Hitchin's equations and M-theory phenomenology

Tony Pantev [a,*], Martijn Wijnholt [b]

[a] *Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104-6395, USA*
[b] *Max Planck Institute (Albert Einstein Institute), Am Mühlenberg 1, D-14476 Potsdam-Golm, Germany*

ABSTRACT

Phenomenological compactifications of M-theory involve seven-manifolds with G_2 holonomy and various singularities. Here we study local geometries with such singularities, by thinking of them as compactifications of $7d$ supersymmetric Yang–Mills theory on a three-manifold Q_3. We give a general discussion of compactifications of $7d$ Yang–Mills theory in terms of Higgs bundles on Q_3. We show that they can be constructed using spectral covers, which are Lagrangian branes with a flat connection in the cotangent bundle T^*Q_3. We explain the dictionary with ALE fibrations over Q_3 and conjecture that these configurations have G_2 holonomy. We further develop tools to study the low energy effective theory of such a model. We show that the naive massless spectrum is corrected by instanton effects. Taking the instanton effects into account, we find that the massless spectrum and many of the interactions can be computed with Morse theoretic methods.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

String theory vacua are explicit realizations of the ideas of Kaluza and Klein on extra dimensions. As such, geometric structures inevitably play a central role in studying such vacua. Geometric techniques are widely used for phenomenological model building in the heterotic string.

In the past few years, we have learnt to start applying techniques from geometric engineering to model building problems in type II settings. Using ideas about exceptional collections, this led to the construction of 'local models' in which the force of gravity may be treated as a small perturbation [1–3]. More recently, it has led to the first new class of models since the appearance of the heterotic string that can successfully explain unification, namely F-theory GUTs [4–6]. The F-theory models allow for scenarios not available in the heterotic string, such as gauge mediation, and hence may display some strikingly different signatures. Since they have been relatively little studied, it is desirable to develop the phenomenology and mathematics of these models in more detail. A number of interesting papers have recently appeared on this topic.

In this project, we follow a slightly different direction. In light of the line of research mentioned above, it is natural to ask if a similar set of ideas based on geometric engineering may also be used to address some of the difficulties encountered in constructing phenomenological models in other type II settings. Here we would like to address the construction of GUT models in M-theory. Although such models were expected to exist and qualitative features have been studied assuming their existence, there are currently no examples or techniques for explicitly constructing them. For a recent summary, see [7].

One of the main ideas that we employ, in physical terms, is the central role played by the BPS equations of a 'worldvolume' supersymmetric Yang–Mills theory, which lives in seven dimensions in the present case. In mathematical terms we are dealing with Higgs bundles, that is a gauge field and an adjoint Higgs field satisfying a version of Hitchin's equations, which are exactly the BPS equations of the $7d$ Yang–Mills theory. Even though we are in a non-perturbative regime of string theory,

* Corresponding author.
E-mail address: tpantev@math.upenn.edu (T. Pantev).

the gauge theory description can be trusted as long as the gauge and Higgs field are slowly varying. This strategy was successfully used in recent phenomenological constructions in *F*-theory, and we will see how it carries over in *M*-theory.

On the other hand, supersymmetric compactifications of *M*-theory to four dimensions are known to be given by G_2-manifolds. The principle that the worldvolume gauge theory completely determines the local geometry of the brane is well established in other contexts in string theory. Therefore we expect that our approach should establish the existence of a large class of non-compact G_2-manifolds with singularities. We will outline how one recovers the data of a non-compact G_2-manifold with singularities from the data of the compactified 7*d* gauge theory, but we will not give a complete proof of the correspondence in this paper. At any rate, string, *M* or *F*-theory only plays a secondary role in our philosophy. The primary object of interest is the higher dimensional Yang–Mills theory, and the main role of string/*M*/*F*-theory is to provide a UV completion of this Yang–Mills theory.

Spectral covers are a powerful technique for constructing solutions to Hitchin's equations. In the present setting, spectral covers correspond to Lagrangian *A*-branes in the auxiliary Calabi–Yau geometry T^*S^3, and thus much of our intuition about intersecting Lagrangians can be carried over. One might think this is not surprising, because if we do not use exceptional gauge groups, then we should be able to take a perturbative IIa limit and end up with a configuration of *D*6-branes and orientifold planes on a Calabi–Yau. However this picture is misleading for at least two reasons. First, our Lagrangian branes are intrinsic to the 7*d* Yang–Mills theory and do not assume any string or *M*-theory. Second, we are working at finite string coupling, so we do not expect sharply localized 6-branes. Indeed, our Lagrangian branes are auxiliary objects, and when we go back to the physical *M*-theory space-time they get mapped to the two-forms ω^i discussed in Section 2.2, which in general are not sharply localized. Analogous statements are also familiar in the *F*-theory context, where the spectral cover branes should not be confused with the (p, q) 7-branes [8].

For phenomenological purposes it is important to understand the spectrum and interactions in such models. Some qualitative results have already been obtained in the literature. Here we will find that these results may be better understood and extended using Morse theory as a principal tool. As a result, we find that the massless spectrum and many of the interactions reflect topological properties of the configuration, and can be computed without any knowledge of the solutions of the *D*-term equations. This is a remarkable simplification which should be of great help in understanding the phenomenological signatures of these models.

In this paper we mostly focus on abelian examples, although we will make statements that apply more generally. Work on non-abelian examples is still in progress [9].

2. Local models in *M*-theory

As mentioned in the introduction, our goal will be to construct compactifications of the 7*d* supersymmetric Yang–Mills theory. Such compactifications are mathematically described by Higgs bundles. In order to explain the relevance of such compactifications to *M*-theory, we have to explain how this data is related to G_2-manifolds with singularities. The main purpose of this section is to set up the dictionary between ALE fibrations in *M*-theory, Higgs bundles, and spectral covers in an auxiliary Calabi–Yau geometry.

2.1. General properties of G_2-manifolds

We are interested in engineering effectively four-dimensional models from *M*-theory. Since *M*-theory lives in 11*d*, this means we must compactify on a seven-dimensional internal space X_7. Furthermore, if we require $N = 1$ supersymmetry in four dimensions, then X_7 should admit a Killing spinor. As is well known, the existence of a Killing spinor implies that X_7 must admit a metric of G_2-holonomy. Such metrics are hard to find explicitly. However, much like Calabi–Yau metrics which are also hard to find, one may reformulate the problem of finding G_2-metrics in terms of anti-symmetric tensor fields, which are much easier to work with. For G_2-manifolds, the relevant tensor field is a three-form Φ.

Given a smooth seven-manifold, a three-form Φ is said to be stable if it lies in an open orbit of $Gl(7)$. In terms of a 7-bein, Φ may be written as

$$\Phi = e^1 \wedge e^2 \wedge e^3 + e^i \wedge \Sigma_i \tag{2.1}$$

where

$$\Sigma_1 = e^4 \wedge e^5 - e^6 \wedge e^7, \qquad \Sigma_2 = e^4 \wedge e^6 - e^7 \wedge e^5, \qquad \Sigma_3 = e^4 \wedge e^7 - e^5 \wedge e^6. \tag{2.2}$$

From such a three-form, one may reconstruct a metric through the following formulae [10]:

$$g_{ij} = \det(s_{ij})^{-1/9} s_{ij}$$

$$s_{ij} = -\frac{1}{144} \Phi_{i\mu_1\mu_2} \Phi_{j\mu_3\mu_4} \Phi_{\mu_5\mu_6\mu_7} \epsilon^{\mu_1 \cdots \mu_7} \sim -\frac{1}{144} \Phi_i \wedge \Phi_j \wedge \Phi. \tag{2.3}$$

In terms of Φ, the condition that the metric has G_2 holonomy is equivalent to [11]

$$d\Phi = 0 \quad \text{(F-term)}$$

$$d *_\Phi \Phi = 0 \quad \text{(D-term).} \tag{2.4}$$

The $*$-operator depends on the metric and hence implicitly on Φ, as we have indicated.

These two equations can be obtained as the critical points of two functionals which have a natural four-dimensional interpretation. The first equation in (2.4) is the equation of motion of a Chern–Simons functional $W \sim \int \Phi \wedge d\Phi$. In fact, M-theory not only yields a metric but also a three-form tensor field C_3, and in $N = 1$ SUSY compactifications it is natural to combine them in a single complex three-form field $C_3 + i\Phi$. A quick way to see this is by reducing on a circle to type IIa, in which case we get the complexified Kähler form $B + iJ$ as required by supersymmetry. Including the dynamics of C_3, the BPS equations are generalized to

$$d(C_3 + i\Phi) = 0$$
$$d *_\Phi \Phi = 0 \tag{2.5}$$

and the Chern–Simons functional may be generalized to:

$$W = \frac{1}{16\pi^2} \int_{X_7} (C_3 + i\Phi) \wedge d(C_3 + i\Phi). \tag{2.6}$$

This expression combines both the action of [12–14] and that of Hitchin [15]. It is interpreted as a term in the four-dimensional superpotential.

The second equation in (2.5) can be interpreted as a moment map condition. The three-form field in M-theory transforms under a group $\mathscr{G}$ of gauge transformations as $C_3 \to C_3 + d\Lambda$. We also have a natural Kähler form associated to the metric

$$\int_{X_7} (C'_3 + i\Phi') \wedge \overline{*_\Phi (C''_3 + i\Phi'')} \tag{2.7}$$

on the space of solutions of the F-terms. Then $i\, d^\dagger \Phi$ is the moment map associated to $\mathscr{G}$, and the second equation in (2.5) describes the critical points of a D-term potential. Under a suitable stability condition, we would expect that for every solution to the F-terms there exists a unique solution to the D-terms in the same $\mathscr{G}^c$ orbit. This is often guaranteed by the Kempf–Ness theorem, but the standard version of this theorem does not apply here.

The linearized deformations of (2.5) modulo gauge transformations by $\mathscr{G}$ are counted by harmonic three-forms, and these deformations are unobstructed [16]. Therefore the number of complex deformations of the G_2 structure is given by $h^3(X_7)$. The Kähler potential on moduli space turns out to be given by [12]

$$\mathcal{K} = -3 \log \frac{1}{2\pi^2} \frac{1}{7} \int |C_3 + i\Phi|^2. \tag{2.8}$$

In this paper we will be taking a slightly different point of view however, and we will not be using these expressions explicitly.

To summarize, the main point of this subsection is that the three-form Φ is an equivalent but more useful variable than the G_2 metric itself. Moreover, in terms of these variables the equations naturally split up into a set of first order equations which can be interpreted as F-terms and D-terms.

2.2. ALE fibrations

Now consider a local G_2-manifold X_7 which is an ALE fibration over a three-manifold Q_3. We will usually assume that the ALE fibration has a section, and also use Q_3 to denote this section. On each ALE fibre, there is a natural set of two-cycles $\alpha_i \in H_2(ALE, \mathbf{Z})$ which intersect according to the Cartan matrix associated to the ALE, generating an ADE root lattice Λ. There is also a dual set of two-forms ω^i.

The moduli space $\mathcal{M}(\Lambda)$ of M-theory on an ALE surface is described as follows. Given a hyperkähler structure $\{\mathcal{I}, \mathcal{J}, \mathcal{K}\}$ on the ALE, we can construct a triplet of two-forms $\vec{\Omega} = (\mathcal{I}^\nu_\mu g_{\nu\lambda}, \mathcal{J}^\nu_\mu g_{\nu\lambda}, \mathcal{K}^\nu_\mu g_{\nu\lambda})$. Their periods over the α_i are the parameters

$$\int_{\alpha_i} \vec{\Omega} = \vec{\phi}_i, \tag{2.9}$$

which describe complex structure and Kähler moduli of the ALE. They are often called the FI parameters because they appear as such in the hyperkähler quotient construction of the ALE. They naturally transform as a vector under an $SO(3)_R$ symmetry. In addition, we may expand the M-theory three-form in terms of the ω^i, yielding n vectors in seven dimensions, where n is the rank of the lattice Λ. Since the ALE preserves half of the 32 supersymmetry generators, we are guaranteed to recover their fermionic superpartners as well. In fact they are given by the same internal wave functions on the ALE. So for large sizes of the vanishing cycles we get a supersymmetric $7d$ gauge theory with gauge group $U(1)^n$. But what happens when the vanishing cycles are small?

There are additional supersymmetric states obtained from wrapping $M2$-branes on the vanishing cycles of the ALE. Their masses are given by $m \sim m_{pl}|N^i \vec{\phi}_i|$, i.e. they are proportional to the size of the vanishing cycle $N^i \alpha_i$ that the membrane is wrapping. Quantizing such a particle yields a vector multiplet, since this is the only non-gravitational multiplet available in a $7d$ theory with 16 supercharges. Since membranes couple to C_3 in $11d$, the $U(1)^n$ charges of these states are precisely

those of the W-bosons of a non-abelian ADE gauge theory with root lattice Λ. Therefore the effective low energy dynamics of M-theory on an ALE surface with small periods should be described by the corresponding $7d$ supersymmetric non-abelian ADE gauge theory. Additional evidence for this statement can be obtained through heterotic/type II duality.

When we further fibre the ALE over Q_3, additional supersymmetries will be broken. In G_2 compactifications, supersymmetry requires that C_3 and Φ are paired into a complex three-form. Expanding in a basis of exceptional cycles of the ALE, locally we get n complex one-forms on Q_3:

$$\int_{\alpha_j} C_3 + i\Phi = A_j + i\,\phi_j. \tag{2.10}$$

Since Φ describes zero modes of the metric, the one-form ϕ_j must be identified with the triplet of adjoint scalars of the $7d$ gauge theory encountered above. One can see this more explicitly from the canonical expression of Φ in terms of a 7-bein. Therefore the three adjoint scalars associated to each α_j must be twisted to a one-form on Q_3 [17]. In other words, in order to preserve $N = 1$ supersymmetry in four dimensions, the $SO(3)_R$-symmetry acting on the $\vec{\phi}$ is identified with the (dual of the) $SO(3)_Q$ structure group of Q_3. Equivalently, there exists a covariantly constant tensor J_μ^ν on Q_3 where μ transforms as a vector under $SO(3)_Q$ and ν transforms as a vector of $SO(3)_R$. In suitable coordinates we may write $J_\mu^\nu = i\delta_\mu^\nu$.

Thus locally the ALE fibration may be described by a set of n complex one-forms. There is an additional symmetry however which may be used in gluing local patches together in a global model. Namely the ALE has a diffeomorphism symmetry group which acts on the cycles as the ADE Weyl group. Eg. for A_{n-1} ALEs this is the symmetric group S_n on n letters. This symmetry may be thought of as a residual gauge symmetry from the non-abelian gauge theory. Thus altogether we see that an ALE fibration may be described by n one-forms on Q_3, with branch points across which the one-forms may be permuted. Said differently, locally on Q_3 we have a map from Q_3 into the parameter space $\mathfrak{h}_\mathbb{R}^{\otimes 3}/W$ of the universal unfolding of our ADE singularity. This is the essence of a Higgs bundle, as we will explain in more detail in Section 2.4.

Conversely, given a configuration for the ϕ_i on Q_3, we may try to reconstruct an ALE fibration over Q_3 with G_2 holonomy. To first order we get

$$\Phi = \Phi_0 + \phi_i \wedge \omega^i + \dots \tag{2.11}$$

where Φ_0 corresponds to the three-form for a constant ALE fibration, which certainly exists (it may be written down explicitly). The equations $d\Phi = d * \Phi = 0$ put constraints on the ϕ_i and on the higher order terms. By analogy with Kodaira–Spencer theory for Calabi–Yau manifolds [18,19], we conjecture that if the ϕ_i satisfy certain first order equations discussed below as well as tadpole constraints, and if Q_3 has non-negative curvature (so as to avoid curvature singularities at finite distance from the zero section that one would otherwise likely have), then the above series may be uniquely completed and has a finite radius of convergence.

So our main point here is that the data of the ALE fibration may be described as a field configuration in a supersymmetric $7d$ Yang–Mills theory living on Q_3. This naturally leads to a set of equations which we would expect the data to satisfy, in the limit that the fields are slowly varying.

2.3. Hitchin system

The BPS conditions in the $7d$ gauge theory are given by reducing the BPS conditions for ten-dimensional YM theory to seven dimensions. In $10d$ we have

$$\delta\lambda = \mathsf{F}_{\mu\nu}\,\Gamma^{\mu\nu}\epsilon = 0. \tag{2.12}$$

To preserve $3 + 1d$ Poincaré invariance, we assume only field configurations on the $6d$ internal space are turned on. On a Calabi–Yau manifold, this yields the well-known Hermitian Yang–Mills equations

$$\mathsf{F}^{2,0} = 0, \qquad g^{\mu\bar\nu}\mathsf{F}_{\mu\bar\nu} = 0. \tag{2.13}$$

Next we reduce these equations to $3d$. This is well understood [20,21].

Let us work on a local patch of Q_3. Then we can put a Calabi–Yau metric on the tangent bundle which is semi-flat. That is, the metric on the tangent bundle is expressed as $ds^2 = K_{ij}(x)(dx^i dx^j + dy^i dy^j)$, where K is a real Kähler potential which satisfies a real Monge–Ampère equation, and x and y are coordinates along Q_3 and along the bundle directions respectively. There is a natural complex structure in which the complex coordinates are given by $z^j = x^j + iy^j$. To perform the reduction, we assume that the gauge field is independent of the y coordinates, and we write the gauge field on the tangent bundle as

$$\mathsf{A} = A_i(x)dx^i + \phi_j(x)dy^j = (A + \phi_J)_i dx^i \tag{2.14}$$

where we used $dy^j = J_i^j dx^i$ and wrote $\phi_{Ji} = \phi_j J_i^j \sim \sqrt{-1}\,\phi_i$. Thus A naturally defines a complexified gauge field on Q_3. We have

$$\mathsf{F}^{2,0} = \mathsf{F}_{jk}dz^j \wedge dz^k \tag{2.15}$$

so the condition $\mathsf{F}^{2,0} = 0$ simply becomes the condition that the curvature of the complex connection A vanishes. Decomposing in real and imaginary parts, we get the *F*-terms

$$0 = \operatorname{Re} \mathsf{F}^{0,2} = F - [\phi_J, \phi_J]$$
$$0 = \operatorname{Im} \mathsf{F}^{0,2} = D_A \phi_J. \tag{2.16}$$

Further, we have the *D*-terms

$$0 = g^{\mu\bar{\nu}} \mathsf{F}_{\mu\bar{\nu}} = i K^{ij} D_i \phi_{lj} = i D_A^\dagger \phi_J. \tag{2.17}$$

These equations are precisely Hitchin's equations [20,22].

The equations for the Yang–Mills–Higgs fields on Q_3 are the primary objects for our purposes. However if Q_3 admits an integral affine structure, then the above isomorphism between solutions of Hitchin's equations on Q_3 and the Hermitian Yang–Mills equations on TQ_3 may be extended globally over Q_3. In order to do this on S^3 we would need to excise a suitable graph. Since the Higgs field takes values in the cotangent bundle T^*Q_3, it will be useful to define dual coordinates:

$$d\tilde{y}_j = K_{jk} dy^k. \tag{2.18}$$

The Kähler form on TQ_3 naturally gets identified with the standard symplectic form on T^*Q_3:

$$\omega = \frac{i}{2} K_{ij} dz^i \wedge d\bar{z}^j = K_{ij} dx^i \wedge dy^j = dx^i \wedge d\tilde{y}_i. \tag{2.19}$$

For later use, we can also dualize $d\tilde{x}_j = K_{jk} dx^k$. Then there is a natural complex structure also on T^*Q_3, in which the complex coordinates are given by $\tilde{z}_j = \tilde{x}_j + i\tilde{y}_j$, and a Calabi–Yau metric given by

$$ds^2 = K^{ij}(d\tilde{x}_i d\tilde{x}_j + d\tilde{y}_i d\tilde{y}_j). \tag{2.20}$$

Furthermore, we will see in the next section that solutions of Hitchin's equations can be interpreted as Lagrangian branes in T^*Q_3. This mapping between $\mathbf{R}^3$-invariant solutions of the Hermitian Yang–Mills equations and solutions of the Yang–Mills–Higgs equations is a real version of the Fourier–Mukai transform.

There are two basic type of solutions of the above Yang–Mills–Higgs equations. Solutions with $[\phi, \phi] = 0$ are said to be flat, and with this restriction the above data describes a Higgs bundle. One can also have solutions with $[\phi, \phi] \neq 0$. These will not describe Higgs bundles or ALE fibrations, but rather (in the picture described in Section 2.4) they will describe coisotropic branes in T^*Q_3. Such configurations were proposed to be relevant for moduli stabilization in [23]. In this paper we consider $[\phi, \phi] = 0$.

We would like to elaborate a bit on the interpretation of the *D*-terms. Let us say that A is an $Sl(n, \mathbf{C})$ connection. Given a solution to the *F*-term equations, in order to write the *D*-term equation we need to split A up into a real part A and an imaginary part ϕ_J. In general this cannot be done canonically, but depends on a choice of Hermitian metric h, which can be thought of as an equivariant map from the universal cover

$$h : \widetilde{Q}_3 \to \mathbf{H}_n, \qquad \mathbf{H}_n = Sl(n, \mathbf{C})/SU(n). \tag{2.21}$$

Given such a metric h, the covariant derivative D_{A} will generally not preserve it, but it can be split up as

$$D_{\mathsf{A}} = D_A + \phi_J \tag{2.22}$$

where D_A preserves the metric. Furthermore ϕ is locally identified with the derivative $\nabla h : T\widetilde{Q}_3 \to T\mathbf{H}_n$, and D_A is the pull-back of the Levi-Civita connection on $\mathbf{H}_n$. The *D*-term equation

$$D_A^\dagger \phi = D_A^\dagger \nabla h = 0 \tag{2.23}$$

is precisely the requirement of harmonicity of h as a map $\widetilde{Q}_3 \to \mathbf{H}_n$. Hence the solution of the *D*-term equation is also called the harmonic metric [22,24].

Let us spell this out in a little bit more detail in the abelian case. In this case $\mathbf{C}^*/U(1) = \mathbf{R}_{\geq 0}$ is the one-dimensional version of hyperbolic space, $\nabla h = h^{-1} dh$ and $A = -\frac{1}{2} h^{-1} dh$. The Hermitian metric can be written as $h = e^f$ where $f : \widetilde{Q}_3 \to \mathbf{R}$. The *D*-term equation says that

$$D_A^\dagger \nabla h = d^\dagger df = 0 \tag{2.24}$$

which says that f is a harmonic function on $\widetilde{Q}_3$, or that $\phi = df$ is a harmonic one-form on Q_3. Of course we can also argue more directly that setting A to zero locally and solving $d\phi = 0 \Rightarrow \phi = df$ implies that the *D*-term can be written as $d^\dagger df = 0$. But the above point of view is useful for establishing the existence of solutions in the non-abelian case [22,24].

The *F*-term part of the Yang–Mills–Higgs equations are the critical points of a Chern–Simons functional with complex gauge group, so we identify this as the four-dimensional superpotential

$$W = \frac{1}{16\pi^2} \int_{Q_3} \operatorname{Tr} \left(\mathsf{A} \wedge d\mathsf{A} + \frac{2}{3} \mathsf{A} \wedge \mathsf{A} \wedge \mathsf{A} \right). \tag{2.25}$$

In fact we could have gotten this more directly by applying the real Fourier–Mukai transform to the holomorphic Chern–Simons functional [21]. Further the *D*-term is the moment map for real gauge transformations, with respect to the Kähler form associated to the metric

$$g(\mathsf{A}, \mathsf{A}) = \int_{Q_3} |\mathsf{A}|^2 = \int_{Q_3} |A + i\phi|^2. \tag{2.26}$$

Thus it can be obtained as the critical point of the *D*-term potential:

$$V_D \sim \frac{1}{2} \int |D_A^\dagger \phi|^2 \tag{2.27}$$

when varied over possible hermitian metrics *h*. These are the precise analogues of the quantities we wrote earlier for G_2-manifolds in Section 2.1, but they also incorporate non-perturbative states that arise at ALE singularities.

2.4. Spectral cover picture

Suppose we are given a configuration for the adjoints ϕ satisfying $[\phi, \phi] = 0$. Then the three components of ϕ may be diagonalized simultaneously, and we may associate to ϕ its spectral data (i.e. its eigenvalues). Conversely, the Higgs field ϕ may be reconstructed from its spectral data. This picture yields *A*-model branes in an auxiliary Calabi–Yau three-fold $X = T^*Q_3$. In this subsection we discuss this construction in more detail. There is a completely analogous construction in *F*-theory [8,25] involving spectral covers in the canonical bundle over a complex surface. For convenience we will take $Q_3 = S^3$.

In the case of A_n-fibrations, the spectral cover picture is more than just an auxiliary construction, since it describes the *D*6-branes that we see in the weak coupling IIa limit. This is not a coincidence, because in the limit $vol(S^3) \to \infty$ the worldvolume theory on the *D*6-branes is well approximated by the maximally supersymmetric 7*d* YM theory.

Let us briefly recap the general structure. For the moment we will restrict to A_{n-1}-fibrations, i.e. 7*d* gauge theory with gauge group $G = SU(n)$. The twisted adjoint scalars of the 7*d* gauge theory give a section of

$$\phi \in T^*S^3 \otimes Ad(\mathcal{G}) \tag{2.28}$$

where $\mathcal{G}$ is the principle bundle with gauge group *G*. Consider the Hitchin map, which takes the Higgs field to the coefficients of *s* in

$$\det(sI - \phi) \in \operatorname{Sym}^n T^*S^3. \tag{2.29}$$

Here *s* is a local coordinate on the bundle direction of T^*S^3, and Sym^n is the *n*th symmetric power. The eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ of ϕ correspond to the zero set of the above section. Each λ_i is a one-form, so has three components. Thus the zero set defines an *n*-fold covering of the zero section in T^*S^3. This is called the spectral cover for the fundamental representation. We will denote it as $C_{(E,\phi)}$, or simply C_E for brevity, and the covering map by $p_C : C_E \to S^3$. Since ϕ lives in the adjoint of $SU(n)$, the eigenvalues add to zero on each fibre of T^*S^3:

$$\lambda_1 + \ldots + \lambda_n = 0. \tag{2.30}$$

The gauge field *A* gives a flat connection on a bundle *E* associated to the fundamental representation of $SU(n)$. We can think of ϕ as a map

$$\phi : E \to E \otimes T^*S^3. \tag{2.31}$$

Then on each fibre *E* decomposes into a sum of eigenspaces $\oplus_i \mathbf{C} |i\rangle$ under the action of ϕ. The assignment $\lambda_i \to \mathbf{C} |i\rangle$ gives a line bundle L_E on *C* called the spectral line bundle, and since $D_A \phi = 0$, D_A commutes with the action of ϕ on *E* and therefore *A* gives a flat connection on this line bundle. Conversely, given a spectral cover C_E together with a flat line bundle L_E, we can reconstruct the Higgs bundle on S^3 by

$$E = p_{C*}L_E, \qquad \phi = p_{C*}s \tag{2.32}$$

which yields a rank *n* bundle *E* and a map $\phi : E \to E \otimes T^*S^3$. In order for this to be an $SU(n)$ bundle, rather than a $U(n)$ bundle, we must have

$$\det(p_{C*}L_E) = 1 \tag{2.33}$$

where 1 denotes the trivial line bundle. This puts a topological constraint on the allowed line bundles on the spectral cover.

We claim that the spectral cover yields an *A*-type brane in T^*S^3. In order to see this, we may analyse the *F*- and *D*-terms locally. Since the gauge field *A* is flat, it may locally be gauged away. Further, as we discussed the equation $D\phi = 0$ splits up into $n - 1$ abelian equations $d\phi = 0$, where in the last equation we used ϕ to denote an abelian Higgs field. It is well known that the condition that ϕ be closed is equivalent to the section being Lagrangian with respect to the standard symplectic form on the cotangent bundle. To see this, writing $\phi = \phi_i dx^i$, the equation $\omega|_C = 0$ gives

$$0 = dx^i \wedge d\tilde{y}_i|_C = dx^i \wedge d\phi_i(x) = -d(\phi_i dx^i) = -d\phi. \tag{2.34}$$

Thus the *F*-terms equations correspond precisely to the condition that the spectral cover is a Lagrangian submanifold of T^*S^3, together with a flat connection. It is also worth noting that $d\phi = 0$ implies that $\phi = df$ for some function f on a local patch of S^3. Thus each sheet of the spectral cover may locally be represented as the graph of the differential of a real-valued function.

Naively one might expect these branes to be special Lagrangian also, since this is the usual requirement for supersymmetric *D*6 branes in IIa string theory. However, this is not quite the case. Locally on each sheet we have

$$\Omega^{3,0}|_C = d\tilde{z}_1 \wedge d\tilde{z}_2 \wedge d\tilde{z}_3|_C = \det(I + i\text{Hess}(f))d\tilde{x}_1 \wedge d\tilde{x}_2 \wedge d\tilde{x}_3 \tag{2.35}$$

where we used $d\tilde{z}_i = d\tilde{x}_i + id\tilde{y}_i$ and $\tilde{y}_i(x)dx^i = \phi = df$. Requiring that the imaginary part vanishes identically leads to the non-linear equation

$$\Delta f = \det(\text{Hess}(f)). \tag{2.36}$$

It is not hard to see where this apparent discrepancy comes from. The right hand side comes from a higher derivative correction to the two-derivative SYM theory. Indeed one may write a similar non-linear correction term for 10*d* SYM theory, which is related to (2.36) by the Fourier–Mukai transform[1] [21]:

$$\omega \wedge \omega \wedge F = F \wedge F \wedge F. \tag{2.37}$$

In the large volume limit in which we are working, such higher derivative corrections are parametrically small and can be neglected to first approximation. Thus in our approximation, *D*6-branes in type IIa and more generally spectral covers of ALE fibrations/Higgs bundles are described by 'harmonic' Lagrangian branes in T^*S^3, rather than special Lagrangian branes.

In the IIa context it is natural to conjecture that the 'harmonic' Lagrangian flows to a unique special Lagrangian under mean curvature flow [26]. In the *M*-theory context however it seems inappropriate to look at special Lagrangians. The reason is that in the IIa context we have two expansion parameters in 4*d*, namely g_s and ℓ_s/R_{KK}, but in the present context there is only a single expansion parameter (namely $1/vol(S^3)$ in Planck units), and so the higher order corrections discussed above can compete at the same order in this expansion parameter with other corrections such as KK loops. It would be physically incorrect to include only one type of correction and ignore the other contributions at the same order in the expansion parameter.

For applications to *M*-theory phenomenology we are interested in E_8 Higgs bundles. The spectral cover in this case can get quite complicated, but fortunately phenomenological considerations dictate that we only consider non-trivial configurations for an $SU(n)$ sub-bundle (so that the Higgs field breaks the E_8 gauge group to the commutant of $SU(n)$; the unbroken part of the gauge group is called the GUT group). In particular we would like to consider the case $n = 5$, which yields an $SU(5)$ GUT group. The E_8 spectral cover has 248 sheets and decomposes into several pieces, according to the decomposition[2]

$$\mathbf{248} = (\mathbf{24}, \mathbf{1}) + (\mathbf{1}, \mathbf{24}) + (\mathbf{5}, \mathbf{10}) + (\mathbf{\bar{5}}, \mathbf{\overline{10}}) + (\mathbf{10}, \mathbf{\bar{5}}) + (\mathbf{\overline{10}}, \mathbf{5}) \tag{2.38}$$

of E_8 under $SU(5)_{GUT} \times SU(5)_H$. The most important is the spectral cover for the fundamental representation of $SU(5)_H$, which determines all the others uniquely. This cover intersects each fibre of T^*S^3 in five points $\lambda_1, \ldots, \lambda_5$, with

$$\lambda_1 + \ldots + \lambda_5 = 0. \tag{2.39}$$

Here addition is defined in the obvious way in each fibre. In the language of ALE fibrations, the λ_i's correspond to certain *FI* parameters of the E_8 ALE using the dictionary described in Section 2.2. Using the labelling in Fig. 1, we may take the λ_i to describe the size parameters of the following exceptional cycles:

$$\begin{aligned}
|1\rangle &= \alpha_4 & |4\rangle &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 \\
|2\rangle &= \alpha_3 + \alpha_4 & |5\rangle &= \alpha_{-\theta} + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 \\
|3\rangle &= \alpha_2 + \alpha_3 + \alpha_4.
\end{aligned} \tag{2.40}$$

The sizes of the cycles $\{\alpha_5, \ldots, \alpha_8\}$ are taken to be zero, generating an $SU(5)$ GUT group, and all other cycles are obtained as linear combinations. As we will discuss in more detail later, when one of the λ_i's goes to zero, i.e. when the cover intersects the zero section, one may get a chiral or anti-chiral field in the **10** localized here. Another piece of the E_8 cover corresponds to the anti-symmetric representation of $SU(5)_H$. The cover $C_{\Lambda^2 E}$ intersects each fibre in

$$\lambda_i + \lambda_j, \qquad i < j \tag{2.41}$$

and we will see later that when this cover intersects the zero section, we may get a chiral or anti-chiral field in the **$\bar{5}$** localized there.

---

[1] This is for the abelian case; the non-abelian case is apparently not yet completely understood.

[2] The cover corresponding to (**24**, 1) further splits into two pieces, a cover of degree 20 and a four-fold multiple of the zero section.
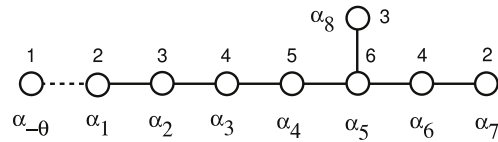
Fig. 1. The extended E_8 Dynkin diagram and Dynkin indices.

To summarize, we have gone through the following chain:

$$
\boxed{G_2 \text{ metric and flat } C_3}
$$

$$\updownarrow$$

$$
\boxed{d(C_3 + i\Phi) = d^*\Phi = 0}
$$

$$\wr$$

$$
\boxed{\text{Higgs bundle}}
$$

$$\updownarrow$$

$$
\boxed{\text{h-Lag branes in } T^*Q_3}
$$

(2.42)

The conjecture is that we can also go in the reverse, i.e. given solutions to Hitchin's equations with $[\phi, \phi] = 0$ one may reconstruct solutions to $d\Phi = d^*\Phi = 0$. This must be correct if the 7*d* gauge theory is to give an accurate description of *M*-theory dynamics on an ALE singularity, as is expected physically for large S^3.

2.5. Non-compact branes

We claim that in order to get interesting solutions, we have to allow for certain source terms in the YMH equations. To see this, let us assume that we do not have any source terms. Now flat connections are characterized by their monodromies, and since $\pi_1(S^3) = 1$ any flat connection is equivalent to the trivial connection. This is in accord with the statement that branes wrapped on the minimal three-sphere in T^*S^3 do not form bound states [27].

One may get non-trivial solutions by instead quotienting the S^3 by a freely acting discrete group Γ, so that $\pi_1(S^3/\Gamma)$ will be non-trivial. However the non-trivial bound states one can make are still not of the type we want. We need the discrete group Γ to be freely acting, and in this case must be finite (basically Γ is a product of ADE subgroups of $SU(2)_L \times SU(2)_R$ acting on S^3). Therefore the monodromies must be contained in the compact part of the complexified gauge group. This means that Wilson lines for the gauge field can be turned on but the Higgs field has zero expectation value. In order to get more interesting solutions with a non-trivial Higgs field we need to do something else.

In order to get interesting solutions we need to allow for non-compact flavour branes, i.e. source terms in the Yang–Mills–Higgs equations. This is completely analogous to the meromorphic Higgs bundles appearing in local *F*-theory models [8,25]. One may also see how this arises by applying heterotic/*M*-theory duality to heterotic models. Let us take for instance a heterotic model with the spin connection embedded in the gauge connection. In a suitable limit, the heterotic Calabi–Yau three-fold admits a Lagrangian T^3-fibration over S^3. The Wilson lines of the bundle along the T^3 fibres form a covering of the S^3, which consist of three-points on the generic dual T^3 fibre. This is the heterotic picture of the spectral cover that we discussed, for the special case of the tangent bundle. However over special subsets in S^3 the cover may wrap some of the circles of the T^3-fibration. Eg. over a graph in S^3 it may contain a $T^2 \subset T^3$, and over special points in S^3 it may contain the whole T^3. As in *F*-theory/heterotic duality, we expect that in a suitable limit this data is equivalent at the level of *F*-terms to a degenerate $K3$ fibration over S^3 with a section of $E_6 ALE$ singularities, although for *M*-theory/heterotic duality this is of course not established. Taking the local limit, we should take the size of the dual T^3 fibres to infinity in the heterotic picture. This gives the picture advocated above of a Higgs field which is generically finite, but may blow up over a special subset of S^3.

3. The effective theory

In the previous section, we reformulated the problem of constructing phenomenological M-theory compactifications in terms of Higgs bundles and spectral covers. In this section we explain how the low energy degrees of freedom and their interactions arise from the compactified Yang–Mills theory. Qualitatively this is already largely understood in the literature, but in order to construct models and carry out the computations explicitly we need some new tools. Therefore we will reformulate some old results in our present language and introduce Morse theory in order to relate the spectrum to more readily computable quantities.

3.1. Chiral matter

Intuitively, chiral matter will be localized on some kind of solitonic configuration of the Higgs field. At the centre of such a soliton, one of the eigenvalues of the Higgs field is going to zero. Thus we would like to analyse the Dirac equation in such a solitonic background. In non-degenerate situations, there is only a single eigenvalue or combination of eigenvalues of the Higgs field going to zero. Therefore in the non-degenerate case it is sufficient to consider abelian Higgs fields only, and we will assume this through much of the discussion below. As we will discuss in the next subsection however, there are some global effects which may require us to look at the full non-abelian field.

Recall that locally we can set the gauge field to zero, and write $\phi = df$. Therefore the zeroes of ϕ are sometimes also called 'critical points'. We will now describe how this gives rise to chiral matter. This is essentially already discussed in the literature, particularly [28,29] for the present setting, but will be reformulated somewhat to fit our purposes.

In order to calculate the spectrum, we need to solve a Dirac equation with a background Higgs field turned on. Following [30] we will rescale the Higgs field by a positive real number $t \sim 1/\hbar$ and calculate the spectrum in a $1/t$ expansion.

It is convenient to think of the spinors in the $7d$ Yang–Mills theory from a ten-dimensional point of view. The ten-dimensional **16** decomposes under $SO(3, 1) \times SO(6)$ as

$$(\mathbf{2}, \mathbf{1}, \mathbf{4}) + (\mathbf{1}, \mathbf{2}, \bar{\mathbf{4}}) \tag{3.1}$$

which are further related by the Majorana condition. The **4** and $\bar{\mathbf{4}}$ of $SO(6)$ may be identified with $(0, p)$ forms, with p even odd for **4** and p even for $\bar{\mathbf{4}}$. Each breaks up under $SO(3)_Q \times SO(3)_R$ as a $(\mathbf{2}, \mathbf{2})$, but they have different eigenvalues under Γ_6. We can denote them as $(\mathbf{2}, \mathbf{2})_\pm$ according to their Γ_6 eigenvalue.

As we discussed, due to twisting needed to maintain $N = 1$ supersymmetry, we can identify $SO(3)_R$ with $SO(3)_{Q_3}$. This diagonal $SO(3)$ may be identified with the real subgroup $SO(3) \subset SU(3)$ of the Calabi–Yau holonomy group fixed by an anti-holomorphic involution, which gives another way to see the twisting. In any case, our spinors are functions of x_i and transform as spinor bilinears on $Q_3 = S^3$. As is well known, such bilinears can be identified with differential forms on Q_3, i.e. they may be identified with sections of

$$\psi \in \mathcal{A}^p(Q_3, \mathbf{C}) \otimes Ad(\mathcal{G}), \quad p = 0, 1 \tag{3.2}$$

where $\mathcal{G}$ is the principal bundle with gauge group G. Note that this matches with the bosonic field content. In order to relate this to the description above, given a p-form wave function ψ we may associate with it a $(0, p)$-form by replacing $dx^i \to d\bar{z}^i$ in ψ, and a $(p, 0)$-form by replacing $dx^i \to dz^i$ in ψ^*. By Serre duality (i.e. taking the complex conjugate and contracting with $\Omega^{3,0}$) we may relate the $(p, 0)$-forms to $(0, 3 - p)$-forms which transform in the same representation. The reason for replacing ψ^* by its Serre dual is that the Dirac operator acts more naturally in this basis. In terms of the original p-form ψ this is just the real Hodge $*$-operator $\psi \to *\psi$ on Q_3 without complex conjugation. So it is natural to allow differential forms (3.2) for all values of p, with equivalent degrees of freedom related by the $*$-operator. Tracing back to (3.1) we see that chiral fermions are naturally paired with odd p-forms and anti-chiral fermions are paired with even p-forms.

We assume that we have a gauge group G which is broken to a subgroup H by turning on an abelian component of the Higgs field. We decompose the adjoint representation of G under $H \times U(1)$ as

$$Ad(G) = Ad(H)_0 + R_q(H) + \bar{R}_{-q}(H) + \mathbf{1}_0. \tag{3.3}$$

The worldvolume gauge fields can be set to zero locally. The Dirac operator acting on the spinors in the R-representation is then given by

$$i\mathcal{Q}_t = i \sum_{j=1,2,3} (\partial_j + t\partial_j f)(a_j^\dagger + a^j). \tag{3.4}$$

Here we set $q \to 1$ because its precise value is inconsequential, only its sign is important. The Dirac operator acting on spinors in the $\bar{R}$-representation is given by interchanging $f \to -f$. When identifying the spinors with forms, the Clifford algebra may be represented as $a_i^\dagger = \wedge dx_i$ and $a^i = \iota_{\partial/\partial x_i}$, so we get

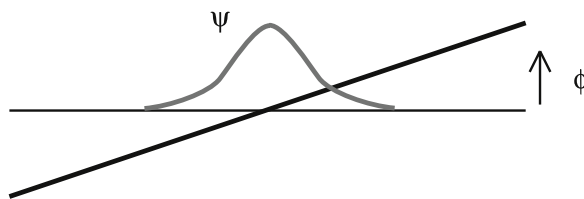$$\mathcal{Q}_t = d_t + d_t^\dagger \tag{3.5}$$

Fig. 2. Fermion zero mode localized on a defect.

where

$$d_t = d + t\, df\, \wedge\,. \tag{3.6}$$

The operator Q_t is exactly the operator discussed at length in [30], so we will borrow from the discussion there.

The operator Q_t may be thought of as a supercharge for supersymmetric quantum mechanics with target space M. The Hamiltonian of this system is given by

$$Q_t^2 = H_t = \Delta + t^2 (df)^2 + \sum_{i,j} t \frac{D^2 f}{Dx_i Dx_j} [a^{\dagger i}, a^j] \tag{3.7}$$

where Δ denotes the usual Laplace–Beltrami operator. For large t, the Hamiltonian is dominated by the potential energy $|df|^2$. In order to minimize the potential energy for large t, the eigenfunctions must be peaked around the critical points of f. Therefore we may focus on a single critical point and approximate f by a quadratic potential. Up to coordinate transformations, f may locally be written as:

$$f = \frac{1}{2} \sum_{i=1,2,3} p_i x_i^2. \tag{3.8}$$

The p_i, $i = 1, 2, 3$ are real constants. They are all non-zero because we assumed that the critical points are non-degenerate. Then the Higgs field near the critical point is given by

$$\phi = df = \sum_{i=1,2,3} p_i x_i\, dx_i. \tag{3.9}$$

The D-term equation put a restriction on the p_i:

$$\text{Tr Hess}(f) = p_1 + p_2 + p_3 = 0. \tag{3.10}$$

Using coordinates in which f takes the diagonal form (3.8), we may clearly use separation of variables. Thus we concentrate on one variable x_i temporarily and finally tensor the wave functions together. Then we have a standard domain wall set-up (see Fig. 2). The Dirac equation becomes

$$\left[\frac{\partial}{\partial x_1} + t p_1 x_1 \right] \psi_1^+ = 0, \qquad \left[\frac{\partial}{\partial x_1} - t p_1 x_1 \right] \psi_1^- = 0. \tag{3.11}$$

The local solution is

$$\psi_1^{\pm}(x) = e^{\pm t p_1 x_1^2/2} \epsilon^{\pm} \tag{3.12}$$

where $\epsilon^+ = 1$ and $\epsilon^- = dx_1$. Inspecting the exponential factor, we see that either ψ_1^+ is normalizable and ψ_1^- is not, or vice versa. The normalizable solution is physically sensible and the non-normalizable one should be discarded. Which solution is normalizable clearly depends only on the sign of p_1.

Tensoring together with ψ_2 and ψ_3 and a four-dimensional chiral or anti-chiral spinor $\chi^{\pm}$, we get the full wave function:

$$\chi^{\pm} \psi_1^{\pm} \psi_2^{\pm} \psi_3^{\pm} \otimes R(\mathcal{H}) \in \mathcal{A}^*(Q_3; \mathbf{C}) \otimes R(\mathcal{H}). \tag{3.13}$$

As we discussed, not all of these combinations are allowed. The four-dimensional chirality is correlated with the degree of the form, which now becomes the number of negative eigenvalues. The number of negative eigenvalues of the Hessian at an isolated critical point is called the Morse index. The cases of Morse index zero or three (i.e. $(+, +, +)$ and $(-, -, -)$) are ruled out by the D-terms, since a harmonic function cannot have local minima or maxima. Therefore up to permutations, we have Morse index one $(+, +, -)$ which gives a chiral fermion, and Morse index two $(+, -, -)$ which gives an anti-chiral fermion.

In addition, we should analyse the Dirac equation for the remaining pieces in the decomposition of $Ad(G)$, namely $Ad(H)_0$ and 1_0. The corresponding zero modes are not localized at the critical points and we need some global information. In this case the Dirac operator is just given by the exterior differential, and the zero modes are in one-to-one correspondence
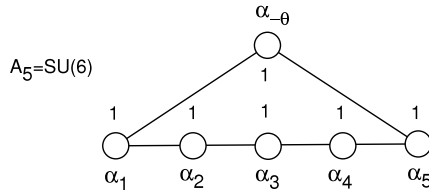
Fig. 3. The extended A_5 Dynkin diagram and Dynkin indices.

with Betti numbers. From $Ad(H)_0$ we get a gaugino transforming in $Ad(H)$, and $b_1(Q)$ adjoint chirals. From the 1_0 we get $H^1(C)$ moduli (where C is the spectral cover). This concludes the derivation of the massless spectrum to all orders in the $1/t$ expansion.

As a simple example, consider the local unfolding of an $SU(6)$ singularity to an $SU(5)$ singularity, considered in [29,31]. That is, we will consider an A_5 ALE surface fibred over $\mathbf{R}^3$, such that for $\vec{x} = \vec{0}$ all the vanishing cycles are zero size and we have an A_5 singularity, and such that for $\vec{x} \neq 0$ the ALE is partially resolved, but we still have an A_4 singularity. Then we need to turn on an abelian Higgs field. Under $SU(5) \times U(1)_Q$, the adjoint of $SU(6)$ decomposes as

$$Ad(SU(6)) = Ad(SU(5)) + \mathbf{5} + \bar{\mathbf{5}} + \mathbf{1} \tag{3.14}$$

and so we expect a chiral fermion at $\vec{x} = 0$ which transforms as a $\mathbf{5}$ or $\bar{\mathbf{5}}$ under the unbroken $SU(5)$ gauge symmetry. Let us first phrase the configuration in our current language, and then compare with [29].

Our Higgs field will be proportional to a Cartan generator $U(1)_Q$ which is embedded in $SU(6)$ as $\omega_Q = $ diag $(1, 1, 1, 1, 1, -5)$. In terms of the canonical basis ω_k satisfying

$$[\omega_k, \alpha_l] = \delta_{kl} \tag{3.15}$$

this corresponds to $\omega_Q = 6\omega_5$. To each node we can associate an abelian Higgs field $\phi_k\omega_k$, whose three components are the three FI parameters for the corresponding cycle α_k. They satisfy the constraint

$$d_{-\theta}\phi_{-\theta} + \cdots + d_5\phi_{\alpha_5} = 0 \tag{3.16}$$

where d_k are the Dynkin indices (see Fig. 3). In the present case, the d_k are all equal to one. (This description is redundant because we can always use this relation to eliminate $\phi_{-\theta}$, but it becomes quite convenient for non-abelian Higgs field VEVs). Now we set

$$\phi_{\alpha_5} = -\phi_{-\theta} = df, \qquad f \propto \sum_{i=1}^{3} p_i x_i^2. \tag{3.17}$$

Since $\{\phi_{\alpha_1}, \ldots, \phi_{\alpha_4}\}$ are all kept zero, the corresponding cycles α_k are all kept at zero size, and an $SU(5)$ singularity is preserved. This satisfies the F-terms, and using the metric $ds^2 = \sum dx_i^2$ on $\mathbf{R}^3$ it also satisfies the D-terms provided $p_1 + p_2 + p_3$ = 0. By our previous analysis, we get a chiral fermion localized at $\vec{x} = \vec{0}$, in the $\mathbf{5}$ if $p_1 p_2 p_3 > 0$ or in the $\bar{\mathbf{5}}$ if $p_1 p_2 p_3 < 0$.

This description agrees with the hyperkähler quotient construction of [29]. Their D-terms are given by $\vec{D} = (a, \bar{b})/U(1)$, where the $U(1)$ acts with charge one on a and b. The authors of [29] choose the unfolding $\vec{x} = (a, b)/U(1)$, i.e. we changed the sign of one of the components of $\vec{D}$ (changing the Morse index from zero to one) and then identified the image with $\mathbf{R}^3$. Thus this agrees with our claims above except for an inconsequential rescaling in the metric on $\mathbf{R}^3$. Additional constructions along the lines of [29] can be found in [32,33].

Of course we want a gap between the massless modes and the KK modes of the $7d$ gauge theory, so we are not interested in fibering over $\mathbf{R}^3$ but over a compact three-manifold like S^3 or S^3/Γ. Then there will necessarily be higher order terms in f and additional critical points. Our calculation was not exact, and in fact there are corrections to H_t exponentially suppressed in t which may lift some of the zero modes we found. These corrections are the topic of Section 3.3.

We can also state the formula for chiral matter in terms of the spectral cover in the auxiliary Calabi–Yau T^*S^3. Locally we can describe a sheet by the graph of df. A critical point of f corresponds to an intersection point between the graph of df and the zero section, and the sign of the intersection is just $(-1)^p$ where p is the Morse index. Thus the statement is that one must count with sign the number of intersection points of a harmonic Lagrangian brane C with the zero section C_0:

$$N_\chi(R) = \#(C_0 \cap C)_-, \qquad N_\chi(\bar{R}) = \#(C_0 \cap C)_+. \tag{3.18}$$

Consider for instance E_8 models broken to $SU(5)$ by an $SU(5)$ Higgs field. The $SU(5)$ Higgs field can be encoded in a Lagrangian brane C_E in T^*S^3 which is a five-fold covering of the zero section, or in a ten-fold covering $C_{\Lambda^2 E}$ associated to the anti-symmetric representation. When C_E intersects the zero section, we have $\lambda_i \to 0$ for some i and the symmetry is locally enhanced to $SO(10)$. Using the decomposition

$$Ad(SO(10)) = Ad(SU(5)) + \mathbf{10} + \overline{\mathbf{10}} + \mathbf{1} \tag{3.19}$$

the chirals in the **10** are counted by the number of negative intersections between C_E and the zero section. Similarly when $C_{\Lambda^2 E}$ intersects the zero section, $\lambda_i + \lambda_j \to 0$ for some $i \neq j$ and the symmetry is locally enhanced to $SU(6)$, i.e. it locally looks like the example discussed above. Thus the number of **5** or $\overline{\bf 5}$'s is counted by the intersection points of the cover $C_{\Lambda^2 E}$ with the zero section. However pairs of chirals and anti-chirals may still be lifted through instanton effects, leading to the quantum intersection theory of Lagrangian branes in T^*S^3.

3.2. Abelian solutions

Recall that in order to get non-trivial solutions to Hitchin's equations on S^3 or S^3/Γ, we have to allow for non-compact branes, i.e. we have to allow for singularities in the Higgs field. We will generally assume that these singularities are located on a graph Δ in S^3, although one could consider more general situations. Then locally we may choose a coordinate r transverse to the graph, and an angle θ in the plane transverse to the graph. In the abelian case, the local behaviour as $r \to 0$ is

$$A \sim \alpha d\theta$$
$$\phi \sim \beta d\log r + \gamma d\theta. \tag{3.20}$$

The parameters α, β and γ are boundary data that we have to specify. In this section we discuss the special case when there is no monodromy of the gauge and Higgs fields, i.e. $\alpha + i\gamma = 0$, and we return to the general case later.

We excise a small tubular neighbourhood of Δ from S^3, which therefore becomes a manifold with boundary which we will denote by M. Then the F-terms on M simply read

$$d\phi = 0 \tag{3.21}$$

and since the Higgs field carries no monodromy ($\gamma = 0$), we may express $\phi = df$ for a globally defined function f. The D-terms read

$$d^* df = 0 \tag{3.22}$$

on the complement of Δ. In fact it is simpler in this case to think of β as a charge density along Δ, and write the D-terms as a Poisson equation on S^3

$$d^* df = \beta. \tag{3.23}$$

The above equations are of course very familiar. We are simply dealing with an electro-statics problem, with the Higgs field ϕ playing the role of the (dual of the) electric field, β playing the role of a charge density along the graph, and $f = \log h$ playing the role of the electro-static potential. Thus we can solve this problem in the standard way, by using the Green's function for the Laplacian. There is a single consistency constraint that needs to be satisfied: using the divergence theorem, we get

$$0 = \int_M d^* \phi = \int_{\partial M} \phi \cdot \nu \tag{3.24}$$

and hence the total flux through the boundary must vanish, which means that the total 'charge' on S^3 must vanish (Gauss's law). Accordingly there must be positive and negative charges, and we can split up the graph Δ as $\Delta^+ \cup \Delta^-$ which are positively and negatively charged respectively. Let us denote the Green's function for the Laplacian as $G(x, y)$. The solution to the BPS equations is given by

$$\phi = df, \qquad f(y) = \int_{S^3} d^3 x \sqrt{g} \, \beta(x) G(x, y). \tag{3.25}$$

3.3. Instanton corrections and Morse cohomology

In the above we saw that one may construct abelian solutions by solving a simple electro-statics problem. However even though we know there exists a solution satisfying both the F-terms and the D-terms, one can learn much by imposing only the F-term equations. That is, in the following we will assume that ϕ is closed but not necessarily harmonic. Recall that the spectrum of massless charged chiral and anti-chiral matter is related to the critical points of ϕ. As we will now discuss, the massless spectrum may actually be deduced using purely combinatorial methods and is common to all solutions of the F-terms, independent of whether ϕ is harmonic.

Let us denote the set of critical points of index i by $\mathrm{Crit}_i(f)$. It is a well-known fact that on a compact manifold the number of critical points of Morse index i is bounded below by the i-th Betti number:

$$\# \, \mathrm{Crit}_i(f) \geq h^i(M). \tag{3.26}$$

This is the weak form of the Morse inequalities, one of the central results of Morse theory (a brief review may be found in Appendix). In the present setting we also have boundaries, and we really get a bound in terms of Betti numbers for relative cohomology. For simplicity let us temporarily ignore this issue. The point is that we get a lower bound on the massless spectrum in terms of topological data.

Now it turns out the situation is actually much better than that. The calculation of the massless spectrum in Section 3.1 was exact to all orders in a $1/t$ expansion, but may still be corrected by instantons. Once we take these quantum effects into account, we will find that the massless spectrum is in fact *exactly* computed by the Betti numbers. In other words, we may read of the massless spectrum just from the topology of M, which may be computed by purely combinatorial methods. Following [30] (see also [34]), we will now briefly explain how instantons correct the computation.

The potential energy function $V \sim |df|^2$ has multiple critical points. However it is not generally true that the states we found at each critical point are all true ground states. We have not yet accounted for the possibility of tunnelling. To see this, we consider two critical points p, q and their associated ground states $|p\rangle$, $|q\rangle$ and compute the amplitude $\langle q|Q_t p\rangle$. The effective Lagrangian of the 7d SYM describing excitations along M in the representation R is that of supersymmetric quantum mechanics on M:

$$\mathcal{L} = \frac{1}{2}(dx)^2 - \frac{1}{2}t^2(df)^2 + \frac{i}{2}\bar{\psi}(D + tD^2f)\psi + \frac{1}{4}R\bar{\psi}\psi\bar{\psi}\psi. \tag{3.27}$$

The fermions may be thought of as the operators $a^i \sim \wedge dx^i$ encountered earlier. Ignoring the fermions, the Euclidean action is given by:

$$2\,S_E = \int d\lambda\,(dx)^2 + (tdf)^2 = \int d\lambda(dx \pm tdf)^2 \mp t\int d\lambda df$$
$$\geq t\,|f(q) - f(p)| \tag{3.28}$$

where λ denotes Euclidean time. The potential is then turned upside down, and there are instanton solutions:

$$\frac{\partial x_i}{d\lambda} = \pm tg^{ij}\frac{\partial f}{\partial x_i} \tag{3.29}$$

with $x(\lambda \to -\infty) = q$, $x(\lambda \to +\infty) = p$. These are gradient flow trajectories that connect critical points p and q. For instanton contributions to d_t we want the minus sign above. Each such instanton contributes

$$\langle q|d_t|p\rangle \sim \int dx_0 d\bar{\psi}_0 \frac{\det_F}{\det_B^{1/2}}\,d_t|_{class}e^{-\frac{1}{2}t(f(q)-f(p))}. \tag{3.30}$$

Here x_0 and $\bar{\psi}_0$ denote the bosonic and fermionic zero modes of the instanton; $\det_B$ and $\det_F$ denote the bosonic and fermionic fluctuation determinants; and we evaluate d_t on the instanton solution.

To get the amplitude, we need to compute the one loop fluctuation determinant. In supersymmetric theories, the determinants for the non-zero modes of the bosons and fermions cancel, but there are zero modes. For an instanton connecting two critical points whose Morse index differs by one, there is one bosonic zero mode (for the broken translation invariance in Euclidean time) and one fermionic zero mode (for the broken supersymmetry generator). More generally, the number zero modes is given by the Morse index of p minus the Morse index of q. Since the operator $d_t = \bar{\psi}(\partial x + \partial f)$ soaks up exactly one bosonic and one fermionic zero mode, the instanton only contributes if the difference in the Morse indices is equal to one.

We further need to know the coefficient, in particular the sign of the coefficient. This is slightly subtle since the coefficient is proportional to $1/t$, so we simply state the result. Consider the subspace V_p of the tangent space at p on which the Hessian is negative definite. Its dimension is given by the Morse index of p and it carries an orientation induced by the differential form $|p\rangle$. Let $\vec{v}$ denote a vector tangent to the gradient flow trajectory, and $V_p^\perp$ the subspace of V_p orthogonal to $\vec{v}$. It carries an orientation induced by the contraction of $|p\rangle$ with $\vec{v}$. At q we have an analogous subspace V_q with orientation induced by $|q\rangle$. Now we transport $V_p^\perp$ along the gradient flow trajectory to q. If the orientation agrees with V_q we use the plus sign, and if they disagree we use the minus sign.

Hence we deduce that non-perturbatively in $1/t$, the action of d_t on the would-be zero modes has the following correction:

$$d_t\,|p\rangle = \sum_{C^{i+1}(M)} n(p, q)e^{-\frac{1}{2}t(f(q)-f(p))}\,|q\rangle \tag{3.31}$$

where $n(p, q)$ counts with sign the number of trajectories. By rescaling $|p\rangle \to e^{\frac{1}{2}tf(p)}\,|p\rangle$, $|q\rangle \to e^{\frac{1}{2}tf(q)}\,|q\rangle$ we can get rid of the exponential factors, and hence we simply get

$$d_t\,|p\rangle = \sum_{C^{i+1}(M)} n(p, q)\,|q\rangle. \tag{3.32}$$

Since $d_t^2 = 0$, it is a boundary operator. The actual zero modes are those which are also annihilated by d_t and $d_t^\dagger$, so the true zero modes are in one–one correspondence with the cohomology of d_t. This is the Morse complex.

It is not hard to recover these instantons in the other pictures we have been using. In the ALE fibration picture, the gradient flow trajectories connecting critical points can be interpreted as membrane instantons. Let us assume for instance

that we have an A_1 ALE space fibred over S^3, and our line bundle is embedded as the diagonal generator in $SU(2)$, generically breaking the $SU(2) \rightarrow U(1)$. Then in the ALE sitting over the critical point of f, an S^2 shrinks to zero size and the $U(1)$ gets enhanced to $SU(2)$ due to a massless $M2$-brane wrapping the S^2. Now let us take this S^2 and transport it along the gradient flow trajectory to the other critical point. In this way we trace out an S^3 inside the ALE fibration which projects to the gradient flow trajectory in the base Q_3, and there is a membrane instanton obtained by wrapping an $M2$-brane on this S^3. The action of this instanton is proportional to the area of the S^3, which we can find because the S^3 is a calibrated cycle. Recalling that $\Phi \sim \Phi_0 + tdf \wedge \omega$, we get:

$$\text{vol} = \int_{S^3} \Phi \sim \int t \partial_\lambda f \, d\lambda \sim t \left(f(q) - f(p) \right). \tag{3.33}$$

This agrees with the action of the gradient flow instanton. Clearly this generalizes to more complicated ALE fibrations. Note that we should not interpret these $M2$ instantons as saying that our gauge theory breaks down, we are still describing a tunnelling effect in the 7d gauge theory.

We can also interpret these instantons in the spectral cover picture. It will not come as a surprise that they lift to disc instantons. To see this, consider a disc stretching between the zero section and the Lagrangian brane defined by the function f, which projects to the gradient flow trajectory on S^3 between the critical points p and q. Its area, measured with respect to the standard symplectic form on T^*S^3, is given by

$$\int_{-\infty}^{\infty} d\lambda \int_0^1 dx \, x |t \nabla f| = \frac{1}{2} t (f(q) - f(p)) \tag{3.34}$$

and therefore the action of the disc instanton also agrees with the action of the gradient flow instanton.

Now the beauty of the Morse complex is that it reconstructs the ordinary cohomology of the underlying manifold, i.e. we have

$$H^*_{Morse}(M, f) = H^*(M, \mathbf{R}). \tag{3.35}$$

One way to understand this isomorphism is by noticing [30] that our differential operator is related to the ordinary exterior differential by a similarity transformation

$$d_t = e^{-tf} de^{tf} = d + t \, df \wedge . \tag{3.36}$$

Thus the Betti numbers are independent of t. In the limit $t \rightarrow \infty$ we obtain the Morse complex, and in the limit $t \rightarrow 0$ we recover the definition of the ordinary de Rham cohomology.

There is also a version of this isomorphism for manifolds with boundary, which is the case relevant for us. Recall that we effectively have boundaries where the Morse function (or electro-static potential) becomes infinite. We will assume the boundary may be split up into disjoint positively charged and negatively charged pieces

$$\partial M = \partial^+ M \cup \partial^- M. \tag{3.37}$$

Then the Morse complex reconstructs the relative cohomology

$$H^*_{Morse}(M, f) = H^*(M, \partial^+ M). \tag{3.38}$$

Actually in order for Morse theory with boundaries to be well defined and reproduce the relative cohomology, we must ensure that the gradient flow trajectories connecting critical points do not hit the boundary. This is automatically the case for the harmonic solutions that satisfy the D-term equations. In the electro-statics analogy, the critical points are the points where a test charge would experience zero force. The gradient flow trajectories describe the possible trajectories of a positive test charge. A trajectory connecting two critical points cannot hit a boundary — the potential energy is minimized at a boundary and the test charge cannot climb back out.

Therefore we conclude that the massless matter content depends only on topological properties of the configuration and is independent of the explicit solution to the D-term equation. The massless matter content is simply given by the ranks of the relative cohomology groups:

$$N_\chi(R) = h^1(M, \partial^+ M), \qquad N_{\bar\chi}(R) = h^2(M, \partial^+ M) \tag{3.39}$$

where $\partial^+ M$ is the boundary of M where the Morse function is increasing, i.e. where the positive charges are located. Therefore the computation of the spectrum is reduced to a purely *combinatorial* problem involving triangulations or cell complexes. In particular the net number in the representation R is simply given by

$$\text{net chiral}(R) = \chi(M, \partial^+ M). \tag{3.40}$$

As we will discuss later, this formula holds much more generally.

More generally we may have multiple abelian Higgs fields. Suppose there are two, with associated Morse functions f_a, f_b. Let us denote the boundaries as

$$\partial_a = \partial_a^+ - \partial_a^- \tag{3.41}$$

where we used the minus sign to indicate that there are negative charges located on ∂_a^-. We will assume again that positive and negative boundaries do not intersect, and we will also assume for convenience that the charge density along the boundary is uniform. Then it is not hard to see that for a general linear combination $q_a f_a + q_b f_b$, the boundary is given by

$$\partial_{q_a, q_b} = q_a \partial_a^+ - q_a \partial_a^- + q_b \partial_b^+ - q_b \partial_b^-. \tag{3.42}$$

Depending on how the Higgs fields are embedded into a non-abelian group G, one will be interested in the critical points of various linear combinations of f_a and f_b. These linear combinations depend on the $U(1) \times U(1)$ charges (q_a, q_b) of the multiplets appearing in the decomposition of the adjoint of G under $H \times U(1)_a \times U(1)_b$. In each case we have

$$N_\chi(q_a, q_b) = H^1(M, q_a f_a + q_b f_b) = H^1(M, \partial_{q_a, q_b}^+). \tag{3.43}$$

So far, our results were stated as certain properties of the Higgs field. We may also restate some of the results in the spectral cover description, which yields a more geometric picture. We have already seen that critical points correspond to intersections of components of the spectral cover with the zero section. We have also seem that gradient flow trajectories lift to disc instantons in T^*S^3. Therefore the instanton corrected spectrum is computed by the Floer cohomology groups

$$HF^*(C_0, C_E) \tag{3.44}$$

where we used C_0 to denote the zero section. Indeed it is well known in the literature that Floer cohomology in the cotangent bundle coincides with Morse/Novikov cohomology on the base manifold [35–37].

The analysis of the charged chiral spectrum implies that if $\# \operatorname{Crit}_i(f) > h^i(M, \partial^+ M)$, then there are chiral fields in the spectrum whose masses are exponentially suppressed compared to the GUT or KK scale:

$$M^2 \sim e^{-\frac{1}{\alpha_{GUT}}} M_{GUT}^2. \tag{3.45}$$

These massive modes modify the running of the gauge couplings below the GUT scale (which may be identified with the KK scale up to threshold corrections), and may provide channels for proton decay. The GUT breaking mechanism of [38] using discrete Wilson lines in fact *requires* the existence of such relatively light massive modes, namely the Higgsino triplets, and depending on the model there could be additional modes. The GUT breaking mechanism of [38] in principle also allow one to eliminate dimension five operators leading to proton decay. However with such a low mass for the triplets, even dimension six proton decay could lead to trouble. Therefore it was suggested in [39] that for the GUT breaking mechanism in [38] to be viable, one should also modify the running so as to get $\alpha_{GUT} \sim 0.2 - 0.3$. This could be engineered by having additional $\mathbf{5}\,\overline{\mathbf{5}}$ pairs below the GUT scale. We will briefly suggest a different GUT breaking mechanism in Section 3.7.

3.4. Comments on anomaly cancellation

In type IIa string theory, non-abelian anomalies due to chiral fermions at brane intersections can be cancelled by anomaly inflow. To see this, we have an interaction

$$S_{CS} = \int C^{(1)} \wedge \operatorname{Tr}(F^3) = -\int F^{(2)} \wedge \omega_5(A). \tag{3.46}$$

Under a gauge transformation Λ we have

$$\delta_\Lambda S_{CS} = -\int F^{(2)} \wedge \delta_\Lambda \omega_5(A) = \int dF^{(2)} \wedge I_4^1(A, \Lambda). \tag{3.47}$$

Further, 6-branes are monopole configurations for $C^{(1)}$:

$$dF^{(2)}/2\pi \sim \sum_a N_a \delta^3(P_a). \tag{3.48}$$

Therefore the term $\delta_\Lambda S_{CS}$ above is of precisely the right form to cancel non-abelian anomalies due to chiral fermions at brane intersections.

In M-theory, the RR gauge field $C^{(1)}$ is generally massive, and so is not included as a propagating degree of freedom in the effective action. However after integrating it out, there must be a residual interaction which is not invariant under gauge transformations, for otherwise the anomalies above could not be cancelled. Ref. [40] explained what this residual interaction looks like in a local model. The geometry near a 6-brane locally looks like an A_n ALE fibration over S^3. The ALE has a natural $U(1)$ isometry that commutes with the holonomy, and as we fibre over S^3 we may get a non-trivial $U(1)$ bundle. The curvature of the corresponding $U(1)$ bundle over S^3 is denoted by K; it becomes $F^{(2)}$ in a IIa limit. Then in [40] it is essentially argued that the coupling (3.46) survives in M-theory in the following form:

$$\int d^7x \frac{K}{2\pi} \wedge \omega_5(A), \qquad dK = \sum n_a \delta_{P_a} \tag{3.49}$$

where P_a are the locations of chiral and anti-chiral matter, and n_a their multiplicities. As written, this interaction only makes sense for local geometries which are fibred by A_n ALE spaces and not for more general G_2-manifolds, but presumably there is a more general expression that reduces to the above one in a scaling limit. This is an additional coupling in the action beyond the terms we have considered so far, which is needed for consistency. The associated tadpole constraint expresses cancellation of the non-abelian anomalies.

One may also consider abelian anomalies. Let us briefly review some arguments in [40] (see also [41]). Chiral fields are localized at codimension 7 singularities, and locally the G_2 metric may be written as

$$ds^2 \sim dr^2 + r^2 ds^2_{Y_\alpha} \tag{3.50}$$

where r is a radial coordinate. We cut such a local neighbourhood around each codimension 7 singularity. Then our G_2-manifold X_7 becomes a manifold with boundary X'_7, with $\partial X'_7 = \cup_\alpha Y_\alpha$. We expand C_3 in harmonic forms on X'_7:

$$C_3 \sim A_i \wedge \omega^i + a_j \wedge \chi^j. \tag{3.51}$$

Here ω^i are harmonic two-forms and χ^j are harmonic three-forms, so the A_i are our four-dimensional abelian gauge fields and the a_j are four-dimensional axions. Then under a gauge transformation $C_3 \to C_3 + d\Lambda$, the Chern–Simons term varies as

$$\delta_\Lambda S_{CS} \sim -\int_{R^4 \times \partial X} \Lambda \wedge G \wedge G \sim -\sum_\alpha \int_{R^4} \epsilon_i F_j \wedge F_k \int_{Y_\alpha} \omega^i \wedge \omega^j \wedge \omega^k \tag{3.52}$$

where we decomposed Λ as $\Lambda \sim \epsilon_i \omega^i$. This is cancelled if there are fermion zero modes ψ_σ localized at the singularity, with $U(1)$ anomalies given by

$$\sum_\sigma q_\sigma^i q_\sigma^j q_\sigma^k = \int_{Y_\alpha} \omega^i \wedge \omega^j \wedge \omega^k. \tag{3.53}$$

Moreover from Stokes' theorem we get

$$0 = \sum_{\sigma,\alpha} q_\sigma^i q_\sigma^j q_\sigma^k \tag{3.54}$$

which one interprets as cancellation of the cubic $U(1)$ anomalies. Similarly one may discuss mixed abelian-gravitational anomalies and mixed abelian/non-abelian anomalies.

At first sight there is one puzzling aspect about this derivation. It implies that in M-theory compactifications on G_2-manifolds, the chiral spectrum is such that the light $U(1)$'s are always non-anomalous, and no Green–Schwarz mechanism is ever needed. On the other hand, type IIa with $D6$ branes lifts to M-theory on G_2, and there can also be heterotic duals. In both of these contexts, there can be light anomalous $U(1)$'s and there is a Green–Schwarz mechanism for their cancellation.

The likely resolution to this puzzle is as follows. In these other settings, the anomalous $U(1)$'s obtain a mass of order $\sim g_{YM,4}\ell_s^{-1}$ through the Green–Schwarz terms, and this mass is always parametrically lighter than the KK scale. Thus it makes sense to have anomalous $U(1)$'s below the KK scale and a Green–Schwarz mechanism for cancelling their anomalies. This is true even in F-theory. However the lift from IIa to M-theory is a little more subtle, and it seems a priori possible that such anomalous $U(1)$'s in type IIa will lift to massive $U(1)$'s in M-theory with masses scaling like $1/R_{KK}$. In that case we should treat these massive $U(1)$'s on the same footing as other massive $U(1)$ gauge bosons, and there will be no anomalous $U(1)$'s in the effective action below the KK scale, in agreement with the arguments of [40].

3.5. Techniques from algebraic topology

In this section we would like to apply some simple techniques from algebraic topology in order to compute the relative Betti numbers $h^i(M, \partial^+ M)$ in terms of the topological properties of the positive and negative boundaries. We actually work with the homology, which can easily be dualized to cohomology. We assume that the positive charge density is smeared along a graph Δ^+ with n_+ components and ℓ_+ loops, and similarly the negative charge density is smeared along a graph Δ^- with n_- components and ℓ_- loops. For some examples see Fig. 5 in Section 3.6. The open manifold M is identified with $S^3 \setminus (\Delta^+ \cup \Delta^-)$.

In order to compute the relative Betti numbers and the Euler character, there are two relevant long exact sequences. The first is the Mayer–Vietoris sequence. Suppose a manifold X is covered by two open sets U, V. Then we have the long exact sequence

$$\cdots \to H_i(U \cap V) \to H_i(U) \oplus H_i(V) \to H_i(X) \to H_{i-1}(U \cap V) \to \cdots. \tag{3.55}$$

In particular, the Euler characters are related as

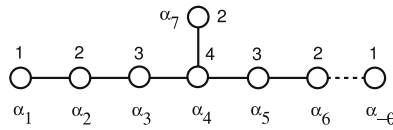$$\chi(X) = \chi(U) + \chi(V) - \chi(U \cap V). \tag{3.56}$$

Fig. 4. The extended E_7 Dynkin diagram and Dynkin indices.

In the application we have in mind, $X = S^3$, U is S^3 with a small tubular neighbourhood of the negative boundary excised, and V is itself a tubular neighbourhood of the negative boundary. Suppose the graph has n_- components and has ℓ_- loops. Then $U \cap V$ is topologically a collection of higher genus Riemann surfaces, so we have $b_i(U \cap V) = \{n_-, 2\ell_-, n_-, 0\}$. Furthermore we have $b_i(V) = b_i(\Delta^-)$. It follows from the long exact sequence that

$$b_3(S^3 \setminus \Delta^-) = 0, \qquad b_2(S^3 \setminus \Delta^-) = n_- - 1, \qquad b_1(S^3 \setminus \Delta^-) = \ell_-, \qquad b_0(S^3 \setminus \Delta^-) = 1. \tag{3.57}$$

In particular

$$\chi(S^3 \setminus \Delta^-) = n_- - \ell_-. \tag{3.58}$$

The second sequence we need is the one for relative homology. If A is a subset of X, then we have

$$\ldots \to H_i(A) \to H_i(X) \to H_i(X, A) \to H_{i-1}(A) \to \ldots. \tag{3.59}$$

In particular

$$\chi(X, A) = \chi(X) - \chi(A). \tag{3.60}$$

For our application, we would like to take X to be $M \sim S^3 \setminus (\Delta^+ \cup \Delta^-)$, and A to be a small tubular neighbourhood of $\Delta^+ \sim \partial^+ M$. In fact by the excision axiom, because $\Delta^+ \subset A$, it is equivalent to take $X = S^3 \setminus \Delta^-$. The precise Betti numbers $b_1(X, A)$ and $b_2(X, A)$ depend on the details of how the graphs are linked. From the long exact sequence, we find that they are given by

$$\begin{aligned} b_2(M, \partial^+ M) &= n_- - 1 + \ell_+ - r \\ b_1(M, \partial^+ M) &= n_+ - 1 + \ell_- - r \end{aligned} \quad 0 \le r \le \min(\ell_+, \ell_-). \tag{3.61}$$

Here r is the rank of the inclusion map

$$H_1(\partial^+ M) \to H_1(S^3 \setminus \Delta^-). \tag{3.62}$$

That is, any loop in Δ^+ is naturally embedded in $S^3 \setminus \Delta^-$, and r counts the number of loops that remain independent in homology after embedding. Thus we see that r indeed describes the linking between the positive and negative graphs. If the graphs are unlinked then $r = 0$. Moreover the Euler character is easily calculated and given by

$$\chi(M, \partial^+ M) = \chi(M) - \chi(\partial^+ M) = n_- - \ell_- - n_+ + \ell_+ \tag{3.63}$$

which is independent of r. Thus the net number of generations is easily computed from the topology of the positive and negative boundaries.

3.6. Example

In this section we would like to consider a toy GUT model, obtained by compactifying a gauge theory with gauge group E_7 on the three-manifold $Q_3 = S^3$. We will use the labelling of the roots shown in Fig. 4. The group E_7 contains a maximal $SO(10) \times U(1)_a \times U(1)_b$ subgroup, and we can take the $SO(10)$ to be generated by $\{\alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_7\}$. Under this subgroup, the adjoint representation of E_7 decomposes as

$$\begin{aligned} \mathbf{133} &= \mathbf{45}_{0,0} + \mathbf{1}_{0,0} + \mathbf{1}_{0,0} + \mathbf{r} + \bar{\mathbf{r}} \\ \mathbf{r} &= \mathbf{16}_{-2,1} + \mathbf{16}_{0,-3} + \mathbf{10}_{2,2} + \mathbf{1}_{-2,4}. \end{aligned} \tag{3.64}$$

In terms of our canonical basis ω_i dual to α_j, the two $U(1)$'s correspond to

$$\omega_a = 2\omega_1, \qquad \omega_b = 2\omega_1 - 3\omega_6. \tag{3.65}$$

We would like to have an unbroken $SO(10)$ gauge group after compactification. In order to achieve this breaking pattern, we need to turn on a profile for the Higgs fields corresponding to $U(1)_a$ and $U(1)_b$. The Morse functions will be denoted as $\log h_1 = f_1$ and $\log h_6 = f_6$. That is, the FI parameters are given by

$$\lambda_1 = df_1, \qquad \lambda_6 = df_6, \qquad \lambda_\theta = -df_1 - 2df_6. \tag{3.66}$$
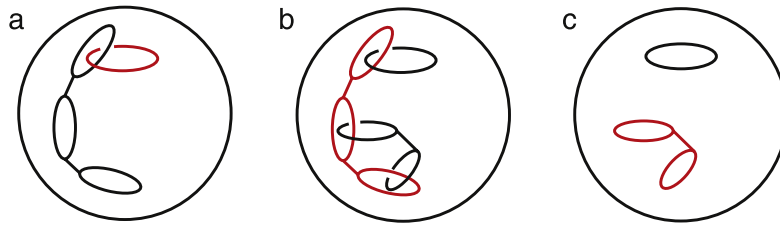
Fig. 5. Positively charged (red) and negatively charged (black) boundaries of Morse functions on S^3. Case (a) yields an electro-static potential that we call f_6, case (c) yields the potential f_1, and case (b) corresponds to the linear combination $-f_1 - f_6$. Note that we took one of the pieces of the boundary common to f_1 and f_6, but with the charge density for f_6 dominant there. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Then the various chiral fields are counted by the following Morse cohomology groups:

$$
\begin{aligned}
&\mathbf{16}_{0,-3} \to h^1(M, f_6) \quad &&\mathbf{16}_{-2,1} \to h^1(M, -f_1 - f_6) \\
&\mathbf{10}_{2,2} \to h^1(M, f_1) \quad &&\mathbf{1}_{-2,4} \to h^1(M, -f_1 - 2f_6).
\end{aligned}
\tag{3.67}
$$

Similarly the chiral fields in the conjugate representations are counted by $h^2(M, f)$.

Now we will define the electro-static potentials f_1 and f_6 by specifying suitable positive and negative charges on S^3. An example is shown in Fig. 5. To compute the spectrum, we use the formulae derived in Section 3.5, dualized to cohomology:

$$
\begin{aligned}
h^1(M, \partial^+ M) &= n_+ - 1 + \ell_- - r \\
h^2(M, \partial^+ M) &= n_- - 1 + \ell_+ - r
\end{aligned}
\quad 0 \le r \le \min(\ell_+, \ell_-).
\tag{3.68}
$$

Using the charge configurations in Fig. 5, we find

$$
\begin{aligned}
&h^1(M, f_6) = 2 \quad &&h^1(M, -f_6 - f_1) = 1 \quad &&h^1(M, f_1) = 1 \\
&h^2(M, f_6) = 0 \quad &&h^2(M, -f_6 - f_1) = 0 \quad &&h^2(M, f_1) = 2.
\end{aligned}
\tag{3.69}
$$

Therefore we have precisely three chiral generations in the $\mathbf{16}$, three Higgs fields, and also one chiral field in the representation $\mathbf{1}_{-2,4}$. The Yukawa coupling

$$
\mathbf{16}_{0,-3} \times \mathbf{16}_{-2,1} \times \mathbf{10}_{2,2}
\tag{3.70}
$$

is in principle allowed by the symmetries.

This example is unrealistic in a number of ways. In our non-compact set-up the $U(1)_a \times U(1)_b$ is not dynamical, but it imposes some strong selection rules. Since the $U(1)$'s are anomalous under the spectrum of the local model, we expect them to become massive upon compactification, but interactions violating the selection rules would then still be suppressed by the compactification scale. A nicer way would be to break the extra $U(1)$'s by using non-abelian Higgs fields. Or one could start with an abelian configuration and try to turn on an expectation value for the $\mathbf{1}_{-2,4}$. In the present example such a VEV would still leave an unbroken $U(1)$, so it would be better to start with E_8 instead since this can be broken to $SO(10)$ by turning on an $SU(4)$ valued Higgs field. Finally we have not yet incorporated a mechanism for breaking the GUT group to the Standard Model.

3.7. Breaking the GUT group with an abelian Higgs field

In order to break the GUT group, we can in principle proceed in two ways. One method is to engineer a four-dimensional Higgs field that can do the breaking, like an adjoint field. This is not possible if we work with $Q_3 = S^3/\Gamma$ since $h_1(S^3/\Gamma) = 0$, and even if it were one would end up with a conventional four-dimensional GUT model with the associated problems like doublet–triplet splitting.

The other method is to give a VEV to a charged field in the higher dimensional gauge theory in the process of compactification. The available charged fields which can get a VEV are the $7d$ gauge field and the $7d$ Higgs field.

We can turn on a VEV for the gauge field in two ways. If $\pi_1(Q_3)$ is non-zero we can use discrete Wilson lines with a hypercharge component. This is the conventional mechanism used in the heterotic string and introduced in the M-theory context in [38]. Unfortunately as we already discussed, this mechanism leads to light Higgsino triplets and looks somewhat less than desirable. The other possibility is to turn on a flux with a hypercharge component. However we have $h^2(S^3/\Gamma) = 0$ so this is not available for us [42].

There is then still one other option: we could turn on an abelian hypercharged Higgs field to break $SU(5) \to SU(3) \times SU(2) \times U(1)$. This corresponds to a reducible spectral cover. Apart from the usual Standard Model fields which descend from the $\mathbf{10}$ and $\overline{\mathbf{5}}$ however, there can be additional unwanted matter. Let us denote the pair (C, L) simply by $\mathcal{L}$. The spectral cover moduli decompose as

$$
HF^1(\mathcal{L}, \mathcal{L}) = \sum_{m,n} HF^1(\mathcal{L}_m, \mathcal{L}_n)
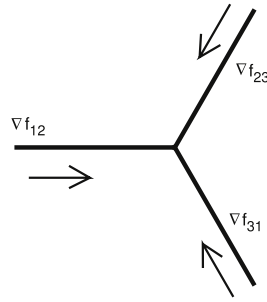\tag{3.71}
$$

Fig. 6. Graph of gradient flow trajectories relevant for Yukawa couplings.

where m, n run over the irreducible components. The off-diagonal components are charged under the extra $U(1)$'s, which includes hypercharge in this case. We saw an example of this in Section 3.6, where we had a scalar $\mathbf{1}_{-2,4}$. So in this scenario one has to make sure that such hypercharged moduli are massive, i.e. there is no deformation to a smooth spectral cover, because there are no light hypercharged scalars in the real world. This can likely be arranged by a suitable boundary condition or by turning on flat spectral line bundles on the irreducible components of the cover which cannot be obtained as a limit of a smooth line bundle after deformation. The crucial part is to check that this can be done while still satisfying the D-terms.

Another possible issue is that in the local set-ups we have been discussing, the $U(1)$ would not be dynamical. It can become dynamical when embedded in a compact model but it may have some UV sensitivity. Nevertheless this would lead to very different signatures than breaking by discrete Wilson lines, and so may be worth pursuing.

3.8. Superpotential terms

In this section we would like to explain how superpotential couplings are recovered from Morse theory. In the $1/t$ expansion we can think of such couplings as generated by membrane instantons or disc instantons, which map to trees of gradient flow trajectories on S^3. The prescriptions were originally found by [37,43]. We focus on the Yukawa couplings and quartic couplings. This is all that is needed for practical purposes. The Yukawa couplings correspond to the cup product on cohomology, and the quartic couplings to the Massey product, and so once more we see that the F-term data reflects the underlying topology and may be computed using alternative, combinatorial methods.

To define a three-point function we need three functions f_i, $i = 1, 2, 3$ such that the differences $f_{ij} = f_i - f_j$ are Morse functions. In practice one of the f_i will correspond to the zero section and so is taken constant. Our chiral fields correspond to certain linear combinations of index one critical points of the f_{ij}. We assume that a chiral field can be associated to a definite critical point. One can extend the definition of the three-point function using linearity.

The three-point function is given by the classical overlap of the wave functions of the chiral zero modes. In the $1/t$ approximation, it receives contributions from minimal area membrane instantons, which map to graphs of gradient flow trajectories in S^3 along the lines we discussed in Section 3.3. Namely we consider embedding graph as in Fig. 6 into S^3, such that the edges get mapped to gradient flows of f_{ij} as indicated, and the ends get mapped to an index one critical point of f_{ij}. The moduli space of such graphs is denoted by $M(p_{12}, p_{23}, p_{31})$. When non-empty, it is a manifold of dimension

$$\dim M(p_{12}, p_{23}, p_{31}) = m(p_{12}) + m(p_{23}) + m(p_{31}) - 3. \tag{3.72}$$

In particular when $m(p_{12}) = m(p_{23}) = m(p_{31}) = 1$, it is a finite set of points counted with signs. The action of such an instanton is given by

$$\exp -|f_{12}(p_{12}) + f_{23}(p_{23}) + f_{31}(p_{31})| \tag{3.73}$$

and so the Yukawa coupling is given by

$$\lambda_{123} = \sum_{graphs} n(p_{12}, p_{23}, p_{31}) \, e^{-|f_{12}(p_{12}) + f_{23}(p_{23}) + f_{31}(p_{31})|}. \tag{3.74}$$

Although the exponential factors may be scaled out by field redefinitions and are therefore ignored in the mathematics literature, they are physically relevant because the field redefinitions would change the normalization of the kinetic terms.

Similarly we may recover the four-point function. In this case use four functions f_i, $i = 1, 2, 3, 4$ and we look for graphs of gradient trajectories as in Fig. 7. The virtual dimension of the moduli space of graphs of type (a) and (b) is given by

$$\dim M(p_{12}, p_{23}, p_{34}, p_{41}) = m(p_{12}) + m(p_{23}) + m(p_{34}) + m(p_{41}) - 4. \tag{3.75}$$

These two moduli spaces are glued along the moduli space of graphs of type (c), whose dimension is one lower. Again we are interested in the case when all $m(i) = 1$, in which case we can count a discrete number of graphs of type (a) and (b). The quartic coupling is defined as

$$\lambda_{1234} = \sum_{graphs} \#M(p_{12}, p_{23}, p_{34}, p_{41}) \, e^{-|\sum_i f_{i,i+1}(p_{i,i+1})|}. \tag{3.76}$$
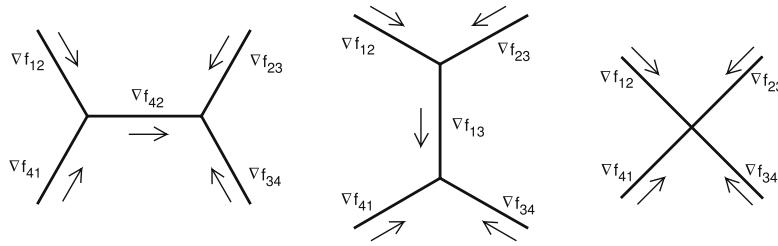
Fig. 7. Graphs of gradient flow trajectories generating quartic couplings.

There exist alternative ways to compute the four-point function. Once the boundary operator and the three-point functions have been obtained, one may also construct higher point functions through a recursion relation [36,44]. This may be thought of as a Feynman diagram expansion of our Chern–Simons theory (2.25), however it is more general in that it does not necessarily require that we use the Morse–Smale complex. In case of the four-point function this computes the (length 3) Massey product. If the boundary operator and the three-point function have been normalized correctly, these two constructions should be equivalent up to field redefinition of the form $\Phi_i \to \Phi_i + c_{ijk}\Phi_j\Phi_k+$ higher order. There are generalizations to higher point functions along the same lines.

3.9. Unsuppressed couplings and degenerate critical points

So far we have assumed that chiral fields are localized at non-degenerate critical points. In this case the interactions between chiral fields are described as being generated by instanton effects. Interactions are therefore generically small, and this will lead to phenomenological problems. In particular, this would lead to the prediction that the top quark Yukawa coupling is rather small, whereas in fact it turns out to be of order one.

In order to get unsuppressed couplings, we must tune the boundary conditions on the flavour branes so that the endpoints of the trivalent graph for the Yukawa coupling live close together. In this limit, the instanton approximation is not good, but we can get another weakly coupled description by analysing the Dirac equation and couplings directly for a degenerate (non-Morse) critical point. Although we have not carried out the analysis, one expects to be able to get a non-zero classical contribution to the Yukawa couplings this way, at least for the down-type or off-diagonal up-type Yukawa couplings.

This however may still not be enough for the top quark, as was pointed out in [45]. In our present language, if the Q_L and U_R come from the same critical point, then the trivalent graph for the Yukawa coupling must have two legs on that critical point. This seems impossible to achieve if the graph must live in a small local neighbourhood of the critical point, because the two **10**'s in the **10** · **10** · **5** coupling have different weights under the holonomy group of the Higgs bundle (i.e. they live on different sheets of the spectral cover), and so one of the legs of the graph will have to pass through a branch point or wrap around a boundary. Thus no matter how close the **10** and **5**$_h$, the diagonal up-type Yukawa couplings would still seem to be subject to some exponential suppression. Overcoming this issue is one of the key points that needs to be addressed in M-theory phenomenology.

It is also interesting to point out that one needs exceptional gauge groups in $7d$ to describe the top quark Yukawa coupling in M-theory, just as in F-theory or the heterotic string, and essentially for the same reason [45]. In the language of the present paper, this is because the fermion zero modes live in non-trivial representations of the holonomy group of the Higgs bundle, which is itself a subgroup of the (complexified) gauge group in $7d$. Interaction terms must be invariant however, so it must be possible to make the Yukawa coupling a singlet under this holonomy group. In the case of the top quark Yukawa coupling in an $SU(5)$ GUT model, this singles out the exceptional groups for purely group theoretical reasons.

In F-theory one may take a scaling limit to map models to a IIb description [8]. The top quark Yukawa is then described as a $D1$ instanton effect. An analogous limit is not yet known for G_2-manifolds, but we know that membrane instantons get mapped to worldsheet disc instantons and $D2$-instantons. As usual, the top quark Yukawa is charged under the $U(1) \subset U(5)$ in the IIa description. This $U(1)$ is anomalous but it is parametrically lighter than the KK scale in the IIa limit. The instanton action will then have to shift under the $U(1)$ gauge transformation in order to compensate for the lack of invariance of the up-type Yukawa coupling, or it should be neutral for the case of the down-type coupling. Since the Ramond three-form shifts under such a gauge transformation and the NS two-form does not, we expect that membrane instantons get mapped to $D2$ instantons if they generate an up-type Yukawa, and to disc instantons if they generate a down-type Yukawa.

3.10. Novikov cohomology

Now we discuss the more general possibilities for the boundary behaviour of abelian Higgs fields:

$$A \sim \alpha d\theta$$
$$\phi \sim \beta d \log r + \gamma d\theta. \tag{3.77}$$

First we consider $\beta, \gamma \neq 0$ but still $\alpha = 0$. In this case we can no longer define an electro-static potential, since it would have to be multivalued. Related to this, there may be an infinite number of gradient flow trajectories connecting two critical points, and hence the boundary operator is no longer well defined. However we can still define a potential on a multiple cover of M, define critical points and flow lines, and define a equivariant boundary operator and a cohomology theory. This more general cohomology is called Novikov cohomology. We will next explain how this arises from the supersymmetric quantum mechanics discussed previously, which can actually be defined by replacing df by an arbitrary closed one-form ϕ.

The Novikov homology depends on the cohomology class $[\phi]$ of the closed one-form, which specifies the monodromies. The D-terms imply that ϕ is harmonic. By Hodge theory arguments, there should be a unique harmonic form in the class whose periods are specified by $[\phi]$. Given a cohomology class, we may construct the minimal cover $\tilde{M}_\phi$ on which ϕ becomes exact, $\pi^*\phi = df$. Generically this is the universal cover. Denote by Γ the group of covering transformations. Then f satisfies $f(gx) = f(x) + \xi(g)$, $g \in \Gamma$, where $\xi(g)$ are the 'periods' of $[\phi]$. Note that $\xi(g)$ is independent of x.

Since locally the situation is exactly the same as when the periods of ϕ vanish, we still get a chiral fermion for each critical point of ϕ in the $1/t$ expansion. Thus the generators of the Novikov cohomology are still given by the critical points. However the coefficients will no longer be real-valued but rather valued in a power series in order to keep track of additional information. As before the action of d_t defines a boundary operator

$$d_t |a\rangle = \sum_{b \in \mathrm{Cr}_{i+1}(M), g \in \Gamma} n(a, gb) e^{-t(f(gb)-f(a))} |b\rangle . \tag{3.78}$$

Note that $n(a, gb)$ is only non-zero when $f(gb) > f(a)$. By rescaling $|a\rangle$, $|b\rangle$ we can write this as

$$d_t |a\rangle = \sum_{|b\rangle \in \mathrm{Cr}_{i+1}(M), g \in \Gamma} n(a, gb) q^{\xi(g)} |b\rangle \tag{3.79}$$

where $q = e^{-t}$ and $q^{\xi(g)}$ corresponds to the part of the instanton action that cannot be scaled out.

We see that the 'numbers' multiplying $|b\rangle$ are no longer integers, but live in a ring of power series called the Novikov ring $Nov(\Gamma)$. It is defined as the ring of formal power series

$$Nov(\Gamma) = \left\{ \sum_{i=0}^{\infty} n_i q^{\gamma_i} \,|\, \gamma_i \in \mathbf{R}, \gamma_i < \gamma_{i+1}, \gamma_i \to -\infty \right\} \tag{3.80}$$

i.e. γ_i takes values in a discrete set which tends to minus infinity, and the n_i are integers. The cohomology of d_t is called the Novikov cohomology:

$$H^*_{Nov}(M, [\phi]). \tag{3.81}$$

Since the chain complex is a module over $Nov(\Gamma)$, hence so is the cohomology. It is known that $H^*_N(M, [\phi])$ decomposes as a finite sum of free and torsion modules over $Nov(\Gamma)$.

The zero modes we are after are annihilated by both d_t and $d_t^\dagger$. This number may jump for a finite number of values of t, but excluding those values it is independent of t. Hence the generic number of zero modes corresponds to the number of generators of the cohomology of d_t over the Novikov ring, i.e. by the Novikov–Betti numbers:

$$h^i(M, [\phi]) = \mathrm{rank}_{Nov(\Gamma)} H^i_N(M, [\phi]). \tag{3.82}$$

If $[\phi] = 0$, i.e. if $\phi = df$ for some f, this reduces to the usual Betti numbers. Since our Dirac equation is defined over the real numbers, we can ignore the torsion. Thus $h^i(M, [\phi])$ counts the massless chiral matter.

It is natural to expect that the effect of turning on additional monodromy is to reduce the total amount of massless matter, i.e. we expect

$$h^i(M, [\phi]) \leq h^i(M, [\phi] = 0). \tag{3.83}$$

In the case without boundary, a proof can be found in [46]. Furthermore, if the periods can be turned off continuously (as would certainly be the case if we work on a simply connected manifold like S^3), then the net amount of chiral matter can clearly not change. Thus

$$h^2(M, [\phi]) - h^1(M, [\phi]) = \chi(M, \partial^+ M). \tag{3.84}$$

As we vary $[\phi]$ in $H^1(M, \mathbf{R})$, the Betti numbers $h^i(M, [\phi])$ are generically constant, but may jump up on algebraic subsets of $H^1(M, \mathbf{R})$. However they must always satisfy the identities above.

Finally we would like to also allow for $\alpha \neq 0$, i.e. we turn on a flat spectral line bundle. In this case we also need to take into account the holonomy of the gauge field, and the action of d_t is modified to

$$d_t |a\rangle = \sum_{b \in \mathrm{Cr}_{i+1}(M), g \in \Gamma} n(a, gb) \, e^{i \int_a^{gb} A} \, e^{-\frac{1}{2}t(f(gb)-f(a))} |b\rangle . \tag{3.85}$$

Again after scaling out a common piece, we can write this as

$$d_t \, |a\rangle = \sum_{b \in \mathrm{Cr}_{i+1}(M), g \in \Gamma} n(a, gb) \, e^{i\rho(g)} q^{\xi(g)} \, |b\rangle \, . \tag{3.86}$$

Here $\rho(g) = \int_b^{gb} A$ is the representation describing the holonomies of the flat line bundle. This yields a twisted version of Novikov cohomology. Not much seems to be known about it. However when $\gamma \to 0$ it reduces to the cohomology with values in L, i.e. $H^i(M, L)$ [46]. Since we would expect that turning on γ generally only decreases the Betti numbers, this can be used to get even stronger constraints on the spectrum. If we can continuously set α to zero, then the net number of chirals is not affected.

So all in all, in the abelian case we have a fair amount of control. The Betti numbers are computable when $\gamma = 0$ and should only decrease when we turn on Higgs field monodromies starting from a configuration with $\gamma = 0$. Further under such continuous deformations, the net number of chirals is unchanged. In all cases, thinking in the spectral cover picture, we can 'close the loop' of the boundary of a disc instanton by adding a segment along the zero section, recovering the definition of Floer cohomology on the cotangent bundle.

3.11. Non-abelian Higgs bundles

In this paper we have focused primarily on abelian configurations. In order to make realistic models, one is naturally interested in non-abelian configurations. In this case, there are two challenges to be addressed. First, for the configurations of interest the Higgs field is not finite everywhere, but a generalization of Corlette's theorem in such situations has not been established. Thus when we try to write a model by specifying boundary data, we do not know if a solution to Hitchin's equations exists. Second, even if such a solution exists, we have only presented formal expressions for the spectrum and the interactions, given by the cohomology groups of the Higgs bundle and the product structure on this cohomology. We still need to generalize the tools we have developed for abelian Higgs bundles in order to explicitly compute them.

Both these issues may be addressed by employing a Fourier–Mukai transform to turn the problem into an algebro-geometric one. Writing down holomorphic brane configurations and computing their spectrum and interactions is much more straightforward. This is the subject of forthcoming work [9].

3.12. Knot invariants

Given a knot, link or graph Δ in S^3 with specified $Gl(n, \mathbf{C})$ monodromies around it, there should be Lagrangian three-cycle in T^*S^3 with boundary specified by Δ. As discussed in this paper, such a configuration defines an effective $4d$ gauge theory, in fact it defines a whole ensemble of gauge theories since we can embed the holonomy in different gauge groups. Similar configurations have been considered before but focus on the Wilson line correlation functions of the CS theory as knot invariants. Here we see that there is another natural set of observables which can be used as invariants, namely the $4d$ massless spectrum of the compactified gauge theory and its superpotential (as a function of the massless chiral fields). Mathematically speaking, the invariants are given by the hypercohomology groups of the Higgs bundle, and by the structure constants of the Yoneda pairing and Massey products on these hypercohomology groups. We have already seen earlier in this paper that these invariants know about simple properties like the linking number.

The information in these invariants is in some sense stronger than what we get from Chern–Simons theory, because they yield generators of a Hilbert space and an algebra on these generators. It would be interesting to understand their properties in more detail.

Acknowledgements

Appendix. Some elements of Morse theory

We only summarize some of the basics here. See [47] for a classic account and [48,30] for a more modern perspective.

Consider a compact manifold M and a function $f : M \to \mathbf{R}$. The function f is said to be Morse if all its critical points are non-degenerate, i.e. the Hessian at the critical point has no zero eigenvalues. We pick an auxiliary metric g and define the gradient flow

$$\frac{\partial \vec{x}}{\partial \lambda} = -\nabla f. \tag{A.1}$$

We denote by $\phi_\lambda(x)$ the solution of this equation (the gradient flow trajectory) which starts at x at time $\lambda = 0$. Since any critical point is non-degenerate, up to coordinate transformations the only invariant data of the Hessian is the number of

positive and the number of negative eigenvalues of the Hessian. The number of negative eigenvalues is called the Morse index of the critical point.

Morse functions can be used to deduce the homotopy type of the manifold M. Given a critical point p, we define the unstable manifold of p to be the set of points on M that lie on a gradient flow trajectory starting at p:

$$W_u(p) = \{x \in M | \phi(x, -\infty) = p\}. \tag{A.2}$$

Similarly one may define the stable manifold $W_s(p)$, which consists for the gradient flow trajectories ending at p. Then $W_u(p)$ has the topology of a cell of dimension i, where i is the Morse index of p, and M is the union of these cells

$$M = \bigcup_p W_u(p). \tag{A.3}$$

To learn about the homotopy type, we have to understand how the cells are glued together. We define the manifolds-with-boundary

$$M_a = \{p \in M | f(p) \le a\}. \tag{A.4}$$

The manifolds M_a and M_b are diffeomorphic if there are no critical points in $f^{-1}[a, b]$. However if there is a critical point, then the topology changes. For a critical point c of Morse index i, the manifold $M_{c+\epsilon}$ is homotopy equivalent to $M_{c-\epsilon} \cup W_s(c)$. The proof is a local argument near each critical point, see [47] for details.

Unfortunately the above decomposition of M does not necessarily define a CW complex, and so the relation with the homology of the manifold remains unclear. The missing condition was introduced by Smale. The pair (f, g) is said to be Morse–Smale if the stable and unstable manifolds intersect each other transversally. This can be shown to be satisfied for generic (f, g).

Using such Morse–Smale pairs (f, g), one may construct a homology called the Morse homology. We define

$$C_i = \sum_{\mathrm{Crit}_i(f)} \mathbf{Z}p \tag{A.5}$$

to be the free abelian group generated by the critical points of Morse index i. The boundary operator is defined as follows. Given two critical points p, q, we denote the moduli space of gradient flow trajectories connecting p and q by $M(p, q)$:

$$M(p, q) = W_u(p) \cap W_s(q). \tag{A.6}$$

When non-empty, the dimension of $M(p, q)$ is $m(p) - m(q)$, the difference between the Morse indices. There is a natural action of the real line $\mathbf{R}$ on this moduli space, given by rescaling the parameter λ that parametrizes the gradient flow trajectory. Modding out by this rescaling, we define

$$n(p, q) = M(p, q)/\mathbf{R} \tag{A.7}$$

whose dimension is $m(p) - m(q) - 1$. If the Morse indices differ by one, then $n(p, q)$ is zero dimensional and counts the number of trajectories connecting p and q. The moduli spaces $M(p, q)$ and $n(p, q)$ come with natural orientations, and thus $n(p, q)$ counts trajectories with a sign. We can define a boundary operator

$$\partial : C_i \to C_{i-1}, \qquad \partial p = \sum_{q \in \mathrm{Crit}_{i-1}(f)} n(p, q)q. \tag{A.8}$$

The fact that $\partial^2 = 0$ relies on an analysis of broken flow lines, which lie on the boundary of the moduli space $M(p, r)$ with $m(p) - m(r) = 2$ and inherit their orientation from $M(p, r)$.

The correspondence with the ordinary homology of the underlying manifold can be understood by establishing an isomorphism with cellular homology, which is built from the free abelian groups generated by the cells of a cell decomposition. We have already seen how a Morse function gives rise to a cellular decomposition. To each critical point of index i we can associate an i-cell, namely the associated unstable manifold, and the original manifold is precisely the union of these cells. For Morse–Smale pairs, the boundary of such an i-cell is contained in the skeleton built from the k-cells with $k < i$. The unstable manifold $W_u(p)$ for $p \in C_i$ gets attached to the unstable manifolds $\bigcup_q W_u(q)$ where $q \in C_{i-1}$ runs over the critical points connected to p by gradient flow. The boundary operator of the cellular complex maps each cell to its boundary, with an induced orientation. Similarly the boundary operator of the Morse complex induces a boundary map on the cell complex, since there is a one–one correspondence between critical points and generators of the cell complex. It will probably not come as a surprise that these two boundary operators are equivalent.

Dually we can also define the Morse cohomology. This is more natural than homology for physicists, because it describes properties of functions on the underlying manifold, rather than the underlying manifold itself (although these two points of view are of course related by Poincaré duality). We write generators of C_i as bra-vectors, and define the dual as

$$C^i = \mathrm{Hom}(C_i, \mathbf{R}), \qquad \delta : C^i \to C^{i+1}, \qquad \langle r | \delta p \rangle = \langle \partial r | p \rangle \quad \forall \langle r | \in C_{i+1} \tag{A.9}$$
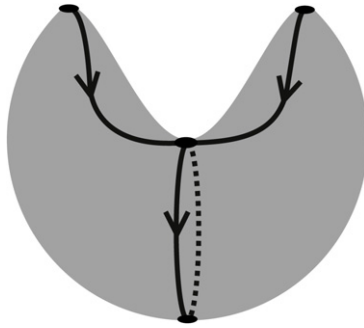
Fig. 8. Gradient flow trajectories connecting critical points of the height function on a deformed two-sphere.

or more explicitly, up to a similarity transformation,

$$\delta \, |p\rangle = \sum_{q \in \mathrm{Crit}_{i+1}(f)} n(q, p) \, |q\rangle .$$

(A.10)

Intuitively one may think of a generator $|p\rangle$ of C^i as a delta-function differential form localized at the critical point p, with i indices all of which lie along the unstable directions at p. We are using the same numbers $n(p, q)$ as before, but the Morse index is increasing instead of decreasing. Thus this is isomorphic to replacing negative gradient flow with positive gradient flow, or replacing $f \to -f$, which is the incarnation of Poincaré duality in this context.

Another way to think about Morse cohomology was introduced by Witten, and is closer to the M-theory/Yang–Mills picture. The conjugation

$$d \to e^{tf} d e^{-tf}$$

(A.11)

is a similarity transformation and induces an isomorphism between the ordinary De Rham cohomology at $t = 0$ and a deformed cohomology as $t \to \infty$. By the arguments reviewed in Section 3.3, the deformed cohomology is identified with the Morse cohomology. However since the two are related by a similarity transformation, the Betti numbers do not depend on t.

A well-known example is shown in Fig. 8. Let us apply Morse theory to the height function of the deformed two-sphere shown in this figure. There are four critical points, two of them with Morse index two, one with index one and one more with index zero. However the boundary operator maps the saddle point to the difference between the two maxima. Therefore the Morse cohomology is given by

$$H^2(S^2, f) = \mathbf{R}, \qquad H^1(S^2, f) = 0, \qquad H^0(S^2, f) = \mathbf{R}$$

(A.12)

which is of course the same as the ordinary cohomology of S^2.

References

[1] G. Aldazabal, L.E. Ibanez, F. Quevedo, A.M. Uranga, D-branes at singularities: A bottom-up approach to the string embedding of the standard model, JHEP 0008 (2000) 002. arXiv:hep-th/0005067.
[2] D. Berenstein, V. Jejjala, R.G. Leigh, The standard model on a D-brane, Phys. Rev. Lett. 88 (2002) 071602. arXiv:hep-ph/0105042.
[3] H. Verlinde, M. Wijnholt, Building the Standard Model on a D3-brane, JHEP 0701 (2007) 106. arXiv:hep-th/0508089.
[4] R. Donagi, M. Wijnholt, Model Building with F-Theory. arXiv:0802.2969 [hep-th].
[5] C. Beasley, J.J. Heckman, C. Vafa, GUTs and exceptional branes in F-theory - I, JHEP 0901 (2009) 058. arXiv:0802.3391 [hep-th].
[6] H. Hayashi, R. Tatar, Y. Toda, T. Watari, M. Yamazaki, New aspects of Heterotic–F theory duality, Nuclear. Phys. B 806 (2009) 224. arXiv:0805.1057 [hep-th].
[7] B.S. Acharya, K. Bobkov, G.L. Kane, J. Shao, P. Kumar, The G_2-MSSM - An M theory motivated model of particle physics, Phys. Rev. D 78 (2008) 065038. arXiv:0801.0478 [hep-ph].
[8] R. Donagi, M. Wijnholt, Higgs Bundles and UV Completion in F-Theory. arXiv:0904.1218 [hep-th].
[9] T. Pantev, M. Wijnholt, (in progress).
[10] N. Hitchin, The geometry of three-forms in six and seven dimensions. arXiv:math/0010054.
[11] M. Fernández, A. Gray, Riemannian manifolds with structure group G_2, Ann. Mat. Pura Appl. 132 (1982) 19–45.
[12] C. Beasley, E. Witten, A note on fluxes and superpotentials in M-theory compactifications on manifolds of G(2) holonomy, JHEP 0207 (2002) 046. arXiv:hep-th/0203061.
[13] S. Gukov, Solitons, superpotentials and calibrations, Nuclear. Phys. B 574 (2000) 169. arXiv:hep-th/9911011.
[14] B.S. Acharya, B.J. Spence, Flux, supersymmetry and M theory on 7-manifolds. arXiv:hep-th/0007213.
[15] N. Hitchin, Stable forms and special metrics. arXiv:math/0107101.
[16] D.D. Joyce, Compact Riemannian 7-manifolds with Holonomy G_2, I, J. Differential Geom. 43 (1996) 291–328.
[17] B.S. Acharya, M theory, Joyce orbifolds and super Yang–Mills, Adv. Theor. Math. Phys. 3 (1999) 227. arXiv:hep-th/9812205.
[18] G Tian, Smoothness of the universal deformation space of compact Calabi–Yau manifolds and its Petersson-Weil metric, in: S.-T. Yau (Ed.), Mathematical Aspects of String Theory, in: Advanced Series in Mathematical Physics, vol. 1, World Scientific Publishing Co, Singapore, 1987, pp. 629–646.
[19] A.N. Todorov, The Weil-Petersson geometry of the moduli space of $SU (n \geq 3)$ (Calabi–Yau) manifolds. I, Comm. Math. Phys. 126 (1989) 325–346.
[20] N.J. Hitchin, The self-duality equations on a riemann surface, Proc. Lond. Math. Soc. (3) 55 (1987) 59–126.

[21] N.C. Leung, S.T. Yau, E. Zaslow, From special Lagrangian to Hermitian–Yang–Mills via Fourier–Mukai transform, Adv. Theor. Math. Phys. 4 (2002) 1319. arXiv:math/0005118.
[22] S.K. Donaldson, Twisted harmonic maps and the self-duality equations, Proc. Lond. Math. Soc. (3) 55 (1987) 127–131.
[23] B.S. Acharya, A moduli fixing mechanism in M theory. arXiv:hep-th/0212294.
[24] K. Corlette, Flat *G*-bundles with canonical metrics, J. Differential Geom. 28 (1988) 361–382.
[25] H. Hayashi, T. Kawano, R. Tatar, T. Watari, Codimension-3 singularities and Yukawa couplings in F-theory. arXiv:0901.4941 [hep-th].
[26] R.P. Thomas, S.T. Yau, Special Lagrangians, stable bundles and mean curvature flow. arXiv:math/0104197.
[27] A. Strominger, Massless black holes and conifolds in string theory, Nuclear. Phys. B 451 (1995) 96. arXiv:hep-th/9504090.
[28] M. Berkooz, M.R. Douglas, R.G. Leigh, Branes intersecting at angles, Nuclear. Phys. B 480 (1996) 265. arXiv:hep-th/9606139.
[29] B. Acharya, E. Witten, Chiral fermions from manifolds of G(2) holonomy. arXiv:hep-th/0109152.
[30] E. Witten, Supersymmetry and Morse theory, J. Differential Geom. 17 (1982) 661.
[31] M. Atiyah, E. Witten, M-theory dynamics on a manifold of G(2) holonomy, Adv. Theor. Math. Phys. 6 (2003) 1. arXiv:hep-th/0107177.
[32] P. Berglund, A. Brandhuber, Matter from G(2) manifolds, Nuclear. Phys. B 641 (2002) 351. arXiv:hep-th/0205184.
[33] J.L. Bourjaily, Local models in F-theory and M-theory with three generations. arXiv:0901.3785 [hep-th].
[34] P. Salomonson, J.W. van Holten, Fermionic coordinates and supersymmetry in quantum mechanics, Nuclear. Phys. B 196 (1982) 509.
[35] A. Floer, Morse theory for Lagrangian intersections, J. Differ. Geom. 28 (3) (1988) 513–547.
[36] E. Witten, Chern–Simons gauge theory as a string theory, Progr. Math. 133 (1995) 637. arXiv:hep-th/9207094.
[37] K. Fukaya, Morse homotopy and its quantization, in: Geometric topology (Athens, GA, 1993), in: AMS/IP Stud. Adv. Math., vol. 2.1, Amer. Math. Soc, Providence, RI, 1997, pp. 409–440.
[38] E. Witten, Deconstruction, G(2) holonomy, and doublet-triplet splitting. arXiv:hep-ph/0201018.
[39] T. Friedmann, E. Witten, Unification scale, proton decay, and manifolds of G(2) holonomy, Adv. Theor. Math. Phys. 7 (2003) 577. arXiv:hep-th/0211269.
[40] E. Witten, Anomaly cancellation on G(2) manifolds. arXiv:hep-th/0108165.
[41] A. Bilal, S. Metzger, Anomalies in M-theory on singular G(2)-manifolds, Nuclear. Phys. B 672 (2003) 239. arXiv:hep-th/0303243.
[42] R. Donagi, M. Wijnholt, Breaking GUT groups in F-theory. arXiv:0808.2223 [hep-th].
[43] M. Betz, R.L. Cohen, Moduli spaces of graphs and cohomology operations, Turkish J. Math. (18) (1994) 23–41.
[44] S.A. Merkulov, Strong homotopy algebras of a Kähler manifold, Internat. Math. Res. Notices (3) (1999) 153–164. arXiv:math/9809172.
[45] R. Tatar, T. Watari, Proton decay, Yukawa couplings and underlying gauge symmetry in string theory, Nuclear. Phys. B 747 (2006) 212. arXiv:hep-th/0602238.
[46] M. Farber, Topology of Closed One-forms, AMS.
[47] J. Milnor, Morse Theory, Princeton University Press, 1963.
[48] M. Schwarz, Morse Homology, Birkhäuser Verlag, 1993.