_p^{-\delta}\nabla^A\xi\|_4 + \|r_p^{-\delta}\nabla^A\nabla^A\xi\|_2\\ &\quad\lesssim \|r_p^{-\delta}(\mathcal{D}^A)^*\omega\|_2 + \lambda^{\delta'-1}\|r_A^{-\delta-\delta'}\omega\|_2 + \lambda^{\delta'-2}\|r_A^{1-\delta-\delta'}\omega\|_2.\end{aligned} \tag{10.15}$$

As a corollary of Lemma 10.1, (10.12)–(10.15), *if* $\delta_0 > 0$ *is sufficiently small and* $0 < \delta < \delta_0$, *then*

$$|G^A_\oplus \omega|(p) \le c(\delta) \cdot \{\textit{RHS of}\,(10.11)\} \quad \textit{and} \quad |\xi(p)| \lesssim c(\delta) \cdot \{\textit{RHS of}\,(10.15)\}. \tag{10.16}$$

We remark that in (10.11), (10.14) and (10.15) it is important that r_A appears where it does rather than r_p, or we would not get strong enough estimates in our applications. The fact that both r_A and r_p appear together in Proposition 10.2 complicates its proof.

Proof. A slightly less general set of bounds was derived in [6, Lemma 3.3] for Δ^A_+, the Laplacian on SD 2-forms only, but for the reasons mentioned prior to stating the proposition, essentially the same proof works here. The only differences are that (i) in [6] the decay (6.11) was true on all of N, not merely in $B(p_A, 2\epsilon)$, and (ii) Ref. [6] dealt only with the case $p = p_A$. Since the cited proof is rather long, we will not repeat the parts that require only minor modifications, and will jump to the points of departure.

To establish (10.10), the proof in [6, Lemma 3.3a] works verbatim to show that

$$\|r_p^{-\delta-1} G^A_\oplus \omega\|_2 + \|r_p^{-\delta} G^A_\oplus \omega\|_4 + \|r_p^{-\delta} \nabla G^A_\oplus \omega\|_2 \lesssim \|r_p^{-\delta+1} \omega\|_2. \tag{10.17}$$

Note that δ need not be positive here. Since $|G^A_\oplus \omega| \lesssim |r_p^{\delta-1} G^A_\oplus \omega|$, (10.10) follows.

Moving to (10.11), let $\eta \in \Omega^*(Ad\,P)$ be a form of arbitrary degree. The procedure in [6] for proving its Lemma 3.3b,c — squaring, integrating by parts, commuting a covariant derivative past a trace Laplacian $(\nabla^A)^*\nabla^A = \nabla^*\nabla$, and juggling terms — leads to

$$\begin{aligned} \|r_p^{-\delta} \nabla\eta\|_4 + \|r_p^{-\delta} \nabla\nabla\eta\|_2 &\le c(\|r_p^{-\delta} \nabla^*\nabla\eta\|_2 + \|r_p^{-\delta}\eta\|_2 \\ &\quad + \|r_p^{-\delta}\nabla\eta\|_2 + \|r_A^{\delta+\delta'} r_p^{-\delta} F_A\|_4 \|r_A^{-\delta-\delta'} \nabla\eta\|_2); \end{aligned} \tag{10.18}$$

here the smallness of $|\delta|$ has also been used to ensure that the term $|\delta| \|r_p^{-\delta-1}\nabla\eta\|_2$ that initially comes up on the right-hand side is $\lesssim |\delta| \|r_p^{-\delta}\nabla\nabla\eta\|_2$; see [6, Lemma 3.2]. (In [6], there was no need to insert $r_A^{\pm(\delta+\delta')}$.) First consider the case $\eta = G^A_\oplus \omega$, where $\omega \in \Omega^0(Ad\,P) \oplus \Omega^2_+(Ad\,P)$. The Weitzenböck formula gives $\Delta^A \eta = \omega + \mathcal{R}(G^A_\oplus \omega)$, where $\mathcal{R}$ is an endomorphism proportional to the Riemann tensor. Moreover, we will see in Lemma 10.5(b) that for δ, δ' sufficiently small, $\|r_A^{\delta+\delta'} r_p^{-\delta} F_A\|_4 \lesssim \lambda^{\delta'-1}$. Inserting these facts into (10.18), one can continue the argument as in [6] and arrive at an extended version of (10.11):

$$\begin{aligned} &\|r_p^{-\delta-1}\nabla G^A_\oplus \omega\|_2 + \|r_p^{-\delta}\nabla G^A_\oplus \omega\|_4 + \|r_p^{-\delta}\nabla\nabla G^A_\oplus \omega\|_2 \\ &\quad \lesssim \|r_p^{-\delta}\omega\|_2 + \lambda^{\delta'-1} \|r_A^{1-\delta-\delta'}\omega\|_2. \end{aligned} \tag{10.19}$$

As for (10.12), since $|\mathcal{D}^*\eta| \le c|\nabla\eta|$, the desired estimate follows from (10.17).

By similar manipulations, one can also establish that

$$\|r_A^{-\delta+1}\nabla G^A_\oplus \omega\|_4 + \|r_A^{-\delta+1}\nabla\nabla G^A_\oplus \omega\|_2 \lesssim \|r_A^{1-\delta}\omega\|_2. \tag{10.20}$$

Since $\mathcal{D}$ is ∇^A followed by a covariantly constant projection, the same bounds hold with $\nabla G^A_\oplus\omega$ replaced by $\mathcal{D}^*G^A_\oplus\omega = \xi$ yielding (10.13). For the same reason, (10.14) follows from (10.19).

Finally, to establish (10.15), return to (10.18) and use the Weitzenböck formula for 1-forms,

$$\nabla^*\nabla\xi = \Delta_1^A\xi + \mathcal{F}(\xi) + \mathcal{R}(\xi). \tag{10.21}$$

Here $\mathcal{F}$ is an endomorphism proportional to F. Since $\Delta_1^A\xi = \mathcal{D}^*\mathcal{D}(\mathcal{D}^*G^A_\oplus\omega) = \mathcal{D}^*\omega$ (see (10.9)), we have

$$\|r^{-\delta}\nabla^*\nabla\xi\|_2 \le c(\|r^{-\delta}\mathcal{D}^*\omega|_2 + \|r^{-\delta}\xi\|_2 + \|r_A^{\delta+\delta'}r^{-\delta}F\|_4\|r_A^{-\delta-\delta'}\xi\|_4). \tag{10.22}$$

Hence

$$\begin{aligned}\|r^{-\delta}\nabla\xi\|_4 + \|r^{-\delta}\nabla\nabla\xi\|_2 &\le c(\|r^{-\delta}(\mathcal{D}^A)^*\omega\|_2 + \|r^{-\delta}\xi\|_2 + \|r^{-\delta}\nabla\xi\|_2\\ &\quad + \|r_A^{\delta+\delta'}r^{-\delta}F^A\|_4(\|r_A^{-\delta-\delta'}\xi\|_4 + \|r_A^{-\delta-\delta'}\nabla\xi\|_2)).\end{aligned} \tag{10.23}$$

Once again $\|r_A^{\delta+\delta'}r^{-\delta}F\|_4 \lesssim \lambda^{\delta'-1}$, and (10.19) (with $\nabla G^A_\oplus\omega$ replaced by ξ) implies

$$(\|r_A^{-\delta-\delta'}\xi\|_4 + \|r_A^{-\delta-\delta'}\nabla\xi\|_2) \le c(\|r_A^{-\delta-\delta'}\omega\|_2 + \lambda^{-1+\delta''}\|r_A^{-\delta-\delta'-\delta''}\omega\|_2). \tag{10.24}$$

Using (10.19) and (10.20) to bound the other terms in (10.23), the bound (10.15) follows. □

Proposition 10.2 gives the same bounds for all $p \in N$; to obtain (9.6), we need estimates that show decay as $d = r_A(p)$ grows. The following proposition provides these estimates. We separate the estimates into cases (a) and (b) below because for many purposes the only ω's for which we need to estimate the quantities in Proposition 10.2 are compactly supported in a 2ϵ-ball around p_A, and we get sharper estimates in this case. Part (a) will thus be used to bound the terms $G_0^A R''(X, Y)$ and $G_0^A\{\tilde{X}, \xi_Y\}$ in $Rem_2(X, Y)$; part (b) will be used to bound $G_0^A\{\xi_X, \xi_Y\}$.

Proposition 10.3. *Notation as in Proposition* 10.2. *There exists* $\delta_0 > 0$ *such that the following are true.*

(a) *Suppose that for some* ϵ_0 (*not necessarily related to* $\epsilon = c\lambda^{1/2}$, *and allowed to depend on* ω), (i) $\mathrm{supp}(\omega) \subset B(p_A, \epsilon_0)$, (ii) $d = r_A(p) = \mathrm{dist}(p, p_A) \ge 2\epsilon_0$, *and* (iii) $|F^A| \le B$ *on the complement of* $\mathrm{supp}(\omega)$. *Let* $\tilde{\beta}$ *be a cut-off function of the form* $\beta_{\mathrm{std}}(4r_p/d)$ (*so that* $\mathrm{supp}(\tilde{\beta}) \subset B(p, d/2)$). *Then for any* δ' *with* $|\delta'| \le \delta_0$, *and any* $\delta \in (0, \delta_0)$,

$$|G^A_\oplus\omega|(p) \le c(\delta)d^{-1-\delta-\delta'}(1 + B^{1/2})\|r_A^{1+\delta'}\omega\|_2, \tag{10.25}$$

$$|\xi|(p) \le c(\delta)d^{-\delta-\delta'}(d^{-2} + B)\|r_A^{1+\delta'}\omega\|_2. \tag{10.26}$$

Thus if $\mathrm{supp}(\omega) \subset B(p_A, 2\epsilon)$ *and* $d \ge 4\epsilon = c\lambda^{1/2}$, *then using* (6.10),

$$|G^A_\oplus\omega|(p) \le c(\delta)d^{-1-\delta-\delta'}\|r_A^{1+\delta'}\omega\|_2, \tag{10.27}$$

$$|\xi|(p) \le c(\delta)d^{-2-\delta-\delta'}\|r_A^{1+\delta'}\omega\|_2. \tag{10.28}$$

(b) *Suppose only that* $|F^A| \leq B$ *on* $B(p, d/2)$, *where* $d = \mathrm{dist}(p, p_A) > 0$; *suppose nothing about the support of* ω. *Let* $\tilde{\beta}$ *be as in* (a). *Then for all* δ' *with* $|\delta'| \leq \delta_0$, *we have*

$$|G^A_\oplus \omega|(p) \leq c(\delta)(1 + B^{1/2})(d^{-1-\delta-\delta'} \|r_A^{1+\delta'}\omega\|_2 + \|r_p^{-\delta}\tilde{\beta}\omega\|_2). \tag{10.29}$$

Thus if $d \geq c\lambda^{1/2}$, *then*

$$|G^A_\oplus \omega|(p) \leq c(\delta)(d^{-1-\delta-\delta'} \|r_A^{1+\delta'}\omega\|_2 + \|r_p^{-\delta}\tilde{\beta}\omega\|_2). \tag{10.30}$$

Proof. (a) We will apply the Sobolev inequality (10.1), but first we must bound $\|r_p^{-\delta}\nabla\nabla(\tilde{\beta}G^A_\oplus\omega)\|_2$, $\|\tilde{\beta}G^A_\oplus\omega\|_2$, and similar expressions with $G^A_\oplus\omega$ replaced by ξ.

(i) First we will show that

$$\|r_p^{-\delta}\nabla\nabla(\tilde{\beta}G^A_\oplus\omega)\|_2 \lesssim d^{-1-\delta-\delta'}(1 + B^{1/2})\|r_A^{1+\delta'}\omega\|_2. \tag{10.31}$$

Let $\eta \in \Omega^0(Ad\,P) \oplus \Omega^2_+(Ad\,P)$. Proceed as in the proof of (10.11) — squaring, integrating by parts, etc. — but this time leave the term proportional to F (which arises from commuting ∇^A past a trace-Laplacian) in integrated form. One arrives at

$$\|r_p^{-\delta}\nabla\nabla\eta\|_2^2 \lesssim \|r_p^{-\delta}\Delta\eta\|_2^2 + \|r_p^{-\delta}\eta\|_2^2 + \|r_p^{-\delta}\nabla\eta\|_2^2 \int r_p^{-2\delta}|F||\nabla\eta|^2, \tag{10.32}$$

where $\Delta = (\nabla^A)^*\nabla^A$. Now replace η by $\tilde{\beta}\eta$. In the integral we have $|F| \leq B$, so

$$\|r_p^{-\delta}\nabla\nabla(\tilde{\beta}\eta)\|_2^2 \leq c(\|r_p^{-\delta}\Delta(\tilde{\beta}\eta)\|_2^2 + \|r_p^{-\delta}\tilde{\beta}\eta\|_2^2 + (1 + B)\|r_p^{-\delta}\nabla(\tilde{\beta}\eta)\|_2^2). \tag{10.33}$$

An integration by parts plus various steps already seen in the proof of Lemma 10.2 gives

$$\|r_p^{-\delta}\nabla(\tilde{\beta}\eta)\|_2^2 \lesssim \|r_p^{-\delta}\Delta(\tilde{\beta}\eta)\|_2\|r_p^{-\delta}\tilde{\beta}\eta\|_2 \leq c(k\|r_p^{-\delta}\Delta(\tilde{\beta}\eta)\|_2^2 + k^{-1}\|r_p^{-\delta}\tilde{\beta}\eta\|_2^2) \tag{10.34}$$

for arbitrary k. Inserting this into (10.33) with $k \lesssim (1 + B)^{-1}$, we find

$$\|r_p^{-\delta}\nabla\nabla(\tilde{\beta}\eta)\|_2^2 \lesssim \|r_p^{-\delta}\Delta(\tilde{\beta}\eta)\|_2^2 + (1 + B)\|r_p^{-\delta}\tilde{\beta}\eta\|_2^2. \tag{10.35}$$

Using the Weitzenböck formula as in the proof of Proposition 10.2, we can replace Δ by $\Delta^A_\oplus$, absorbing the zeroth-order term into $(1 + B)\|r_p^{-\delta}\tilde{\beta}\eta\|_2^2$. Additionally, by (10.17) we have $\|r_p^{-\delta}\tilde{\beta}\eta\|_2 \lesssim \|r_p^{-\delta}\Delta^A_\oplus(\tilde{\beta}\eta)\|_2$. Hence

$$\|r_p^{-\delta}\nabla\nabla(\tilde{\beta}\eta)\|_2 \lesssim (1 + B^{1/2})\|r_p^{-\delta}\Delta^A_\oplus(\tilde{\beta}\eta)\|_2. \tag{10.36}$$

Next, note that for any function f,

$$|\Delta^A_\oplus(f\eta) - f\Delta^A_\oplus\eta| \lesssim |\nabla\nabla f||\eta| + |\nabla f||\nabla\eta|. \tag{10.37}$$

Apply this with $f = \tilde{\beta}$ and $\eta = G^A_\oplus\omega$, noting that by the hypothesis on the support of ω we have $\tilde{\beta}\Delta^A_\oplus\eta = \tilde{\beta}\eta \equiv 0$. Since $|\nabla^j\tilde{\beta}| \leq cd^{-j}$, and since on the support of $\nabla\tilde{\beta}$ we have

both $\frac{1}{2}d \le r \le d$ and $\frac{1}{2}d \le r_A \le \frac{3}{2}d$, we obtain

$$|r_p^{-\delta}\Delta_{\oplus}^A(\tilde{\beta}\eta)| \le cr_p^{-\delta}\tilde{\chi}(d^{-2}|\eta| + d^{-1}|\nabla\eta|) \le cd^{-1-\delta-\delta'}\tilde{\chi}(r_A^{-1+\delta'}|\eta| + r_A^{\delta'}|\nabla\eta|), \tag{10.38}$$

where $\tilde{\chi}$ denotes the characteristic function of the annulus $\frac{1}{4}d \le r \le \frac{1}{2}d$. Inserting this into (10.36), we have

$$\begin{aligned}\|r_p^{-\delta}\nabla\nabla(\tilde{\beta}\eta)\|_2 &\le (1+B^{1/2})cd^{-1-\delta-\delta'}(\|\tilde{\chi}r_A^{-1+\delta'}\eta\|_2 + \|\tilde{\chi}r_A^{\delta'}\nabla\eta\|_2)\\ &\le (1+B^{1/2})cd^{-1-\delta-\delta'}(\|r_A^{-1+\delta'}\eta\|_2 + \|r_A^{\delta'}\nabla\eta\|_2).\end{aligned} \tag{10.39}$$

Now apply (10.17) to obtain

$$\|r_p^{-\delta}\nabla\nabla(\tilde{\beta}\eta)\|_2 \le (1+B^{1/2})cd^{-1-\delta-\delta'}\|r_A^{1+\delta'}\omega\|_2, \tag{10.40}$$

which leads to (10.31).

Moving on to $\|\tilde{\beta}G_{\oplus}^A\omega\|_2$, and repeating some of the steps in the proof of (a) with $\delta = 0$, we have

$$\begin{aligned}\|\tilde{\beta}G_{\oplus}^A\omega\|_2 &\lesssim \|\Delta_{\oplus}^A(\tilde{\beta}G_{\oplus}^A\omega)\|_2\\ &\lesssim d^{-1-\delta'}(\|\tilde{\chi}r_A^{-1+\delta'}G_{\oplus}^A\omega\|_2 + \|\tilde{\chi}r_A^{\delta'}\nabla G_{\oplus}^A\omega\|_2) \lesssim d^{-1-\delta'}\|r_A^{1+\delta'}\omega\|_2.\end{aligned} \tag{10.41}$$

This is smaller than the bound (10.31), so (10.1) gives (10.25).

The bound (10.26) is derived by methods similar to preceding ones and those used in Proposition 10.2. We leave the details to the reader.

(b) Proceed as in (a); the only change is that now we no longer have $\tilde{\beta}\omega \equiv 0$. The first effect of this change occurs in (10.38), where we have to add $|r^{-\alpha}\tilde{\beta}\omega|$ to the RHS. The effect of this term is to add $(1+B^{1/2})\|r^{-\alpha}\tilde{\beta}\omega\|_2$ to the RHS of (10.39) and (10.40), hence to (10.31). There is a similar change in the bound on $\|\tilde{\beta}G_{\oplus}^A\omega\|_2$, but its effect is smaller than the preceding one. □

To apply Propositions 10.2 and 10.3 to estimate Rem_2, we need to estimate expresssions of the form $\|r_p^m\omega\|_2$ for various m, where $\omega = Rem'_{2,\mathrm{loc}}$, $Rem'_{2,\mathrm{semiloc}}$, or $Rem'_{2,\mathrm{nonloc}}$ (see (10.3)). First we deal with the purely local object $Rem'_{2,\mathrm{loc}}(X,Y) = R''(X,Y)$. To start, we need a pointwise estimate given by the next lemma. The conclusion of the lemma is deceptively simple; the way in which the derivatives of X and Y are coupled to each other and to ∇F in the definition of Rem''_2 is crucial.

Lemma 10.4. *For $X, Y \in \mathfrak{h}_A$,*

$$|R''(X,Y)| \lesssim |\hat{X}||\hat{Y}|(\beta + \epsilon^{-2}\chi)(|F| + r_A|\nabla F|). \tag{10.42}$$

(*Here $\hat{X}$, $\hat{Y}$ are the un-cut-off versions of X, Y; see* (6.8).)

We remark that for general vector fields, this lemma would be false.

Proof. Let $\phi = R''(X, Y)$. From (8.4) we have $R''(X, Y) = \beta^2 R''(\hat{X}, \hat{Y}) +$terms involving the derivative of β. The latter are easily dealt with, giving the terms proportional to χ in (10.42). For $R''(\hat{X}, \hat{Y})$, the first three terms in (8.4) have norm bounded by $|F|(|\hat{X}||\hat{Y}| + |\Delta\hat{X}||\hat{Y}| + |\hat{X}||\Delta\hat{Y}|)$, and an easy computation shows that for $X \in \mathfrak{h}_A$, $|\Delta\hat{X}| \lesssim |\hat{X}|$. Furthermore, because F is an ASD (and hence Yang–Mills as well) and the "rotational" parts of $\hat{X}$, $\hat{Y}$ are SD, the remaining three terms in $R''(\hat{X}, \hat{Y})$ would vanish if the metric on N were Euclidean. When we do the bookkeeping necessary for the $O(r_A^2)$ difference between the metric coefficients g_{ij} and δ_{ij}, we obtain contributions bounded by $|\hat{X}||\hat{Y}|$ $(|F| + r_A|\nabla F|)$. □

Thus, bounding $|G_0^A R''(X, Y)|$ pointwise boils down to estimates of the form in the following lemma.

Lemma 10.5. *Let $p \in N$ be arbitrary and let $d = \mathrm{dist}(p, p_A)$. Then we have the following estimates.*

(a) *Assume $0 \le \delta < 2$ and $n > -2 + \delta$. Then*

$$\|\beta r_p^{-\delta} r_A^n F\|_2 + \|\beta r_p^{-\delta} r_A^{n+1} F\|_4 + \|\beta r_p^{-\delta} r_A^{n+1} \nabla F\|_2 \lesssim \begin{cases} \lambda^{n-\delta}, & n - \delta < 2, \\ \lambda^2 |\log \lambda|^{1/2}, & n - \delta = 2, \\ \lambda^{1+(n-\delta)/2}, & n - \delta > 2, \end{cases} \tag{10.43}$$

and for all n,

$$\|\chi r_p^{-\delta} r_A^n F\|_2 + \|\chi r_p^{-\delta} r_A^{n+1} \nabla F\|_2 \lesssim \lambda^{1-\delta+n/2}. \tag{10.44}$$

(b) *Let $\epsilon_0 > \lambda_0^{1/2}$ be some fixed number. For $0 < \delta < 1$ and $-1 + \delta < n < 3$, we have*

$$\|r_p^{-\delta} r_A^n F\|_4 \lesssim 1 + \begin{cases} \lambda^{n-1-\delta} & \text{if } d \lesssim \epsilon_0, \\ \lambda^{n-1} & \text{if } \epsilon_0 \lesssim d. \end{cases} \tag{10.45}$$

Proof. (a) First consider the case $\delta = 0$. From (6.11) one quickly finds

$$\|\beta r_A^n F\|_2 + \|\beta r_A^{n+1} F\|_4 \lesssim \begin{cases} \lambda^n, & -2 < n < 2, \\ \lambda^2 |\log \lambda|^{1/2}, & n = 2, \\ \lambda^{1+n/2}, & n > 2, \end{cases} \tag{10.46}$$

and for all n,

$$\|\chi r_A^n F\|_2 \lesssim \lambda^{1+n/2}. \tag{10.47}$$

As for $\|\beta r_A^{n+1} \nabla F\|$, the same argument as in the proof of [5, Lemma 3.3b] shows that

$$\|\beta r_A^{n+1} \nabla F\|_2^2 \lesssim n\|\beta r_A^n F\|_2^2 + \|\beta r_A^{n+1} F\|_2^2 + \||\mathrm{d}\beta| r_A^{n+1} F\|_2^2 + \int \beta^2 r_A^{2n+2} |F|^3. \tag{10.48}$$

Since $|\mathrm{d}\beta| \lesssim \epsilon^{-1}\chi$, we have $|\mathrm{d}\beta| r_A^{n+1} \leq c\epsilon^n \chi$. Thus

$$\|\beta r_A^{n+1}\nabla F\|_2 \lesssim \|\beta r_A^n F\|_2 + \lambda^{n/2}\|\chi F\|_2 + \left(\int \beta^2 r_A^{2n+2}|F|^3\right)^{1/2}. \tag{10.49}$$

>From (6.11), one can deduce that

$$\left(\int \beta^2 r_A^{2n+2}|F|^3\right)^{1/2} \lesssim \begin{cases} \lambda^n, & n<3, \\ \lambda^3|\log\lambda|^{1/2}, & n=3, \\ \lambda^{3n/2-3/2}, & n>3. \end{cases} \tag{10.50}$$

Combining this with our previous bounds, we find that

$$\|\beta r_A^{n+1}\nabla F\|_2 \lesssim \text{RHS of (10.46)}. \tag{10.51}$$

To bound $\|\chi r_A^{n+1}\nabla F\|_2$, again use the analysis leading to (10.48), but with β replaced by a smooth extension of χ of the form $f(r_A/\epsilon)$ with f supported in $[\frac{1}{2}, 3]$. (It is simplest first to note that since $r_A \leq 2\epsilon$ on $\mathrm{supp}(\chi)$, $\|\chi r_A^{n+1}\nabla F\|_2 \lesssim \lambda^{(n+1)/2}\|\chi\nabla F\|_2$.) Then analysis similar to the above leads to

$$\|\chi r_A^{n+1}\nabla F\|_2 \lesssim \text{RHS of (10.47)}. \tag{10.52}$$

This completes the case $\delta = 0$ and we move on to the general case.

We first bound $\|\beta r_p^{-\delta} r_A^n F\|_2$; the method for bounding $\|\beta r_p^{-\delta} r_A^{n+1} F\|_4$ is identical. Break the ball $B(p_A, 2\epsilon)$ into two pieces: an inner region $B_{\text{in}} = B(p, \frac{1}{2}d) \cap B(p_A, 2\epsilon)$ and an outer region $B_{\text{out}} = B(p_A, 2\epsilon) - B_{\text{in}}$. On B_{out}, we have $r_p \geq \frac{1}{2}d$, and hence $r_A/r_p \leq (r_p + d)/r_p \leq 3$. Thus

$$r_p^{-\delta} r_A^n |F| = (r_A/r_p)^\delta r_A^{n-\delta}|F| \lesssim r_A^{n-\delta}|F|, \tag{10.53}$$

implying

$$\begin{aligned} \|\beta r_p^{-\delta} r_A^n F\|_{L^2(B_{\text{out}})} &\lesssim \|\beta r_A^{n-\delta} F\|_{L^2(B_{\text{out}})} \lesssim \|\beta r_A^{n-\delta} F\|_{L^2(B(p_A,2\epsilon))} \\ &\lesssim \begin{cases} \lambda^{n-\delta}, & -2 < n-\delta < 2, \\ \lambda^2|\log\lambda|^{1/2}, & n-\delta = 2, \\ \lambda^{1+(n-\delta)/2}, & n-\delta > 2. \end{cases} \end{aligned} \tag{10.54}$$

For the integral over B_{in}, first suppose $n - \delta \leq 2$ and separately consider the cases $d \leq \lambda$, $d \geq \lambda$. In both cases note that $\frac{1}{2}d \leq r_A \leq \frac{3}{2}d$ in this region. When $d \leq \lambda$, we then have $r_A \lesssim \lambda$ and $|F| \lesssim \lambda^{-2}$ on B_{in}, so

$$\|\beta r_p^{-\delta} r_A^n F\|_{L^2(B_{\text{in}})} \lesssim \lambda^{n-2}\|r_p^{-\delta}\|_{L^2(B_{\text{in}})} \lesssim \lambda^{n-2} d^{2-\delta} \lesssim \lambda^{n-\delta}. \tag{10.55}$$

On the other hand, if $d \geq \lambda$, then since r_A/d is bounded above and below on B_{in}, (6.11) implies $r_A^n|F| \lesssim \lambda^2 r_A^{n-4} \lesssim \lambda^2 d^{n-4}$. Hence

$$\|\beta r_p^{-\delta} r_A^n F\|_{L^2(B_{\text{in}})} \lesssim \lambda^2 d^{n-4}\|r_p^{-\delta}\|_{L^2(B_{\text{in}})} \lesssim \lambda^2 d^{n-\delta-2} \leq \lambda^{n-\delta}, \tag{10.56}$$

since $n-\delta \le 2$. Combining this with the estimate for B_{out}, we obtain the top two lines of (10.43).

If $n-\delta > 2$, separately consider the cases $d \le 4\epsilon$ and $d \ge 4\epsilon$. If $d \le 4\epsilon$, the procedure for the case $\delta \ge \lambda$ above yields

$$\|\beta r_p^{-\delta} r_A^n F\|_{L^2(B_{\text{in}})} \lesssim \lambda^{n-2} \|r_p^{-\delta}\|_{L^2(B_{\text{in}})} \lesssim \lambda^2 d^{n-\delta-2} \lesssim \lambda^{1+(n-\delta)/2}, \tag{10.57}$$

the same bound as on B_{in}. If $d \ge 4\epsilon$ then on the support of β we have $r_p \ge \epsilon$, so $r_p^{-\delta} r_A^n |F| \lesssim \lambda^{-\delta/2} r_A^n |F|$. Thus (10.46) yields the remaining case of (10.43) for the bound on $\|r_p^{-\delta} r_A^n F\|_2$.

The method for bounding $\|\beta r_p^{-\delta} r_A^{n+1} \nabla F\|$ is essentially identical to the method for bounding $\|\beta r_p^{-\delta} r_A^{n+1} \nabla F\|_2$, except that for the estimates over B_{in}, first multiply by a cut-off function of the form $\beta_{\text{std}}(2r_p/d)$, and then integrate by parts as in (10.48).

To bound $\|\chi r_p^{-\delta} r_A^n F\|_2$, note that on $\text{supp}(\chi)$ we have $|F| \le$ const. and $r_A \le \epsilon$, so

$$\|\chi r_p^{-\delta} r_A^n F\| \lesssim \lambda^{n/2} \|\chi r_p^{-\delta}\|_2 \lesssim \lambda^{n/2} \|r_p^{-\delta}\|_{L^2(B(p,2\epsilon))} \lesssim \lambda^{1+(n-\delta)/2}. \tag{10.58}$$

Similarly $\|\chi r_p^{-\delta} r_A^n \nabla F\| \lesssim \lambda^{n/2} \|\chi r_p^{-\delta} \nabla F\|_2$, and the same procedure as for $\delta = 0$ completes the work.

(b) First write

$$\|r_p^{1-\delta} r_A^n F\|_4 \le \|(1-\tilde{\beta}) r_p^{1-\delta} r_A^n F\|_4 + \|\tilde{\beta} r_p^{1-\delta} r_A^n F\|_4, \tag{10.59}$$

where $\tilde{\beta}$ is a cut-off of scale ϵ_0 centered at p_A. On the support of $1-\tilde{\beta}$ we have $|F| \le$ const., so the first term on the RHS is bounded by a constant. The second term can be estimated as in the proof of (b). □

We are now in a position to bound $|G_0^A Rem'_{2,\text{loc}}|$ pointwise, but we postpone this until we have collected the estimates needed to bound the semi-local and nonlocal contributions to $G_0^A Rem'_2$. These require bounds on norms of $\xi_X = \mathcal{D}^* G_\oplus^A \mathcal{D}\tilde{X}$, which in turn require pointwise bounds on $\mathcal{D}\tilde{X}$.

Lemma 10.6. *For any vector field X on N, and any ASD connection A, we have the pointwise formulas*

$$(d^A)^* \iota_X F_A = \langle d_+ X^*, F_A \rangle, \tag{10.60}$$

$$d_+^A(\iota_X F_A) = \text{Sym}_0^2(\nabla X^*) \sharp F_A, \tag{10.61}$$

where X^* is the metric dual of X, $d_+ X^*$ is the self-dual part of dX^*, $\text{Sym}_0^2(T)$ denotes the traceless symmetric part of a rank-two tensor field $T \in \Gamma(T^*N \otimes T^*N)$, and in a local orthonormal basis θ^i of the cotangent bundle, $\langle T, F \rangle = \frac{1}{2} T_{ij} F_{ij} \in$ and $T \sharp F = T_{ij} F_{jk} \theta^i \wedge \theta^k$. Hence

$$|\mathcal{D}^A(\iota_X F_A)| \lesssim (|d_+ X^*| + |\text{Sym}_0^2(\nabla X^*)|) |F_A|, \tag{10.62}$$

$$|\nabla^A \mathcal{D}^A(\iota_X F_A)| \le c((|\nabla(d_+ X^*)| + |\nabla(\mathrm{Sym}_0^2(\nabla X^*))|)|F_A| + (|d_+ X^*| + |\mathrm{Sym}_0^2(\nabla X^*)|)|F_A|). \tag{10.63}$$

Hence if $X \in \mathfrak{h}_A$, *then*

$$|\mathcal{D}^A \tilde{X}| \lesssim (r_A \beta + \epsilon^{-1}\chi)|\hat{X}||F_A|, \tag{10.64}$$

$$\|(\mathcal{D}^A)^* \mathcal{D}^A \tilde{X}| \le c|\nabla^A \mathcal{D}^A \tilde{X}| \lesssim (\beta + \epsilon^{-2}\chi)|\hat{X}|(|F_A| + r_A|\nabla^A F_A|). \tag{10.65}$$

Proof. Using the facts that $d^* = - * d *$, $d^A F = 0$, and $*F = -F$, we have

$$(d^A)^*(\iota_X F) = - * d^A(*\iota_X(*F)) = *d^A(X^* \wedge F) = *(dX^* \wedge F) = (dX^*, F); \tag{10.66}$$

this gives (10.60). Now fix $p \in N$. Calculating in a local orthonormal frame $\{e_i\}$ of TN and dual coframe $\{\theta^i\}$ with $\nabla e_i|_p = 0$,

$$\begin{aligned} d_+^A(i_X F) &= p_+\left(\sum \theta^i \wedge \iota_{\nabla_i X} F\right) \\ &= \sum (\nabla_i X_j) p_+(\theta^i \wedge \iota_{e_j} F) = p_+(\mathrm{Sym}^2(\nabla X) \sharp F) \end{aligned} \tag{10.67}$$

by Lemma 2.3 of [5]. Since for any symmetric 2-tensor T, the pure-trace part of T yields a *self*-dual 2-form under the operation $\sharp F$, we may replace Sym^2 by Sym_0^2 in (10.67), and by simple representation theory, the p_+ in (10.67) is redundant.

Eqs. (10.64) and (10.65) follow from Lemma 10.6 and a pointwise computation of $d_+ X^*$, $\mathrm{Sym}_0^2(\nabla X^*)$ that we leave to the reader. □

Corollary 10.7. *For all* $X \in \mathfrak{h}_A$, *and all* $p \in N$, *the elements* $\tilde{X} \in \mathcal{H}_A$ *satisfy the following integral bounds:*

(a) *If* $-1 < m < 2$, *then*

$$\|r_A^m \tilde{X}\|_4 \le c(m)\lambda^{m-1}. \tag{10.68}$$

(b) *If* $-1 < m \le 0$, *or if* $d = \mathrm{dist}(p, p_A) \lesssim \epsilon$ *and* $-1 < m < 2$, *then*

$$\|r_p^m \mathcal{D}^A \tilde{X}\|_2 \le c(m)\lambda^{m/2}(b \cdot \lambda^{1/2} + a). \tag{10.69}$$

(c) *For all* $\delta \in (0, 1)$,

$$\|r_p^{-\delta}(\mathcal{D}^A)^* \mathcal{D}^A \tilde{X}\|_2 \le c\|r_p^{-\delta} \nabla^A \mathcal{D}^A \tilde{X}\|_2 \lesssim \lambda^{-\delta}(b + a \cdot \lambda^{-1/2}). \tag{10.70}$$

Proof. Using (10.64) and (10.65) plus $|X| \le b + a\lambda^{-1} r_A$, most of these bounds follow directly from Lemma 10.5. The exception is (10.69) in the case $m > 0$, for which one must also use the triangle inequality $r_p \le r_A + d \lesssim r_A + \lambda^{m/2}$. □

We are now in a position to derive our final estimates on the norms of ξ needed to bound $|G_0^A Rem'_{2,\mathrm{semiloc}}|$ and $|G_0^A Rem'_{2,\mathrm{nonloc}}|$ pointwise. We also use the opportunity to prove (6.13).

Proposition 10.8. *There exists $\delta_0 > 0$ such that if $0 < \delta < \delta_0$ and $0 \le \delta' < \delta_0$, then for $\xi = (\mathcal{D}^A)^* G^A_\oplus \mathcal{D}^A \tilde{X}$ (with $X \in \mathfrak{h}_A$) we have the following.*

(a) *If $0 < \delta < \delta_0$, then*

$$|\xi(p)| \lesssim \lambda^{\delta'}(b \cdot \lambda^{-1} + a \cdot \lambda^{-3/2}). \tag{10.71}$$

If furthermore $d = r_A(p) \ge 4\epsilon = c\lambda^{1/2}$, then

$$|\xi(p)| \le c(\delta) r_A(p)^{-2-\delta-\delta'} \lambda^{\delta'}(b \cdot \lambda + a \cdot \lambda^{1/2}). \tag{10.72}$$

(b)

$$\|\xi\|_2 \lesssim \lambda^{\delta'/2}(b \cdot \lambda + a \cdot \lambda^{1/2}). \tag{10.73}$$

Since $\xi_X = \tilde{X} - \pi_A \tilde{X}$, this implies (6.13).

(c) *If $0 \le \delta \le \delta_0$, then*

$$\|r_p^{-\delta} \xi\|_4 \lesssim \lambda^{\delta'}(b + a \cdot \lambda^{-1/2}). \tag{10.74}$$

(d) *If $|\delta| < \delta_0$, then*

$$\|r_A^{1+\delta} \xi\|_4 \lesssim \lambda^{\delta/2}(b \cdot \lambda + a \cdot \lambda^{1/2}). \tag{10.75}$$

Proof. (a) We will omit writing the δ-dependence of the constants. From (10.15) given δ, δ_0 as above, there exists $\delta' > 0$ such that

$$|\xi(p)| \lesssim r_p^{-\delta}(d_+^A)^* \omega\|_2 + \lambda^{2\delta'-1}\|r_A^{-\delta-2\delta'}\omega\|_2 + \lambda^{2\delta'+\delta-2}\|r_A^{1-2\delta-2\delta'}\omega\|_2, \tag{10.76}$$

where $\omega = \mathcal{D}^A \tilde{X}$. Using Corollary 10.7, we compute

$$\lambda^{2\delta'-1}\|r_A^{-\delta-2\delta'}\omega\|_2 + \lambda^{2\delta'+\delta-2}\|r_A^{1-2\delta-2\delta'}\omega\|_2 \lesssim \lambda^{\delta'}(b \cdot \lambda^{-1} + a \cdot \lambda^{-3/2}). \tag{10.77}$$

The bound on $\|r_p^{-\delta}(d_+^A)^*\omega\|_2$ from Corollary 10.7 is smaller than this, so we obtain (10.71).

For (10.72), apply (10.28) and Corollary 10.7.

(b)–(d). Apply Proposition 10.2 and Corollary 10.7. □

We remark that by using the pointwise decay estimate (10.76) one can obtain the weighted L^4 decay

$$\|r_p^{-\delta}\|_4 \lesssim \lambda^{-\delta} r_A(p)^{-\delta}(b + a \cdot \lambda^{-1/2}) \tag{10.78}$$

for $d \ge c\lambda^{1/2}$, but this is of no help to us.

We are now ready to collate all the estimates needed to prove Proposition 9.2.

Corollary 10.9. (a) *There exist $\delta_0 > 0$, $\delta' > 0$ such that for $0 \le \delta < \delta_0$, the following are true.*

$$\|r_p^{-\delta} Rem'_{2,\mathrm{loc}}(X, Y)\|_2 \lesssim \lambda^{-\delta/2}(b^2 + ba \cdot \lambda^{-1/2} + a^2 \cdot \lambda^{-1}), \tag{10.79}$$

$$\|r_A^{1\pm\delta} Rem'_{2,\mathrm{loc}}(X,Y)\|_2 \lesssim \lambda^{1/2\pm\delta/2}(b^2 + ba\cdot\lambda^{-1/2} + a^2\cdot\lambda^{-1}), \tag{10.80}$$

$$\|r_p^{-\delta} Rem'_{2,\mathrm{semiloc}}\|_2 \lesssim \lambda^{-1+\delta'}(b^2 + ba\cdot\lambda^{-1/2} + a^2\cdot\lambda^{-1/2}), \tag{10.81}$$

$$\|r_A^{1\pm\delta} Rem'_{2,\mathrm{semiloc}}\|_2 \lesssim \lambda^{\delta'}(b^2 + ba\cdot\lambda^{-1/2} + a^2\cdot\lambda^{-1/2}), \tag{10.82}$$

$$\|r_p^{-\delta} Rem'_{2,\mathrm{nonloc}}(X,Y)\|_2 \lesssim \lambda^{\delta'}(b^2 + ba\cdot\lambda^{-1/2} + a^2\cdot\lambda^{-1}), \tag{10.83}$$

$$\|r_A^{1\pm\delta} Rem'_{2,\mathrm{nonloc}}(X,Y)\|_2 \lesssim \lambda^{1+\delta'}(b^2 + ba\cdot\lambda^{-1/2} + a^2\cdot\lambda^{-1}). \tag{10.84}$$

(b) *There exists $\delta' > 0$ such that for all $p \in N$, the following are true.*

$$|G_0^A Rem'_{2,\mathrm{loc}}(X,Y)|(p) \lesssim \lambda^{-1/2+\delta'}(b^2 + ba\cdot\lambda^{-1/2} + a^2\cdot\lambda^{-1}), \tag{10.85}$$

$$|G_0^A Rem'_{2,\mathrm{semiloc}}(X,Y)|(p) \lesssim \lambda^{-1+\delta'}(b^2 + ba\cdot\lambda^{-1/2} + a^2\cdot\lambda^{-1/2}), \tag{10.86}$$

$$|G_0^A Rem'_{2,\mathrm{nonloc}}(X,Y)|(p) \lesssim \lambda^{\delta'}(b^2 + ba\cdot\lambda^{-1/2} + a^2\cdot\lambda^{-1}). \tag{10.87}$$

(c) *There exists $\delta_0 > 0, \delta' > 0$ such that if $0 < \delta < \delta_0$ and $d \geq 4\epsilon = c\lambda^{1/2}$, the following are true.*

$$|G_0^A Rem'_{2,\mathrm{loc}}(X,Y)|(p) \lesssim d^{-1-\delta-\delta'}\lambda^{\delta'}(b^2\cdot\lambda^{1/2} + ba + a^2\cdot\lambda^{-1/2}), \tag{10.88}$$

$$|G_0^A Rem'_{2,\mathrm{semiloc}}(X,Y)|(p) \lesssim d^{-1-\delta-\delta'}\lambda^{\delta'}(b^2 + ba\cdot\lambda^{-1/2} + a^2\cdot\lambda^{-1/2}), \tag{10.89}$$

$$|G_0^A Rem'_{2,\mathrm{nonloc}}(X,Y)|(p) \lesssim d^{-1-\delta-\delta'}\lambda^{\delta'}(b^2\cdot\lambda + ba\cdot\lambda^{1/2} + a^2). \tag{10.90}$$

Proof. (a) These bounds follow directly from Lemma 10.4, Corollary 10.7, the L^4 bounds in Proposition 10.8, and Hölder's inequality.

(b) Use part (a) and Proposition 10.3.

(c) Since $Rem'_{2,\mathrm{loc}}(X,Y)$ and $Rem'_{2,\mathrm{semiloc}}(X,Y)$ are supported in $B(p_A, 2\epsilon)$, for these terms we can apply (10.27) and the corresponding bounds in (a). As $Rem'_{2,\mathrm{semiloc}}(X,Y)$ is not locally supported, we appeal instead to (10.30):

$$|G_0^A\{\xi,\xi\}|(p) \lesssim (d^{-1-\delta-\delta'}\|r_A^{1+\delta'}\{\xi,\xi\}\|_2 + \|r_p^{-\delta}\tilde{\beta}\{\xi,\xi\}\|_2), \tag{10.91}$$

where $\tilde{\beta}$ is a cut-off of scale $\frac{1}{2}d$ as in Lemma 10.4.

If we estimate $\|r_A^{1+\delta'}\{\xi,\xi\}\|_2$ using (10.84), we obtain the right-hand side of (10.90). Were we next to estimate $\|r_p^{-\delta}\tilde{\beta}\{\xi,\xi\}\|_2$, the resulting bound would be too large to be of use. Instead, since $d \leq r_A \leq 3d$ on the support of $\tilde{\beta}$, we can use the pointwise bound (10.72) to find

$$\begin{aligned}\|r_p^{-\delta}\tilde{\beta}\{\xi,\xi\}\|_2 &\lesssim \|r_p^{-\delta}\tilde{\beta}\|_2(d^{-2-\delta-\delta'}\lambda\delta'(b\cdot\lambda + a\cdot\lambda^{1/2}))^2\\ &\lesssim d^{-2-3\delta-2\delta'}\lambda^{2\delta'}(b^2\cdot\lambda^2 + ba\cdot\lambda^{3/2} + a^2\cdot\lambda)\\ &\lesssim d^{-1-\delta-\delta'}\lambda^{3/2-\delta''}(b^2 + ba\cdot\lambda^{-1/2} + a^2\cdot\lambda^{-1}),\end{aligned} \tag{10.92}$$

which is much smaller than our bound on $\|r_A^{1+\delta'}\{\xi,\xi\}\|_2$. Thus (10.90) follows. □

Finally, we have the following proof.

Proof of Proposition 9.2. (a) Add the bounds (10.85)–(10.87) to obtain (9.5). If we add the bounds (10.88)–(10.90) we obtain a stronger bound than (9.6):

$$|Rem_2(X, Y)| \lesssim r_A^{-1-\delta'} \lambda^{\delta'} (b^2 + ba \cdot \lambda^{-1/2} + a^2 \cdot \lambda^{-1/2}). \tag{10.93}$$

(b) In the proof of part (a), the only way in which ξ entered was through the L^4 bounds on $\|r_p^{-\delta}\xi_4\|_4$, $\|r_A^{1+\delta'}\xi_4\|_4$, and the pointwise decay (10.72). Hence our assertion follows from the hypothesis (Z5) of Section 7. □

Acknowledgements

The authors thank the 1994 Park City Mathematics Institute, where this work was begun, the National Science Foundation, and the Texas Advanced Research Program. We also thank Dan Freed, Tom Parker, Cliff Taubes, and Karen Uhlenbeck for helpful insights and criticism, and Margaret Combs for assistance with the figures.

Appendix A

The point of the following weighted Sobolev inequality is that on an m-dimensional manifold there is no Sobolev embedding $L_1^m \hookrightarrow L^\infty$, but the failure is borderline. Thus by introducing an arbitrarily small weight into the Sobolev norm, we are able to obtain an embedding.

Lemma A.1. *Let $E \to N$ be a Riemannian vector bundle with metric-compatible connection ∇, where N is compact, Riemannian, and m-dimensional ($m > 1$). Given $p \in N$ and $R_2 > R_1 > 0$, let $\Omega(p; R_1, R_2)$ denote the annulus $\{R_1 \le r_p \le R_2\}$, where r_p is the distance to p. There exists a constant c, independent of ∇, such that for any $\delta > 0$, $R_2 > R_1 > 0$ (but smaller than the injectivity radius), any $p \in N$, and any $\phi \in \Gamma(E)$, we have*

$$|\phi(p)| \le c\delta^{-(1-1/m)} R_2^\delta \left(\frac{1}{R_2 - R_1} \|r_p^{-\delta}\phi\|_{L^m(\Omega(p;R_1,R_2))} + \|r_p^{-\delta}\nabla\phi\|_{L^m(B_{R_2}(p))} \right). \tag{A.1}$$

Consequently,

$$|\phi(p)| \le \delta^{-(1-1/m)} (\|\phi\|_{L^m(N)} + \|r_p^{-\delta}\nabla\phi\|_{L^m(N)}), \tag{A.2}$$

$$\|\phi\|_{L^\infty(N)} \le c\delta^{-(1-1/m)} (\|\phi\|_{L^m(N)} + \sup_{p\in N} (\|r_p^{-\delta}\nabla\phi\|_{L^m(N)})). \tag{A.3}$$

Proof. By Kato's inequality, it suffices to prove this for the trivial real line bundle, i.e. for functions on N.

First replace N by $\mathbf{R}^m$ and consider a compactly supported function $f \in C_0^\infty(B(0,R))$. Let $\theta \in S^{m-1}$. Then, using polar coordinates on $B(0,R)$, we have

$$|f(0)| = \left| \int_0^R \frac{\partial f}{\partial r}(r,\theta)\,\mathrm{d}r \right| \leq \int_0^R |\nabla f|(r,\theta)\,\mathrm{d}r, \tag{A.4}$$

implying

$$\mathrm{Vol}(S^{m-1})|f(0)| \leq \int_{S^{m-1}} \mathrm{d}\theta \left(\int_0^R |\nabla f|(r,\theta)\,\mathrm{d}r \right) = \int_{B(0,R)} |\nabla f| \cdot r^{1-m}\,\mathrm{d}vol. \tag{A.5}$$

Applying the same argument on a normal-coordinate ball $B(p,R)$ in N (where $f \in C_0^\infty(B(p,R))$), using the compactness of N to get uniformity in the constants below, we obtain

$$\begin{aligned} |f(p)| &\leq c\int_{B(p,R)} |\nabla f| r_p^{1-m}\,\mathrm{d}vol = c\int_{B(p,R)} r^{-\delta}|\nabla f| r_p^{1-m+\delta}\,dvol \\ &\leq c\|r_p^{-\delta}\nabla f\|_{L^m(B(p,R))}\|r_p^{1-m+\delta}\|_{L^{m/(m-1)}(B(p,R))} \\ &\leq c\delta^{-(1-1/m)}R^\delta \|r_p^{-\delta}\nabla f\|_{L^m(B(p,R))}. \end{aligned}$$

Now remove the assumption that f is supported inside a normal coordinate ball. Replace f in the preceding argument by $\beta(r)f$, where β is a cut-off function identically 1 for $r \leq R_1$ and vanishing for $r \geq R_2$; thus $|\nabla\beta| \leq c/(R_2 - R_1)$. We then have

$$\begin{aligned} |f(p)| &\leq c\delta^{-(1-1/m)}R_2^\delta \|r_p^{-\delta}\nabla(\beta f)\|_{L^m} \\ &\leq c\delta^{-(1-1/m)}R_2^\delta (\|r_p^{-\delta}(\nabla\beta)f\|_{L^m} + \|\beta r_p^{-\delta}\nabla f\|_{L^m}) \\ &\leq c\delta^{-(1-1/m)}R_2^\delta \left(\frac{1}{R_2 - R_1}\|r_p^{-\delta} f\|_{L^m(\Omega(R_1,R_2,p))} + \|r_p^{-\delta}\nabla f\|_{L^m(B_{R_2}(p))} \right), \end{aligned} \tag{A.7}$$

yielding (A.1). Taking $R_2 = 2R_1$ to be, say, half the injectivity radius of N, we obtain (A.3). □

As a corollary, we have the following.

Corollary A.2. *Let E, N, ∇ be as in Lemma* A.1, *and assume* $\dim N = 4$. *Then for all $\delta \in (0,1)$, there exist constants $c(\delta)$,* independent of ∇, *such that for all $\phi \in \Gamma(E)$,*

$$|\phi(p)| \leq c(\delta)(\|\phi\|_{L^2(N)} + \|r_p^{-\delta}\nabla\nabla\phi\|_{L^2(N)}), \tag{A.8}$$

and hence

$$\|\phi\|_\infty \leq c(\delta)\sup_{p\in N}(\|\phi\|_2 + \|r_p^{-\delta}\nabla\nabla\phi\|_2). \tag{A.9}$$

Proof. Applying Lemma (A.1) with $m = 4$, we have

$$|\phi(p)| \leq c(\delta)(\|\phi\|_4 + \|r_p^{-\delta}\nabla\phi\|_4). \tag{A.10}$$

Using the Sobolev embedding $L_1^2(N) \hookrightarrow L^4(N)$, we then find

$$|\phi(p)| \leq c(\delta)(\|\phi\|_2 + \|r_p^{-\delta-1}\nabla\phi\|_2 + \|r_p^{-\delta}\nabla\nabla\phi\|_2). \tag{A.11}$$

But since $\delta < 1$, we also have the weighted Sobolev inequality of Lemma 3.1 of [6]:

$$\|r^{-1-\delta}\psi\|_2 \leq c(\|r^{-\delta}\psi\|_2 + \|r^{-\delta}\nabla\psi\|_2) \tag{A.12}$$

(the proof is again a polar-coordinate computation). Using this we can bootstrap (A.11) into the form (A.8). □

References

[1] S.K. Donaldson, Connections, cohomology and the intersection forms of four manifolds, J. Diff. Geom. 24 (1986) 275–341.
[2] S.K. Donaldson, Polynomial invariants for smooth 4-manifolds, Topology 29 (1990) 257–315.
[3] S.K. Donaldson, The orientation of Yang–Mills moduli spaces and 4-manifold topology, J. Diff. Geom. 26 (1987) 397–428.
[4] S.K. Donaldson, P. Kronheimer, The Geometry of Four-Manifolds, Oxford University Press, Oxford, 1990.
[5] D. Groisser, Curvature of Yang–Mills moduli spaces near the boundary. I, Commun. Anal. Geom. 1 (1993) 139–216.
[6] D. Groisser, Totally geodesic boundaries of Yang–Mills moduli spaces, Houston J. Math. 24 (1998) 221–276.
[7] L. Göttsche, Modular forms and Donaldson invariants for 4-manifolds with $b_+ = 1$, J. Am. Math. Soc. 9 (1996) 827–843.
[8] D. Groisser, T.H. Parker, The geometry of the Yang–Mills moduli space for definite manifolds, J. Diff. Geom. 29 (1989) 499–544.
[9] D. Groisser, T.H. Parker, Sharp decay estimates for Yang–Mills fields, Commun. Anal. Geom. 5 (1997) 439–474.
[10] D. Groisser, T.H. Parker, Differential forms on the Yang–Mills moduli space, Unpublished notes.
[11] D. Kotschick, P. Lisca, Instanton invariants of $\mathbf{C}P^2$ via topology, Math. Ann. 303 (1995) 345–371.
[12] P. Kronheimer, T. Mrowka, Embedded surfaces and the structure of Donaldson's polynomial invariants, J. Diff. Geom. 41 (1995) 573–734.
[13] L. Sadun, A simple geometric representative for μ of a point, Commun. Math. Phys. 178 (1996) 107–113.
[14] L. Sadun, Simple type is not a boundary phenomenon, in: Apanasov, et al. (Eds.), Geometry, Topology and Physics, Proceedings of the First Brazil–USA Workshop, Walter de Gruyter, Berlin, 1997.
[15] N. Seiberg, E. Witten, Electric–magnetic duality, monopole condensation and confinement in $N = 2$ supersymmetric Yang–Mills theory, Nucl. Phys. B 426 (1994) 19.
[16] C.H. Taubes, Metrics, Connections and Gluing Theorems, CBMS Regional Conference Series in Mathematics, Vol. 89, American Mathematical Society, Providence, RI, 1996.
[17] E. Witten, Topological quantum field theory, Commun. Math. Phys. 117 (1988) 353.
[18] E. Witten, Monopoles and 4-manifolds, Math. Res. Lett. 1 (1994) 769–796.