Gauge Theory Enters Topology

In this rather brief section we attempt to provide some sense of how the analysis of a moduli space of gauge equivalence classes of connections on a principal bundle over a smoonth 4-manifold M can yield information about M itself. The details have been pruned away in order that the view of the forest not be obscured by myriad trees. We will have a somewhat closer look at some of these details in the following section, but for the full story one must turn to [ ], [ ], [ ], or [ ].

We will consider compact, simply connected, oriented, smooth (i.e., C^∞) 4-dimensional manifolds, a typical specimen of which will be denoted M. Examples include the 4-sphere S^4, the product $S^2 \times S^2$ of two 2-spheres, the complex projective planes $\mathbb{CP}^2$ (with its canonical orientation as a complex manifold) and $\overline{\mathbb{CP}}^{\,2}$ (with the opposite of the canonical orientation), and the Kummer surface $K3$ (the complex algebraic surface in $\mathbb{CP}^3$ whose homogeneous coordinates z_1, z_2, z_3, z_4 satisfy $z_1{}^4 + z_2{}^4 + z_3{}^4 + z_4{}^4 = 0$). As the need arises we will make various assumptions about the intersection form

$$Q_M : H_2(M; \mathbb{Z}) \oplus H_2(M; \mathbb{Z}) \longrightarrow \mathbb{Z}$$

of M. This is a unimodular, symmetric, $\mathbb{Z}$-valued, bilinear form on the second homology $H_2(M; \mathbb{Z})$ of M and can be thought of in the following way: Any two homology classes α_1 and α_2 in $H_2(M; \mathbb{Z})$ can be represented by smoothly embedded, compact, oriented surfaces $\sum_1$ and $\sum_2$ which can be assumed to intersect transversally and therefore in a finite set of isolated points.

Figure 1

An intersection point p is assigned the value $+1$ if an oriented basis for the tangent space $T_p(\sum_1)$ followed by an oriented basis for $T_p(\sum_2)$ gives an oriented basis for $T_p(M)$, and -1 otherwise. $Q_M(\alpha_1, \alpha_2)$ is then the sum of these values over all of the intersection points. Assuming that $H_2(M; \mathbb{Z}) \neq 0$, the maximal dimension of a subspace of $H_2(M; \mathbb{Z})$ on which Q_M is positive definite is denoted $b_2^+(M)$ and happens to agree with the dimension of the space of harmonic, self-dual 2-forms on M (for any choice of Riemannian metric on M). In the following brief list of examples, the intersection form is represented by a matrix relative to some particular basis for $H_2(M; \mathbb{Z})$ over $\mathbb{Z}$.

M	$H_2(M;\mathbb{Z})$	Q_M	$b_2^+(M)$
S^4	0	$\emptyset$	$-$
$\mathbb{C}\mathbb{P}^2$	$\mathbb{Z}$	(1)	1
$\overline{\mathbb{C}\mathbb{P}}^2$	$\mathbb{Z}$	(-1)	0
$S^2 \times S^2$	$\mathbb{Z} \oplus \mathbb{Z} = 2\mathbb{Z}$	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$	1
$K3$	$22\mathbb{Z}$	$3\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \oplus 2(-E_8)$	3

Here E_8 is given by

$$\begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 2 \end{pmatrix}.$$

Additional examples arise from connected sums of such manifolds, for which one can show that

$$H_2(M_1 \# M_2;\mathbb{Z}) \cong H_2(M_1;\mathbb{Z}) \oplus H_2(M_2;\mathbb{Z})$$

and

$$Q_{M_1 \# M_2} = Q_{M_1} \oplus Q_{M_2}.$$

Thus, for example,

$$Q_{p\,\mathbb{C}\mathbb{P}^2 \# q\,\overline{\mathbb{C}\mathbb{P}}^2} = \operatorname{diag}\left(1, \overset{p}{\cdots}, 1, -1, \overset{q}{\cdots}, -1\right)$$

and so

$$b_2^+\left(p\,\mathbb{C}\mathbb{P}^2 \# q\,\overline{\mathbb{C}\mathbb{P}}^2\right) = p.$$

The intersection form can also be defined for topological (as opposed to smooth) 4-manifolds, but we will not enter into this here.

It has been known for some time that the intersection form is a basic invariant for compact 4-manifolds. In 1949, Whitehead [ ] proved that two compact, simply connected 4-manifolds M_1 and M_2 have the same homotopy type if and only if their intersection forms are equivalent (i.e., there exist bases for $H_2(M_1;\mathbb{Z})$ and $H_2(M_2;\mathbb{Z})$ relative to which Q_{M_1} and Q_{M_2} have the same matrix). In 1982, Freedman [ ] showed that *every* unimodular, symmetric, $\mathbb{Z}$-valued, bilinear form on a finitely generated, free Abelian group is the intersection form of at least one (and at most two) compact, simply connected, oriented *topological* 4-manifold(s). In particular, this is true of the vast, impenetrable maze of (positive or negative) definite forms (if you do not believe that this is a "vast, impenetrable maze", consider the fact that, when the rank is 40, there are at least 10^{51} equivalence classes of definite forms). In 1983, Donaldson [ ] showed that the differential topologist need not venture into this maze because only *one* positive/negative definite unimodular, symmetric, $\mathbb{Z}$-valued, bilinear form can arise as the intersection form of a compact, simply connected, oriented, *smooth* 4-manifold.

Donaldson's 1983 Theorem: *Let M be a compact, simply connected, oriented, smooth 4-manifold with positive (respectively, negative) definite intersection form. Then there is a basis for $H_2(M;\mathbb{Z})$ relative to which the matrix of Q_M is the identity matrix (respectively, minus the identity matrix).*

Donaldson's Theorem is remarkable, but still more remarkable is its proof, which is a byproduct of the analysis of an instanton moduli space for M. We (very, very) briefly sketch the idea in the negative definite case, i.e., when $b_2^+(M) = 0$. Begin by considering the principal $SU(2)$-bundle over M with second Chern class 1, which we denote

$$SU(2) \hookrightarrow P_1 \xrightarrow{\pi_1} M$$

(an $SU(2)$-bundle over M is characterized up to equivalence by its second Chern class which can be identified with an integer k that we will describe shortly).

Next we choose some Riemannian metric g on M. Both the bundle and the metric are to be regarded as auxiliary structures to facilitate the study of M. From g and the given orientation of M one obtains a Hodge star operator $*$. Now, any connection ω on $SU(2) \hookrightarrow P_1 \xrightarrow{\pi_1} M$ has a curvature which can be identified with a 2-form $F_\omega \in \Omega^2(M, \operatorname{ad} P_1)$ on M taking values in the adjoint bundle $\operatorname{ad} P_1$ (this is the vector bundle associated to $SU(2) \hookrightarrow P_1 \xrightarrow{\pi_1} M$ by the adjoint action of $SU(2)$ on its Lie algebra $su(2)$). As such, F_ω has a Hodge dual ${}^*F_\omega \in \Omega^2(M, \operatorname{ad} P_1)$ and we will say that ω is g-**anti-self-dual** (g-**ASD**) if

$${}^*F_\omega = -F_\omega\ . \tag{1}$$

Remarks: It is a deep result of Taubes [ ] that, because $b_2^+(M) = 0$, such connections do, in fact, exist on $SU(2) \hookrightarrow P_1 \xrightarrow{\pi_1} M$. The **anti-self-dual equations** (1) also provide us with our first glimpse of physics lurking in the background since it is not difficult to show that any solution ω to them (they are, after all, partial differential equations for ω) also satisfies the so-called Yang-Mills equations which are a nonlinear generalization of Maxwell's equations used to describe interactions between elementary particles. Next, we remark that the discussion to follow could equally well be phrased in terms of self-dual (SD) connections (${}^*F_\omega = F_\omega$), but, in that case, one must begin with the Chern class -1 bundle $SU(2) \hookrightarrow P_{-1} \xrightarrow{\pi_{-1}} M$ since ASD connections can exist only on bundles with non-negative Chern class, whereas SD connections can exist only if the Chern class is non-positive. To see this we recall that, given a Riemannian metric g on M, the curvature F_ω of any connection ω on any $SU(2)$-bundle over M can be written as the sum $F_\omega = F_\omega^+ + F_\omega^-$ of its self-dual ($F_\omega^+ = \frac{1}{2}(F_\omega + {}^*F_\omega)$) and anti-self-dual ($F_\omega^- = \frac{1}{2}(F_\omega - {}^*F_\omega)$) parts and then

$$k = \frac{1}{8\pi^2}\int_M \operatorname{tr}(F_\omega \wedge F_\omega) = \frac{1}{8\pi^2}\int_M \Big(|F_\omega^-|^2 - |F_\omega^+|^2 \Big) \operatorname{Vol}_g . \qquad (2)$$

Thus, if ω is g-ASD ($F_\omega^+ = 0$), then $k \geq 0$, but if ω is g-SD, $k \leq 0$. Also note that, if $k = 0$, any g-ASD (or g-SD) connection if flat ($F_\omega = 0$).

A g-ASD connection ω is called an **instanton** and we now introduce an equivalence relation (called **gauge equivalence**) on the set of instantons on $SU(2) \hookrightarrow P_1 \xrightarrow{\pi_1} M$. For this we define the **gauge group** $\mathcal{G}(P_1)$ of our bundle to be the group (under composition) of all diffeomorphisms $f : P_1 \to P_1$ of P_1 onto itself which respect the group action of the bundle ($f(p \cdot g) = f(p) \cdot g$ for all $p \in P_1$ and all $g \in SU(2)$) and cover the identity on M ($\pi_1 \circ f = \pi_1$). The elements of $\mathcal{G}(P_1)$ are called (**global**) **gauge transformations** and they act on g-ASD connections by pullback ($\omega \cdot f = f^*\omega$). We will say that ω and ω' are **gauge equivalent** if $\omega' = f^*\omega$ for some $f \in \mathcal{G}(P_1)$.

Remark: Mathematically, this is the natural notion of equivalence for connections on a fixed bundle, but the real motivation actually lies in physics. There connections ω are referred to as **gauge fields** and their curvatures Ω are called **field strengths**. A local section s of the principal bundle is called a choice of **gauge** and the pullback $s^*\omega$ is a **gauge potential**. The **gauge principle** asserts that the laws of physics should be invariant under an arbitrary change of gauge so that quantities with the same set of gauge representations are to be regarded as physically equivalent. But if s is a section and $f \in \mathcal{G}(P_1)$, then $f \circ s$ is also a section and $s^*(f^*\omega) = (f \circ s)^*\omega$ so ω and $f^*\omega$ have the same set of gauge potentials.

Now, the collection of all gauge equivalence classes of g-ASD connections on $SU(2) \hookrightarrow P_1 \xrightarrow{\pi_1} M$ is denoted

$$\mathcal{M} = \mathcal{M}(P_1, g)$$

and called the **moduli space of g-ASD connections** (**instantons**) on $SU(2) \hookrightarrow P_1 \xrightarrow{\pi_1} M$. Donaldson has shown that, for a "generic" choice of the Riemannian metric g, $\mathcal{M}(P_1, g)$ looks like this:
Somewhat more precisely:

Figure 2

1. If m is half the number of $\alpha \in H_2(M\,;\,\mathbb{Z})$ with $Q_M(\alpha\,,\,\alpha) = -1$, then there exist points $p_1\,,\,\ldots\,,\,p_m$ in $\mathcal{M}$ such that $\mathcal{M} - \{p_1\,,\,\ldots\,,\,p_m\}$ is a smooth, orientable, 5-manifold.

2. Each $p_i\,,\ i = 1,\ldots\, m$, has a neighborhood in $\mathcal{M}$ homeomorphic to a cone over $\mathbb{CP}^2$ with p_i at the vertex (the cone over $\mathbb{CP}^2$ is obtained from the cylinder $\mathbb{CP}^2 \times [\,0\,,\,1\,]$ by identifying $\mathbb{CP}^2 \times \{\,1\,\}$ to a point p, called the vertex of the cone).

3. There is a compact subspace K of $\mathcal{M}$ such that $\mathcal{M} - K$ is an open submanifold of $\mathcal{M} - \{p_1\,,\,\ldots\,,\,p_m\}$ diffeomorphic to the cylinder $M \times (\,0\,,\,1\,)$.

Remark: Proving all of this is, to say the least, arduous. In the next section we will sketch, in a more general context, some of the ideas behind the arguments.

Now construct from $\mathcal{M}$ another space $\mathcal{M}_0$ by cutting off the top half of each cone and the bottom half of the cylinder.

Figure 3

Then $\mathcal{M}_0$ is compact (because K is compact) and oriented. It is also a smooth manifold with boundary whose boundary consists of a disjoint union of a copy of M and m copies of $\mathbb{CP}^2$ (or $\overline{\mathbb{CP}}^2$). Thus, $\mathcal{M}_0$ is a cobordism from M to a disjoint union $p\,\mathbb{CP}^2 \bigsqcup q\,\overline{\mathbb{CP}}^2$, where $p+q = m$. As it happens, the signature of the intersection form is a cobordism invariant. From this, the known intersection forms of $\mathbb{CP}^2$ and $\overline{\mathbb{CP}}^2$, and the assumption that $b_2^+(\,M\,) = 0$, a bit of integer linear algebra suffices to produce a basis for $H_2(\,M\,;\,\mathbb{Z}\,)$ relative to which the matrix of Q_M is minus the identity. Here is the argument, with b_2^- denoting the maximal dimension of a subspace on which the intersection form is negative definite, $\sigma = b_2^+ - b_2^-$ the signature and $b_2 = b_2^+ + b_2^-$ the second Betti number.

$$\begin{aligned}
\sigma(\,M\,) \;&=\; \sigma\left(p\,\mathbb{CP}^2 \# q\,\overline{\mathbb{CP}}^2\right)\\
b_2^+(\,M\,) \,-\, b_2^-(\,M\,) &= p \,-\, q\\
b_2^-(\,M\,) &= q \,-\, p\\
b_2(\,M\,) &= q \,-\, p \,\leq\, q \,+\, p \,=\, m\,.
\end{aligned}$$

We show next that $b_2(\,M\,) \geq m$ as well. Select $x_1 \in H_2(\,M;\mathbb{Z}\,)$ with $Q_M(\,x_1\,) = -1$ (there must be at least one such because we have assumed $b_2^+(\,M\,) = 0$, but

$b_2(M) \neq 0$). Then there is a Q_M-orthogonal decomposition

$$H_2(M;\mathbb{Z}) \cong \mathbb{Z}\,x_1 \oplus G_1 .$$

Now consider any $x_2 \in H_2(M;\mathbb{Z})$ with $Q_M(x_2, x_2) = -1$ and $x_2 \neq \pm x_1$ (if such a thing happens to exist). The Schwartz Inequality gives

$$\big(Q_M(x_1, x_2)\big)^2 < Q_M(x_1, x_1)\, Q_M(x_2, x_2) = 1 .$$

But $Q_M(x_1, x_2)$ is an integer so $Q_M(x_1, x_2) = 0$ and so $x_2 \in G_1$. Now repeat the argument inside G_1 and continue inductively until you run out of $x \in H_2(M;\mathbb{Z})$ for which $Q_M(x, x) = -1$ (which you will because $H_2(M;\mathbb{Z})$ is finitely generated). The result is a Q_M-orthogonal decomposition

$$H_2(M;\mathbb{Z}) \cong \mathbb{Z}\,x_1 \oplus \cdots \oplus \mathbb{Z}\,x_m \oplus G$$

where G is either empty or the orthogonal complement of $\mathbb{Z}\,x_1 \oplus \cdots \oplus \mathbb{Z}\,x_m$. In particular, $m \leq b_2(M)$. Since we showed earlier that $m \geq b_2(M)$ we have

$$b_2(M) = m = p + q .$$

Thus, in fact, G must be empty and so

$$H_2(M;\mathbb{Z}) \cong \mathbb{Z}\,x_1 \oplus \cdots \oplus \mathbb{Z}\,x_m ,$$

where

$$Q_M(x_i, x_i) = -1 , \qquad i = 1, \ldots, m .$$

The matrix of Q_M relative to the basis $\{x_1, \ldots, x_m\}$ is minus the identity.