2. $\pi_1(\mathcal{Q})$ and theta sectors

We have seen in the previous chapter that when the configuration space of a theory is not connected, there is a conserved topological charge and, if the dynamics is properly chosen, topological solitons. In this chapter we will assume that the configuration space is connected but not simply connected. This leads again to a splitting of the Hilbert space in superselection sectors, but the physical interpretation is different. The paradigm of this phenomenon is the Aharonov–Bohm effect. We will give several examples of this phenomenon in quantum mechanics and quantum field theory.

2.1 The Aharonov Bohm effect

Consider an experiment where electrons emerge from a source, graze a solenoid carrying a magnetic flux Φ and thereafter form an interference pattern on a screen. As the magnetic flux is varied, the interference pattern is observed to vary. It is found that the interference pattern repeats itself when the flux is changed by $\frac{2\pi\hbar}{e}$, where e is the charge of the electron. †

We will now give a theoretical interpretation of this phenomenon. Consider the idealized situation of an infinite perfect solenoid lying along the z axis. The core of the solenoid is assumed to be totally impenetrable to the electrons (the core of the solenoid is made e.g. of iron, and we neglect the probability of an electron tunnelling through the solenoid). When the current flows, there is a constant magnetic field inside the solenoid but the magnetic field is zero outside (a real solenoid is not infinitely long and the distance between the coils is not zero, so the magnetic field has a weak "tail" outside the solenoid; this we also neglect). As a result of these approximations, the electrons move in a configuration space $\mathcal{Q}$ which is all of $\mathbf{R}^3$ with the solenoid removed and the magnetic field vanishes on $\mathcal{Q}$. The space $\mathcal{Q}$ is multiply connected, with $\pi_1(\mathcal{Q})=\mathbf{Z}$. Consider the magnetic potential

$$\mathcal{A} = \theta\frac{\hbar}{2\pi e}d\varphi \ , \tag{2.1.1}$$

where θ is an arbitrary real parameter, and φ is the azimuthal cylindrical coordinate around the z axis. The magnetic field corresponding to $\mathcal{A}$ is zero, so $\mathcal{A}$ is a good gauge potential on $\mathcal{Q}$. To find the meaning of the parameter θ, consider the line integral of $\mathcal{A}$ along a loop encircling once the z axis: $\oint\mathcal{A}=\theta\frac{\hbar}{e}$. On the other hand, using Stokes' theorem, the line integral is equal to the integral of $\mathcal{F}=d\mathcal{A}$ on a surface bounded by the loop; such a surface cuts through the solenoid, so the integral is equal to the magnetic flux through the solenoid, Φ. So we find $\theta=\frac{e}{\hbar}\Phi$. We conclude that $\mathcal{A}$ is the potential seen by an electron travelling outside the solenoid when the flux in the solenoid is $\frac{\hbar}{e}\theta$.

The interference pattern on the screen arises from the phase difference between waves that travel above and below the solenoid. Consider first the case when there is no flux, $\theta=0$. The wave function satisfies the free Schrödinger equation $H_0\psi_0=E\psi_0$, with the free hamiltonian $H_0=-\frac{\hbar^2}{2m}\partial_i\partial_i$. Let us now turn the flux on. The Hamiltonian becomes

$$H = -\frac{\hbar^2}{2m}\mathcal{D}_i\mathcal{D}_i \tag{2.1.2}$$

where $\mathcal{D}_i=\partial_i-\frac{ie}{\hbar}\mathcal{A}_i$ is the covariant derivative with respect to $\mathcal{A}$. It is immediate to check that

$$\psi(q) = \psi_0(q)e^{\frac{ie}{\hbar}\int^q\mathcal{A}} \ , \tag{2.1.3}$$

obey the Schrödinger equation with Hamiltonian (2.1.2) and the same energy eigenvalue E. The phase difference between waves that travel above and below the solenoid in the presence of the magnetic flux is equal to the phase difference in the absence of magnetic flux, plus $\frac{e}{\hbar}\oint\mathcal{A}=\theta$. This phase, and hence the interference pattern, varies linearly with flux. When θ changes by 2π, the phase repeats itself. So the

† Y.Aharonov and D. Bohm "Significance of electromagnetic potentials in quantum theory", Phys. Rev. **115** 485491 (1959). "Further Considerations on Electromagnetic Potentials in the Quantum Theory", Phys. Rev. **123** 15111524 (1961).

interference pattern has to be periodic in Φ with period $\frac{2\pi\hbar}{e}$, as observed. This concludes the theoretical explanation of the Aharonov-Bohm effect.

Let us now discuss the mathematical meaning of the angle θ. The effect of a gauge transformation on the wave functions and potentials is

$$\mathcal{A}' = g^{-1}\mathcal{A}g - \frac{\hbar}{ie}g^{-1}dg = \mathcal{A} - \frac{\hbar}{e}d\alpha \ , \qquad \psi' = g^{-1}\psi \ , \tag{2.1.4}$$

where $g(x) = e^{i\alpha(x)}$ is a function from $\mathcal{Q}$ into $U(1)$. We assume that the wavefunctions ψ are periodic both before and after the gauge transformation, and therefore also g has to be a well-defined, single valued function into $U(1)$. † We say that two gauge potentials $\mathcal{A}$ and $\mathcal{A}'$ are gauge related only if the function g in (2.1.4) is single valued.

Now consider two gauge potentials $\mathcal{A} = \theta\frac{\hbar}{2\pi e}d\varphi$ and $\mathcal{A}' = \theta'\frac{\hbar}{2\pi e}d\varphi$ corresponding to different values $\Phi = \frac{\hbar}{e}\theta$ and $\Phi' = \frac{\hbar}{e}\theta'$ of the flux. Are they gauge related in the strict sense defined above? We have

$$\mathcal{A}' - \mathcal{A} = (\theta' - \theta)\frac{\hbar}{2\pi e}d\varphi \ , \tag{2.1.5}$$

and comparing with (2.1.4) we see that $\alpha(\varphi) = \frac{\theta-\theta'}{2\pi}\varphi$. The gauge potentials $\mathcal{A}$ and $\mathcal{A}'$ are gauge related if $e^{i\alpha}$ is single valued, which is equivalent to $\theta - \theta' = 2\pi n$, with n integer. There follows that the gauge equivalence classes of $U(1)$ gauge potentials on $\mathcal{Q}$ are parameterized by the values of θ belonging to some fundamental domain, for example $0 \le \theta < 2\pi$. The Aharonov–Bohm effect shows that experiments in quantum physics are sensitive to the gauge equivalence class of the potential.

One way of interpreting this effect is to say that there is an ambiguity in the quantization of the electron on the multiply connected space $\mathcal{Q} = \mathbf{R}^3 \backslash \{\text{line}\}$. In classical physics a charged particle feels the effect of the electromagnetic field only through the Lorentz force and since in the present case the magnetic field on $\mathcal{Q}$ vanishes, the *classical electron* is completely insensitive to the flux. The Aharonov-Bohm experiment shows that the *quantum electron* is sensitive to the flux, *i.e.* it is sensitive to changes of the gauge potentials $\mathcal{A}$, even when the field strength $\mathcal{F}$ remains the same (in this case, zero). One concludes that to a single classical theory there correspond infinitely many quantum theories, parametrized by the angle θ.

These considerations can be extended to arbitrary configuration spaces. Consider a particle with mass m, electric charge e, moving on a manifold $\mathcal{Q}$ with metric $g_{ij}(q)$, potential $V(q)$, magnetic field $\mathcal{F}_{ij}(q)$. Also, let $\mathcal{A}_i(q)$ be a gauge potential such that $\mathcal{F}_{ij} = \partial_i\mathcal{A}_j - \partial_j\mathcal{A}_i$. Everything that follows is true also in the case when $\mathcal{Q}$ is infinite dimensional (see Appendix E). The most general lagrangian quadratic in time derivatives of q is

$$L = \frac{1}{2}mg_{ij}(q)\dot{q}^i\dot{q}^j + e\mathcal{A}_i(q)\dot{q}^i - V(q) \ . \tag{2.1.6}$$

The momentum conjugate to q^i is

$$p_i = mg_{ij}(q)\dot{q}^j + e\mathcal{A}_i(q) \ , \tag{2.1.7}$$

and the canonical hamiltonian is

$$H = \frac{1}{2m}g^{ij}(q)\big(p_i - e\mathcal{A}_i\big)\big(p_j - e\mathcal{A}_j\big) + V(q) \ , \tag{2.1.8}$$

where $g^{ij}g_{jk} = \delta^i_k$. In the Schrödinger picture, coordinate representation, quantization is achieved by replacing q^i with the multiplicative operator $\hat{q}_i = q^i$ and p_i with the derivative operator $\hat{p}_i = -i\hbar\frac{\partial}{\partial q^i}$. Then we have

$$\widehat{p_i - e\mathcal{A}_i} = -i\hbar\left(\frac{\partial}{\partial q^i} - \frac{ie}{\hbar}\mathcal{A}_i\right) = -i\hbar\mathcal{D}_i \ , \tag{2.1.9}$$

† It is not strictly necessary to assume that ψ is periodic. See the discussion in section 3.2. However we will see there that one does not lose any generality by making this assumption.

where $\mathcal{D}_i$ is the covariant derivative with respect to $\mathcal{A}_i$, acting now on wavefunctions $\psi(q)$. Under local phase transformations (2.1.4) we have $\mathcal{D}'_i\psi' = e^{-i\alpha}(\mathcal{D}_i\psi)$. The hamiltonian becomes the operator

$$\hat{H} = -\frac{\hbar^2}{2m}\frac{1}{\sqrt{g}}\mathcal{D}_i\sqrt{g}g^{ij}\mathcal{D}_j + V \tag{2.1.10}$$

where $g = \det(g_{ij})$. (We have chosen a certain factor ordering in the first term which makes it equal to the covariant laplacian in the metric g_{ij}. This will be of no relevance in what follows.)

Now let us consider the special case $\mathcal{F} = 0$. In this case $\mathcal{A}$ is called a flat connection. There exists at least locally a function Λ such that

$$\mathcal{A}_i = -\frac{\hbar}{e}\partial_i\Lambda \tag{2.1.11}$$

so the second term in (2.1.6) is a total derivative:

$$L_T = e\dot{q}^i\mathcal{A}_i(q) = -\hbar\frac{d\Lambda}{dt}\ . \tag{2.1.12}$$

This term does not affect the equations of motion and therefore can be neglected in the classical theory. It is called a "topological term".

Looking at (2.1.11) one may be tempted to say that $\mathcal{A}$ is a pure gauge. This would be correct if $\mathcal{A}$ was a gauge field for the group of translations, $\mathbf{R}$. However, in quantum mechanics, $\mathcal{A}$ has to be interpreted as a gauge field for the gauge group $U(1)$. This is because in first quantization the gauge transformations of electromagnetism are phase transformations of the wave function (see (2.1.4)). The abelian groups $U(1)$ and $\mathbf{R}$ have the same Lie algebra and since $\mathcal{A}$ is a Lie-algebra-valued one-form, there is no way to discriminate locally between the two cases. The distinction appears at the global level, in the allowed gauge transformations. In fact, a function $\alpha : \mathcal{Q} \to \mathbf{R}$ defines a true (i.e. single-valued) $U(1)$ gauge transformation $g = e^{i\alpha}$ only if α, restricted to any loop in $\mathcal{Q}$, has a polydromy which is an integral multiple of 2π. Note that if the loop is homotopic to a constant, α is necessarily single valued. Therefore, nontrivial $U(1)$ flat connections can only exist if $\mathcal{Q}$ is multiply connected.

Classically, the Lagrangian (2.1.12) is completely immaterial. However, at the quantum level, the topological term matters: gauge inequivalent potentials give rise to different quantum theories, as proven by the Aharonov-Bohm effect. Thus, for a given classical theory there will be as many inequivalent quantum theories as there are gauge equivalence classes of flat $U(1)$ connections on $\mathcal{Q}$.

The set of flat $U(1)$ connections modulo gauge transformations is $\mathrm{Hom}\big(\pi_1(\mathcal{Q}), U(1)\big)$, the set of homomorphisms from $\pi_1(\mathcal{Q})$ into $U(1)$ $\big($these homomorphisms are also called "characters" of $\pi_1(\mathcal{Q})\big)$. To show this we recall that all the gauge invariant information about a connection is contained in its holonomies ("Wilson loops")

$$\chi(\ell) = e^{\frac{ie}{\hbar}\oint_\ell \mathcal{A}}\ . \tag{2.1.13}$$

In the case of a flat connection, these holonomies are invariant under continuous deformations of the loop (homotopies). Thus they only depend on the homotopy class of the loop. It is easy to see, using the definition of product of homotopy classes given in Appendix A, that $\chi(\ell_1 \cdot \ell_2) = \chi(\ell_1)\chi(\ell_2)$, so χ defines a homomorphism from $\pi_1(\mathcal{Q})$ into $U(1)$. Conversely, given any character χ, it can be shown that there exists a flat connection $\mathcal{A}$ such that (2.1.13) holds.

To summarize: if $\pi_1(\mathcal{Q}) \neq 0$, the correspondence between classical and quantum theories is not unique. All classical theories with lagrangian (2.1.6) and $\mathcal{F} = d\mathcal{A} = 0$ are equivalent. Quantum theories with lagrangian (2.1.6) and $\mathcal{F} = d\mathcal{A} = 0$ are only equivalent when the potentials are $U(1)$-related. The set of inequivalent quantum theories with the same classical limit is parameterized by the characters of $\pi_1(\mathcal{Q})$.

In the next four sections we shall consider increasingly complicated systems whose configuration spaces will have $\pi_1(\mathcal{Q}) = \mathbf{Z}$. Since $\mathrm{Hom}(\mathbf{Z}, U(1)) = U(1)$ † these theories can be quantized in inequivalent ways parametrized by an angle $0 \leq \theta < 2\pi$. These inequivalent quantum theories are called "theta sectors".

† Consider the homomorphisms $h_\theta : \mathbf{Z} \to U(1)$ defined by $h_\theta(n) = e^{in\theta}$. Clearly $h_{\theta+2\pi} = h_\theta$.

2.2. Quantum mechanical examples

The prototype of all theories admitting theta vacua is the pendulum. Its configuration space is $\mathcal{Q}=S^1$, and since $\pi_1(S^1)=\mathbf{Z}$, we expect to find inequivalent quantizations labelled by an angle θ. The usual lagrangian for the pendulum is, in suitable units,

$$L_0=\frac{1}{2}\dot{\varphi}^2-V(\varphi)\ , \tag{2.2.1}$$

where $0\leq\varphi<2\pi$ is the coordinate on S^1 and $V(\varphi)=1-\cos\varphi$ is the gravitational potential. The explicit form of the kinetic and potential terms will not enter in the considerations of this section, but will become relevant later. In particular, the presence of the gravitational potential will be necessary in Section 3.7 for the application of the WKB method.

In order to recognize the existence of inequivalent quantum theories, we use the freedom of adding to the lagrangian a total time derivative $\frac{d\Lambda}{dt}$ where Λ is a given function of φ. So we add to L_0 a term

$$L_T=\theta\frac{\hbar}{2\pi}\frac{d\varphi}{dt} \tag{2.2.2}$$

where θ is an arbitrary real parameter. This does not change the equations of motion, so the classical theory is independent of the value of θ. Assuming that for $|t|\to\infty$, $\varphi(t)\to 0$, this corresponds to adding to the action the term

$$S_T(\varphi)=\theta\frac{\hbar}{2\pi}\int dt\frac{d\varphi}{dt}=\theta\hbar W(\varphi)\ , \tag{2.2.3}$$

where $W(\varphi)$ is the winding number of the history $\varphi(t)$, counting the total number of times the pendulum rotates about its center in the course of the time evolution. Because of this topological significance, the term S_T is known as a "topological term".

From a physical point of view, the term L_T represents the interaction of the particle (carrying charge $e=1$) with the magnetic potential $\mathcal{A}_{(\theta)}=\theta\frac{\hbar}{2\pi}d\varphi$. This is the same as the Aharonov-Bohm potential (2.1.1). In fact, apart from giving the explicit form of L_0, we have simply restricted the motion of the particle by fixing the value of z (the axial coordinate along the solenoid) and r (the distance from the center of the solenoid). So the discussion in the previous section goes through unchanged and we are led to conclude that values of θ differing by integers multiples of 2π correspond to gauge equivalent potentials and therefore give equivalent quantum theories.

For each of these theories the Hilbert space is $\mathcal{H}=L^2(S^1)$, the space of complex functions $\psi(\varphi)$ such that

$$\psi(\varphi+2\pi)=\psi(\varphi) \tag{2.2.4}$$

and $\int_0^{2\pi}d\varphi\psi^*\psi<\infty$. The Hamiltonian operator is

$$\hat{H}_\theta=-\frac{\hbar^2}{2}\mathcal{D}_\varphi\mathcal{D}_\varphi+V(\varphi)\ , \tag{2.2.5}$$

where

$$\mathcal{D}_\varphi=\frac{\partial}{\partial\varphi}-i\frac{\theta}{2\pi}\ . \tag{2.2.6}$$

(Note that since the metric on S^1 is independent of φ in this case there are no ordering ambiguities).

Could one reveal the existence of theta sectors without adding the topological term to the lagrangian? This is possible if we observe that on a multiply connected space such as S^1 the wave functions need not satisfy the strict periodicity condition (2.2.4) but could satisfy instead a condition of periodicity up to a phase:

$$\psi(\varphi+2\pi)=e^{-i\theta}\psi(\varphi). \tag{2.2.7}$$

(In geometrical language one would say that the wave functions are not ordinary complex functions but rather sections of a complex line bundle. We will not discuss this type of geometrical refinements here).

Such wave functions form a Hilbert space $\mathcal{H}_\theta$ and it is clear that values of θ differing by integer multiples of 2π correspond to the same quantum theory, so it is only θ mod 2π that counts. Since the lagrangian is given only by L_0, in this case the hamiltonian operator is

$$\hat{H}_0 = -\frac{\hbar^2}{2}\partial_\varphi^2 + V(\varphi) \tag{2.2.8}$$

For every value of $\theta \mathrm{mod} 2\pi$ one has a different Hilbert space and therefore a different quantum theory.

How is this description related to the previous one? We note that $\mathcal{A}_{(\theta)} = \theta\frac{\hbar}{2\pi}d\varphi = -\frac{\hbar}{i}g^{-1}dg$ where $g(\varphi) = e^{-i\frac{\theta}{2\pi}\varphi}$. We have seen that unless θ is an integral multiple of 2π this is not a $U(1)$ gauge transformation in a strict sense. However, we can try to implement this transformation in the quantum theory. Formally, it is given by an operator $\mathcal{U} = e^{-i\frac{\theta}{2\pi}\hat{\varphi}}$. It is easy to see that if ψ is periodic, then $\psi' = \mathcal{U}\psi$ satisfies (2.2.7) and $\hat{H}_0 = \mathcal{U}\hat{H}_\theta\mathcal{U}^{-1}$. Therefore $\mathcal{U}$ is an isomorphism of $\mathcal{H}$ to $\mathcal{H}_\theta$ mapping $\hat{H}_\theta$ to $\hat{H}_0$.

It is a general fact that the theta sectors always admit two descriptions: either with a topological term in the lagrangian and single-valued wave functions or without topological term and with multiple-valued wave functions. In the first description the θ dependence is in the Hamiltonian, in the second in the states. (In this sense, the relation between these descriptions is similar to the relation between the Heisenberg and the Schrödinger picture of quantum mechanics.) The transformation between the two descriptions has the form of a gauge transformation with multiple-valued gauge function. Thus it is not a $U(1)$ gauge transformation in the strict sense. Except for this section, in these notes we will stick to the first description.

Before coming to the field theoretic examples it is worthwile mentioning that quantum spin and statistics can also be seen as a manifestation of the same type of ambiguity that leads to the existence of theta sectors.

A classical model for a particle with spin is the rigid rotator. The configuration space of this system is $\mathcal{Q} = \mathbf{R}^d \times SO(d)$, where d is the dimension of space. The group $SO(d)$ has fundamental group $\mathbf{Z}$ for $d = 2$ and $\mathbf{Z}_2$ for $d > 2$. Thus one would expect inequivalent quantizations labelled by an angle in two dimensions and by $Hom(\mathbf{Z}_2, U(1)) = \mathbf{Z}_2$ in higher dimensions. This is indeed what happens. We have seen that the inequivalent quantizations can be described by choosing the periodicity conditions on the wave function:

$$\psi(t + 2\pi) = e^{i\theta}\psi(t) \ , \tag{2.2.9}$$

where t is some parameter along the loop. In the case of the rotation group, the fundamental noncontractible loop consists of a rotation of the body by an angle 2π about some axis. Therefore (2.2.9) describes the behaviour of the wave function under a 2π rotation. This can be compared with the definition of spin in quantum mechanics. The wave function of a system with spin s acquires a phase $e^{2\pi i s}$ when the system is rotated by 2π. So we learn that θ is equal to $2\pi s \bmod 2\pi$. In $d > 2$ the spin can be integer, corresponding to single-valued wave functions, if $\theta = 2\pi n$, or half integer, corresponding to wave functions that change sign under 2π rotations, if $\theta = \pi n$ with n odd. In two dimensions the spin can take any real value and the corresponding particles are called *anyons*.

For a multiparticle system, the statistical parameter σ is defined by

$$\psi(\ldots, \vec{x}_i, \ldots, \vec{x}_j, \ldots) = e^{2\pi i\sigma}\psi(\ldots, \vec{x}_j, \ldots, \vec{x}_i, \ldots) \ . \tag{2.2.10}$$

The usual Bose–Einstein and Fermi–Dirac statistics correspond to σ integer and half-integer respectively. To see the connection between statistics and inequivalent quantizations, consider the classical configuration space of two identical particles in d dimensions. Let us also assume that the particles cannot be at the same point in space, because in this case the statistics could only be bosonic ((2.2.10) is compatible with $\vec{x}_1 = \vec{x}_2$ only for integer σ). The configuration space is then $\mathcal{Q} = (\mathbf{R}^{2d} \setminus \Delta)/S_2$, where Δ is the subset of $\mathbf{R}^{2d}$ for which the particle positions coincide, and $S_2 = \mathbf{Z}_2$ is the permutation group of two objects. Passing from the coordinates $(\vec{x}_1, \vec{x}_2)$ to the center-of-mass coordinates $(x_{\mathrm{CM}}, \Delta\vec{x}) = (\frac{\vec{x}_1+\vec{x}_2}{2}, \frac{\vec{x}_2-\vec{x}_1}{2})$ shows that the topology of the space $\mathbf{R}^{2d} \setminus \Delta$ is $\mathbf{R}^d \times \mathbf{R}^+ \times S^{d-1}$ (here $\mathbf{R}^d$ is parametrized by $\vec{x}_{\mathrm{CM}}$, $\mathbf{R}^+$ is parametrized by $|\Delta\vec{x}|$ and S^{d-1} is parametrized by the angular variables of $\Delta\vec{x}$). For $d > 2$ this space is simply connected; the group S_2 acts on it by $(x_{\mathrm{CM}}, \Delta\vec{x}) \to (x_{\mathrm{CM}}, -\Delta\vec{x})$ and therefore acts antipodally on S^{d-1}; the quotient has topology $\mathbf{R}^d \times \mathbf{R}^+ \times RP^{d-1}$ where $RP^{d-1} = S^{d-1}/\mathbf{Z}_2$ is a real projective space, whose fundamental group is $\mathbf{Z}_2$. The system of two particles can therefore be quantized in two inequivalent ways, corresponding to

bosonic and fermionic statistics. For $d=2$, already $\mathbf{R}^{2d}\setminus\Delta$ has a nontrivial fundamental group, equal to $\mathbf{Z}$, and $\pi_1(\mathcal{Q})=\mathbf{Z}$ too. In this case the inequivalent quantizations are labelled by the angle σ; one then speaks of fractional statistics. These considerations can be generalized to the case of N indistinguishable particles.

2.3. Spherical sigma models

Let us now consider the S^2 nonlinear sigma model in 1+1 dimensions. This is perhaps the simplest field theoretic example showing the existence of theta sectors. It is easier to discuss than gauge theories, because one can work directly with the true, unconstrained degrees of freedom of the theory and there are no complications due to gauge invariance. We work in the "intrinsic" formulation, in terms of two fields φ^α which have the meaning of coordinates on S^2. We choose a metric $h_{\alpha\beta}(\varphi)$ on S^2 and write the action as

$$S_0=-\frac{f^2}{2}\int d^2x\,\partial_\mu\varphi^\alpha\partial^\mu\varphi^\beta h_{\alpha\beta}(\varphi) \tag{2.3.1}$$

The canonical configuration space of this model is $\mathcal{Q}=\Gamma_*(S^1,S^2)$, where the constant time spacelike surfaces $\mathbf{R}$ have been compactified to S^1 due to the requirement of finiteness of the energy [†]. This space is called the loop space of S^2. Its fundamental group is $\pi_1(\mathcal{Q})=\pi_2(S^2)=\mathbf{Z}$ (see Appendix F). So this theory will admit theta sectors, labelled by an angle $0\le\theta<2\pi$. The fundamental non-contractible loop in $\mathcal{Q}$ (i.e. the loop whose homotopy class generates $\pi_1(\mathcal{Q})$) can be described as follows. Points on $\mathcal{Q}$ are loops in S^2 beginning and ending at some basepoint y_0, i.e. maps $c:[0,1]\to S^2$ such that $c(0)=c(1)=y_0$. The basepoint of $\mathcal{Q}$ itself is the constant loop which maps all of $[0,1]$ into y_0. Consider the one-parameter family of loops c_t depicted if fig.**XXX**. When $t=0$ we have the constant loop. For growing t, the loops sweep out the whole sphere, and for $t\to1$ it shrinks back to the constant loop. Clearly c_t is a non-contractible loop of loops. More formally, the isomorphism between $\pi_0(\mathcal{Q})$ and $\pi_2(S^2)$ can be described as follows: if $c:I\to\mathcal{Q}$ is a loop in $\mathcal{Q}$ we define $\hat c:I\times I\to S^2$ by $\hat c(t,s)=\big(c(t)\big)(s)$, where $c(t)$, for fixed t, is regarded as a map $I\to S^2$. We have $\hat c(t,s)=y_0$ whenever t or s are equal to 0 or 1, so $\hat c$ defines a map $S^2\to S^2$. Clearly homotopies of c correspond to homotopies of $\hat c$. So the desired isomorphism correspond to mapping $[c]$ to $[\hat c]$.

In order to make the theta sectors manifest, we add to the action a topological term $S_T=\theta W(\varphi)$, where

$$W(\varphi)=\frac{1}{4\pi}\int d^2x\,\varepsilon^{\mu\nu}\partial_\mu\varphi^\alpha\partial_\nu\varphi^\beta\frac{1}{2!}\sqrt{h}\,\varepsilon_{\alpha\beta} \tag{2.3.2}$$

is the winding number of the map φ (see Appendix A). The addition of W does not change the equations of motion, nor the form of the energy because it is a total derivative. In fact, we have locally $\sqrt{h}\varepsilon_{\alpha\beta}=\partial_\alpha\tau_\beta-\partial_\beta\tau_\alpha$ for some one-form τ. Then $W(\varphi)=\int d^2x\,\partial_\mu\omega^\mu$, where

$$\omega^\mu=\frac{1}{4\pi}\varepsilon^{\mu\nu}\partial_\nu\varphi^\alpha\tau_\alpha(\varphi)\ . \tag{2.3.3}$$

However, the addition of the topological term affects the relation between velocities and momenta:

$$\pi_\alpha=f^2h_{\alpha\beta}\partial_0\varphi^\beta+\mathcal{A}_\alpha\ , \tag{2.3.4}$$

where

$$\mathcal{A}_\alpha(x)=\frac{\theta}{4\pi}\partial_1\varphi^\beta\sqrt{h}\,\varepsilon_{\alpha\beta}\ . \tag{2.3.5}$$

Comparing with equation (2.1.7) we see that $\mathcal{A}_\alpha$ can be regarded as a "functional magnetic potential" on $\mathcal{Q}$. In fact we can write the action $S=S_0+S_T=\int dt(L_0+L_T)$, with

$$L_0=\frac{f^2}{2}g_\varphi(\dot\varphi,\dot\varphi)-V(\varphi)\quad;\quad L_T=\mathcal{A}_\varphi(\dot\varphi) \tag{2.3.6}$$

[†] Strictly speaking finiteness of the energy requires only that $\varphi(x)\to N_\pm$ for $x\to\pm\infty$ where N_+ could be different from N_-. This would not change the following results. We assume $N_+=N_-$ in the following.

This is exactly the form (2.2.1) except for the replacement of the index i with the infinite indexing set (α, x). The potential is $V(\varphi)=\frac{f^2}{2}g_\varphi(\partial_1\varphi,\partial_1\varphi)$, the magnetic potential is the one-form $\mathcal{A}=\int dx\mathcal{A}_\alpha(x)\delta\varphi^\alpha(x)$ and the riemannian metric is $g=\int dx h_{\alpha\beta}\big[\varphi(x)\big]\delta\varphi^\alpha(x)\delta\varphi^\beta(x)$. In these formulae $\delta\varphi^\alpha(x)$ play the role of the differentials dq^i in the finite dimensional case. This terminology is further explained in appendix E.

In this way the theory can be interpreted as the motion of a particle with mass $m=f^2$ and charge $e=1$ on the manifold $\mathcal{Q}$ in a background metric g and background magnetic field with magnetic potential $\mathcal{A}$. Since the topological term (i.e. the magnetic field) does not appear in the equation of motion, we expect that $\mathcal{F}=d\mathcal{A}=0$. This is what one gets from a direct calculation based on formula (E.16)

$$d\mathcal{A}(v,w)=v\big(\mathcal{A}(w)\big)-w\big(\mathcal{A}(v)\big)-\mathcal{A}\big([v,w]\big)\ . \tag{2.3.7}$$

There follows that, at least locally on $\mathcal{Q}$,

$$\mathcal{A}=d\Lambda. \tag{2.3.8}$$

In fact we have

$$\Lambda=\theta\int dx\,\omega^0=\frac{\theta}{4\pi}\int dx\,\partial_1\varphi^\alpha\tau_\alpha\ . \tag{2.3.9}$$

The function Λ is not single valued. The polydromy of Λ on the fundamental loop in $\mathcal{Q}$ is

$$\begin{aligned}\oint d\Lambda=\oint\mathcal{A}=\int d\tau\bigg[\frac{\theta}{4\pi}\int dx\partial_1\varphi^\alpha\frac{d\varphi^\beta}{d\tau}\sqrt{h}\,\varepsilon_{\alpha\beta}\bigg]\\ =\frac{\theta}{4\pi}\int d^2x\varepsilon^{\lambda\mu}\partial_\lambda\hat\varphi^\alpha\partial_\mu\hat\varphi^\beta\frac{1}{2}\sqrt{h}\,\varepsilon_{\alpha\beta}=\theta W(\hat\varphi)=\theta\end{aligned} \tag{2.3.10}$$

Therefore, Λ is single-valued only if $\theta=0$. However, if $\theta=2\pi n$, with $n\in\mathbf{Z}$, $e^{i\Lambda}$ is a single-valued function $\Gamma_*(S^1,S^2)\to U(1)$ and so the gauge potentials $\mathcal{A}_{\theta+2\pi n}$ and $\mathcal{A}_\theta$ are gauge-related in the strict sense. The gauge inequivalent magnetic potentials, and hence the inequivalent quantizations, are labelled by $0\le\theta<2\pi$.

The pendulum and the sigma model discussed in this section are the $d=0$ and $d=1$ cases of an infinite sequence of theories that behave all in the same way. The S^{d+1}-valued sigma model in d space dimensions has configuration space $\mathcal{Q}=\Gamma_*(S^d,S^{d+1})$ and $\pi_1(\mathcal{Q})=\pi_d(S^d)=\mathbf{Z}$. The topological term is given again by the winding number.

2.4. Abelian gauge theory in 1+1 dimensions

We now consider a $U(1)$ gauge field A_μ in one space dimension. The action is

$$S=S_{YM}+S_T \tag{2.4.1}$$

where

$$S_{YM}=-\frac{1}{4}\int d^2x\,F_{\mu\nu}F^{\mu\nu} \tag{2.4.2}$$

is the usual kinetic term and $S_T=\theta c_1$, with

$$c_1=\frac{1}{4\pi}\int d^2x\,\varepsilon^{\mu\nu}F_{\mu\nu} \tag{2.4.3}$$

is a "topological term". The topological significance of this term will be understood better in section 2.8. For the time being we merely observe that $\frac{1}{4\pi}\varepsilon^{\mu\nu}F_{\mu\nu}=\partial_\mu C^\mu$, where

$$C^\mu=\frac{1}{2\pi}\varepsilon^{\mu\nu}A_\nu \tag{2.4.4}$$

is known as the (dual of the) one-dimensional Chern-Simons form. There follows that c_1 is invariant under infinitesimal variations of the field A_μ that vanish at infinity, and therefore does not contribute to the classical equations of motion. However, it does enter the canonical definition of momentum and hamiltonian

$$P^1(x)=\frac{\partial\mathcal{L}}{\partial\partial_0A_1(x)}=E_1(x)+\frac{\theta}{2\pi} \tag{2.4.5}$$

$$H = \int dx \left[\frac{1}{2} \left(P^1 - \frac{\theta}{2\pi} \right)^2 - A_0 \partial_1 E_1 \right] \tag{2.4.6}$$

where $E_1 = F_{01} = \partial_0 A_1 - \partial_1 A_0$. The field A_0 has vanishing conjugate momentum. In fact, its role is that of Lagrange multiplier enforcing the Gauss law constraint $\partial_1 P^1 = 0$. Our discussion will be simplified by choosing now the gauge $A_0 = 0$. This leaves a residual gauge freedom consisting of time-independent gauge transformations. With this choice of gauge $E_1 = \dot{A}_1$, so the energy $E = \int dx \frac{1}{2} E_1^2$ is seen to be of purely kinetic character: the static energy is zero. The configuration space $\mathcal{Q}$ of this theory consists of gauge fields $A_1(x)$ modulo gauge transformations. We denote $\mathcal{C}$ the space of $U(1)$ connections and $\mathcal{G} = \Gamma_*(S^1, U(1))$ the gauge group, consisting of maps $g : \mathbf{R} \to U(1)$ such that $g \to 1$ for $|x| \to \infty$ (hence the possibility of compactifying $\mathbf{R}$ to S^1). So $\mathcal{Q} = \mathcal{C}/\mathcal{G}$. The space $\mathcal{C}$ is essentially a vector space, so it has trivial topology, but $\mathcal{Q}$ is multiply connected. In fact, $\pi_1(\mathcal{Q}) = \pi_0(\mathcal{G}) = \pi_1(S^1) = \mathbf{Z}$. The fact that $\pi_1(\mathcal{Q})$ and $\pi_0(\mathcal{G})$ are isomorphic is proven in Appendix D. Here we merely describe this isomorphism.

The gauge group $\mathcal{G}$ consists of infinitely many connected components $\mathcal{G}_n = \{ g : S^1 \to U(1) \mid W(g) = n \}$. Now choose a basepoint $A_{(0)} = 0 \in \mathcal{C}$ (for definiteness we will take $A_{(0)} = 0$, but this is by no means necessary) and consider the orbit through $A_{(0)}$, i.e. the set of all connections of the form $A^g_{(0)} = g^{-1} dg$ for $g \in \mathcal{G}$. The only gauge transformation that leaves any gauge field invariant is the identity; in mathematical language, the gauge group acts on $\mathcal{C}$ freely. Therefore there is a one-to-one correspondence between points of $\mathcal{G}$ and points of the orbit through $A_{(0)}$. (See Appendix C). So the topology of the orbit is the same as the topology of $\mathcal{G}$. There is a natural projection $p : \mathcal{C} \to \mathcal{Q}$ which associates to A its gauge equivalence class $[A]$. Under this projection all points in the orbit through $A_{(0)}$ are mapped to the same point $[A_{(0)}]$ in $\mathcal{Q}$. It is natural to take $A_{(0)}$ as the basepoint in $\mathcal{C}$, $[A_{(0)}]$ as a basepoint in $\mathcal{Q}$. Now consider a gauge transformation g with $W(g) = 1$. There is no continuous path in $\mathcal{G}$ joining g to the identity, and therefore there is also no path in the orbit through $A_{(0)}$ joining $A^g_{(0)} = g^{-1} dg$ to $A_{(0)}$. However, the space $\mathcal{C}$ is connected and so there is some path $\tilde{\ell}_t$ in $\mathcal{C}$, with $t \in [0, 1]$ such that $\tilde{\ell}_0 = A_{(0)}$ and $\tilde{\ell}_1 = A^g_{(0)}$. For instance one can take $c_t = t\, g^{-1} dg = t\, d\alpha$ (note that although for all t the curvature of $\tilde{\ell}_t$ is zero, there is no continuous path g_t such that $\tilde{\ell}_t = g_t^{-1} dg_t$). The natural projection p maps this path in $\mathcal{C}$ to a path $\ell_t = [\tilde{\ell}_t]$ in $\mathcal{Q}$ beginning and ending at $[A_{(0)}]$. The desired isomorphism $\pi_1(\mathcal{Q}) \to \pi_0(\mathcal{G})$ is obtained by mapping the homotopy class of g to the homotopy class of the loop ℓ_t in $\mathcal{Q}$. See fig. 10.

Returning to equations (2.4.5) and (2.4.6) we see that the topological term θc_1 in the action can be written, in the gauge $A_0 = 0$, as $\int dt \int dx \dot{A}_1 \frac{\theta}{2\pi}$ and hence can be regarded as the interaction of a particle with unit charge and coordinate $A_1(x)$ with a magnetic potential (a one-form on $\mathcal{C}$)

$$\tilde{\mathcal{A}} = \int dx \, \frac{\theta}{2\pi} \delta A_1(x) \ . \tag{2.4.7}$$

Since the components of the vector potential are constant, it is easy to verify that the corresponding magnetic field $\tilde{\mathcal{F}} = d\tilde{\mathcal{A}} = 0$. This is in accordance with the fact that the topological term does not contribute to the equation of motion: if it did, one could interpret the corresponding term in the equation of motion as a Lorentz force due to a nonzero $\tilde{\mathcal{F}}$. Since $d\tilde{\mathcal{A}} = 0$, we can write at least locally $\tilde{\mathcal{A}} = d\tilde{\Lambda}$. The functional $\tilde{\Lambda}$ on $\mathcal{C}$ that has this property is

$$\tilde{\Lambda} = \frac{\theta}{2\pi} \int dx A_1(x) \ . \tag{2.4.8}$$

All this is on the contractible space $\mathcal{C}$.

We would like now to see the corresponding steps being carried out on $\mathcal{Q}$. It is convenient to write a time-independent gauge transformation in the form $g(x) = e^{i\alpha(x)}$, where $\alpha \to 2\pi n_-$, for $x \to -\infty$ and $\alpha \to 2\pi n_+$, for $x \to \infty$. The winding number of g is just $n_+ - n_-$. Infinitesimal gauge transformations are real-valued functions $\epsilon(x)$ such that $\epsilon \to 0$ for $|x| \to \infty$.

We now consider again the function $\tilde{\Lambda}$ and ask whether it is the pullback of a function on $\mathcal{Q}$. This will be the case provided $\tilde{\Lambda}$ is constant on the orbits, *i.e.* if it is gauge invariant. Under a gauge transformation g,

$$\tilde{\Lambda}(A^g) - \tilde{\Lambda}(A) = \frac{\theta}{2\pi i} \int dx \, g^{-1} dg = \frac{\theta}{2\pi} \int dx \, \frac{d\alpha}{dx} = \theta W(g) \ . \tag{2.4.9}$$

Therefore, $\tilde{\Lambda}$ is invariant under gauge transformations which are connected to the identity, but not under "large" gauge transformations, *i.e.* transformations that have winding number different from zero. Under these circumstances, $\tilde{\Lambda}$ does not define a function Λ on $\mathcal{C}/\mathcal{G}$, but only a function which is defined up to integer multiples of θ.

Similarly, we can ask if $\tilde{\mathcal{A}}=p^*\mathcal{A}$ for some one-form $\mathcal{A}$ on $\mathcal{C}/\mathcal{G}$. This is true provided:

1) $\tilde{\mathcal{A}}$ is gauge invariant;
2) $\tilde{\mathcal{A}}(v)=0$ when v is a vertical vector (i.e. v is tangent to the orbit).

(See reference [1] vol. II, p. 294, lemma 1). The first condition is obviously satisfied, and for the second we observe that a vertical vector has the form $v_\epsilon = \int dx\, \partial_1\epsilon \frac{\delta}{\delta A_1}$, where ϵ is an infinitesimal gauge parameter; then

$$\tilde{\mathcal{A}}(v_\epsilon)=\frac{\theta}{2\pi}\int dx\partial_1\epsilon=\frac{\theta}{2\pi}\big(\epsilon(\infty)-\epsilon(-\infty)\big)=0\ . \tag{2.4.10}$$

So there is a one-form $\mathcal{A}$ on $\mathcal{Q}$ such that $\tilde{\mathcal{A}}=p^*\mathcal{A}$. Since p is surjective, $\mathcal{A}$ is entirely determined by $\tilde{\mathcal{A}}$, and since $p^*d=dp^*$, $d\mathcal{A}=0$ and, locally, $\mathcal{A}=d\Lambda=\frac{1}{i}e^{-i\Lambda}de^{i\Lambda}$.

According to the general discussion in section 3.1, inequivalent quantizations correspond to the gauge inequivalent magnetic potentials $\mathcal{A}$. The magnetic potential $\mathcal{A}(\theta)$ will be gauge equivalent to $\mathcal{A}(\theta=0)$ if the function $e^{i\Lambda}$ is single-valued, i.e. if the polydromy of Λ is an integral multiple of 2π. From the construction of the fundamental loop ℓ in $\mathcal{Q}$ we see that the polydromy of Λ on ℓ is equal to $\oint_\ell \mathcal{A}=\int_{\tilde{\ell}}\tilde{\mathcal{A}}$, where $\tilde{\ell}$ is a lift of ℓ, i.e. a path joining $A_{(0)}$ to $A^g_{(0)}$, with $W(g)=1$. But then $\int_{\tilde{\ell}}\tilde{\mathcal{A}}=\tilde{\Lambda}(A^g)-\tilde{\Lambda}(A)=\theta$, by equation (2.4.9). So, whenever $\theta=2\pi n$, $\mathcal{A}(\theta)$ is a pure gauge. The classes of gauge inequivalent $\mathcal{A}$'s are parameterized again by $0\le\theta<2\pi$.

2.5. Nonabelian Yang–Mills theory in 3+1 dimensions

Except for algebraic complications, the discussion of a nonabelian Yang– Mills theory in 3+1 dimensions is very similar to that of the abelian theory in 1+1 dimensions. It is convenient to use the rescaled, geometrical gauge fields, so that the curvature is given by and the gauge transformations act as in . The total action is $S = S_{YM} + S_T$ where S_{YM} is given by (with $d=3$) and $S_T = \theta c_2$, where

$$c_2 = \frac{1}{64\pi^2}\int d^4x\, \varepsilon^{\mu\nu\rho\sigma}F^a_{\mu\nu}F^a_{\rho\sigma} \tag{2.5.1}$$

is a topological term, known as the second Chern class. This term does not modify the classical equations of motion since

$$\frac{1}{64\pi^2}\varepsilon^{\mu\nu\rho\sigma}F^a_{\mu\nu}F^a_{\rho\sigma} = \partial_\mu C^\mu\ , \tag{2.5.2}$$

where

$$C^\mu = \frac{1}{16\pi^2}\varepsilon^{\mu\nu\rho\sigma}\left(A^a_\nu\partial_\rho A^a_\sigma + \frac{1}{3}f_{abc}A^a_\nu A^b_\rho A^c_\sigma\right) \tag{2.5.3}$$

is known as the (dual of the) three dimensional Chern-Simons form. Thus c_2 is invariant under infinitesimal variations of A^a_μ. However, it changes the relation between velocities and momenta. We have

$$\begin{aligned} P^0_a &= \frac{\partial L}{\partial\partial_0 A^a_0} = 0 \\ P^i_a &= \frac{\partial L}{\partial\partial_0 A^a_i} = \frac{1}{e^2}E^a_i + \frac{\theta}{8\pi^2}B^a_i \end{aligned} \tag{2.5.4}$$

where $E^a_i = F^a_{0i} = \partial_0 A^a_i - D_i A^a_0$ and $B^a_i = \frac{1}{2}\varepsilon_{ijk}F^a_{jk}$. The hamiltonian is

$$H = \int d^3x\, \left[\frac{e^2}{2}\left(P^a_i - \theta\frac{e^2}{8\pi^2}B^a_i\right)^2 + \frac{1}{2e^2}\left(B^a_i\right)^2 - A^a_0 D_i E^i_a\right]\ . \tag{2.5.5}$$

We now choose the gauge $A_0 = 0$. In this case the last term in H drops out, while the first and the second are recognized as kinetic and static energy respectively (in this gauge $E_i^a = \partial_0 A_i^a$). Let $\mathcal{C}$ be the space of all gauge potentials A_i^a with finite static energy, i.e. such that $\int d^3x (B_i^a)^2$ is finite. Let $\mathcal{G}$ be the residual gauge group, consisting of time-independent gauge transformations such that $g(x) \to \mathbf{1}$ for $|\vec{x}| \to \infty$. With these boundary conditions, $\mathbf{R}^3$ can be compactified to S^3 and $\mathcal{G} = \Gamma_*(S^3, G)$. As in the previous section, $\mathcal{G}$ acts freely on $\mathcal{C}$. † The physical configuration space of the theory is the orbit space $\mathcal{Q} = \mathcal{C}/\mathcal{G}$.

Since $\mathcal{C}$ is topologically trivial we have, following again the arguments of Appendix D, $\pi_1(\mathcal{Q}) = \pi_0(\mathcal{G}) = \pi_3(G) = \mathbf{Z}$. The isomorphism between $\pi_1(\mathcal{Q})$ and $\pi_0(\mathcal{G})$ is described again by fig. 10. The homotopy class $[g]$ of a gauge transformation corresponds to the homotopy class of the loop ℓ which is obtained by projecting to $\mathcal{Q}$ a curve $\tilde{\ell}$ joining $A = 0$ to $A^g = g^{-1}dg$. Comparing equations (2.5.4) and (2.5.5) with (2.1.10) and (2.1.9) we see that the topological term has given rise to a magnetic potential $\tilde{\mathcal{A}}$ on $\mathcal{C}$ defined by

$$\tilde{\mathcal{A}}(A) = \frac{\theta}{8\pi^2} \int d^3x \, B_i^a \delta A_i^a \ . \tag{2.5.6}$$

A direct calculation shows that $d\tilde{\mathcal{A}} = 0$. In fact, we have $\tilde{\mathcal{A}} = d\tilde{\Lambda}$, with

$$\tilde{\Lambda} = \theta \int d^3x \, C^0 = \frac{\theta}{16\pi^2} \int d^3x \, \varepsilon^{ijk} \left(A_i^a \partial_j A_k^a + \frac{1}{3} f_{abc} A_i^a A_j^b A_k^c \right) \ . \tag{2.5.7}$$

See Exercise 2.5.1. As in the previous section, one would like to describe the theory as a particle moving in $\mathcal{Q}$, rather than $\mathcal{C}$, so the question arises again whether the function $\tilde{\Lambda}$ and the form $\tilde{\mathcal{A}}$ can be projected onto a function Λ and a form $\mathcal{A}$ on $\mathcal{Q}$. Under a gauge transformation g, one finds

$$\tilde{\Lambda}(A^g) - \tilde{\Lambda}(A) = \theta W(g) \ . \tag{2.5.8}$$

So $\tilde{\Lambda}$ is invariant under gauge transformations connected to the identity, but not under "large" transformations: it projects to a function Λ on $\mathcal{Q}$ which is only defined modulo integral multiples of θ.

To see if $\tilde{\mathcal{A}}$ projects, we have to verify whether the conditions given in the preceding section are satisfied. Given an infinitesimal gauge transformation parameter ϵ, a map from $\mathbf{R}^3$ to the Lie algebra of $SU(2)$ which goes to zero at infinity, we construct the corresponding vertical vectorfield in $\mathcal{C}$

$$v_\epsilon = \int d^3x \, D_i \epsilon^a \frac{\delta}{\delta A_i^a} \ .$$

Then we have:

1) $\tilde{\mathcal{A}}$ is gauge invariant (B_i^a and δA_i^a both transform homogeneously);
2) $\tilde{\mathcal{A}}(v_\epsilon) = \frac{\theta}{8\pi^2} \int d^3x B_i^a D_i \epsilon^a = 0$ upon integrating by parts, using Bianchi's identity and the fact that $\epsilon \to 0$ for $|\vec{x}| \to \infty$.

So $\tilde{\mathcal{A}}$ satisfy the two conditions which are needed for it to be the pullback of a one-form $\mathcal{A}$ on $\mathcal{Q}$. The relation between $\mathcal{A}$ and Λ is again, locally, $\mathcal{A} = d\Lambda = \frac{1}{i} e^{-i\Lambda} de^{i\Lambda}$. The polydromy of Λ on the loop ℓ which generates $\pi_1(\mathcal{Q})$ is $\oint_\ell \mathcal{A} = \int_{\tilde{\ell}} \tilde{\mathcal{A}} = \tilde{\Lambda}(A^g) - \tilde{\Lambda}(A) = \theta$. So we come again to the conclusion that there is a $U(1)$'s worth of quantum Yang-Mills theories, parameterized by the angle $0 \le \theta < 2\pi$.

Before closing this section we note for future reference the following interpretation of the Gauss law of the theory. Let $G_\epsilon = \int d^3x \, \epsilon^a G_a$. If the theory is quantized before eliminating all unphysical degrees of freedom, the wave functions are complex functionals on $\mathcal{C}$ and Gauss' law has to be imposed as a constraint on the physical states: $G_\epsilon \psi_{\text{phys}} = 0$ for all ϵ. Upon using the quantization rule $P_i^a = -i \frac{\delta}{\delta A_i^a}$, we find

$$\begin{aligned} G_\epsilon \psi =& e^2 \int d^3x \, \epsilon^a D_i \left(P_i^a - \frac{\theta}{8\pi^2} B_i^a \right) \psi \\ =& ie^2 \int d^3x \, D_i \epsilon^a \left(\frac{\delta \psi}{\delta A_i^a} - i \frac{\theta}{8\pi^2} B_i^a \psi \right) \\ =& ie^2 \left(v_\epsilon \psi + i \tilde{\mathcal{A}}(v_\epsilon) \psi \right) = ie^2 v_\epsilon \psi \ , \end{aligned} \tag{2.5.9}$$

† If A is a fixed point for a gauge transformation g, we have $g^{-1}Ag + g^{-1}dg = A$. Thus g satisfies the equation $dg + [A, g] = 0$. Together with the boundary condition that $g(\infty) = 1$, this implies $g = 1$.

Therefore, Gauss' law states that the physical wave functions are precisely those which are locally constant along the orbits. Since the orbits are not connected, they need not be globally constant, as the preceding discussion shows.

2.6. Path integrals and instantons

We have discussed various examples of physical systems with a multiply connected configuration space $\mathcal{Q}$ and in each case we have described the appearance of a "magnetic potential" $\mathcal{A}$ on $\mathcal{Q}$; inequivalent quantizations correspond to gauge inequivalent magnetic potentials, and these were labelled by an angle $0 \leq \theta < 2\pi$. No approximation was used. The result of the existence of inequivalent quantum theories ("theta sectors") is therefore exact. Now we would like to find the ground state in each theta sector and to see whether the vacuum expectation value of the energy and other observables depends on θ or not. We will do this using the method of path integrals, which can be easily generalized from quantum mechanics to quantum field theory.

Our first task will be to understand the appearance of theta sectors from the path integral point of view. Recall that for a system with configuration space $\mathcal{Q}$ and action $S_0(q)$, the transition amplitude to go from position q_1 at the time t_1 to position q_2 at the time t_2 can be written as

$$K(q_2, t_2 \,|\, q_1, t_1) = \int_{q_1, t_1}^{q_2, t_2} (dq) e^{\frac{i}{\hbar} S_0(q)} \ , \tag{2.6.1}$$

where the integral is performed over all paths joining q_1 to q_2. We assume that the action has the form

$$S_0 = \int dt \ \left[\frac{1}{2} m g_{ij}(q) \dot{q}^i \dot{q}^j - V(q) \right] \ . \tag{2.6.2}$$

Forgetting all that has been said in the preceding sections, suppose we want to compute the amplitude (2.6.1) in a theory with action (2.6.2). We observe that the paths from q_1 to q_2 fall into homotopy classes. Clearly there are as many homotopy classes of paths from q_1 to q_2 as there are homotopy classes of loops beginning and ending at the basepoint q_0, i.e. elements of $\pi_1(\mathcal{Q})$. However, the correspondence between homotopy classes of paths and elements of $\pi_1(\mathcal{Q})$ is not unique. To construct one such correspondence, choose two paths c_1 and c_2 joining q_0 to q_1 and q_2 respectively (fig. 11). Then we associate the homotopy class of the path $q(t)$ to the homotopy class of the loop $c_2^{-1} \cdot q \cdot c_1$. Having chosen this correspondence, we can consider the partial amplitude

$$K_\alpha(q_2, t_2 \,|\, q_1, t_1) = \int_{q_1, t_1}^{q_2, t_2} (dq)_\alpha e^{\frac{i}{\hbar} S_0(q)} \ , \tag{2.6.3}$$

where the subscript α in the measure means that the integral is performed over all paths such that $c_2^{-1} \cdot q \cdot c_1$ is in the class $\alpha \in \pi_1(\mathcal{Q})$. Since paths in different homotopy classes form disjoint sets, we can weigh differently the contribution of each homotopy class and write the total amplitude as

$$K(q_2, t_2 \,|\, q_1, t_1) = \sum_{\alpha \in \pi_1(\mathcal{Q})} \chi(\alpha) K_\alpha(q_2, t_2 \,|\, q_1, t_1) \ . \tag{2.6.4}$$

The complex weights $\chi(\alpha)$ have to be chosen so that the following requirements are satisfied:
1) the total amplitude must be independent of the choice of the paths c_1 and c_2
2) the total amplitude must satisfy the factorization property

$$K(q_2, t_2 \,|\, q_1, t_1) = \int dq \, K(q_2, t_2 \,|\, q, t) K(q, t \,|\, q_1, t_1) \tag{2.6.5}$$

for $t_1 < t < t_2$.

It can be shown that these conditions imply that $\chi \in U(1)$ and $\chi(\alpha \cdot \beta) = \chi(\alpha)\chi(\beta)$, where $\alpha \cdot \beta$ is the product in the fundamental group defined in Appendix A. Thus χ has to be a character of $\pi_1(\mathcal{Q})$. Each choice of χ defines an inequivalent quantum theory, so we have reached again the conclusion that inequivalent quantizations are labelled by $\mathrm{Hom}\big(\pi_1(\mathcal{Q}), U(1)\big)$.

How is this related to the previous discussion based on the addition of topological terms to the action? As mentioned in Section 3.1 there is a one-to-one correspondence between the characters of $\pi_1(\mathcal{Q})$ and the gauge equivalence classes of flat $U(1)$ connections on $\mathcal{Q}$. The correspondence can be described as follows: if $\mathcal{A}$ is a flat connection, the corresponding character is given by

$$\chi(\alpha) = e^{\frac{ie}{\hbar}\oint_\ell \mathcal{A}} \ , \tag{2.6.6}$$

where ℓ is a loop in the homotopy class α (χ depends only on the homotopy class of ℓ since $\mathcal{A}$ is flat). There follows that if we define a "topological term"

$$S_T = e \int_{t_1}^{t_2} dt\, \dot{q}^i \mathcal{A}_i \ ,$$

this term has the same value for all curves joining the point q_1 at the time t_1 to the point q_2 at the time t_2 and such that $c_2^{-1} \cdot q \cdot c_1$ is in a fixed homotopy class α. Thus we can absorb this term in the functional integral and write:

$$\sum_\alpha \chi(\alpha) K_\alpha(q_2, t_2; q_1, t_1) = \sum_\alpha \int_{q_1,t_1}^{q_2,t_2} (dq)_\alpha e^{\frac{i}{\hbar}(S_0+S_T)} = \int_{q_1,t_1}^{q_2,t_2} (dq) e^{\frac{i}{\hbar}(S_0+S_T)} \ . \tag{2.6.7}$$

So the effect of performing the functional integral with the action S_0 and weighting the partial amplitudes with characters of $\pi_1(\mathcal{Q})$ is exactly the same as performing the functional integral with the action $S_0 + S_T$.

Our main aim in the following section will be to compute the θ-dependence of the ground state energy (or other gauge invariant observables). This can be extracted from the vacuum-to-vacuum amplitude using the following trick. To perform the functional integral one has first to perform a Wick rotation to imaginary time $\tau = it$. Defining the euclidean action

$$S_{0E} = -iS_0(t = -i\tau) = \int d\tau \left[\frac{1}{2} m g_{ij} \left(\frac{dq^i}{d\tau}\right)\left(\frac{dq^j}{d\tau}\right) + V(q)\right] \ , \tag{2.6.8}$$

the euclidean amplitude is

$$K_E(q_2, \tau_2 \,|\, q_1, \tau_1) = \int_{q_1,\tau_1}^{q_2,\tau_2} (dq) e^{-\frac{1}{\hbar} S_{0E}(q)} \ . \tag{2.6.9}$$

On the other hand, denoting $\hat{H}$ the hamiltonian, the (euclidean) evolution operator is $e^{-\frac{i}{\hbar}\hat{H}t} = e^{-\frac{1}{\hbar}\hat{H}\tau}$, and

$$K_E(q_2, \tau_2 \,|\, q_1, \tau_1) = \langle q_2 | e^{-\frac{1}{\hbar}\hat{H}(\tau_2 - \tau_1)} | q_1 \rangle = \sum_n \langle q_2 | E_n \rangle \langle E_n | q_1 \rangle e^{-\frac{1}{\hbar} E_n (\tau_2 - \tau_1)} \ , \tag{2.6.10}$$

where $|q_i\rangle$ are position eigenstates and $\{|E_n\rangle\}$ is a complete set of eigenstates of the hamiltonian with eigenvalues E_n. For $\tau_2 - \tau_1 \to \infty$ the lowest energy eigenstate dominates the sum. Putting $\tau_1 = -T/2$ and $\tau_2 = T/2$ we find

$$\lim_{T\to\infty} \langle q_2 | E_0 \rangle \langle E_0 | q_1 \rangle e^{-\frac{1}{\hbar} E_0 T} = \lim_{T\to\infty} \int_{q_1,-T/2}^{q_2,T/2} (dq) e^{-\frac{1}{\hbar} S_E(q)} \ . \tag{2.6.11}$$

In this way one can compute the lowest energy eigenvalue E_0 performing the functional integral on the right hand side and isolating the T dependence. For example in the case of a harmonic oscillator, with $m = 1$ and $V(q) = \frac{1}{2}\omega^2 q^2$, the vacuum to vacuum amplitude turns out to be equal to

$$\left(\frac{\omega}{\pi\hbar}\right)^{1/2} e^{-\frac{\omega T}{2}} \ . \tag{2.6.12}$$

Comparing with (2.6.11) one finds the ground state energy $E_0 = \frac{1}{2}\hbar\omega$. In the next Section we will compute the effect of multiconnectedness.

2.7. The dilute instanton gas approximation

We begin by discussing the example of the pendulum $\mathcal{Q}=S^1$ as presented in section 2.2. As a preliminary, we observe that this problem is very similar to that of a particle in a periodic potential. This problem is well-known in solid-state physics. In a "zeroth-order" approximation one would expand the potential around a minimum $2\pi n$ and the lowest energy eigenfunction, with energy $E_0=\frac{1}{2}\hbar\omega$, would be the one of the harmonic oscillator centered around $2\pi n$. There would be one such eigenfunction for each minimum, so the ground state would consist of infinitely degenerate states with energy $\frac{1}{2}\hbar\omega$. However, this approximation neglects tunnelling between neighbouring minima. When taken into account, this breaks the degeneracy and one gets a continuous band of states whose energy depends on the parameter θ (see Exercise 2.7.1). However, there is an important physical difference. Although the pendulum and the particle in the periodic potential have the same classical Lagrangian, they are different because in the former case all points on the line are identified mod 2π, whereas in the latter they are not. We will see that this will lead to a different physical interpretation of the results.

We are going to study the vacuum energy of the pendulum as a function of θ $\big($where θ is the coefficient which appears in (2.2.2)$\big)$ using the method of path integrals. We begin by observing that the classical "vacuum state" of the pendulum is $\varphi=0 \bmod 2\pi$ (independent of time). The vacuum-to-vacuum transition amplitude is

$$K(0\, mod\, 2\pi, T/2\,|\,0,-T/2) = \sum_{n=-\infty}^{\infty} e^{in\theta} K_n(2\pi n, T/2\,|\,0,-T/2)\ , \tag{2.7.1}$$

where $n\in\mathbf{Z}=\pi_1(S^1)$ labels the homotopy class of the loop $\varphi(t)$ and we assume without loss of generality that $\varphi = 0$ for $T\to-\infty$. The partial amplitudes are computed here with the action S_0 corresponding to the Lagrangian (2.2.1). Since K_n is a path integral over loops in a fixed homotopy class, we can bring the character inside the path integral and write

$$\begin{aligned} e^{in\theta}K_n(2\pi n, T/2\,|\,0,-T/2) &= \int_{0,-T/2}^{2\pi n, T/2} (d\varphi)_n e^{\frac{i}{\hbar}S_0 + i\theta n} \\ &= \int_{0,-T/2}^{2\pi n, T/2} (d\varphi)_n e^{\frac{i}{\hbar}(S_0+\theta\hbar W)} = \tilde{K}_n(2\pi n, T/2\,|\,0,-T/2)\ . \end{aligned} \tag{2.7.2}$$

In this way the topological term introduced in section 3.2 has made its reappearance.

Next we perform the Wick rotation. Due to the fact that it contains only one time derivative, the euclidean topological term is imaginary:

$$S_{T,E} = -iS_T(t=-i\tau) = -i\theta\hbar W\ . \tag{2.7.3}$$

The euclidean amplitude is

$$\tilde{K}_E(0 \bmod 2\pi, T/2\,|\,0,-T/2) = \sum_n \tilde{K}_{E,n}(2\pi n, T/2\,|\,0,-T/2)\ , \tag{2.7.4}$$

$$\tilde{K}_{E,n}(2\pi n, T/2\,|\,0,-T/2) = \int_{0,-T/2}^{2\pi n, T/2} (d\varphi)_n\, e^{-\frac{1}{\hbar}S_E}\ , \tag{2.7.5}$$

where $S_E = S_{0E} + S_T = S_{0E} - i\hbar\theta W$. Note that the topological term retains oscillatory character also in the Euclidean regime. The partial amplitudes can be evaluated using the WKB, or saddle point approximation: we will now compute the contribution of fields which are near a stationary point of the euclidean action.

Let us begin by evaluating certain contributions to $\tilde{K}_{E1}(2\pi, T/2\,|\,0, -T/2)$, i.e. the sum over paths with winding number one. Such paths can be regarded as maps from the interval $[-T/2, T/2]$ to $\mathbf{R}$, and with the given boundary conditions they form a functional space which is diffeomorphic to the space $\mathcal{Q}_{01}$ that we considered in section 2.3 for the sine-Gordon theory. Furthermore, the euclidean action S_{0E} for the pendulum is equal to the static energy for the sine-Gordon theory. Therefore the absolute minimum of S_{0E} is the trajectory given by (1.1.11), with ϕ and x replaced by φ and τ. Within the context of the present theory, a trajectory of this type is called an *instanton* with topological number one centered around τ_0 (the reason for the name is that for most of the time the trajectory lies near the classical vacuum $\varphi = 0 \bmod 2\pi$, except for a brief "instant" of time around τ_0).

Next we expand the action around this particular classical trajectory $\varphi_{\rm cl}(\tau)$. We get

$$S_E(\varphi) = S_E(\varphi_{\rm cl}) + \frac{1}{2}\int d\tau d\tau' \eta(\tau)\mathcal{O}(\tau,\tau')\eta(\tau') \ , \tag{2.7.6}$$

where $\eta = \varphi - \varphi_{\rm cl}$ and

$$\mathcal{O}(\tau,\tau') = \left.\frac{\delta^2 S_E}{\delta\varphi(\tau)\delta\varphi(\tau')}\right|_{\varphi_{\rm cl}} = \delta(\tau-\tau')\left(-\frac{d^2}{d\tau^2} + V''(\varphi_{\rm cl})\right) \ . \tag{2.7.7}$$

In the WKB approximation

$$K_{E1}(2\pi, T/2\,|\,0,-T/2) = e^{-\frac{1}{\hbar}S_E(\varphi_{\rm cl})}\int (d\eta)\, e^{-\frac{1}{2}\int \eta\mathcal{O}\eta} = e^{-\frac{1}{\hbar}S_E(\varphi_{\rm cl})} B(T)\,[{\rm Det}\mathcal{O}]^{-1/2} \tag{2.7.8}$$

where $B(T)$ is a measure factor .

The operator $-\frac{d^2}{dt^2} + V''(\varphi_{\rm cl})$ has a translational zero mode, corresponding to the fact that the position of the instanton is arbitrary (see Exercise 2.7.2). The integration on the zero mode is replaced by an integration on the "collective coordinate" τ_0, with some jacobian factor J. So (2.7.8) can be rewritten

$$e^{-\frac{1}{\hbar}S_E(\varphi_{\rm cl})} B(T) J T \left[{\rm Det}'\mathcal{O}\right]^{-1/2} \ , \tag{2.7.9}$$

where Det$'$ is the product of the nonzero eigenvalues. The evaluation of the determinant is difficult because $\varphi_{\rm cl}$, which appears in the operator (2.7.7) depends explicitly on time. However, the size of the instanton was fixed by the form of the potential and is independent of T, so if we are only interested in the limit of large T, we see that "most of the time" $\varphi_{\rm cl} = 0 \bmod 2\pi$ and therefore $V''(\varphi_{\rm cl}) = \omega^2$. We can write

$$\left[{\rm Det}'\mathcal{O}\right]^{-1/2} = K\left[{\rm Det}\left(-\frac{d^2}{dt^2} + \omega^2\right)\right]^{-1/2} \tag{2.7.10}$$

where K, the ratio of the determinants, becomes a constant independent of T for large T. The determinant on the r.h.s. together with the factor $B(T)$ is the partition function of a harmonic oscillator, which is given by (2.6.12). We thus find

$$K_{E,1}(2\pi, T/2\,|\,0,-T/2) = e^{-\frac{1}{\hbar}S_{0E}+i\theta} K J T e^{-\frac{\omega T}{2}}\left(\frac{\omega}{\pi\hbar}\right)^{1/2} \tag{2.7.11}$$

where we have written $S_E(\varphi_{\rm cl}) = S_{0E}(\varphi_{\rm cl}) - i\theta\hbar W(\varphi_{\rm cl}) = S_{0E} - i\theta\hbar$. This is the contribution of the one-instanton sector to the total amplitude. The one anti-instanton sector gives

$$K_{E,-1}(-2\pi, T/2\,|\,0,-T/2) = e^{-\frac{1}{\hbar}S_{0E}-i\theta} K J T e^{-\frac{\omega T}{2}}\left(\frac{\omega}{\pi\hbar}\right)^{1/2} \tag{2.7.12}$$

In principle we should now evaluate the contributions of paths with higher winding numbers and then sum over the winding numbers. However, we have already observed in Exercise 1.1.1 that there are no classical solutions to the equation $-\frac{d^2\varphi}{dt^2} + \frac{dV}{d\varphi} = 0$ in the sectors $\mathcal{Q}_{0i}$ with $|i| > 1$, i.e. solutions interpolating between

nonadjacent minima. This means that there are no exact multi-instanton solutions around which to expand the action. Thus, we cannot directly apply the WKB method to compute the contribution of paths with winding number greater than one. In practice the calculation can still be done, but in a different way.

We observe that a configuration consisting of m_1 instantons and m_2 anti-instantons, all widely separated, will provide an approximate solution to the classical equation of motion with $W = m_1 - m_2$. Such a configuration will contribute to the partial amplitude $\tilde{K}_E\big((m_1 - m_2)2\pi, T/2 \,|\, 0, -T/2\big)$. The evaluation of the functional integral for this case proceeds much as in the one-instanton case, with the following changes: every instanton gives a contribution to $S_E(\varphi_{\rm cl})$ equal to $S_{0E} - i\theta\hbar$ and each anti-instanton gives a contribution $S_{0E} + i\theta\hbar$; every instanton and anti-instanton has a translational zero mode contributing a factor TJ; as long as they are widely separated, every instanton and anti-instanton contributes a factor K when $V''(\varphi_{\rm cl})$ is replaced by ω^2 in the determinant. Altogether the contribution to the total amplitude due to configurations containing m_1 instantons and m_2 anti-instantons is

$$\frac{1}{m_1! m_2!} \exp\left[-\frac{1}{\hbar}(m_1 + m_2)S_{0E} + i(m_1 - m_2)\theta\right] (KJT)^{m_1+m_2} \left(\frac{\omega}{\pi\hbar}\right)^{1/2} e^{-\frac{\omega T}{2}} \tag{2.7.13}$$

The factor $\frac{1}{m_1! m_2!}$ is due to the indistinguishability of the instantons and anti-instantons (in the integral over the collective coordinates, the situation when instanton 1 is in position τ_1 and instanton 2 is in position τ_2 is physically the same as when instanton 1 is in position τ_2 and instanton 2 is in position τ_1). The total amplitude is obtained by summing over m_1 and m_2. This automatically includes a sum over winding numbers. The sums can be performed explicitly and we get

$$\begin{aligned} \tilde{K}(0 \bmod 2\pi, T/2 \,|\, 0, -T/2) &= \exp\left(KJTe^{-\frac{1}{\hbar}S_{0E} + i\theta}\right) \exp\left(KJTe^{-\frac{1}{\hbar}S_{0E} - i\theta}\right) \left(\frac{\omega}{\pi\hbar}\right)^{1/2} e^{-\frac{\omega T}{2}} \\ &= \left(\frac{\omega}{\pi\hbar}\right)^{1/2} \exp\left[-\frac{1}{\hbar}T\left(\frac{1}{2}\hbar\omega - 2\hbar KJe^{-\frac{1}{\hbar}S_{0E}} \cos\theta\right)\right] \ . \end{aligned} \tag{2.7.14}$$

Comparing with (2.6.11) we find that the energy of the vacuum in the presence of the θ-term in the action is

$$E_\theta = \frac{1}{2}\hbar\omega - 2\hbar KJe^{-\frac{1}{\hbar}S_{0E}} \cos\theta \ . \tag{2.7.15}$$

This is much the same result that one obtains for a particle in a periodic potential but with an important difference: there all states in the band belong to the same Hilbert space and therefore transitions between states with different values of θ are permitted. Here every value of θ defines a different theory and no transition between different θ-states can occur.

This way of computing the functional integral for a theory with multiply connected $\mathcal{Q}$ is known as the dilute instanton gas approximation. We will see that it can be easily generalized to the case of fields theories.

The preceding discussion shows that instantons are related to tunnelling. The static energy E_S, as a function on $\mathcal{Q}$, has its minimum at some point φ_0 that we call the classical vacuum. We take φ_0 as "basepoint" in $\mathcal{Q}$. Without loss of generality we can assume the vacuum energy to be zero; elsewhere it is positive. Therefore if we consider a non-contractible loop in $\mathcal{Q}_0$ parameterized by $-\infty \le t \le \infty$, we see that the energy as a function of t has the shape shown in fig. 12. If the system is in a low energy state, it cannot classically follow such a trajectory, but it can do it in the quantum theory by tunnelling. In the WKB approximation the tunnelling amplitude is evaluated as a sum over trajectories which are near a classical solution of the equations of motion. No classical solutions exists in the real time, minkowskian section, but as we have seen, solutions exist in the imaginary time, euclidean section. Thus it is the WKB approximation that requires performing the Wick rotation. In the end the amplitude is analytically continued back to real time.

We conclude with some general remarks on instantons. In general one will call an *instanton* a classical solution of the euclidean equations of motion which is smooth, localized in spacetime has finite euclidean action and represents a history of the system that traverses a noncontractible loop in configuration space. The condition of finite action implies that for $t \to \pm\infty$ the instanton must approach classical vacuum solutions of the theory. Thus it is only for a certain interval of time (an "instant") that it differs from the vacuum solution. In fact, in field theories, instantons are localized both in time and space. Note that the euclidean

action of the theory is equal to the static energy of the same theory in one more dimension. Therefore the instanton of a theory is the same as the soliton of the same theory in one more dimension. We have seen this in the case of the pendulum (a 0+1-dimensional field theory) and the Sine-Gordon model (a 1+1-dimensional field theory). We will see further explicit examples of this fact later.

2.8. Theta vacua in the abelian Higgs model

We have shown in section 2.4. that an abelian gauge theory in 1+1 dimensions has a multiply connected configuration space and therefore has theta sectors. In order to apply the dilute instanton gas approximation to this model one has first of all to find the single instanton solution. This would be a solution of the euclidean field equations with finite euclidean action $S_E = \int dx d\tau \frac{1}{4} F_{\mu\nu} F_{\mu\nu}$. We observe that this euclidean action is equal to the energy of a static abelian gauge field in $2+1$ dimensions in the gauge $A_0 = 0$. But we have shown in section 2.6 that this functional does not have nontrivial minima. Therefore, the abelian gauge theory in 1+1 dimensions does not admit instantons and we cannot apply the WBK approximation. For this reason we are going to consider instead the abelian Higgs model in $1+1$ dimensions, with euclidean action $S_E(A,\phi) = S_{0E}(A,\phi) - i\theta c_1(A)$, where

$$S_{0E} = \int dx d\tau \left[\frac{1}{4} F_{\mu\nu} F_{\mu\nu} + \frac{1}{2} \left(D_\mu \phi \right)^* \left(D_\mu \phi \right) + \frac{\lambda}{4} \left(|\phi|^2 - f^2 \right)^2 \right] \tag{2.8.1}$$

and c_1 is given by (2.4.3). Note again that the topological term retains oscillatory character also in the euclidean path integral. In the gauge $A_0 = 0$, the configuration space is $\mathcal{Q} = (\mathcal{C} \times \Gamma)/\mathcal{G}$, where Γ is the space of scalar fields ϕ, with appropriate boundary conditions imposed on A_i and ϕ. Since Γ is a vectorspace, $\mathcal{C} \times \Gamma$ has trivial homotopy groups just like $\mathcal{C}$. Thus the topology of $\mathcal{Q}$ is unaffected by the presence of the scalars and the argument in section 3.4 proving the existence of the theta sectors remains valid.

We now have to understand better in what sense c_1 is a "topological number". We restrict our attention to spacetime fields with finite euclidean action. This demands that when $r = \sqrt{x^2 + \tau^2} \to \infty$, $A_i \to \frac{i}{e} g_\infty^{-1} dg_\infty$ and $\phi \to g_\infty^{-1} f$, where $g_\infty(\theta)$ is a map from S^1_∞ to $U(1)$. Such maps are classified by their winding number, so the fields with finite action fall into disjoint classes, characterized by different asymptotic behaviour. These classes are usually called the topological sectors. (This is the same term that we used in Chapter 1 for a theory admitting topological solitons. The mathematical similarity of the two situations should not obscure the profoundly different physical implications). We can now evaluate the quantity c_1 on such a field. Using (2.4.4) and the asymptotic form of A we get

$$c_1 = \frac{1}{4\pi} \int d^2x \, \varepsilon^{\mu\nu} F_{\mu\nu} = \frac{1}{2\pi} \int_{S^1_\infty} A = W(g_\infty) \ . \tag{2.8.2}$$

Thus c_1 is a measure of the nontriviality of the asymptotic behaviour of the fields.

The one-instanton of this theory is going to be a solution of the euclidean field equations describing the tunnelling of the system through the fundamental non-contractible loop in $\mathcal{Q}$. It follows from the discussion of section 3.4 that this loop in $\mathcal{Q}$ is the projection of a path in $\mathcal{C} \times \Gamma$ joining the classical vacuum $(A_{(0)}, \phi_{(0)}) = (0, f)$ to $(A^{g_1}_{(0)}, \phi^{g_1}_{(0)}) = (\frac{i}{e} g_1^{-1} dg_1, g_1^{-1} f)$, where $g_1 = e^{i\alpha}$ is a time-independent gauge transformation with winding number one, i.e. $\alpha(x \to -\infty) = 0$, $\alpha(x \to +\infty) = 2\pi$. Next we observe that the euclidean action (2.8.1) is equal to the static energy of the abelian Higgs model in $2+1$ dimensions. We have found in section 2.7 a stationary point of this functional with the boundary condition that when $r = \sqrt{x^2 + \tau^2} \to \infty$, $A_i \to \frac{i}{e} g_1^{-1} dg_1$ and $\phi \to f g_1$, where $g_1(\theta)$ is a map from S^1_∞ to $U(1)$ with winding number one: it was called the vortex. These are exactly the boundary conditions that we need (fig. 13). The explicit form of the solution was given in ; it can be rewritten in the gauge $A_0 = 0$ that we are using here (remember that one of the spatial coordinates of section 1.7 should be reinterpreted as Euclidean time. See exercise 2.8.1). Therefore, the vortex solution of the abelian Higgs model in $2+1$ dimensions with unit flux is the desired instanton solution of the same model in $1+1$ dimensions.

The functional integral

$$Z_\theta(T) = \int (dA\, d\phi\, d\phi^*)\, e^{-S_{0E}+i\theta c_1} = \lim_{T\to\infty} e^{-TE_\theta} \tag{2.8.3}$$

can be evaluated as in the previous section using the dilute instanton gas approximation. This is well justified since the instantons have a fixed finite size which is negligible in the limit of large T and L, where L is the spatial extension of a box in which the system is enclosed. The main novelty is that now there are two translation zero modes for each instanton and anti-instanton, so the integration over the corresponding collective coordinates yields a factor LT for each instanton and anti-instanton. Thus we find

$$\lim_{T\to\infty} Z_\theta(T) = Ae^{-LT\left(C-e^{-S_{0E}}2B\cos\theta\right)} \tag{2.8.4}$$

for some constants A,B,C, where S_{0E} denotes the action for the single instanton solution. From here one reads off the energy density

$$\frac{E_\theta}{L} = C - e^{-S_{0E}}2B\cos\theta \tag{2.8.5}$$

analogous to the result (2.7.15).

The physical meaning of the parameter θ can be further clarified by considering the vacuum expectation value of the electric field $\langle E_1\rangle_\theta = i\langle F_{01}\rangle_\theta$. Due to translational invariance

$$\langle F_{01}(x,\tau)\rangle_\theta = \frac{1}{LT}\left\langle \int dx d\tau\, F_{01}\right\rangle_\theta = \frac{1}{2LT}\left\langle \int dx d\tau\, \varepsilon^{\mu\nu}F_{\mu\nu}\right\rangle_\theta = \frac{2\pi}{LT}\langle c_1\rangle_\theta \tag{2.8.6}$$

We have

$$\langle c_1\rangle_\theta = i\frac{d}{d\theta}\ln Z_\theta = -i\frac{d}{d\theta}(E_\theta T) = -iLTe^{-S_0}2B\sin\theta\ . \tag{2.8.7}$$

Therefore

$$\langle E_1(x,\tau)\rangle_\theta = 4\pi e^{-S_0}B\sin\theta\ . \tag{2.8.8}$$

Therefore, in the theta vacuum, there is a uniform background electric field. This fact leads us to suspect the existence of long range forces, in spite of the fact that at tree level, due to the occurrence of the Higgs phenomenon, we would expect only short range forces. We will now prove that instantons lead to long range forces and confinement in this model.

Consider two (nondynamical, external) charges q and $-q$ at a fixed distance $\tilde{L}$. The potential energy between these charges is given by the difference of the energy of the system in the presence and in the absence of the charges. If the system is quasi static, these energies in turn can be evaluated as the effective actions divided by the time.

More precisely, suppose that the pair of charge and anticharge is created at some instant, brought to distance $\tilde{L}$, then left there for a large time $\tilde{T}$ and finally annihilated again. The classical contribution to the action due to the presence of the charges is

$$\int d^2x\, j^\mu A_\mu = q\oint A$$

where $j^\mu(x) = q\delta^{(2)}(x-x(t))\frac{dx^\mu}{dt}$ is the current generated by the charges. The quantity $W = e^{iq\oint A}$ is called the Wilson loop. As before, we enclose the system in a spacetime volume of sides $L \gg \tilde{L}$ and $T \gg \tilde{T}$. In the limit $T\to\infty$ the Euclidean functional integral gives the exponential of the energy in the presence of the charges:

$$\int (dAd\phi d\phi^*)e^{-S_{0E}+i\theta c_1}W = \exp(-TE_\theta - \tilde{T}\Delta E_\theta(\tilde{L}))\ . \tag{2.8.9}$$

We have then

$$\lim_{\tilde{T}\to\infty}\langle W\rangle = \frac{1}{Z_\theta(T)}\int (dAd\phi d\phi^*)e^{-S_{0E}+i\theta c_1 - iq\oint A} = e^{-\tilde{T}\Delta E_\theta(\tilde{L})}$$

Therefore we can compute the interaction between the charges from the Wilson loop:

$$\Delta E_\theta(\tilde{L}) = -\lim_{\tilde{T}\to\infty} \frac{1}{\tilde{T}} \ln\langle W\rangle_\theta \tag{2.8.10}$$

We are going to perform a dilute instanton gas approximation. It is convenient to write $n_\pm = n_\pm^{(\mathrm{in})} + n_\pm^{(\mathrm{out})}$, where we count separately instantons and anti-instantons that lie inside or outside the spacetime loop traced by the charges. The reason for this is that the Wilson loop can be rewritten:

$$W = e^{iq\oint A} = e^{\frac{iq}{2}\int_U d^2x\epsilon^{\mu\nu}F_{\mu\nu}} = e^{2\pi iq(n_+^{(\mathrm{in})}-n_-^{(\mathrm{in})})}$$

where U is the region enclosed by the loop. Then, the functional integral (2.8.9) can be evaluated as follows:

$$\begin{aligned} Ae^{-LTC} \sum_{n_+^{(\mathrm{in})},n_-^{(\mathrm{in})},n_+^{(\mathrm{out})},n_-^{(\mathrm{out})}} & \frac{1}{n_+^{(\mathrm{in})}!n_-^{(\mathrm{in})}!n_+^{(\mathrm{out})}!n_-^{(\mathrm{out})}!} \\ & \times \exp(-(n_+^{(\mathrm{in})}+n_-^{(\mathrm{in})}+n_+^{(\mathrm{out})}+n_-^{(\mathrm{out})})S_{0E} + i\theta(n_+^{(\mathrm{in})}-n_-^{(\mathrm{in})}+n_+^{(\mathrm{out})}-n_-^{(\mathrm{out})})) \\ & \times (B(LT-\tilde{L}\tilde{T}))^{n_+^{(\mathrm{out})}+n_-^{(\mathrm{out})}} (B\tilde{L}\tilde{T})^{n_+^{(\mathrm{in})}+n_-^{(\mathrm{in})}} \exp(2\pi iq(n_+^{(\mathrm{in})}-n_-^{(\mathrm{in})})) \\ = & A\exp\Big(-LTC + 2Be^{-S_{0E}}[\tilde{L}\tilde{T}\cos(\theta+2\pi q) + (LT-\tilde{L}\tilde{T})\cos\theta]\Big) \end{aligned} \tag{2.8.11}$$

Using (2.8.4) and (2.8.11) in (2.8.10) we get

$$\Delta E_\theta(\tilde{L}) = 2Be^{-S_{0E}}\tilde{L}(\cos\theta - \cos(\theta+2\pi q)) \ . \tag{2.8.12}$$

From this formula we see that the potential grows with distance, leading to confinement of the charges. From the factor $e^{-S_{0E}}$ we also see that the result is strictly nonperturbative (the numerator contains a hidden factor $1/\hbar$) and that the force vanishes exponentially in the classical limit. Finally, if the charges are integer, the force vanishes. This is interpreted as a complete screening due to intervening pairs of quanta of the scalar field. If the external charges are not integers, the scalar particle-antiparticle pairs cannot completely screen the electric field, leaving a residual force which is independent of distance.

2.9. The BPST instanton

We have proven in section 3.5 that a pure Yang-Mills theory has a multiply connected configurations space and hence theta sectors. Again, to compute the dependence of the ground state energy from the parameter θ, we need to know explicitly a time-dependent solution of the Euclidean equations of motion, describing the motion of the system through the fundamental, non-contractible loop in $\mathcal{Q}$. Such a solution will be called a Yang-Mills instanton. We observe that the Euclidean Yang-Mills action $S_E = \frac{1}{4}\int d^4x F^a_{\mu\nu}F^{\mu\nu a}$ is equal to the static energy of a pure Yang-Mills theory in 4+1 dimensions in gauge $A_0 = 0$. The solution we are looking for could therefore also be regarded as a static soliton in five dimensions. The scaling argument of section 2.6 did not rule out the existence of such solutions. We are now going to derive it explicitly.

We begin by giving a topological classification of four dimensional Yang-Mills fields. From the fact that the time evolution traces a continuous curve in $\mathcal{Q}$ and from the multiple connectedness of $\mathcal{Q}$, there follows that four-dimensional Yang-Mills fields must fall into disjoint classes labelled by the integers. These classes can be described more explicitely as follows. We impose that $A^a_\mu(\vec{x},\tau)$ has finite Euclidean action. This requires that at spacetime infinity, i.e. for $|x| = \sqrt{|\vec{x}|^2+\tau^2} \to \infty$, $F^a_{\mu\nu} \to 0$. This in turn implies

$$A_\mu \to g_\infty^{-1}\partial_\mu g_\infty \ , \tag{2.9.1}$$

where g_∞ is a function of the angles or equivalently a function from the sphere at infinity S^3_∞ to the gauge group $SU(2)$. Since $\pi_3\big(SU(2)\big) = \mathbf{Z}$, we find that the finite action gauge potentials A^a_μ fall into topologically

distinct classes distinguished by their asymptotic behaviour. The topological invariant c_2 precisely measures these classes. In fact using (2.5.2) we can write

$$c_2 = \int\limits_{\mathbf{R}^4} d^4x\, \partial_\mu C^\mu = \frac{1}{16\pi^2} \int_{S^3_\infty} d^3x\, \varepsilon^{ijk} \left(A^a_i \partial_j A^a_k + \frac{1}{3} f_{abc} A^a_i A^b_j A^c_k \right) = W(g_\infty) \ . \tag{2.9.2}$$

The last equality is obtained by noting that on S^3_∞ we can replace A^a_i by its asymptotic form (2.9.1); the result then follows from Exercise 2.5.2.

This calculation gives the precise meaning of the statement that c_2 is a topological invariant. An even more sophisticated understanding of the topology of Yang–Mills fields requires the use of fiber bundles and characteristic classes.

The instanton has to represent the motion of the system through the fundamental noncontractible loop in $\mathcal{Q}$. We recall from section 2.5 that in terms of the spatial gauge potential A_i this means a path joining, for example $A_i = 0$ to $A_i = g^{-1}\partial_i g$, where g is a time-independent gauge transformation with winding number one. We take the parameter on this path to be euclidean time τ. In this way the instanton will be represented by a gauge potential $A_\mu(x,\tau)$ with $A_\tau = 0$ and A_i a pure gauge both at spatial and temporal infinity. In fact the gauge function g_∞ for this field will be one everywhere except for $\tau \to \infty$, where it coincides with g_1. Such a field will therefore have $c_2 = 1$.

To find the explicit form of the instanton we consider the inequality

$$0 \le \int d^4x \, \left(F^a_{\mu\nu} \pm {}^*F^a_{\mu\nu} \right) \left(F^{\mu\nu a} \pm {}^*F^{\mu\nu a} \right) = 2 \int d^4x \, F^a_{\mu\nu} F^{\mu\nu a} \pm 2 \int d^4x \, F^a_{\mu\nu} {}^*F^{\mu\nu a} \ , \tag{2.9.3}$$

which implies

$$S_E \ge \frac{8\pi^2}{e^2} |c_2| \ . \tag{2.9.4}$$

The absolute minima of the action in each sector are the gauge fields for which F is either self-dual or anti-self-dual

$$F^a_{\mu\nu} = \pm {}^*F^a_{\mu\nu} \ . \tag{2.9.5}$$

These fields are automatically solutions of the Yang-Mills equations. So we have succeeded in replacing the second order Yang-Mills equation by the simpler first order equations (2.9.5).

To find the explicit form of the instanton in the sector $c_2 = \pm 1$ we make an ansatz of the form

$$A_\mu(x) = f(r^2) g^{-1} \partial_\mu g \tag{2.9.6}$$

Here g, a function of the angles only, has the following explicit representation:

$$g(x) = \begin{pmatrix} \hat{x}^4 + i\hat{x}^3 & \hat{x}^2 + i\hat{x}^1 \\ -\hat{x}^2 + i\hat{x}^2 & \hat{x}_4 - i\hat{x}^3 \end{pmatrix} = \hat{x}^\mu \tau_\mu \ ,$$

where $\tau_k = i\sigma_k$ for $k = 1,2,3$ and $\tau_4 = \mathbf{1}$. This function clearly has $W(g_\infty) = 1$. Note that we can write

$$g(x)^{-1} = g(x)^\dagger = \hat{x}^\mu \bar{\tau}_\mu \ ,$$

where $\bar{\tau}_k = -i\sigma_k$ for $k = 1,2,3$ and $\tau_4 = \mathbf{1}$. From here one finds

$$g^{-1}\partial_\mu g = -2i\bar{\Sigma}_{\mu\rho} \frac{\hat{x}^\rho}{r} \ ; \qquad g \partial_\mu g^{-1} = -2i \Sigma_{\mu\rho} \frac{\hat{x}^\rho}{r} \ ,$$

where

$$\Sigma_{\mu\nu} = \frac{1}{2} \begin{pmatrix} 0 & \sigma_3 & -\sigma_2 & \sigma_1 \\ -\sigma_3 & 0 & \sigma_1 & \sigma_2 \\ \sigma_2 & -\sigma_1 & 0 & \sigma_3 \\ -\sigma_1 & -\sigma_2 & -\sigma_3 & 0 \end{pmatrix} \ ; \qquad \bar{\Sigma}_{\mu\nu} = \frac{1}{2} \begin{pmatrix} 0 & \sigma_3 & -\sigma_2 & -\sigma_1 \\ -\sigma_3 & 0 & \sigma_1 & -\sigma_2 \\ \sigma_2 & -\sigma_1 & 0 & -\sigma_3 \\ \sigma_1 & \sigma_2 & \sigma_3 & 0 \end{pmatrix} \ .$$

These matrix-valued tensors are self-dual and anti-self-dual respectively.

The function f in (2.9.6) must satisfy $f(r^2) \to 1$ for $r^2 \to \infty$ and $f(0) = 0$ to avoid singularities in A (the form $g^{-1}dg$ is ill-defined in the origin). In order to determine the function f we compute the curvature of (2.9.6)

$$F_{\mu\nu} = 4i \left(\bar{\Sigma}_{\mu\rho}\hat{x}^\rho\hat{x}^\nu - \bar{\Sigma}_{\nu\rho}\hat{x}^\rho\hat{x}^\mu\right)\left(f' + \frac{1}{r^2}f(f-1)\right) - 41\bar{\Sigma}_{\mu\nu}\frac{1}{r^2}f(f-1) \ .$$

Here f' denotes the derivative of f with respect to r^2. In order to compute the dual we use

$$\varepsilon_{\mu\nu\alpha\beta}\bar{\Sigma}_{\rho\beta} = -\delta_{\mu\rho}\bar{\Sigma}_{\nu\alpha} + \delta_{\nu\rho}\bar{\Sigma}_{\mu\alpha} - \delta_{\alpha\rho}\bar{\Sigma}_{\mu\nu}$$

and find

$${}^*F_{\mu\nu} = 4i \left(\bar{\Sigma}_{\mu\rho}\hat{x}^\rho\hat{x}^\nu - \bar{\Sigma}_{\nu\rho}\hat{x}^\rho\hat{x}^\mu\right)\left(f' + \frac{1}{r^2}f(f-1)\right) - 4i\bar{\Sigma}_{\mu\nu}f' \ .$$

The anti-self-duality equation $0 = F_{\mu\nu} +^* F_{\mu\nu}$ implies $f' + \frac{1}{r^2}f(f-1) = 0$, which it is solved by

$$f(r^2) = \frac{r^2}{\lambda^2 + r^2}$$

where λ is an arbitrary constant. A similar procedure starting from the map $\hat{x}^\mu\bar{\tau}_\mu$ with winding number -1, and solving the self-duality equation $0 = F_{\mu\nu} -^* F_{\mu\nu}$ leads to the same function f. Altogether the regular unit instanton and anti-instanton solutions can be written in the form

$$A_\mu = -2i\frac{\bar{\Sigma}_{\mu\nu}(x-x_0)^\nu}{\lambda^2 + (x-x_0)^2} \ ; \qquad A_\mu = -2i\frac{\Sigma_{\mu\nu}(x-x_0)^\nu}{\lambda^2 + (x-x_0)^2} \tag{2.9.7}$$

where we have allowed the center of the solutions to be located at an arbitrary position x_0. The respective field strengths are

$$F_{\mu\nu} = -4i\frac{\bar{\Sigma}_{\mu\nu}\lambda^2}{(\lambda^2 + (x-x_0)^2)^2} \ ; \qquad F_{\mu\nu} = -4i\frac{\Sigma_{\mu\nu}\lambda^2}{(\lambda^2 + (x-x_0)^2)^2} \ . \tag{2.9.8}$$

Note that the solution depends of 5 free parameters: the location of the center x_0, reflecting translation invariance of the action, and the scale λ reflecting the conformal invariance of the action.

Note also that if we impose self-duality on the $W = -1$ configuration or anti-self-duality on the $W = 1$ configuration we are led to the equation $f' - \frac{1}{r^2}f(f-1) = 0$, which it is solved by

$$f(r^2) = \frac{\lambda^2}{\lambda^2 + r^2} \ .$$

This gives rise to solutions that are singular in the origin. They are related to the regular solutions given above by a singular gauge transformation with parameter g.

Much work has been done to find all self-dual and anti-self-dual solutions with $|c_2| > 1$. This line of research has led to important developments in mathematics, such as Donaldson theory. While mathematically of the greatest interest, this work has not had much impact in the present context since these exact solutions would give a negligible contribution to the path integral compared to the approximate multi-instanton solutions which we use in the dilute instanton gas approximation.

One can proceed again to evaluate the vacuum-to-vacuum amplitude using the method of the dilute instanton gas approximation. This time one can anticipate problems on the basis of the fact that, unlike the case of the theories discussed in the previous two sections, here the size of the instantons is undetermined. Clearly large instantons can invalidate the hypothesis that the instantons are widely separated. In fact, in the functional integral one now has to integrate over five collective coordinates x_0 and λ for each instanton and anti-instanton. While the primed determinant of small fluctuations around a given instanton is independent

Exercise. Work out explicitly the $d = 2$ case. Write explicit formulae when the coordinates on S^3 are given by the Euler angles.

of x_0, so that the integration over x_0 only gives a volume factor $VT = \int d^4x_0$, it does depend on the scale λ. Therefore, the result of the functional integration in the dilute instanton gas approximation has the form

$$\frac{E_\theta}{V} = C - 2e^{-S_0}\cos\theta \int_0^\infty d\lambda\, B(\lambda)\ . \tag{2.9.9}$$

The λ integration is known to be convergent at the lower end, but is not well under control at the upper end (large instantons). So, while formally the energy density still varies with θ in the by now familiar way, its coefficient may be infrared divergent. For recent results on this issue see XXX.

Exercise. Derive equation (2.4.6) *from the constraint Hamiltonian analysis of electromagnetism in the presence of the topological term.*

Exercise. Prove conditions 1 and 2 above.

Exercise 2.5.1. Prove that $\frac{\delta C^0}{\delta A_i^a(x)} = \frac{1}{8\pi^2} B_i^a(x)$.

Exercise 2.5.2. Prove equation (2.5.8).

Exercise. Compute the functional integral leading to Eq. (2.6.12).

Exercise. Prove that if χ is a homomorphism from $\pi_1(\mathcal{Q})$ to $U(1)$, then the amplitude (2.6.4) *satisfies the conditions (1) and (2) given above. See FORTE.*

Exercise 2.8.1. Write the vortex field in the gauge $A_y = 0$