3. $H^2(\mathcal{Q})$ and the quantization of parameters

Let us consider again the motion of a charged particle on a manifold $\mathcal{Q}$, in a background magnetic field $\mathcal{F}$, with a potential $\mathcal{A}$, and lagrangian . In chapter 2 we have discussed at length the case in which $\mathcal{Q}$ is multiply connected and $\mathcal{F}=0$. We will now consider situations in which $\mathcal{Q}$ is simply connected and $\mathcal{F}\neq 0$. As discussed in section 2.1, quantization demands that $\mathcal{A}$ be a $U(1)$-connection, rather than an $\mathbf{R}$-connection. How can we tell whether this is the case?

A necessary and sufficient condition for $\mathcal{F}$ to be the field strength of a $U(1)$ connection is that

$$\int_m \mathcal{F} = \frac{2\pi\hbar}{e} n \qquad\qquad n \in \mathbf{Z} \ , \tag{3..1}$$

for any two-dimensional submanifold m of $\mathcal{Q}$ without boundary. In mathematical terms, $\mathcal{F}$ has to define an integral cohomology class in $H^2(\mathcal{Q})$ (see Appendix **???**). A general proof of this result is given in Appendix **???**. We will give a proof only in the simplest case, namely when $\mathcal{Q} = S^2$. This leads to the famous quantization condition of the monopole charge given by Dirac. We then move on to discuss some field theoretic analogues of this phenomenon: nonlinear sigma models with Wess–Zumino–Witten terms, and odd dimensional gauge theories with Chern–Simons terms. These are all terms in the action that give a nontrivial contribution to the equations of motion, just like the monopole field enters in the equations of motion of a charged particle. Nevertheless, because of their topological origin, we will still call them "topological terms".

The common denominator of all these theories is given by the integrality condition (3..1). In the case of a charged particle moving in the monopole field, it leads to the quantization of the monopole charge; in the field theoretic examples it leads to a certain parameter in the lagrangian taking quantized values.

As is clear from the preceding discussion, the proper topological setting for these phenomena is cohomology rather than homotopy. However if $\mathcal{Q}$ is simply connected, Hurewicz' theorem (Appendix B) states that $H^2(\mathcal{Q},\mathbf{Z}) = \pi_2(\mathcal{Q})$, so one could loosely say that these phenomena are related to a nontrivial second homotopy group.

3.1 The Dirac quantization condition

Let us consider the magnetic field $B_i = \frac{Q_M}{r^2}\hat{x}^i$. We will regard it as a fixed background, and seek consistency conditions for the quantization of a charged particle moving in this background. As discussed in section 1.7, this field is singular in the origin and therefore should be regarded as a vacuum solution of Maxwell's equations on $\mathcal{Q} = \mathbf{R}^3 \setminus \{0\}$. On this manifold, $d\mathcal{F} = 0$ and so one expects that there exists a magnetic potential $\mathcal{A}$. It turns out that there is no magnetic potential for the monopole which is regular everywhere on $\mathbf{R}^3 \setminus \{0\}$. To see this, suppose a magnetic potential $\mathcal{A}$ is given and consider the line integral $\phi(\theta) = \oint_{\ell(\theta)} \mathcal{A}$, where $\ell(\theta)$ is a parallel at colatitude θ on a sphere of radius r (fig.**???**). Clearly $\phi(0)=\phi(\pi)=0$. On the other hand using Stokes' theorem $\phi(\theta) = \oint_{\ell(\theta)} \mathcal{A} = \int_{U(\theta)} \mathcal{F}$, where $U(\theta)$ is the cap bounded by $\ell(\theta)$. Thus $\phi(\theta)$ is the flux through the cap. This can be easily computed to be $\phi(\theta)=2\pi Q_M(1-\cos\theta)$. For $\theta=\pi$ it is equal to $4\pi Q_M$. Thus we get a contradiction.

In order to understand more clearly what happens we can try to look for explicit forms of the magnetic potential. Using a natural basis in spherical coordinates the field strength reads

$$\mathcal{F} = Q_M \sin\theta \, d\theta\wedge d\varphi \ . \tag{3.1.2}$$

A solution of the equation $\mathcal{F}=d\mathcal{A}$ is given by

$$\mathcal{A}=\mathcal{A}^{(+)}=Q_M(1-\cos\theta)d\varphi \ . \tag{3.1.3}$$

This potential is singular on the negative z-axis ($\theta=\pi$). In fact, the form $d\varphi$ is singular on the whole z-axis but its coefficient $(1-\cos\theta)$ vanishes along the positive z-axis ($\theta=0$). This singularity of the magnetic

potential is known as the Dirac string. Its does not correspond to any singularity of the field, however, since it can be moved by gauge transformations. For example, another choice of magnetic potential is

$$\mathcal{A}^{(-)} = -Q_M(1+\cos\theta)d\varphi \ , \tag{3.1.4}$$

which is singular on the positive axis. Let U^+ and U^- be the subsets of $\mathcal{Q}$ with $\theta \neq \pi$ and $\theta \neq 0$ respectively. Even though one cannot introduce a magnetic potential everywhere on $\mathcal{Q}$, it is still possible to give a satisfactory description of the monopole field by giving the potential $\mathcal{A}^+$, which is regular on U^+ and the potential $\mathcal{A}^-$, which is regular on U^-. Together, these two open sets cover all of $\mathcal{Q}$. On the intersection $U^+ \cap U^- = \mathbf{R}^3 \setminus \{z\text{-axis}\}$, the two potentials are related by a gauge transformation:

$$\mathcal{A}^+ - \mathcal{A}^- = 2Q_M d\varphi \ . \tag{3.1.5}$$

As emphasized in section 2.1, quantization of the particle with lagrangian demands that $\mathcal{A}$ be a $U(1)$ gauge field.

In the present case, to quantize the particle one would introduce wavefunctions $\psi^\pm$ which need only be well-defined on $U^\pm$ respectively, and are related by a gauge transformation $g = e^{i\alpha}$ in the intersection:

$$\psi^+(\theta,\varphi) = g(\theta,\varphi)^{-1}\psi^-(\theta,\varphi) \qquad\qquad \text{on}\, U^+ \cap U^- \ . \tag{3.1.6}$$

Note that $U^+ \cap U^-$ is multiply connected and g is required to be a single-valued function from $U^+ \cap U^-$ to $U(1)$. The corresponding transformation of the gauge potential is

$$\mathcal{A}^+ = \mathcal{A}^- - \frac{\hbar}{ie} g^{-1} dg = \mathcal{A}^- - \frac{\hbar}{e} d\alpha \ , \tag{3.1.7}$$

so comparing with (3.1.5) we see that the appropriate gauge transformation is

$$g(\theta,\varphi) = e^{-i\frac{e}{\hbar}2Q_M\varphi} \ . \tag{3.1.8}$$

This will be single-valued if the magnetic charge satisfies the following Dirac quantization condition:

$$Q_M = \frac{\hbar}{2e} n \ . \tag{3.1.9}$$

Noting that $Q_M = \frac{1}{4\pi}\int_{S_2}\mathcal{F}$, this is precisely the same as (3..1).

One can give a path integral argument leading to (3.1.9). The action of a particle moving in a background magnetic field is

$$I = \int_{-\infty}^{\infty} dt \left[\frac{m}{2}\left(\frac{dq^i}{dt}\right)^2 + e\frac{dq^i}{dt}\mathcal{A}_i \right] \ . \tag{3.1.10}$$

This action suffers from two related problems. First, in the case of the monopole, $\mathcal{A}$ has singularities, as we have seen. This form of the action is therefore only appropriate for those histories of the particle that do not cross the Dirac string. On the other hand we know that the Dirac string is not a physical singularity, so this must be a shortcoming of our description of the system, not of the system itself. Second, the action is not gauge invariant. Under the gauge transformation (3.1.7), $I' = I - \hbar\,(\alpha(\infty) - \alpha(-\infty))$.

There is a way of rewriting the action that avoids both problems. Consider a closed orbit c, with $\vec{x}(\infty) = \vec{x}(-\infty)$ (this is analogous to the choice $\varphi(\infty) = \varphi(-\infty)$ in the discussion of the pendulum in Section 2.2). Then we can apply Stokes' theorem and write

$$e\int_c \mathcal{A} = e\int_U \mathcal{F} \ , \tag{3.1.11}$$

where U is a two dimensional surface having c as boundary. This way of writing the action is gauge invariant and insensitive to the Dirac string, but it makes reference to the surface U, which is not uniquely defined by

the trajectory of the particle. Since $d\mathcal{F}=0$, the integral (3.1.11) is invariant under infinitesimal deformations of the surface that keep the boundary fixed, but it may change for large deformations. In fact, consider two surfaces U_1 and U_2 both having c as boundary, but one passing "above", the other "below" the origin (see Fig. ???). The difference ΔI in the actions $e\int_{U_1}\mathcal{F}$ and $e\int_{U_2}\mathcal{F}$ is equal to the integral of $\mathcal{F}$ on the closed surface formed by joining U_1 and U_2 along the boundary. Since this surface contains the origin, the integral is equal to $4\pi eQ_M$. This arbitrariness in the action will not affect the functional integral if $e^{\frac{i}{\hbar}\Delta I}=1$, which directly implies (3..1).

Finally we observe that the quantization condition can also be seen as an application of the old Bohr–Sommerfeld quantization conditions $\oint pdq = 2\pi\hbar n$. From we have $\oint pdq = \int \left(mg_{ij}\dot{q}^i\dot{q}^j + e\dot{q}^i\mathcal{A}_i\right)dt$. Now consider a very small loop encircling the Dirac string. When the radius of the loop goes to zero, the first term goes to zero but the second becomes $e\oint\mathcal{A} = e\int_{S^2}\mathcal{F}$, having applied Stokes' theorem to a surface bounded by the loop and not containing the string. The Bohr–Sommerfeld rule then gives (3..1).

3.2 Wess–Zumino–Witten terms

Consider a non-linear sigma model with values in $SU(2)\equiv S^3$, in $d=1$ space dimensions. The configuration space is $\mathcal{Q}=\Gamma_*(S^1,S^3)$. Using the, by now familiar, technique of Appendix **???** we find that $\pi_0(\mathcal{Q})=\pi_1\big(SU(2)\big)=0$, $\pi_1(\mathcal{Q})=\pi_2\big(SU(2)\big)=0$ and $\pi_2(\mathcal{Q})=\pi_3\big(SU(2)\big)=\mathbf{Z}$. The generator of $\pi_2(\mathcal{Q})$ is a map $m:S^2\to\mathcal{Q}$ which is defined by $\big(m(t_1,t_2)\big)(t_3)=\hat{m}(t_1,t_2,t_3)$, where $\hat{m}$ is a map of S^3 (a cube $I\times I\times I$ with the boundary identified to a point) to $SU(2)$, sending $\partial(I\times I\times I)$ into the identity element, and with winding number $W(\hat{m})=1$. If the map c in section 2.3 could be referred to as a "loop of loops", the map m defined here could be called a "sphere of loops". By Hurewicz' theorem, together with (B.), one concludes that $H^0(\mathcal{Q},\mathbf{Z})=0$, $H^1(\mathcal{Q},\mathbf{Z})=0$ and $H^2(\mathcal{Q},\mathbf{Z})=\mathbf{Z}$. The low homotopy and cohomology groups of $\mathcal{Q}$ are the same as in the previous section, so one may expect to find some analogue of the Dirac quantization condition in this theory. This is indeed the case. As with theta sectors, in order to reveal the occurrence of topological phenomena, it is necessary to add an appropriate term to the action.

To guess the right term we may look for inspiration in Section 2.3, where we discussed the same theory in one more dimension. We saw that the integrand of the topological term $\theta W(\varphi)$ was a total derivative and therefore W could be written as a surface integral (an integral on a two dimensional space). Suppose now that the boundary conditions on the fields are such that they go to a constant at spacetime infinity (this is the case if we demand that the action be finite), so that spacetime can be compactified to a sphere S^2. We can think of this sphere as the boundary of some (fictitious) three dimensional ball B^3 and regard the fields φ as boundary values of some field $\tilde{\varphi}$ defined on B^3. This is always possible because $\pi_2(SU(2))=0$, so all fields φ are homotopically trivial and therefore have a continuation in the interior of B^3. The topological term we are after is just the topological term of $\tilde{\varphi}$, which depends only on φ and not on the value of the fields in the interior of the ball. We therefore have

$$S_{WZW}=c\int\varphi^*\tau=\frac{c}{2}\int d^2x\,\varepsilon^{\mu\nu}\partial_\mu\varphi^\alpha\partial_\nu\varphi^\beta\tau_{\alpha\beta} \tag{3.2.1}$$

where $\omega=d\tau$ or, in components,

$$\omega_{\alpha\beta\gamma}=3(\partial_\alpha\tau_{\beta\gamma}+\partial_\beta\tau_{\gamma\alpha}+\partial_\gamma\tau_{\alpha\beta})\ . \tag{3.2.2}$$

and ω is the volume form on $SU(2)$, normalized so that $\int_{SU(2)}\omega=1$. We have renamed c the constant that previously was called θ. For example, suppose we choose on $SU(2)$ a coordinate system given by the Euler angles (Θ,Φ,Ψ) (see Appendix G). The volume form is given by

$$\omega=\frac{1}{8\pi}\sin\Theta\,d\Theta\wedge d\Phi\wedge d\Psi\ . \tag{3.2.3}$$

and a choice of τ is

$$\tau=-\frac{1}{8\pi}\cos\Theta\,d\Phi\wedge d\Psi\ . \tag{3.2.4}$$

Since $d\tau \neq 0$, the Wess–Zumino–Witten term (3.2.1)(3.2.3)is not a total derivative term and, as we shall see in a moment, it does contribute to the equations of motion of the theory.

An important consequence of (3.2.2) is that τ is not uniquely defined: if τ satisfy (3.2.2), also $\tau' = \tau + d\beta$ does. This amounts to adding a total derivative to the action (3.2.1). Another fact of the greatest importance is that τ is not globally defined. If it was, ω would define a trivial cohomology class. But we know from Appendix B that the volume-form on a compact manifold always defines a non-trivial cohomology class. This means that the form τ is singular somewhere on $SU(2)$. For example the form τ defined in (3.2.4) is singular for $\Theta = 0$ or π.

Now consider a field $\varphi(x,t)$; we regard it as a map from S^2 (the one-point compactification of spacetime) into $SU(2)$. Generically, φ will not meet the singular points of τ. Thus there will be an open subset $\mathcal{U}$ of $\Gamma_*\big(S^2, SU(2)\big)$ where $\mathrm{Im}\varphi \cap \{\text{singular set}\} = \emptyset$, and the WZW action (3.2.1) will be well defined on $\mathcal{U}$. However, there are also maps φ whose image intersects the singular set of τ. For such maps (3.2.1) is not well defined. We can however use the freedom of adding a total derivative term to the action. The form $\tau' = \tau + d\beta$ will have a different singular set from τ, and the action $c\int \varphi^*\tau' = c\int \varphi^*\tau + c\int d(\varphi^*\beta)$ will be well defined on another open subset $\mathcal{U}'$ of $\Gamma_*\big(S^2, SU(2)\big)$. Since the image of φ has dimensions 2 and the singular set of any form τ has dimension zero, it is clear that for every $\varphi \in \Gamma_*\big(S^2, SU(2)\big)$ one can find a one-form β such that τ' does not have any singularity on the image of φ. In this way one can cover $\Gamma_*\big(S^2, SU(2)\big)$ with open sets, such that on each set there is a well defined function $c\int \varphi^*\tau$, and on the intersection of two sets these functions differ by a total derivative term $c\int d(\varphi^*\beta)$. The collection of these locally defined functions is the WZW term. It is not a functional of φ in the ordinary sense. Instead, it is a section of a certain line bundle over $\Gamma_*\big(S^2, SU(2)\big)$.

Let us consider the non-linear sigma model with the action given by $S = S_0 + S_{WZW}$, where

$$S_0 = -\frac{1}{2}\int d^2x\, \partial_\mu \varphi^\alpha \partial^\mu \varphi^\beta h_{\alpha\beta}\ . \tag{3.2.5}$$

The equation of motion reads

$$h_{\alpha\beta}\partial_\mu\partial^\mu\varphi^\beta + \Gamma_{\alpha,\beta\gamma}\partial_\mu\varphi^\beta\partial^\mu\varphi^\gamma + \frac{c}{2!}\varepsilon^{\mu\nu}\partial_\mu\varphi^\beta\partial_\nu\varphi^\gamma\omega_{\alpha\beta\gamma} = 0 \tag{3.2.6}$$

where $\Gamma_{\alpha,\beta\gamma} = \frac{1}{2}\left(\partial_\beta h_{\alpha\gamma} + \partial_\gamma h_{\alpha\beta} - \partial_\alpha h_{\beta\gamma}\right)$ are the Christoffel symbols of the metric $h_{\alpha\beta}$ on $SU(2)$. The last term is the contribution of the WZW term. It can be interpreted as follows. Note that the WZW term is linear in the time derivative and therefore can be written as $\int dt\, \dot\varphi^\alpha \mathcal{A}_\alpha(\varphi)$, where

$$\mathcal{A} = -c\int dx\, \partial_1\varphi^\alpha \tau_{\alpha\beta}\delta\varphi^\beta \tag{3.2.7}$$

is a one-form on $\mathcal{Q}$. When we think of the sigma model as a particle moving on $\mathcal{Q}$, $\mathcal{A}$ can be interpreted as a "functional vector potential". Unlike the cases discussed in chapter 2, the corresponding "functional magnetic field" is now non-vanishing. A direct calculation using the methods of Appendix E yields

$$\mathcal{F} = d\mathcal{A} = \frac{c}{2}\int dx\, \partial_1\varphi^\alpha \omega_{\alpha\beta\gamma}\delta\varphi^\beta\delta\varphi^\gamma\ . \tag{3.2.8}$$

To confirm our interpretation of $\mathcal{A}$ and $\mathcal{F}$ note that the last term in (3.2.6) can be written $c\dot\varphi^\beta\mathcal{F}_{\alpha\beta}$ and therefore can be interpreted as the Lorentz force due to $\mathcal{F}$. The one- form $\mathcal{A}$ is only well defined on a subset $\mathcal{V}$ of $\mathcal{Q}$ such that the image of $\varphi(x)$ $\big($a loop in $SU(2)\big)$ does not intersect the singular set of τ. By contrast, $\mathcal{F}$ is well defined everywhere on $\mathcal{Q}$. We are therefore in a situation which resembles very closely that of the previous section, with a magnetic field $\mathcal{F}$ that cannot be derived from a globally defined vector potential $\mathcal{A}$. Rather than repeating the discussion of the Dirac quantization condition in the present context, we will apply directly the general result (3..1). Let us therefore compute the integral of $\mathcal{F}$ on the fundamental two-cycle m. We have

$$\int\limits_m \mathcal{F} = \int\limits_0^1 dt_1 \int\limits_0^1 dt_2 \left\{ c\int dx\, \partial_1\varphi^\alpha\omega_{\alpha\beta\gamma}\frac{\partial\varphi^\beta}{\partial t_1}\frac{\partial\varphi^\gamma}{\partial t_2}\right\} = \frac{c}{3!}\int d^3x\, \varepsilon^{\lambda\mu\nu}\frac{\partial\hat\varphi^\alpha}{\partial x^\lambda}\frac{\partial\hat\varphi^\beta}{\partial x^\mu}\frac{\partial\hat\varphi^\gamma}{\partial x^\nu}\omega_{\alpha\beta\gamma} = cW(\hat\varphi) = c\ . \tag{3.2.9}$$

Using (3..1) we find that the theory can be quantized only for

$$c{=}2\pi n \ . \tag{3.2.10}$$

This is the analogue of the Dirac quantization condition.

The quantization of the parameter c can also be proven in the path integral formalism by means of the following argument. The extension $\bar\varphi$ is not unique. Consider two extensions $\bar\varphi_1 : B^3_1 \to SU(2)$ and $\bar\varphi_2 : B^3_2 \to SU(2)$, with $\bar\varphi_1\big|_{S^2} = \bar\varphi_2\big|_{S^2} = \varphi$. Since they coincide on $\partial B_1 = \partial B_2 = S^2$, we can think of them as a single map $\bar\varphi : S^3 \to SU(2)$, where S^3 is obtained by glueing the two balls S^3 along their boundaries (in this picture the two balls are the hemispheres of S^3 and S^2 is the equator of S^3). The maps φ_1 and φ_2 together define a map $\tilde\varphi : S^3 \to SU(2)$. The difference of the WZW actions is therefore equal to $\Delta S_{WZW} = S_{WZW}(\varphi_2) - S_{WZW}(\varphi_1) = cW(\tilde\varphi)$. This arbitrariness will not affect the functional integral if $e^{i\Delta S_{WZW}} = 1$, which again implies (3.2.10).

Can one write a Wess–Zumino–Witten term for a sigma model in 3+1 dimensions? To answer this question, let us review what we have done in this section. We have started from a closed three-form ω representing a nontrivial cohomology class of the target space. This form could be written locally as the exterior differential of a two-form τ. The Wess–Zumino–Witten action was the integral of the pullback of τ. In 3+1 dimensions ω would have to be a closed five form and τ a four-form. There are no five-forms on $SU(2)$, but there are nontrivial five-forms on $SU(N)$ for $N \geq 3$. In fact, $H^5(SU(N), \mathbf{R}) = \mathbf{R}$, and the generator of this group is (the cohomology class of) $-\frac{i}{240\pi^2}\mathrm{tr}R^5$. Therefore, the Wess–Zumino–Witten term can be written in either one of the following two forms:

$$\begin{aligned} S_{WZW} =& c\int d^4x\, \varepsilon^{\mu\nu\rho\sigma}\partial_\mu\varphi^\alpha\partial_\nu\varphi^\beta\partial_\rho\varphi^\gamma\partial_\sigma\varphi^\delta\tau_{\alpha\beta\gamma\delta} \\ =& -\frac{ic}{240\pi^2}\int d^5x\, \varepsilon^{\lambda\mu\nu\rho\sigma}\mathrm{tr}(R_\lambda R_\mu R_\nu R_\rho R_\sigma) \ , \end{aligned}$$

where spacetime has been compactified to a four-sphere and in the last line the integral is over a ball having spacetime as a boundary. This gives rise to the magnetic potential

$$\mathcal{A} = -c\int d^3x\, \varepsilon^{ijk}\partial_i\varphi^\alpha\partial_j\varphi^\beta\partial_k\varphi^\gamma\tau_{\alpha\beta\gamma\delta}\delta\varphi^\delta \tag{3.2.11}$$

on $\mathcal{Q}$. The corresponding field strength is

$$\mathcal{F} = d\mathcal{A} = \frac{c}{2}\int d^3x\, \varepsilon^{ijk}\partial_i\varphi^\alpha\partial_j\varphi^\beta\partial_k\varphi^\gamma\omega_{\alpha\beta\gamma\delta\eta}\delta\varphi^\delta\delta\varphi^\eta \ . \tag{3.2.12}$$

One can now repeat the arguments given above leading to the quantization of the parameter c.

Finally we observe that the relation between ω and $\mathcal{F}$ is a special example of a general construction which relates cohomology classes of N to cohomology classes of $\Gamma(M, N)$. This is discussed in Appendix F.

3.2 Chern–Simons terms

Next we consider an $SU(2)$ gauge theory in $2+1$ dimensions. As in the previous chapter, we will use the rescaled, geometrical gauge fields, with curvature defined by , gauge transformations and action . Instead of writing explicitly the Lie algebra indices, we will use a matrix notation and write $A_\mu = A^a_\mu T_a$, where T_a are matrices satisfying $[T_a, T_b] = \varepsilon_{abc}T_c$ and $\mathrm{tr}(T_aT_b) = -\frac{1}{2}\delta_{ab}$ (for example in the spinor representation, $T_a = -\frac{i}{2}\sigma_a$, where σ_a are the Pauli matrices). In this notation the Yang–Mills action reads

$$S_{YM} = \frac{1}{2e^2}\int d^3x\, \mathrm{tr}\, F_{\mu\nu}F^{\mu\nu} \ . \tag{3.2.1}$$

As in Sections 3.4 and 3.5, we choose the gauge $A_0 = 0$; then the static energy reads

$$E_S = -\frac{1}{e^2}\int d^2x\, \mathrm{tr}\, B^2 \ , \tag{3.2.2}$$

where $B = F_{12}$ is the nonabelian magnetic field. The configuration space is then $\mathcal{Q} = \mathcal{C}/\mathcal{G}$, where $\mathcal{C}$ is the space of connections $A_i(\vec{x})$, $i = 1, 2$, such that E_S is finite, and $\mathcal{G} = \Gamma_*(S^2, SU(2))$ is the residual gauge group consisting of time independent gauge transformations.

This configuration space is connected and furthermore has $\pi_1(\mathcal{Q}) = \pi_0(\mathcal{G}) = \pi_2(SU(2)) = 0$ and $\pi_2(\mathcal{Q}) = \pi_1(\mathcal{G}) = \pi_3(SU(2)) = \mathbf{Z}$. The generator of the group $\pi_2(\mathcal{Q})$ can be described as follows. The gauge group $\mathcal{G}$ is connected but not simply connected. Let $\ell(t)$ be a loop whose homotopy class generates $\pi_1(\mathcal{G})$. Fix a reference point $A_{(0)}$ in $\mathcal{C}$ and consider the loop in the orbit through $A_{(0)}$ given by $A_{(0)}^{\ell(t)}$. This loop cannot be shrunk to a point within the orbit but it can be shrunk to a point in $\mathcal{C}$. Thus there is a map $\tilde{m}$ from a two dimensional ball B^2 to $\mathcal{C}$ which is equal to $A_{(0)}^{\ell(t)}$ on the boundary. Now compose this map with the projection $\mathcal{C} \to \mathcal{Q}$. Since all points on the boundary of the disk are mapped to the same orbit, we get a map m from S^2 to $\mathcal{Q}$ which is not homotopic to a constant (see Fig. ???). The isomorphism between $\pi_1(\mathcal{G}) = \mathbf{Z}$ and $\pi_2(\mathcal{Q})$ is the map that sends (the homotopy class of) ℓ to (the homotopy class of) m.

Once again we have exactly the same homotopy groups as in Section 3.1, so we expect that some parameter will have to be quantized. But what parameter? As in the previous Section, we impose boundary conditions such that spacetime can be compactified to a three dimensional sphere S^3, and regard this sphere as the boundary of a four dimensional ball B^4. Gauge fields A_μ on a three sphere are topologically trivial ($\pi_2(SU(2)) = 0$) and therefore can be thought of as boundary values of gauge fields $\tilde{A}_\mu$ defined on B^4. The $SU(2)$ gauge theory in 3+1 dimensions was discussed in Section 2.5, where, in order to reveal the existence of theta sectors, we added to the action a topological term $S_T = \theta c_2$. With the boundary conditions of Section 2.5, c_2 was an integer; with the boundary conditions used here, the integral $c_2(\tilde{A})$ becomes a function of the boundary values A. Using that the integrand of c_2 is the exterior differential of the Chern–Simons three form , the appropriate topological term to be added to S_{YM} in three dimensions is the Chern–Simons term

$$S_{CS} = \mu \frac{8\pi^2}{e^2} \int d^3x \Omega \ , \tag{3.2.3}$$

where

$$\Omega = -\frac{1}{8\pi^2} \varepsilon^{\lambda\mu\nu} \mathrm{tr} \left(A_\lambda \partial_\mu A_\nu + \frac{2}{3} A_\lambda A_\mu A_\nu \right) \ . \tag{3.2.4}$$

(Note that apart from replacing the coefficient θ by the coefficient $\mu \frac{8\pi^2}{e^2}$, S_{CS} is identical to the functional $\tilde{\Lambda}$ defined in .) The constant μ has dimension of mass. In fact simple manipulations on the equations of motion (Exercise 3.2.1) show that this theory describes spin one particles with mass $|\mu|$. For this reason it was called a "topologically massive gauge theory".

From our previous discussion of the WZW action, we are led to believe that the coefficient of the CS action, the mass μ, has to be quantized in certain units. This is indeed what happens. The proof of this fact turns out to be rather involved at the canonical level, so we will depart from our standard procedure and give first a proof at the level of path integrals.

Let us restrict our attention to field configurations with S_{YM} finite. This implies that spacetime can be compactified to a sphere S^3. Therefore the group of gauge transformations is $\Gamma_*(S^3, SU(2))$, and it consists of infinitely many connected components, labelled by their winding number. The (dual of the) Chern–Simons form transforms as follows

$$\Omega(A^g) = \Omega(A) - \frac{1}{8\pi^2} \varepsilon^{\lambda\mu\nu} \mathrm{tr}\, \partial_\lambda \left(\partial_\mu g g^{-1} A_\nu \right) + \frac{1}{24\pi^2} \varepsilon^{\lambda\mu\nu} \mathrm{tr} \left(g^{-1} \partial_\lambda g^{-1} \partial_\mu g g^{-1} \partial_\nu g \right) \ , \tag{3.2.5}$$

and since we assume g to tend to the identity at infinity, upon integration we find

$$S_{CS}(A^g) = S_{CS}(A) + \mu \frac{8\pi^2}{e^2} W(g) \ . \tag{3.2.6}$$

This is essentially the same calculation that led to (???). Thus, the Chern–Simons action is gauge invariant under gauge transformations which are homotopic to the identity (in particular, it is invariant under infinitesimal gauge transformations), but not under "large" gauge transformations. We demand that the functional integral be insensitive to this ambiguity. This requires that $e^{i\Delta S_{CS}} = 1$, or

$$\mu = \frac{e^2}{4\pi} n \ , \qquad n \in \mathbf{Z} \ . \tag{3.2.7}$$

Note that in the Euclidean path integral one would demand $e^{-\Delta S_{CS,E}} = 1$, where $S_{CS,E}$ is the Euclidean Chern–Simons action. Since S_{CS} is linear in the time derivative, $S_{CS,E} = iS_{CS}$, so we are led again to (3.2.7).

To see what would go wrong if we did not impose the quantization condition (3.2.7), consider the formal procedure for eliminating the volume of the gauge group from the functional integral. Having chosen a gauge condition $f(A) = 0$, one inserts in the functional integral $Z = \int (dA) e^{iS(A)}$ the identity $1 = \Delta_{FP}(A) \int (dg) \delta(f(A^g))$, where $\Delta_{FP}(A)$ is the Faddeev–Popov determinant, a gauge invariant functional of the gauge potential. In the present case, since the gauge group has infinitely many connected components, it is convenient to write the integral over the gauge group as a sum of integrals over the connected components: $\int (dg) = \sum_n \int (dg)_n$. Now we have

$$Z = \sum_n \int (dg)_n \int (dA) \Delta_{FP}(A) \delta(f(A^g)) e^{iS(A)} \ .$$

At this point one usually invokes invariance of the measure, of the Faddeev–Popov determinant and of the action, to rewrite the argument of all functionals on the r.h.s. as A^g (and then A, since it is an integration variable). In the present case the action is not invariant, so taking into account (3.2.6) we find

$$Z = V_0 \sum_n e^{-i\mu \frac{8\pi^2}{e^2} n} \int (dA) \Delta_{FP}(A) \delta(f(A)) e^{iS(A)} \ ,$$

where V_0 is the volume of one connected component of the gauge group. The sum in front of the integral gives zero unless μ satisfies the quantization condition (3.2.7). Thus if (3.2.7) is not satisfied, the functional integral, and similarly the expectation value of any gauge invariant observable, is ill-defined.

Let us now return to the canonical formalism and see how the quantization of the topological mass arises there. The momentum conjugate to A_i is

$$P^i = P^i_a T_a = \frac{1}{e^2} E_i - \frac{\mu}{2e^2} \varepsilon_{ij} A_j \ . \tag{3.2.8}$$

Comparing with , we see that the Chern–Simons term gives rise to a "functional vector potential"

$$\tilde{\mathcal{A}} = \frac{\mu}{e^2} \int d^2x \, \varepsilon_{ij} \, \mathrm{tr} A_j \delta A_i \ , \tag{3.2.9}$$

with the corresponding "functional magnetic field"

$$\tilde{\mathcal{F}} = d\tilde{\mathcal{A}} = -\frac{\mu}{e^2} \int d^2x \, \varepsilon_{ij} \, \mathrm{tr} \, \delta A_i \delta A_j \ . \tag{3.2.10}$$

These forms are defined on $\mathcal{C}$ and, as in Sections 3.4 and 3.5, we ask whether these forms are the pullbacks of forms $\mathcal{A}$ and $\mathcal{F}$ defined on the true configuration space $\mathcal{Q}$. Unfortunately this is not the case. The contraction of $\tilde{\mathcal{F}}$ with a vertical vector

$$v_\epsilon = -2 \int d^2x \, \mathrm{tr} \, D_i \epsilon \frac{\delta}{\delta A_i} \tag{3.2.11}$$

is not zero. Evidently the one-form $\tilde{\mathcal{A}}$ is not the appropriate one. In order to guess what the appropriate form is, we recall that in even dimensional gauge theories the Gauss law could be written as a covariant derivative with respect to $\tilde{\mathcal{A}}$ (see the remarks in the end of Section 2.5). In the theory under consideration this is no longer the case. However, there is another one form, $\tilde{\mathcal{A}}'$, which plays the same role. In the topologically massive gauge theory, the Gauss law reads

$$\begin{aligned} G(x) &= -\frac{1}{e^2} \left(D_i E_i - \mu B \right) \\ &= -\left(D_i P_i + \frac{\mu}{2e^2} \varepsilon_{ij} D_i A_j - \frac{\mu}{e^2} B \right) \end{aligned} \tag{3.2.12}$$

Given a map ϵ from S^2 to the Lie algebra of $SU(2)$ which is equal to zero at the point ∞, we define

$$\begin{aligned} G_\epsilon =& -2\int d^2x \,\mathrm{tr}\,\epsilon(x)G(x) \\ =& 2\int d^2x \,\mathrm{tr}\,\epsilon\Big(D_iP_i + \frac{\mu}{2e^2}\epsilon_{ij}D_iA_j - \frac{\mu}{e^2}D_iD_i\Delta^{-1}B\Big)\ , \end{aligned} \tag{3.2.13}$$

where in the last term we have inserted $1 = D_iD_i\Delta^{-1}$, Δ being the covariant Laplacian. Since the fields A_i and their conjugate momenta P_i have canonical Poisson brackets, G_ϵ is the generator of the infinitesimal gauge transformation ϵ. If the theory is quantized before eliminating the gauge degrees of freedom, as in Dirac's approach, the Gauss law has to be imposed as a condition on physical states:

$$\hat{G}_\epsilon\psi_{\mathrm{phys}} = 0\ , \tag{3.2.14}$$

where $\hat{G}_\epsilon$ is the quantum operator corresponding to the Gauss law constraint. Using the quantization rule $\hat{P}_i = -i\frac{\delta}{\delta A_i}$ this condition can be rewritten as

$$0 = \hat{G}_\epsilon\psi = -i(-2)\int d^2x \,\mathrm{tr}\, D_i\epsilon\left[\frac{\delta\psi}{\delta A_i} - i\left(-\frac{\mu}{2e^2}\epsilon_{ij}A_j + \frac{\mu}{e^2}D_i\Delta^{-1}B\right)\psi\right] = -i\mathcal{D}_{v_\epsilon}\psi\ . \tag{3.2.15}$$

where v_ϵ is given by (3.2.11) and

$$\mathcal{D}'_v\psi = v\psi - i\tilde{\mathcal{A}}'(v)\psi \tag{3.2.16}$$

is the covariant derivative with respect to the functional vector potential

$$\tilde{\mathcal{A}}' = \tilde{\mathcal{A}} - \frac{\mu}{e^2}\int d^2x \,\mathrm{tr}\, D_i\Delta^{-1}B\delta A_i\ . \tag{3.2.17}$$

We have thus succeeded in writing again the quantum Gauss law constraint as the condition that the wave functions be covariantly constant with respect to a certain functional vector potential $\mathcal{A}'$. This is not the functional vector potential that is defined directly by the Chern–Simons term. Rather, it differs from it by the addition of a nonlocal term. One can prove that the field strength $\tilde{\mathcal{F}}' = d\tilde{\mathcal{A}}'$ is the pullback of a two-form $\mathcal{A}$ on $\mathcal{Q}$. (Thus the restriction of $\mathcal{A}'$ to the orbit is the transgression of $\mathcal{F}$).

Having thus obtained, at least implicitly, a connection on $\mathcal{Q}$, we proceed to require that it be a $U(1)$ connection. To this end we have to impose that the curvature $\mathcal{F}$ satisfies the integrality condition (3..1), when m is the fundamental two-sphere in $\mathcal{Q}$ that was described in the beginning of this section. From that discussion we have

$$\int_m \mathcal{F} = \int_{\tilde{m}} \tilde{\mathcal{F}}' = \int_{\tilde{\ell}} \tilde{\mathcal{A}}'\ , \tag{3.2.18}$$

so there remains only to compute the line integral of $\tilde{\mathcal{A}}$ along the noncontractible loop in the orbit. This calculation is described in Exercise 3.2.2 and gives

$$\oint \hat{\mathcal{A}}' = -\frac{\mu}{e^2}8\pi^2 W(\hat{g})\ . \tag{3.2.19}$$

Imposing that this be an integral multiple of 2π we are led again to (3.2.7).

This reasoning was based on a quantization procedure in which one works only with the physical degrees of freedom contained in $\mathcal{Q}$ ("first constrain and then quantize"). This procedure is useful for abstract reasoning, but is to say the least very hard to implement in detail, since the coordinates on $\mathcal{Q}$ are nonlocal fields. In explicit calculations one usually follows the alternative procedure, mentioned above, of constructing wave functionals on $\mathcal{C}$ and imposing the Gauss law as a condition on physical states ("first quantize and then constrain"). If one proceeds in this way, the quantization condition (3.2.7) emerges as an integrability condition for the Gauss law. Since $\tilde{\mathcal{F}}'$ gives zero when contracted with a vertical vector (a vector of the form

(3.2.11)), there follows in particular that the restriction of $\tilde{\mathcal{A}}'$ to the orbits is flat. Therefore the condition (3.2.15) can be integrated to give

$$\psi(A^g) = \psi(A)e^{i\alpha(A,g)} \ , \qquad \alpha(A,g) = \int_A^{A^g} \tilde{\mathcal{A}}' \ , \tag{3.2.20}$$

the integral being performed along any path in the orbit joining A to A^g. Since the orbits are multiply connected, in order to ensure that the physical wave functions are single valued, one has to require that the loop integral $\oint \tilde{\mathcal{A}}'$ be an integral multiple of 2π. Because of (3.2.18), this is exactly the same as the condition of integrality of $\mathcal{F}$.

Exercise 3.2.1: derive the Euler-Lagrange equation that derives from the action $S_{\rm YM} + S_{\rm CS}$. Show that it describes the propagation of a field with mass μ.

Exercise 3.2.2: Let g_t be the loop ℓ in $\mathcal{G}$, $A_t = A^{g_t}_{(0)}$. One can think of the parameter $-\infty < t < \infty$ as time, and the fields $\hat{g}(x,t) = g_t(x)$, $\hat{A}(x,t) = A_t(x)$ as spacetime fields. The line integral is equal to $\int_{\tilde{\ell}} \tilde{\mathcal{A}}' = \int dt \tilde{\mathcal{A}}'(\frac{dA_t}{dt})$, where $\frac{dA_t}{dt} = \hat{D}_i(\hat{g}^{-1}\partial_0\hat{g})$ is the vector tangent to the loop at A_t ($\hat{D}$ is the covariant derivative with respect to $\hat{A}$). So

$$\begin{aligned}\int_{\tilde{\ell}} \tilde{\mathcal{A}}' =& \int dt \int d^2x \, {\rm tr} \, \Big[\frac{\mu}{e^2}\varepsilon_{ij}\hat{A}_j\hat{D}_i(\hat{g}^{-1}\partial_0\hat{g}) - 2\frac{\mu}{e^2}\hat{D}_i(\hat{\Delta}^{-1}\hat{B})\hat{D}_i(\hat{g}^{-1}\partial_0\hat{g})\Big] \\ =& \frac{\mu}{e^2}\int dt \int d^2x \, {\rm tr} \, \Big[-\varepsilon_{ij}\partial_i\hat{A}_j\hat{g}^{-1}\partial_0\hat{g}\Big]\end{aligned} \tag{3.2.21}$$

Now expanding $\partial_i\hat{A}$ in derivatives of A and $\hat{g}$, we can rewrite this as

$$\frac{\mu}{e^2}\int dt \int d^2x \, \varepsilon_{ij} {\rm tr} \Big[-\hat{\partial}_i A_j \partial_0\hat{g}\hat{g}^{-1} - [A_j, \partial_i\hat{g}\hat{g}^{-1}]\partial_0\hat{g}\hat{g}^{-1} + \partial_0\hat{g}\hat{g}^{-1}\partial_i\hat{g}\hat{g}^{-1}\partial_j\hat{g}\hat{g}^{-1}\Big]$$

Since A is independent of t and $A_0 = 0$, inserting the appropriate combinatorial factors we can rewrite this expression in a covariant form:

$$\begin{aligned}&\frac{\mu}{e^2}\int d^3x \, \varepsilon_{\lambda\mu\nu} {\rm tr} \Big[-\partial_\lambda A_\mu \partial_\nu\hat{g}\hat{g}^{-1} - \frac{1}{2}[A_\lambda, \partial_\mu\hat{g}\hat{g}^{-1}]\partial_\nu\hat{g}\hat{g}^{-1} + \frac{1}{3}\partial_\lambda\hat{g}\hat{g}^{-1}\partial_\mu\hat{g}\hat{g}^{-1}\partial_\nu\hat{g}\hat{g}^{-1}\Big] \\ =& -\frac{\mu}{e^2}\int d^3x \, \partial_\lambda \left[\varepsilon_{\lambda\mu\nu}{\rm tr} A_\mu\partial_\nu\hat{g}\hat{g}^{-1}\right] + \frac{1}{3}\frac{\mu}{e^2}\int d^3x \, \varepsilon_{\lambda\mu\nu}{\rm tr} \, \left[\partial_\lambda\hat{g}\hat{g}^{-1}\partial_\mu\hat{g}\hat{g}^{-1}\partial_\nu\hat{g}\hat{g}^{-1}\right]\end{aligned} \tag{3.2.22}$$

For a closed loop starting and ending at the identity the first term is zero, and we find (3.2.19).