Daniel Harlow
Lecture notes for Physics 8.325: Relativistic Quantum field theory III, Spring 2026

Contents

1 Introduction and summary

Let's quickly review where we are so far.[1] In the previous two semesters we did the following:

- We motivated quantum field theory from two points of view:
  - It allows us to combine special relativity and quantum mechanics without violating causality.
  - It arises as the universal long-distance description of many-body quantum systems with local interactions due to the focusing behavior of the renormalization group.
- We showed that particles which arise from relativistic quantum field theories always obey the following rules:
  - Charged particles have antiparticles of the same mass and opposite charge.
  - The number of particles is never conserved unless there are no interactions.
  - Particles of integer/half-integer spin must be bosons/fermions respectively.
  - The $\mathcal{CRT}$ operation exchanging particles with antiparticles and reflecting space and time is always a symmetry.
- We introduced free and interacting field theories constructed from real and complex scalars, Dirac and Majorana spinors, and one-form gauge fields, and we learned how to compute correlation functions and scattering amplitudes in these theories perturbatively using Feynman diagrams. In particular we constructed the (true) Yukawa theory

$$\mathcal{L} = -\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - \frac{m_\phi^2}{2}\phi^2 - i\overline{\psi}(\not{\partial} + m)\psi - g\phi\overline{\psi}\gamma\psi \tag{1.1}$$

  of spinor nucleons interacting with pseudoscalar pions and the spinor electrodynamics

$$\mathcal{L} = -\frac{1}{4}f_{\mu\nu}f^{\mu\nu} - i\overline{\psi}\left(\not{D} + m\right)\psi \tag{1.2}$$

  of protons/electrons interacting with photons. We studied the latter theory in some detail at tree level and one-loop, leading to the famous one-loop calculation of the anomalous magnetic moment of the electron. We also showed that the classical Ising model in 2 spatial dimensions has a long-distance description in terms of a free Majorana fermion, and we used this to explain Onsager's solution of the model.
- We introduced the spontaneous breaking of global symmetries, showing that when the broken symmetry is continuous this leads to massless scalar particles called (Nambu-)Goldstone bosons. We also introduced the Higgs mechanism, whereby a field charged under a continuous gauge symmetry gets an expectation value, and we showed that this causes the gauge boson to acquire a mass. We explained how the Higgs mechanism underlies the remarkable phenomenon of superconductivity.
- Two subtleties which we dwelt on at some length are that spinor fields transform in representations of the spin double cover of the Lorentz group rather than the Lorentz group itself, and that gauge symmetries which vanish at spatial infinity must be viewed as redundancies of description in order to get a well-defined Hamiltonian theory.

[1] This course picks up where my first/second semester QFT lecture notes left off. I recommend looking at them on my website and reviewing anything which isn't familiar. If you took QFT II with another instructor you may have already learned some non-abelian gauge theory, but you likely will not have learned it the way I will teach it.

1.1 Theoretical motivation

In some sense our formal presentation of quantum field theory is complete: we know the general rules, and we know how to do calculations in concrete models. On the other hand you may be wondering to what extent the models we have discussed capture the full set of phenomena which can arise in quantum field theory. Here is an attempt to argue that they do. We have learned how to construct interacting theories for massive and massless particles of spin/helicity 0, 1/2, and 1. If we introduce particles of spin 3/2 or higher, the number of helicity states in the massless case is less than we would get in the massive case. Thus to have massless particles of helicity $\geq 3/2$, some gauge symmetry is needed to eliminate the extra spin states. For helicity 2 this gauge symmetry is diffeomorphism symmetry, so the helicity-2 particle is the graviton. As we discussed last semester we are not including gravity within the framework of quantum field theory, so by definition this is not allowed. For helicity 3/2 the gauge symmetry ends up being a local version of supersymmetry, and it turns out that local supersymmetry is not consistent unless the theory also has a graviton (the resulting theory is called **supergravity**). So helicity 3/2 is also not allowed. For massless particles of helicity greater than 2 there does not seem to be any way to write down gauge-invariant interactions in Minkowski space. Thus we have learned that massless particles of helicity $\geq 3/2$ can't arise in quantum field theory unless they are free. What about massive particles of spin $\geq 3/2$? Interacting theories *can* be written down for these, but their interactions are always irrelevant in the Wilsonian sense so there is no universal predictive theory for an arbitrary set of weakly-interacting higher-spin massive particles.

While there is something to these arguments, they have (at least) two loopholes:

(1) They assume that the particles in the theory arise from weakly-interacting fields. We will see that in the theory of the strong nuclear force, called **quantum chromodynamics** (QCD), this is not the case, and indeed QCD has massive particles with spin $\geq 3/2$ whose interactions can be unambiguously computed.[2]

(2) They assume that the only way to have an interacting one-form gauge field is the way we have already found, i.e. using the Maxwell action with coupling to matter fields via gauge-covariant derivatives. But in fact there is another way: **non-abelian gauge theory**. This is a generalization of quantum electrodynamics where there are multiple gauge fields that are mixed together by gauge transformations. Unlike the $d = 4$ field theories we have constructed so far, we will see that non-abelian gauge fields can flow to strong coupling at long distances (with QCD being the prime example). They thus can lead to interesting new phenomena that we did not have in our ϕ^4, Yukawa, and Maxwell theories.[3]

This semester we will therefore primarily be concerned with constructing non-abelian gauge theories and understanding the remarkable phenomena they can lead to.

1.2 Experimental motivation

In addition to these theoretical motivations, there are also experimental reasons not to be satisfied with the field theories we have considered thus far. We will now discuss several of these, seeing how they eventually lead to a theory of particle physics based on non-abelian gauge fields.[4]

1.2.1 Hadron Proliferation

So far we have given a description of the atomic nucleus as a bound state of protons and neutrons held together by exchanging pseudoscalar pions via the Yukawa Lagrangian (1.1). By the early 1950s however it

[2] On the other hand as far as I know the argument against interacting massless particles of helicity $\geq 3/2$ in Minkowski space is correct.

[3] For $d = 3$ on the other hand we found that ϕ^4 theory flows to strong coupling in the infrared, and in the first semester we studied this using the ϵ-expansion. Unfortunately that method is not strong enough to understand non-abelian gauge theories, so we will need to find something better.

[4] The rest of this section has a substantial amount of particle physics. If you are not so interested in particle physics please bear with me: this will be the highest density of particle physics we encounter all semester, and anyways I at least had some fun thinking through it.

became clear that these particles have many heavier cousins, some of which have spin greater than one as I already mentioned. Together with pions, protons, and neutrons these particles are called **hadrons**. Hadrons come in two types, bosonic **mesons** and fermionic **baryons**. Here is a table of the lightest mesons:[5]

Name	Symbol	Mass (MeV)	Charge	Spin	Lifetime (s)	Quark Content
Neutral Pion	π^0	135	0	0	8×10^{-17}	$\overline{u}u - \overline{d}d$
Charged Pion	$\pi^\pm$	140	± 1	0	3×10^{-8}	$\overline{d}u/\overline{u}d$
Charged Kaon	$K^\pm$	494	± 1	0	1.2×10^{-8}	$\overline{s}u/\overline{u}s$
Neutral Kaon (Short)	K_S^0	498	0	0	9×10^{-11}	$\overline{s}d + \overline{d}s$
Neutral Kaon (Long)	K_L^0	498	0	0	5×10^{-8}	$\overline{s}d - \overline{d}s$
Eta	η	548	0	0	5×10^{-19}	$\overline{u}u + \overline{d}d - 2\overline{s}s$
Eta prime	η'	958	0	0	3×10^{-21}	$\overline{u}u + \overline{d}d + \overline{s}s$
Charged Rho	$\rho^\pm$	775	± 1	1	5×10^{-24}	$\overline{d}u/\overline{u}d$
Neutral Rho	ρ^0	775	0	1	5×10^{-24}	$\overline{u}u - \overline{d}d$
Omega	ω	783	0	1	8×10^{-23}	$\overline{u}u + \overline{d}d$
Charged Vector Kaon	$K^{*\pm}$	892	± 1	1	1×10^{-23}	$\overline{s}u/\overline{u}s$
Neutral Vector Kaon	$K^{*0}/\overline{K^{*0}}$	896	0	1	1×10^{-23}	$\overline{s}d/\overline{d}s$
Phi	ϕ	1020	0	1	2×10^{-22}	$\overline{s}s$

and here is a table of the lightest baryons:

Name	Symbol	Mass (MeV)	Charge	Spin	Lifetime (s)	Quark Content
Proton	p^+	938	1	1/2	$> 3 \times 10^{37}$	uud
Neutron	n	940	0	1/2	9×10^2	udd
Lambda	Λ^0	1116	0	1/2	3×10^{-10}	uds
Sigma (Plus)	Σ^+	1189	+1	1/2	8×10^{-11}	uus
Sigma (Zero)	Σ^0	1193	0	1/2	7×10^{-20}	uds
Sigma (Minus)	Σ^-	1197	-1	1/2	1×10^{-10}	dds
Xi (Zero)	Ξ^0	1315	0	1/2	3×10^{-10}	uss
Xi (Minus)	Ξ^-	1322	-1	1/2	2×10^{-10}	dss
Delta (Plus Plus)	Δ^{++}	1232	2	3/2	6×10^{-24}	uuu
Delta (Plus)	Δ^+	1232	1	3/2	6×10^{-24}	uud
Delta (Zero)	Δ^0	1232	0	3/2	6×10^{-24}	udd
Delta (Minus)	Δ^-	1232	-1	3/2	6×10^{-24}	udd
Sigma (Star Plus)	Σ^{*+}	1383	+1	3/2	2×10^{-23}	uus
Sigma (Star Zero)	Σ^{*0}	1384	0	3/2	2×10^{-23}	uds
Sigma (Star Minus)	Σ^{*-}	1387	-1	3/2	2×10^{-23}	dds
Xi (Star Zero)	Ξ^{*0}	1532	0	3/2	7×10^{-23}	uss
Xi (Star Minus)	Ξ^{*-}	1535	-1	3/2	6×10^{-23}	dss
Omega (Minus)	Ω^-	1672	-1	3/2	8×10^{-11}	sss

Clearly this was a lot for midcentury physicists to process, and a theory which required a separate field for each hadron would not be a very predictive one. In fact these tables are only the beginning, sitting on top of each of the particles in this table is a **Regge trajectory** of additional particles of higher spin and mass, and there are also other heavier hadrons as well. Many of these hadrons are quite long-lived, as the natural timescale associated to hadrons is

$$\tau_{strong} = \frac{\hbar}{1000\text{MeV}} = 6 \times 10^{-25}\,\text{s}. \tag{1.3}$$

[5]These tables are obtained from the Particle Data Group (PDG), but I have suppressed error bars and been somewhat capricious about how many significant figures to include. We are anyways looking for general patterns rather than high precision.

To make sense of these tables we need some kind of new idea, and the key insight that brought order to the madness is indicated in the last column: hadrons are bound states of **quarks**. In other words we can replace these tables by a much simpler one that looks like this:

Name	Symbol	Mass (MeV)	Charge	Spin
Up	u	2.3	2/3	1/2
Down	d	4.8	-1/3	1/2
Charm	c	1.3×10^3	2/3	1/2
Strange	s	95	-1/3	1/2
Top	t	1.7×10^5	2/3	1/2
Bottom	b	4.2×10^3	-1/3	1/2

This table shows the six types of quark, which are conventionally called **flavors**. The previous tables show the lightest hadrons that are bound states of the three lightest flavors: the up, down, and strange quarks (so in fact we only need three rows of the table!). There are a few things which are immediately worth commenting on:

- The masses of the hadrons are substantially larger than the masses of the quarks, suggesting that whatever force binds them together is very strong. As a comparison I will remind you that the binding energy of hydrogen, $-13.6\,\mathrm{eV}$, is much smaller than the electron mass $m_e = 5.1 \times 10^5\,\mathrm{eV}$. Indeed the differ by the square of the fine structure constant $\alpha \approx \frac{1}{137}$, whose smallness you will recall measures the weakness of electromagnetic interactions.

- The mass *differences* among similar hadrons, such as the different pions or the proton and neutron, are of order the quark masses, suggesting that they are at least partly explained by quark masses.

- There are also electromagnetic difference between the hadrons, as the up and down quarks have different electric charges. These should contribute to the hadron masses as

$$q^2/R_H \sim \alpha(1000\mathrm{MeV}) \sim 10\mathrm{MeV}, \tag{1.4}$$

  which again is of order the differences in hadron masses of similar type. So apparently the quark masses and electromagnetism are both important in understanding hadron mass differences.

- The pseudoscalar mesons are substantially lighter than the other hadrons, especially the pions. This is with the notable exception of the η and η', which have a linear combination that does not involve the strange quark but is still much heavier than the seemingly similar π^0.

- Although quarks give a nice way of organizing our hadron tables, no quark has ever been seen in isolation. So the strong interactions which bind the hadrons together must also conspire to prevent free quarks.

Over the course of the semester we will see how QCD is able to explain all of these phenomena. In particular we will see that the pseudoscalars are light because they are approximate Goldstone bosons arising from the spontaneous breaking of an approximate global symmetry, and also that the dynamics of the strong force leads to a potential between quarks that grows linearly with distance. This linear potential causes the phenomenon of **confinement**, whereby quarks must be bound up into hadrons. In figure 1 I show a plot from the PDG review on the quark model, showing remarkable agreement between theory and experiment for the masses of all of the hadrons in the above tables. These theory calculations are done using lattice gauge theory, which is a precise non-perturbative formulation of non-abelian gauge theory that we will study later in the semester.

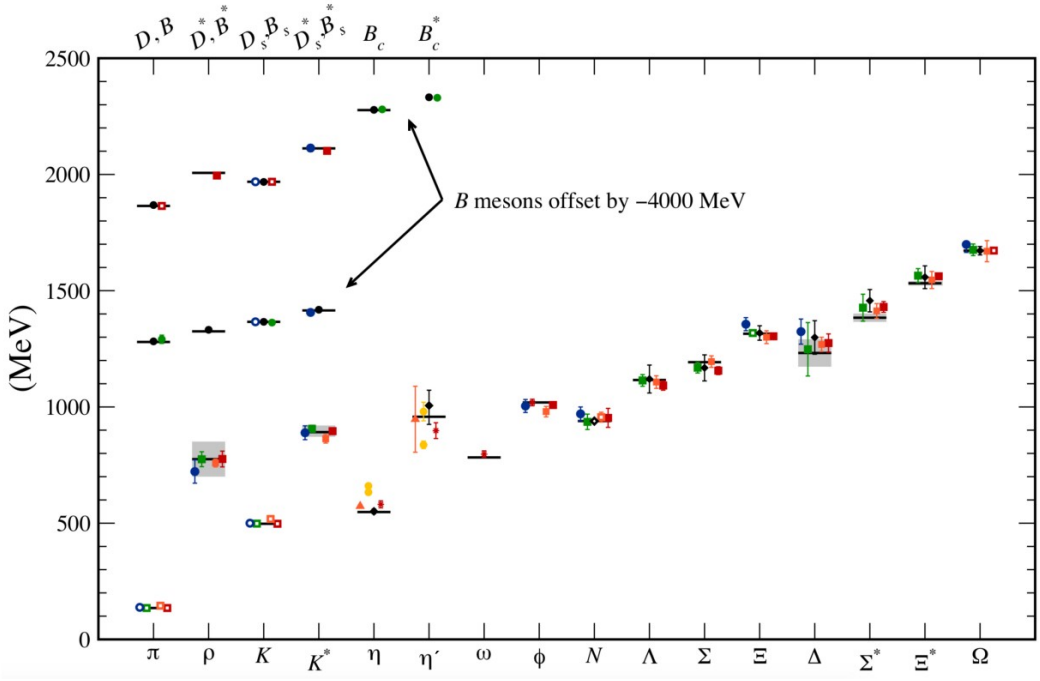

Figure 1: Light hadron masses as computed by lattice QCD. The horizontal black lines are the measured masses, the open points are used to set the Lagrangian parameters (gauge couplings, quark masses), and the closed points are predictions. The different colors are different lattice calculations of the same quantities. Lattice calculations which cannot resolve small hadron mass differences typically set $m_u = m_d$ and ignore electromagnetism, in which case there are three parameters to fit: the gauge coupling of QCD, the up/down quark mass, and the strange mass. Hence there are three open points (for each color) for the low-lying spectrum. To get the heavier D and B mesons two more open points are needed for the charm and bottom quark masses.

1.2.2 Flavor-changing decays and leptons

There are two additional notable features of our hadron tables:

- Each hadron has a definite number of up, down, and strange quarks (here we really mean the number of each flavor minus the number of its antiparticle), with the exception of neutral Kaons where there are a pair of almost degenerate superpositions $K_S^0 \pm K_L^0$, usually called K^0 and $\overline{K}^0$, which obey this property. Moreover if we define a **baryon number** charge under which all quarks have charge $q_B = 1/3$, then each meson has $q_B = 0$ and each baryon has $q_B = 1$ (the corresponding antibaryons have $q_B = -1$).
- There is a rather bimodal distribution of lifetimes. Most of the heavier hadrons have lifetimes which are roughly of order τ_{strong}, which suggests that they decay primarily through the strong force. On the other hand most of the lighter baryons (and a few of the heavier ones) have lifetimes in the $10^{-8}-10^{-10}$s range. Then there are a few oddballs like the π^0 and the η, which are in between these two scales, as well as the neutron, which lives for a geological 15 minutes!

These points are actually related to each other. The first point suggests that the strong force cannot change any of the six flavor numbers or baryon number, and indeed we will see that these are all symmetries of QCD. Moreover if we spend a little time glancing at the table, the hadrons with lifetimes of order τ_{str} all have plausible decay channels that respect all of these symmetries. For example the ρ^+ can decay into a π^+ and a π^0, which is indeed its dominant decay channel, and a Δ^{++} can decay into a proton and a π^+. On the other hand some of our hadrons do not have any plausible decay channels that respect flavor number symmetry. For example a $\pi^\pm$ has no lighter charged hadron to decay into. Similarly a neutron could only decay into a proton, but that has the wrong up, down, and electric charges and we do not have enough energy to fix them by including a π^-. In order to account for these decays we thus need to introduce some additional force that goes beyond the strong and electromagnetic forces: the **weak nuclear force**.

We can get started understanding the weak nuclear force by considering more carefully the decay of the neutron. What could the final state be? It needs to be electrically neutral, and whatever particles are produced their masses need to add up to less than 940 MeV. If one of these particles is a proton (for example to conserve baryon number) then we have only 2 MeV left for whatever else is produced. To cancel the charge of the proton, the only remaining option is an electron:

$$n \to p^+ + e^- + ?\,. \tag{1.5}$$

We have allowed for the possibility of something else in the final state, which we will see in a moment is a good idea. The rest mass of an electron is .5 MeV, so the electron produced by neutron decay will be somewhat relativistic but not extremely so. For historical reasons electrons which are produced by nuclear decay are called β-radiation, so this process is called β-decay.[6] (1.5) is indeed the primarily observed decay channel of the neutron, which we can view as evidence that the weak nuclear force indeed conserves baryon number (otherwise we could have a decay like $n \to e^+ + \pi^-$, and we'd even have to worry about $p^+ \to e^+ + \pi^0$). The lifetime is so long for two reasons: firstly because the weak force is indeed weak, and secondly because there is very little energy for the electron and its mystery partner so the integral over the final state phase space is suppressed.

Let's now turn to the mystery partner in the final state. We'll first conservatively assume that it is not there, so that the final state is just a proton and an electron. In the rest frame of the neutron the momentum p_e of the outgoing electron would then be determined by energy and momentum conservation to solve the equation

$$m_n = \sqrt{m_p^2 + p_e^2} + \sqrt{m_e^2 + p_e^2}. \tag{1.6}$$

After a little algebra this gives

$$\sqrt{p_e^2 + m_e^2} = \frac{m_n^2 + m_e^2 - m_p^2}{2m_n} \approx 1.3\text{MeV} \tag{1.7}$$

[6] When ionizing radiation was discovered in the late 19th century people had no idea what it was made out of, so they classified it into α, β, and γ particles based on how far they could penetrate aluminum. We now know that α particles are helium nuclei, β particles are electrons, and γ particles are photons.

for the energy of the outgoing electron. This however is not what is observed, which is that there is a continuous distribution in electron energy. If that is not bad enough, there is an additional problem with a proton-electron final state. Namely neutrons, protons, and electrons all have spin 1/2, so in a decay $n \to p + e^-$ the angular momentum of the initial state is a half-integer while the angular momentum of the final state is an integer! These problems actually led luminaries such as Bohr to propose that energy and angular momentum are not conserved by the weak nuclear force, but fortunately Pauli in a famous letter proposed an alternative: the true decay of the neutron is

$$n \to p^+ + e^- + \overline{\nu}_e. \tag{1.8}$$

where $\overline{\nu}_e$ is the antipartner of a new particle ν_e called a **neutrino** (more precisely an *electron* neutrino). This removes the tension with energy conservation, as now the electron energy can vary depending on how much energy is carried off by the antineutrino. Let's work out a few properties that the neutrino must have:

- To restore angular momentum conservation it must be a fermion, so conservatively we will take it to have spin (or possibly helicity) 1/2.
- It must be electrically neutral, since otherwise it would have already been seen in the decay products of the neutron.
- It must be neutral under the strong force, since otherwise it again would have been detected (and potentially even be subject to confinement).
- It must *not* be neutral under the weak nuclear force, since it needs to be produced during the decay process.
- It must be very light, since otherwise it could not be produced using our 2 MeV energy budget and moreover it would not have enough possible final momenta to account for the observed range of electron energies.

These properties combine to give a ghostly partner to the electron which is quite difficult to detect, and indeed Pauli described the situation thusly: "I have done a terrible thing, I have postulated a particle that cannot be detected." History proved him wrong in the best possible way: neutrinos were directly observed in 1956 less than thirty years later (and just in time, as Pauli died in 1958).[7]

The electron and the electron antineutrino are our first examples of **leptons**, which are a class spin 1/2 fermions which are neutral under the strong interactions. Just as the process (1.8) preserves baryon number, it also preserves **lepton number**.[8] We saw before that the up and down quarks have heavier cousins with the same quantum numbers, and this turns out to be true for electrons and neutrinos as well:

Name	Symbol	Mass (MeV)	Charge	Spin	Lifetime (s)
Electron	e^-	.51	-1	1/2	$> 10^{35}$
Electron Neutrino	ν_e	$< 10^{-6}$	0	1/2	x
Muon	μ^-	106	-1	1/2	2×10^{-6}
Muon Neutrino	ν_μ	$< 10^{-6}$	0	1/2	x
Tau	τ^-	1777	-1	1/2	3×10^{-13}
Tau Neutrino	ν_τ	$< 10^{-6}$	0	1/2	x

Note in particular that direct measurements of the neutrino masses are consistent with them being zero. This is why the neutrino lifetimes are ambiguous: the lifetime of a massless particle cannot be defined since it does not have a rest frame. Nonetheless no neutrino has every been seen to decay. On the other hand

[7] The way they were detected was via a kind of inverse β decay, where an antineutrino is absorbed by a proton to make a neutron and a positron.

[8] We will eventually see that the weak nuclear force does violate baryon and lepton number symmetry via highly-suppressed non-perturbative processes called instantons, but as far as we know the **difference** of baryon and lepton number, usually called $B - L$, is exactly conserved.

there is by now quite convincing indirect evidence that at least two independent linear combinations of these neutrinos has nonzero masses of order 10^{-8} MeV, and using this value their lifetimes are at least of order months. The muon was discovered in cosmic rays already back in 1936, while the tau was not seen until the 1970s due to its large mass. One simple place where the muon appears is in the dominant decay of the π^+ meson:

$$\pi^+ \to \mu^+ + \nu_\mu. \tag{1.9}$$

1.2.3 $\mathcal{R}/\mathcal{T}$ violation

The theories (1.1) and (1.2) both have the property that addition to $\mathcal{CRT}$ symmetry, they are also invariant under separate spatial and temporal reflection symmetries $\mathcal{R}$ and $\mathcal{T}$ that act on the fields as[9]

$$\begin{aligned} \phi'(x) &= -\phi(\mathcal{R}x) \\ \psi'(x) &= \gamma\gamma^1\psi(\mathcal{R}x) \\ A'_\mu(x) &= \mathcal{R}^\nu{}_\mu A_\nu(\mathcal{R}x) \end{aligned} \tag{1.10}$$

and[10]

$$\begin{aligned} \phi'(x) &= -\phi(\mathcal{T}x) \\ \psi'(x) &= B_1\gamma^0\psi(\mathcal{T}x) \\ A'_\mu(x) &= -\mathcal{T}^\nu{}_\mu A_\nu(\mathcal{T}x), \end{aligned} \tag{1.11}$$

where B_1 is the charge conjugation matrix which in our standard γ-matrix representation in $3+1$ dimensions is given by

$$B_1 = \begin{pmatrix} 0 & -\sigma_2 \\ -\sigma_2 & 0 \end{pmatrix} \tag{1.12}$$

and

$$\mathcal{R} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathcal{T} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{1.13}$$

When we study QCD we will see that these symmetries are also preserved by the strong nuclear force. Quite remarkably however, they are both violated by the weak nuclear force!

The possibility that spatial reflection symmetry might be violated by the weak nuclear force was first proposed by Lee and Yang in 1956, who suggested several possible experiments. They discussed this possibility with Wu, who realized that she was capable of carrying out the first experiment they suggested.[11] The idea was to look at the β decay of Cobalt via the process

$$\mathrm{Co}_{60} \to \mathrm{Ni}_{60} + e^- + \overline{\nu}_e + 2\gamma, \tag{1.14}$$

and in particular at the relationship between the angular distribution of the electron and the initial spin of the Cobalt nucleus. What Wu found is that the electrons are preferentially emitted opposite to the direction of the nuclear spin, which is clearly not consistent with a symmetry that reflects the momentum in the spin

[9] Although $\mathcal{R}$ is the fundamental spatial reflection symmetry, and also the operation that appears naturally in the $\mathcal{CRT}$ theorem, in $3+1$ dimensions it is universal practice to combine it with a spatial rotation to introduce a **parity** transformation $\mathcal{P}$ that inverts all three spatial directions. One thus typically says that it is $\mathcal{P}$ which is shown to be violated by the Wu experiment and $\mathcal{CP}$ which is violated by the decay of K_L^0 to three pions.

[10] The minus sign in the transformation of A_μ arises because there is an i in the covariant derivative so iA_μ needs to transform the same as ∂_μ and time reversal is antiunitary. You can check that this is consistent with the usual statements that the electric field is even under $\mathcal{T}$ while the magnetic field is odd, which you can figure out by thinking about what happens to their sources.

[11] In the early 20th century the romanization of Chinese names was not yet standardized. T.D. Lee, C.S. Wu, and C.N. Yang would today be written as Li Zhengdao, Wu Jianxiong, and Yang Zhenning.

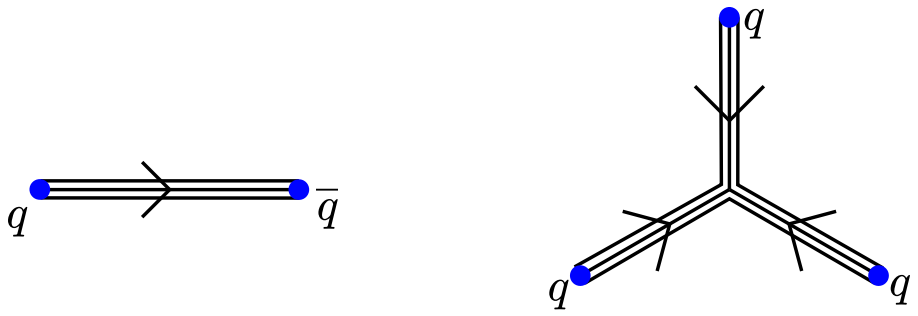

Figure 2: A meson and a baryon, held together by non-abelian electric flux tubes.

direction but does not reflect flip the spin (which is what happens if we act with $\mathcal{R}$ in the direction of the spin). This was an extraordinary discovery, for which Lee and Yang received the Nobel prize in 1957. Wu was outrageously excluded due to her lack of a Y chromosome, which led to substantial criticism of the Nobel committee (including from Pauli). In the next few years it was understood that the origin of the $\mathcal{R}$ violation in the standard model is that the weak force couples only to left-handed particles and right-handed antiparticles.[12] Since $\mathcal{R}$ symmetry mixes the left and right handed components of a Dirac spinor (due to the γ^1 in (1.10)), it therefore must be broken. This breaking of handedness is often described by saying that the weak force is **chiral**: it distinguishes between left and right handedness.

Although the Wu experiment showed that $\mathcal{R}$ symmetry is broken by the weak force, it left open the possibility that $\mathcal{T}$ symmetry is still conserved. This is because the $\mathcal{T}$ transformation (1.11) does not mix left and right-handed spinors. Theorists such as Landau therefore held out the hope for several years that $\mathcal{T}$, or equivalently $\mathcal{CR}$ by the $\mathcal{CRT}$ theorem, would still be a symmetry. This also turned out to be false however: in 1964 Cronin and Fitch showed that the K_L^0 meson, which is odd under $\mathcal{CR}$, can occasionally decay to two pions, which is a $\mathcal{CR}$-even state. This violation of $\mathcal{CR}$ symmetry is more subtle than the violation of $\mathcal{R}$ symmetry, and once we present the standard model in detail we will see that it requires the existence of all 6 flavors of quarks and leptons.

1.3 Towards gauge fields

We have now gathered most of the experimental data we will need for the rest of the semester. What we have not seen however are any non-abelian gauge fields. Why do we need them? We'll discuss first the strong force and then the weak.

1.3.1 Non-abelian gauge fields for the strong force

The main role for non-abelian gauge fields in the strong nuclear force is to provide the mechanism for confinement. For mesons the picture is quite simple: in each meson the quark and antiquark are held together by a **flux tube**, made out of a non-abelian generalization of the electric field. See the left side of figure 2 for an illustration. Since the flux is constant along the tube, the potential energy of the configuration grows linearly with the distance between the quark and antiquark. In particular if we move one of the quarks off to infinity, the remaining configuration has infinite energy. In fact long before we get to this infinite energy, the large energy in the tube would cause the creation of a quark-antiquark pair, leading to a pair of mesons rather than an isolated quark (see figure 3). This picture is appealing, but it has two immediate problems:

- The flux tube picture explains the presence of mesons, but what about baryons? In order to have a baryon we need to have a way for three electric flux lines to come together at a point (see the right

[12] Of course we could always change what we mean by particle and antiparticle, the invariant statement is that there is a choice where this is true.

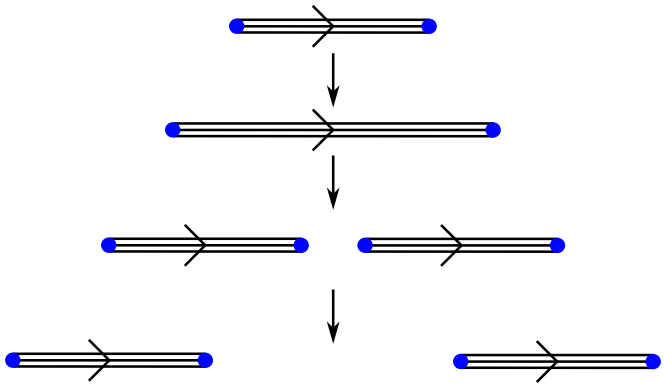

Figure 3: String breaking to prevent formation of an isolated quark: if we pull on the ends of a meson, eventually we get two mesons instead of two quarks.

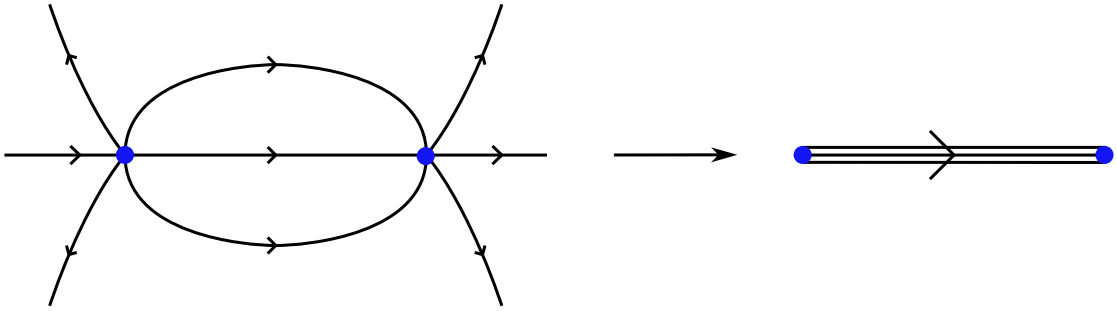

Figure 4: As we increase the gauge coupling the classical electric dipole on the left must adjust into a confining flux tube.

side of figure 2), which is certainly not a solution of Maxwell's equations.

- For the actual electric field, a flux tube is not the ground state configuration of an electric dipole. This is instead given by the classical dipole field shown in figure 4, which has finite energy (provided that we regularize the UV by giving the charges some finite size). Why should non-abelian gauge fields prefer the flux tube configuration instead of the dipole field configuration?

The solution to the first of these problems is that we need to promote the gauge group $U(1)$ of quantum electrodynamics to the non-abelian group $SU(3)$, which allows for gauge-invariant operators of the form

$$b = \sum_{nm\ell} \epsilon_{nm\ell} \psi_n \phi_m \chi_\ell \tag{1.15}$$

where ψ, ϕ, χ are quark fields and n is a three-component index which is mixed under $SU(3)$ gauge transformations. $\epsilon_{nm\ell}$ is the usual completely-antisymmetric tensor with $\epsilon_{123} = 1$. The label n is called **color**; it should not be confused with flavor (e.g. each up quark has three possible colors). This operator is gauge-invariant for the same reason that we can make a pseudoscalar out of three vectors in $\mathbb{R}^3$ using the cross product: $\vec{v} \cdot (\vec{u} \times \vec{w})$. To give a concrete example, a gauge-invariant local operator in QCD with the same quantum numbers as the proton is

$$p = \epsilon_{nm\ell} \left(u_n^T \mathcal{C} \gamma d_m \right) u_\ell, \tag{1.16}$$

where u and d are fields that annihilate up and down quarks and $\mathcal{C} = iB_2\gamma^0$ is the matrix appearing in the action of a Majorana fermion.[13]

The solution of the second of these puzzles is more subtle. The essence is that the classical dipole configuration is preferred when the gauge coupling is small, while the flux tube configuration is preferred when the gauge coupling is large. The latter is clearly not something that we can establish using conventional perturbation theory, but we will be able to show it using lattice gauge theory in an approximation where the coupling is very large instead of very small.

There is a third essential feature of non-abelian gauge theory which is worth commenting on. As we just discussed the flux tube configuration is dominant at strong coupling, but why should the coupling be strong? Indeed from our work on the renormalization group we know that coupling constants in field theory are actually scale-dependent, and in $3+1$ dimensions the couplings we studied so far all get weaker as we go to long distances. More specifically, we found that the β functions for the self coupling λ in $\lambda\phi^4$ theory and the gauge coupling q in spinor electrodynamics are positive. The same is true for the gauge coupling in scalar electrodynamics and the Yukawa coupling g. We will see this semester however that the gauge coupling in non-abelian gauge theory has a negative β function, and thus flows to strong coupling at long distances even if we start with a small coupling in the UV. This is important because it means that confinement is actually a robust feature of non-abelian gauge theory in $3+1$ dimensions: in order to avoid it we need to have some other infrared dynamics such as the Higgs mechanism or a large number of massless charged particles.

1.3.2 Non-abelian gauge fields for the weak force

Let's now return to the problem of the β decay of the neutron. We would like a theory that accounts for the possibility that a neutron can decay into a proton, an electron, and an antineutrino. In 1934 Fermi proposed such a theory, which in modern notation uses a vector interaction[14]

$$\mathcal{L}^{1934}_{Fermi} = -\frac{G_F}{\sqrt{2}} \left(\overline{p}\gamma^\mu n\right) \left(\overline{e}\gamma_\mu \nu_e\right) + \text{h.c.}. \tag{1.17}$$

Here G_F is a coupling constant, which we will estimate in a moment, and p, n, e, and ν_e are Dirac spinors for the proton, neutron, electron, and electron neutrino. After the Wu experiment however this needed to be modified to ensure that only left-handed particles feel the weak force, so the modern version is

$$\mathcal{L}_{Fermi} = -\frac{G_F}{\sqrt{2}} \left[\overline{p}\gamma^\mu(1 - g_A\gamma)n\right] \left[\overline{e}\gamma_\mu(1-\gamma)\nu_e\right] + \text{h.c.} \tag{1.18}$$

The parameter g_A is not one because it is really left-handed quarks that feel the weak force, and the relationship between quark fields and hadrons passes through the complexities of confinement. The chirality of the Fermi interaction is more clear if we write it directly in terms of the quark fields, in which case we get

$$\mathcal{L}^{quarks}_{Fermi} = -\frac{G_F}{\sqrt{2}} V_{ud} \left[\overline{u}\gamma^\mu(1-\gamma)d\right] \left[\overline{e}\gamma_\mu(1-\gamma)\nu_e\right] + \text{h.c.} \tag{1.19}$$

where V_{ud} is an element of the "CKM matrix", which we will learn about later (for now it is enough to know that $V_{ud} \approx 1$). Interactions of this type are called **Fermi interactions**, and the Feynman diagram for the β-decay via the Fermi interaction is shown in figure 5.

[13] See section 5 of QFT II. Quarks are not Majorana fermions; the role of $\mathcal{C}$ here is that $\psi^T\mathcal{C}$ has the same Lorentz transformation as $\overline{\Psi}$, so $u_n^T\mathcal{C}\gamma d_m$ is a Lorentz scalar. You might ask why this scalar did not appear in Yukawa theory or spinor electrodynamics, for example via a Yukawa type interaction $-ig\phi\psi^T\mathcal{C}\psi$. In Yukawa theory this is not allowed because it would violate baryon number symmetry, while in electrodynamics the quantity $\psi^T\mathcal{C}\psi$ wouldn't be gauge-invariant.

[14] Fermi's paper was rejected by Nature on the grounds of being too speculative to be interesting, which apparently caused him to give up on doing theory for a while. This opinion was not shared by his peers, who viewed his paper as a major breakthrough. Perhaps his most audacious idea was that the electron and the antineutrino could be produced "out of nothing" at the cost of converting a neutron to a proton, previous theories had demanded that the electron and the antineutrino be bound inside of the nucleus prior to its decay. Particle nonconservation had been previously accepted for the photon of course, but Fermi was the first to take it seriously for fermions.

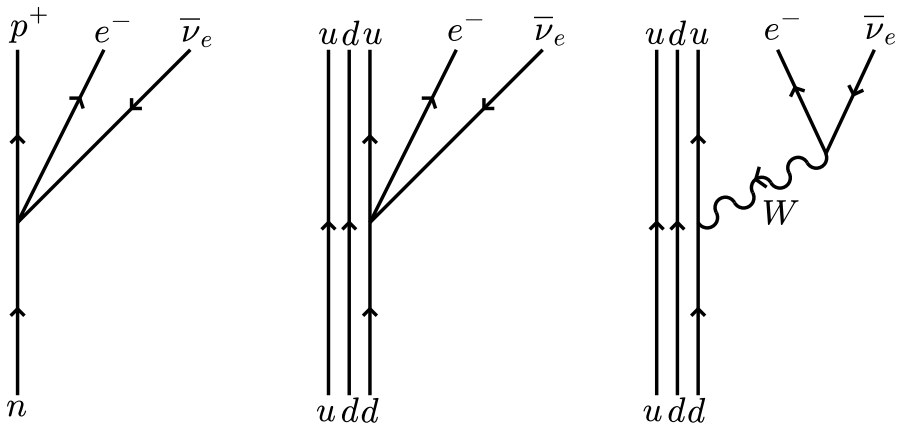

Figure 5: Three perspectives on the β-decay of the neutron. On the left the decay is mediated by the original Fermi interaction between nucleons and leptons. In the middle we have used QCD to resolve the proton and neutron into constituent quarks. On the right we have further resolved the non-renormalizable Fermi interaction using the exchange of a W boson, which is the current theory of this decay. This is a classic illustration of effective field theory, with the renormalization group taking us from right to left.

In $3+1$ dimensions the energy dimension of a spinor field is $3/2$, so the energy dimension of the Fermi coupling is

$$[G_F] = -2. \tag{1.20}$$

In other words the Fermi interaction is not renormalizable, which from the Wilsonian point of view means that it is induced by some kind of short distance physics at an energy scale

$$M_F = \frac{1}{\sqrt{G_F}}. \tag{1.21}$$

We can estimate this scale using our typical weak decay time scales: since the rate should be proportional to the coupling squared, by dimensional analysis the typical hadronic lifetime for a weak decay should be of order

$$\tau \sim \frac{1}{G_F^2 \times Q^5} \tag{1.22}$$

where Q is the available energy for the decay. For the decay of the neutron we have $Q \approx 2\,\mathrm{MeV}$ and $\tau \approx 10^3\,\mathrm{s}$, which gives

$$M_F \sim 10^6\,\mathrm{MeV}. \tag{1.23}$$

This is a bit of an overestimate, actually computing the lifetime and comparing to the observed value gives

$$M_F \approx 3 \times 10^5\,\mathrm{MeV}. \tag{1.24}$$

Thus the Fermi theory suggests that the weak force should arise from interesting physics at an energy scale of order $100\,\mathrm{GeV}$, which is typically called the **electroweak scale**.

What kind of new physics could we expect? The structure of the Fermi interaction (1.19) suggests that it could arise by exchanging a heavy vector boson, conventionally called the **W boson** and denoted W^+, which couples only to left-handed quarks and leptons, and which converts between quark flavors as well as electrons and neutrinos. See figure 5 for an illustration. In order for electric charge to be conserved this vector boson needs to be charged, so it has an antipartner called the W^-. Their mass should be of order M_F. There are a few questions however that need to be answered:

- How can interaction with a vector boson change the flavor of a quark/lepton? This certainly did not happen in quantum electrodynamics.
- Since the W^+ is charged, it must somehow mix together with the photon. What does this interaction look like and how is it gauge-invariant?
- Why is the W^+ massive? Typically massive vector bosons do not have renormalizable interactions, so does this mean there is some further new physics we need to discover?
- If the W^+ only couples to left-handed quarks and leptons, how are we supposed to think about the fact that these are massive particles and thus do not have well-defined handedness?

The answer to the first two of these questions is that the $W^\pm$ need to be gauge bosons that arise from a non-abelian gauge group, whose action on the matter fields mixes quark flavors and electrons with neutrinos. Since they need to mix u and d, and also e and ν_e, this suggests the gauge group $SU(2)$. On the other hand $SU(2)$ cannot be the whole story, since the $W^\pm$ need to be charged under electromagnetism. The resolution of this puzzle is that the full gauge group is

$$G_{EW} = SU(2) \times U(1)_Y, \tag{1.25}$$

where the photon actually arises from a linear combination between the $U(1)_Y$ generator (which is called **hypercharge**) and the σ_z generator of $SU(2)$. To understand why only this linear combination is massless we need to introduce the Higgs mechanism, which gives mass to the other three generators. Two of these generators are the W^+ and W^-, while the third gives yet another massive vector boson, called the **Z boson**. This answers the third question, and actually the fourth as well since we saw last semester that the Higgs mechanism can also give masses to charged particles like quarks and leptons via a Yukawa type interaction. We thus are led to the conclusion that the full gauge group of the standard model is[15]

$$G_{SM} = SU(3) \times SU(2) \times U(1)_Y, \tag{1.26}$$

with the heavy vector bosons $W^\pm$ and Z, as well as a heavy scalar **Higgs boson** H. As of 2012 these have all indeed been observed, with the following masses and lifetimes:

Name	Symbol	Mass (GeV)	Charge	Spin	Lifetime (s)
W boson	W^+	80.4	1	1	3×10^{-25}
Z boson	Z	91.2	0	1	3×10^{-25}
Higgs boson	H	125	0	0	1.6×10^{-22}

We will spend a good part of the rest of the semester filling in the details of this theory of the strong, weak, and electromagnetic forces between elementary particles, which is called the **Standard Model** of particle physics. The standard model has turned out to give an almost complete description of all observed phenomenona in particle physics so far.

1.4 What is missing?

Although the Standard Model has been very successful, it is worth mentioning that there are several observed phenomena that it cannot explain:

- **Gravity:** The most obvious omission is gravity. We can include it as a non-renormalizable interaction, but we do not have any strong candidate for a theory that explains scattering at the Planck scale and above. The evidence so far suggests that a complete theory of quantum gravity would look very different from the field theories we have constructed so far, and in particular that spacetime itself would need to be emergent.

[15] We will eventually see that this is not quite the whole story, as it is possible to have a discrete quotient of this group that mixes the various factors, but that would only be important for subtle non-perturbative questions so it is usually ignored.

- **Dark matter:** The observed distribution of matter in the universe is not compatible with what we would expect from applying Newtonian (or Einstein) gravity. It instead suggests that there is a large amount of mass carried by some kind of massive particle which does not interact with the electromagnetic or strong forces. In the Standard Model there is no such particle, with the only possibility being neutrinos but these are too light to account for the non-relativistic behavior of dark matter.[16]

- **Neutrino mass:** As already mentioned there is by now good evidence that at least two of the neutrinos have small masses of order 10^{-8} MeV. In the Standard Model they are massless, and there are a number of options for how to modify it to fix this which we cannot yet discriminate between experimentally.

- **Baryogenesis:** The observed matter in the universe strongly favors baryons over antibaryons and leptons over antileptons. This is true despite the fact that the expansion of the universe does not break $\mathcal{CR}$ symmetry. The standard model does include processes that break $\mathcal{CR}$ symmetry as we already mentioned, but not enough to explain the observed asymmetry.[17]

- **Structure formation:** The matter overdensities during the big bang which eventually collapsed to form stars and galaxies have a rather specific "nearly scale-invariant" distribution that could not have been produced from only standard model fields starting from simple initial conditions. The leading candidate to explain this is **inflation**, which hypothesizes that in the early universe there was an additional scalar field that slowly rolled down its potential driving a period of rapid expansion.

There are also some more philosophical problems based on the idea that some coupling constants in the standard model or cosmological parameters appear "fine-tuned" without any convincing explanation for why:

- **Cosmological constant problem:** In the standard model loop diagrams naturally generate a cosmological constant which is of order the UV cutoff to the fourth power. If we take the cutoff to be the Planck scale, then this cosmological constant is larger than the observed one by about 10^{123}. To fix this we need to tune the bare cosmological constant to cancel the one that is dynamically generated, to one part in 10^{123}. Even if we lower the cutoff to the scale of 10 TeV, which is the lowest scale that we have not convincingly probed using colliders, there is a tuning to one part in 10^{63}.

- **Electroweak Hierachy problem:** Loop diagrams in the standard model also generate a mass for the Higgs boson which is of order the cutoff squared. If we again take the cutoff to be the Planck scale, the mass squared that we get is of order 10^{34} times the observed mass squared. Thus to fix this we need to tune the bare Higgs mass to cancel this one to one part in 10^{34}. If we lower the cutoff scale to 10 TeV this tuning is still of order one part in 10^4.

- **Strong CP problem:** In QCD we will see that there is a $\mathcal{CR}$-violating term that could be included in the action with a dimensionless coupling called θ. Since $\mathcal{CR}$ is violated in the electroweak sector, we should expect the renormalization group to generate an $O(1)$ value of θ at low energies, which would lead to a sizeable electrical dipole moment for the neutron. In fact no such moment is observed, leading to a bound $\theta < 10^{-10}$. In order to have such a small dipole moment we therefore need to include a bare contribution to θ in the action which cancels the one generated by the weak force to one part in 10^{10}.

- **Cosmological coincidence problem:** As far as we know there are no direct interactions between standard model particles and dark matter. Nonetheless the average energy densities of the two in the universe are comparable, with the dark matter energy density being about five times higher. Moreover

[16] The term "dark matter" is misleading, as dark objects are those which absorb light. A better name would be "transparent matter".

[17] Nonetheless it is tantalizing to speculate that the breaking of $\mathcal{CR}$ symmetry in the standard model is a necessary part of the story, as this could give a compelling reason why it is necessary to have the heavy quarks and leptons (recall that we need all of them in order for $\mathcal{CR}$ to be broken).

these are both comparable to the current energy density in the cosmological constant, with that being about 2.5 times the energy density in dark matter. A priori these three densities have nothing to do with each other, so why are they all similar?

These problems are not as compelling as the previous ones, since in each case the answer could just be "that is the way it is". On the other hand so far the fundamental laws of physics have not tended to involve very large or very small dimensionless parameters, so it would be nice if we can come up with some kind of dynamical explanation. In this class we will for the most part not discuss proposed solutions for either set of problems, as our goal is to understand the laws of physics which currently have strong empirical support.

1.5 Other applications

In the previous two semesters we discussed several applications of quantum field theory to condensed matter physics. So far non-abelian gauge fields seem to be of less importance there than they are in high-energy physics, so our applications this semester will primarily be from particle physics (as we have already seen). That said, a strong candidate for a non-abelian phase in a condensed matter system is the fractional quantum hall effect with filling fraction $\nu = 5/2$, as realized for example with layered GaAs and AlGaAs at low temperature and high magnetic field. This system is particularly interesting from the point of view of quantum computation: a sufficiently tunable implementation would potentially allow us to build a quantum computer with excellent fault-tolerance properties.

Non-abelian gauge fields are also of great importance in quantum gravity and string theory. For example in the most well-understood example of the AdS/CFT correspondence relating quantum gravity in asymptotically anti-de Sitter space to conformal field theory at its asymptotic boundary, the boundary conformal field theory is a supersymmetric version of non-abelian gauge theory with gauge group $SU(N)$. More generally non-abelian gauge fields arise in string theory whenever you have a stack of multiple D-branes, and they can also arise when you compactify string theory on a manifold with a non-abelian isometry group.

Another place where non-abelian gauge fields have led to new insights is mathematics. Indeed the proper mathematical formulation of a non-abelian gauge field is that it is a **connection on a principal bundle over spacetime**, with matter fields that are charged under the gauge field being **sections** of that bundle. This formulation has led to relations with many important topics in mathematics, for example the Atiyah-Singer index theorem has a simple proof based on gauge theory path integrals and Donaldson's theory of classifying smooth four-manifolds relies crucially on instanton moduli spaces in gauge theory. Moreover Edward Witten received his Fields medal for using gauge theory in three dimensions to give a physical interpretation of the Jones polynomial for classifying knots. At the end of the semester we will learn as much about the topology of gauge fields as we have time for.

1.6 Homework

1. Based on our hadronic tables, come up with plausible strong-force decay channels for the ρ^0, the ϕ, the Δ^0, and the Σ^{*+}. Make sure that your proposed decays conserve electric charge, energy, upness, downness, and strangeness.

2. Argue that there is no possible strong decay for the K^+, the Λ^0, the Σ^+, or the Ω^-. Come up with plausible weak-force decay channels for each of these (make sure that you still conserve electric charge, baryon number, and lepton number).

3. Check that the Lagrangians (1.1) and (1.2) are invariant under the $\mathcal{R}$ and $\mathcal{T}$ transformations (1.10) and (1.11). Don't forget that $\mathcal{T}$ is implemented by an antiunitary operator that takes the complex conjugate of c-numbers in the Lagrangian (including γ-matrices). You can refer to problem 5 from section 6 of QFT II for help, as well as section 3.5.

4. Confirm the estimate (1.23) for the electroweak scale starting from the dimensional analysis estimate (1.22).

5. Review anything from the summary on the first page which you are unfamiliar with from QFT I and QFT II.

2 Lie groups and Lie algebras

In the previous section we used experimental data to motivate the idea of having a gauge symmetry that mixes different quark and lepton fields. More abstractly we would like to construct a field theory that is invariant under gauge transformations of the form

$$\phi'_n(x) = D_{nm}(g(x))\phi_m(x), \tag{2.1}$$

where ϕ_n are some collection of M fundamental fields (as in electrodynamics, we will see that the gauge fields themselves have a more complicated transformation rule). Here $g(x)$ is an arbitrary smooth map from spacetime to a Lie group G which is called the **gauge group**. D is a map from G to the set of $M \times M$ matrices that respects the group multiplication

$$D(g_1)D(g_2) = D(g_1 g_2). \tag{2.2}$$

Such a map D is called a representation of G. Quantum electrodynamics was such a theory with $G = U(1)$ or $G = \mathbb{R}$, so what we are doing can be thought of as a generalization of quantum electrodynamics to a broader set of gauge groups. It turns out that performing this generalization will require a more systematic understanding of Lie groups than we have so far discussed. We will thus spend the remainder of this section reviewing some of the foundational theory of Lie groups.

2.1 Basic definitions and examples

A **Lie group** is a smooth manifold G which is also a group, and for which the multiplication and inversion maps are smooth.[18] A Lie group is called **abelian** if $g_1 g_2 = g_2 g_1$ for all $g_1, g_2 \in G$, and otherwise it is called **non-abelian**. $U(1)$ and $\mathbb{R}$ are abelian Lie groups, while the rotation group $O(d-1)$ and the Lorentz group $O(d-1,1)$ are non-abelian. Here are some other well-known Lie groups:

- $GL(N,\mathbb{R})$: the group of $N \times N$ invertible matrices with real components
- $GL(N,\mathbb{C})$: the group of $N \times N$ invertible matrices with complex components.
- $SL(N,\mathbb{R})$: the subgroup of $GL(N,\mathbb{R})$ with determinant one.
- $SL(N,\mathbb{C})$: the subgroup of $GL(N,\mathbb{C})$ with determinant one.
- $U(N)$: the group of $N \times N$ unitary matrices
- $SU(N)$: the subgroup of $U(N)$ with determinant one
- $SO(N)$: the subgroup of $O(N)$ with determinant one

A Lie group is called **connected** if any two points in G can be connected by a continuous path, and it is called **compact** if every sequence of points in G has a convergent subsequence. For example $GL(N,\mathbb{R})$ is disconnected since no path can connect matrices with positive determinant to those with negative determinant, but $GL(N,\mathbb{C})$ is connected since we can have a path where the determinant goes around zero in the complex plane. $U(N)$, $SU(N)$, $O(N)$, and $SO(N)$ are all compact, while $O(d-1,1)$ and the GL and SL groups are all non-compact. For example a sequence of boosts in the x direction with rapidity $\eta = 1, 2, 3, \ldots$ does not have a convergent subsequence in $O(d-1,1)$.

[18] In quantum field theory we can motivate this smoothness requirement by noting that Lagrangians involve derivatives of fields, and thus require derivatives on gauge transformations.

The above examples might give you the idea that all Lie groups are **matrix groups**, meaning groups that can be constructed as subgroups of $GL(N,\mathbb{C})$.[19] This however is not the case, for example the universal covering group of $SL(2,\mathbb{R})$ is not a matrix group. On the other hand the Lie groups that are most often of interest in physics are usually matrix groups.[20] In particular all compact Lie groups are matrix groups, as we will review below. The primary place where this distinction matters is in the discussion of Lie algebras: in a matrix group the Lie algebra is itself a set of matrices, while for a general Lie group the Lie algebra is a set of vector fields on the group manifold G. Our approach will be to first introduce Lie algebras in the context of matrix groups, and then to explain the general definition.

2.2 Matrix Lie Algebras

General Lie groups are complicated beasts. The idea of Lie algebra is that we can simplify our lives by first studying the group multiplication "near the identity". For a matrix group this is easy to formalize: near the identity we can adopt coordinates x^a on the group, with the identity being at $x^a = 0$, and then writing $x^a = \epsilon\theta^a$ we have[21]

$$g = I + i\epsilon\theta^a T_a + O(\epsilon^2), \tag{2.3}$$

where g and T_a are $N \times N$ matrices and the θ^a are real. The Einstein summation convention is in force here, as it will be for a indices from now on. The neglected higher-order terms are matrices whose components are at most of order ϵ^2. If we multiply two such group elements we have

$$g_1 g_2 = I + i\epsilon(\theta_1^a T_a + \theta_2^a T_a) + O(\epsilon^2), \tag{2.4}$$

so at first order in ϵ the group multiplication rule reduces to the addition of matrices. The real vector space of matrices spanned by the T_a is called the **Lie algebra** of the matrix Lie group G, and it is often written as $\mathfrak{g}$. We will denote an arbitrary element of $\mathfrak{g}$ as T, so a general infinitesimal group element has the form

$$g = I + i\epsilon T + O(\epsilon^2). \tag{2.5}$$

The Lie algebras of the above matrix groups are all fairly straightforward to identify, for example the Lie algebra $\mathfrak{gl}(N,\mathbb{R})$ of $GL(N,\mathbb{R})$ is the vector space of imaginary $N \times N$ matrices, the Lie algebra $\mathfrak{sl}(N,\mathbb{R})$ of $SL(N,\mathbb{R})$ is the vector space of imaginary traceless $N \times N$ matrices, and the Lie algebra $\mathfrak{u}(N)$ of $U(N)$ is the vector space of hermitian $N \times N$ complex matrices.

This first-order treatment of the group multiplication cannot be the whole story however, and in particular since addition is always abelian it throws away any non-abelian information about the group. To do better we can consider multiplying together a bunch of copies of an infinitesimal group transformation to get a finite one:[22]

$$\lim_{\epsilon\to 0}\left(I + i\epsilon T + O(\epsilon^2)\right)^{\frac{1}{\epsilon}} = e^{iT}. \tag{2.6}$$

Thus given any element of the Lie algebra $\mathfrak{g}$, we can exponentiate (including a factor of i) to get an element of the Lie group G. In this context the map $T \mapsto e^{iT}$ is called the **exponential map**. Physics textbooks sometimes pretend that we can use it to give a parametrization of the full Lie group, but there are two problems with this:

[19] There is some variance in the mathematics literature on whether or not we should require the subgroup to be closed. Closure is good because it ensures that the subgroup inherits a natural Lie group topology from that of $GL(N,\mathbb{C})$, but it is bad because not all Lie subalgebras generate closed Lie subgroups (a counterexample is the subgroup $\begin{pmatrix} e^{ix} & 0 \\ 0 & e^{i\sqrt{2}x} \end{pmatrix}$ of $GL(2,\mathbb{C})$, with $x \in \mathbb{R}$). For compact Lie groups however, which will be those of primary interest for us, closure is automatic. If we do not require closure, then we should at least require that the inclusion map from G into $GL(N,\mathbb{C})$ be a smooth immersion (this is true for the example just mentioned).

[20] That said, I did recently write a paper which prominently featured the universal cover of $SL(2,\mathbb{R})$!

[21] The factor of i here is conventional in physics; it is there so that if g is unitary then T_a is hermitian. Mathematicians use a different convention that does not include it, in which case a unitary g gives antihermitian T_a. The mathematician convention is arguably superior, as it makes the Lie algebra closed under the commutator instead of i times the commutator, and the exponential map is really the exponential without an annoying extra factor of i. We will stick to the physics convention however, as it is rather entrenched.

[22] If you aren't sure about this limit, try taking the logarithm of both sides.

(1) The exponential map in general is not injective, as there can be be distinct elements of the Lie algebra which map to the same group element. For example when $G = U(1)$, the generators $T = 0$ and $T = 2\pi$ both map to the identity.

(2) The exponential map in general is not surjective, as there can be elements of G which cannot be obtained by exponentiating any element of $\mathfrak{g}$. This is obvious when G is disconnected, as the exponential map can only reach elements in the identity component (imagine "scaling up" T from zero), but even when G is connected there are examples where surjectivity fails.[23]

The statement which *is* true in general is that the restriction of the exponential map to a sufficiently small neighborhood of zero in $\mathfrak{g}$ gives a bijection onto a neighborhood of the identity in G.[24]

Using the exponential map, we can derive a very important property of the Lie algebra of any matrix group. By the surjectivity of the exponential map in a neighborhood of the identity, for sufficiently small ϵ we must have

$$e^{i\epsilon S}e^{i\epsilon T} = e^{i\epsilon R(S,T)} \tag{2.8}$$

for some smooth function $R : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$. We can work out the function R order by order in ϵ. For example writing

$$R(S,T) = R_0(S,T) + \epsilon R_1(S,T) + \dots \tag{2.9}$$

and working to second order in ϵ we need

$$\left(I + i\epsilon S - \frac{\epsilon^2}{2}S^2 + \dots\right)\left(I + i\epsilon T - \frac{\epsilon^2}{2}T^2 + \dots\right) = I + i\epsilon R_0 + i\epsilon^2 R_1 - \frac{\epsilon^2}{2}R_0^2 + \dots. \tag{2.10}$$

This has solution

$$\begin{aligned} R_0 &= S + T \\ R_1 &= \frac{i}{2}[S,T]. \end{aligned} \tag{2.11}$$

Thus we see a new constraint on the Lie algebra: *the commutator of two elements of the Lie algebra times i must also be an element of the Lie algebra.* We will call this the **commutator constraint**. In particular the basis matrices T_a must obey

$$[T_b, T_c] = iC^a{}_{bc}T_a \tag{2.12}$$

for some real constants $C^a{}_{bc}$ that are called the **structure constants** of the Lie algebra. The structure constants capture the "local part" of the non-abelian nature of the group. In the homework you will check the commutator constraint for several of the classical matrix groups mentioned above. The structure constants are clearly antisymmetric in b and c, and from the Jacobi identity

$$[T_a,[T_b,T_c]] + [T_b,[T_c,T_a]] + [T_c,[T_a,T_b]] = 0 \tag{2.13}$$

we see that they must obey

$$C^e{}_{ad}C^d{}_{bc} + C^e{}_{bd}C^d{}_{ca} + C^e{}_{cd}C^d{}_{ab} = 0. \tag{2.14}$$

Having discovered that working at second order leads to a constraint on the Lie algebra, you might be worried that going to higher orders leads to yet more constraints. It turns out however that this is not the case: all of the higher order terms in R can be expressed as nested commutators of S and T with a power of

[23] It is however a theorem that if G is both connected *and* compact, then the exponential map is surjective.

[24] To prove this formally one uses the inverse function theorem of many-variable calculus. See for example theorems 7.6 and 7.10 of "Introduction to Smooth Manifolds" by John Lee. The criterion to check is that the differential of the exponential map at zero is invertible, which is certainly the case here since

$$\frac{d}{ds}e^{isT}|_{s=0} = iT, \tag{2.7}$$

which is zero if and only if T is zero.

i for each commutator! This expression is called the **Baker-Campbell-Hausdorff** formula. It would take us too far afield here to develop it in general, but the first few terms look like this:[25]

$$e^{i\epsilon S}e^{i\epsilon T} = e^{i\left(\epsilon(S+T)+\frac{i\epsilon^2}{2}[S,T]+\frac{\epsilon^3}{12}[T,[S,T]]-\frac{\epsilon^3}{12}[S,[S,T]]+\ldots\right)}. \tag{2.15}$$

Thus the commutator constraint is both necessary and sufficient to ensure that $R(S,T)$ is an element of $\mathfrak{g}$. An important consequence of this result is that given *any* real vector space of matrices which is closed under i times the commutator, we can exponentiate it (with an i) to construct a connected matrix group for which it is the Lie algebra (to get the full group we need to include arbitrary products of exponentials since the exponential map may not be surjective).[26]

As with Lie groups, there is an abstract definition of Lie algebra that does not refer to matrices. Namely a Lie algebra is a vector space V together with a bilinear commutator operation $C : V \times V \to V$, written in mathematics notation (without the i) as $C(x,y) = [x,y]$, such that the Jacobi identity $[x,[y,z]]+[y,[z,x]]+[z,[x,y]] = 0$ holds. The latter is automatic for matrix Lie algebras. It is far from obvious, but a famous theorem due to Ado says that all abstract Lie algebras defined in this way can actually be realized as matrix Lie algebras, so unlike for Lie groups we do not gain anything by using the abstract definition.

2.3 The geometric approach

Having developed some intuition for Lie algebras in the matrix case, we now give the general definition.[27] For any Lie group G and any group element $h \in G$, we an define a smooth **left-multiplication map** $L_h : G \to G$ by

$$L_h(g) = hg. \tag{2.16}$$

As with any smooth map from G to G, this map induces a **pushforward** map

$$L_{h*} : T_gG \to T_{hg}G, \tag{2.17}$$

where T_gG denotes the tangent space of vector fields at g.[28] A vector field X on G is said to be *left-invariant* if it obeys

$$X(hg) = L_{h*}(X(g)) \tag{2.19}$$

for all h and g. These vector fields actually form a finite-dimensional vector space whose dimensionality is the same as the dimensionality of G. This is because the set of left-invariant vector fields is actually isomorphic as a vector space to T_eG the tangent space of G at the identity. Indeed given $X_e \in T_eG$ we can define a left-invariant vector field by simply defining

$$X(g) = L_{g*}(X_e), \tag{2.20}$$

and conversely given a left-invariant vector field X we can get an element of T_eG by simply evaluating $X(e)$.

An important feature of vector fields on manifolds is that we can define their **Lie bracket**, which in components is defined to be

$$[X,Y]^a = X^b\partial_b Y^a - Y^b\partial_b X^a. \tag{2.21}$$

[25] For a nice reference on the BCH formula see this see "Lie Groups, Lie Algebras, and Representations" by Hall.

[26] In case you are worried that products of exponentials might not be enough, I'll mention that it is a theorem that in a connected Lie group G any neighborhood U of the identity generates all of G. The proof is not difficult: we show that any subgroup $H \subset G$ which contains U must be both open and closed. H is open because for each $h \in H$ the neighborhood hU is contained in H, and it is closed because for each $g \notin H$ the neighborhood gU cannot intersect H (otherwise g would be in H).

[27] This section and the following one are somewhat mathematical, hopefully you have taken general relativity and know something about vector fields on manifolds. If not then don't worry, for the rest of the semester the Lie groups we consider will all be matrix groups.

[28] In components, if $f : M \to N$ is a smooth map from a manifold M to a manifold N, then

$$(f_*X)^a(f(x)) = \partial_b f^a(x)X^b(x). \tag{2.18}$$

If we want f_* to be a map from vector fields to vector fields, and not just tangent spaces to tangent spaces, then we should restrict f to be a diffeomorphism (as is the case for L_h).

The Lie bracket has the nice property that if X and Y are vector fields on a manifold M and $f : M \to N$ is a diffeomorphism, then

$$f_*[X, Y] = [f_*X, f_*Y], \tag{2.22}$$

as you will check on the homework. Therefore we see that the Lie bracket of two left-invariant vector fields on G is also a left-invariant vector field:

$$L_{h*}[X, Y] = [L_{h*}X, L_{h*}Y] = [X, Y]. \tag{2.23}$$

Thus we see that the set of left-invariant vector fields on G is a real vector space that is closed under Lie brackets, and that it is isomorphic to the tangent space at the identity. These are precisely the conditions we found for the Lie algebra of a matrix Lie group, with the bracket replaced by i times the matrix commutator, and indeed this is the way to define Lie algebras for general Lie groups: given a Lie group G, its **Lie algebra** is the vector space of left-invariant vector fields on G.

For matrix Lie groups we can check directly that Lie bracket coincides with the Lie algebra as we defined it above. Near the identity we can parametrize the group multiplication in a matrix group as

$$e^{ix_1^a T_a} e^{ix_2^a T_a} = e^{i\widetilde{x}^a(x_1,x_2)T_a}. \tag{2.24}$$

Using the BCH formula, to second order in x_1 and x_2 we have

$$\widetilde{x}^a(x_1, x_2) = x_1^a + x_2^a - \frac{1}{2}x_1^b x_2^c C^a{}_{bc} + \dots. \tag{2.25}$$

In these coordinates we thus have

$$L_g(h)^a = \widetilde{x}^a(x_g, x_h) = x_g^a + x_h^a - \frac{1}{2}x_g^b x_h^c C^a{}_{bc} + \dots. \tag{2.26}$$

Using (2.18) and (2.20) near the identity, we can therefore write a general left-invariant vector field $Y^a(x)$ as

$$Y^a(x) = (\delta_b^a - \frac{1}{2}x^c C^a{}_{cb} + \dots)Y^b(0). \tag{2.27}$$

Thus the Lie bracket is

$$\begin{aligned}[X, Y]^a &= X^c(0)(-\frac{1}{2}C^a{}_{cb}Y^b) - Y^c(0)(-\frac{1}{2}C^a{}_{cb}X^b) \\ &= -C^a{}_{bc}X^b(0)Y^c(0).\end{aligned} \tag{2.28}$$

Comparing this to our matrix Lie algebra

$$i[X^b T_b, Y^c T_c] = -C^a{}_{bc}X^b Y^c T_a, \tag{2.29}$$

we see that the Lie bracket indeed corresponds to i times the matrix commutator (had we used the mathematician convention for the matrix Lie algebra we wouldn't have had the annoying factor of i).

An important tool in our discussion of matrix Lie algebras was the exponential map $T \mapsto e^{iT}$. In a geometric context there is also an exponential map $\exp : \mathfrak{g} \to G$. The idea is that given $Y \in \mathfrak{g}$, we can construct a curve $g(t)$ in G by solving the ordinary differential equation

$$\dot{x}^a(t) = Y^a\big(x(t)\big) \tag{2.30}$$

with the boundary condition that at $x^a(0) = 0$ in coordinates where $x^a = 0$ is the identity. Solutions of (2.30) are called **integral curves** of the vector field Y. The exponential map is then defined by

$$\exp(Y) = g(1). \tag{2.31}$$

We can see that this coincides with our matrix group definition as follows. In our exponential coordinates $e^{ix^a T_a}$, a candidate path that would agree with the matrix exponential is

$$x^a(t) = t\theta^a, \tag{2.32}$$

where $Y = \theta^a T_a$ is the Lie algebra element we are exponentiating. We need to confirm that this path solves the equation (2.30). In other words we need to show that the left-invariant vector field Y^a with $Y^a(0) = \theta^a$ obeys

$$Y^a(t\theta) = \theta^a. \tag{2.33}$$

By definition we have

$$Y^a(t\theta) = \frac{\partial \widetilde{x}(t\theta, x)}{\partial x^b}\Big|_{x=0} \theta^b. \tag{2.34}$$

This quantity is the derivative of $\widetilde{x}(t\theta, x)$ at $x = 0$ as x moves in the θ direction. This is easy to compute: when x_1 and x_2 are parallel to each other we simply have

$$\widetilde{x}^a(x_1, x_2) = x_1^a + x_2^a, \tag{2.35}$$

so

$$Y^a(t\theta) = \frac{d}{ds}(t+s)\theta^a|_{s=0} = \theta^a \tag{2.36}$$

as desired.

One of our main results for matrix Lie groups was that given any matrix Lie algebra we can exponentiate it to generate a matrix Lie group. Our argument above relied on the BCH formula, which is fairly unpleasant to derive. Using geometric methods we can give a more elegant proof. The proof relies on the following famous theorem in differential topology:

Theorem 1. ***(Frobenius theorem)*** *Let M be a manifold and $D \subset TM$ be a smooth distribution of rank k, meaning a choice of k-dimensional subspace $D_p \subset T_pM$ at each $p \in M$ which varies smoothly as we vary p. Then the following statements are equivalent:*

1. *For any two vector fields X and Y such that $X(p), Y(p) \in D_p$ for all p, the same is true for their Lie bracket $[X, Y]$.*
2. *M is **foliated** by immersed submanifolds, meaning that each point $p \in M$ lies within a unique maximal immersed submanifold N such that for each $q \in N$ we have $T_qN = D_q$. Each such manifold N is called a **leaf** of the foliation.*

Essentially the Frobenius theorem gives a necessary and sufficient condition for M to decompose into non-intersecting submanifolds. For example in $\mathbb{R}^d$ the submanifolds where we fix $x^{k+1}, \dots, x^d$ and let $x^1, \dots, x^k$ vary give a foliation of $\mathbb{R}^d$ by k-dimensional submanifolds, whose tangent distribution D_p at each point is the vector space spanned by $\partial_1, \dots \partial_k$. Vector fields that live in this distribution are indeed closed under the Lie bracket. We won't prove the theorem here, see Lee's book. Using it however we can easily prove the following:

Theorem 2. *Let G be a Lie group with Lie algebra $\mathfrak{g}$. For any Lie subalgebra $\mathfrak{h} \subset \mathfrak{g}$, meaning a subspace of $\mathfrak{g}$ which is closed under the Lie bracket, there is a unique connected Lie subgroup $H \subset G$ whose Lie algebra is $\mathfrak{h}$. Here Lie subgroup means a subgroup whose inclusion map is a smooth immersion.*

Proof. The idea is that for each $g \in G$ we choose the subspace $D_g \subset T_gG$ spanned by the Lie algebra vectors in $\mathfrak{h}$. By assumption $\mathfrak{h}$ is closed under the Lie bracket, so by Frobenius's theorem G is foliated by maximal immersed submanifolds whose tangent space at each g is D_g. Denoting by N_g the leaf containing g, we will take $H = N_e$. We need to show that H is a subgroup of G. Indeed say $h, h' \in H$. Then we have

$$hh' = L_h(h') \in L_h(H) = L_h(N_e) = N_h = H, \tag{2.37}$$

with the nontrivial step being to argue that $L_h(N_e) = N_h$. This follows because the distribution D_g is left-invariant since it is spanned by left-invariant vector fields, so acting on any leaf with L_h gives another leaf. In particular acting with L_h on N_e gives the leaf that contains h. Similarly we can show that for any $h \in H$ we have $h^{-1} \in H$:

$$h^{-1} = h^{-1}e \in L_{h^{-1}}(N_e) = L_{h^{-1}}(N_h) = N_e = H. \tag{2.38}$$

Finally to claim uniqueness of H, we need to make sure there is no smaller connected subgroup $\widetilde{H} \subset H$ which still has $\mathfrak{h}$ as its Lie algebra. This however is not possible, as $\widetilde{H}$ would have to contain a neighborhood of the identity in H and any such neighborhood generates all of H by the comment in footnote 26. □

With this theorem in hand, together with the fact mentioned at the end of the previous section that every Lie algebra is isomorphic to a subalgebra of $\mathfrak{gl}(N, \mathbb{C})$, we again see that every abstract Lie algebra is the Lie algebra of a connected matrix group.

2.4 The correspondence between Lie groups and Lie algebras

In the previous two sections we established what is called the **Lie group/Lie algebra correspondence**: given a Lie group G we can construct a Lie algebra $\mathfrak{g}$ of left-invariant vector fields, and conversely given a Lie algebra $\mathfrak{g}$ we can construct a matrix Lie group G with $\mathfrak{g}$ as its Lie algebra. It is natural to ask to what extent this correspondence is one-to-one. To answer this we first need to say a bit more about what it means for two Lie groups or two Lie algebras to be the same. For Lie groups the answer is that G_1 and G_2 are isomorphic, written as $G_1 \cong G_2$, if there is a group isomorphism $\phi : G_1 \to G_2$ which is also a diffeomorphism. Such an isomorphism is called a **Lie isomorphism**. For Lie algebras we instead ask that there be an invertible linear map $L : \mathfrak{g}_1 \to \mathfrak{g}_2$ that preserves the commutator operation. A Lie isomorphism $\phi : G_1 \to G_2$ naturally induces a Lie algebra isomorphism $\phi_* : \mathfrak{g}_1 \to \mathfrak{g}_2$, so in the Lie group → Lie algebra direction the correspondence is one-to-one. The Lie algebra → Lie group direction however is not one-to-one without further restrictions, as we will now discuss.

The simplest way for distinct Lie groups to have the same Lie algebra is if they are disconnected. The exponential map (or the Frobenius theorem) will only ever generate a connected Lie group, so at a minimum if we want a unique correspondence we should restrict to connected Lie groups. Even with this restriction however it is still possible for distinct Lie groups to have isomorphic Lie algebras. The classic example of this is one that we already encountered last semester: $SO(3)$ and $SU(2)$ have the same Lie algebra, but they are different Lie groups. Indeed in section 3.1 we explained how this arose from the fact that $SO(3)$ is not simply-connected: it has a loop consisting of a set of rotations interpolating between no rotation and a rotation by 2π which is not contractible. On the other hand $SU(2)$ is simply-connected, and we used this to argue that any representation of the Lie algebra $\mathfrak{su}(2)$ exponentiates to a representation of $SU(2)$ (which we contrasted with the case of $SO(3)$, which does not have a spin 1/2 representation). This was because we can work out the multiplication law of any two group elements in a connected group by constructing a path in the group corresponding to a sequence of multiplications near the identity, and if the group is also simply-connected then there is no ambiguity arising from which path we choose (see figure 2 from QFT II). Working out the details of this argument establishes the following theorem:

Theorem 3. *Let G_1 and G_2 be two connected and simply-connected Lie groups with isomorphic Lie algebras. Then G_1 and G_2 are also isomorphic.*

We are getting close to the general statement of the Lie group/Lie algebra correspondence, but to finish up we need one more topological result:

Theorem 4. *Every connected Lie group G can be realized as a discrete central quotient of a unique (up to isomorphism) connected and simply-connected **universal covering group** $\widetilde{G}$. In other words we have $G \cong \widetilde{G}/\Gamma$, where Γ is a discrete subgroup of $\widetilde{G}$ that commutes with everything in $\widetilde{G}$. Moreover G and $\widetilde{G}$ have the same Lie algebra.*

You can find the proof of this theorem in Lee's book. The idea is to define $\widetilde{G}$ as the space of topologically-inequivalent paths from the identity to an arbitrary point in G, with the group multiplication being the product of paths. Γ is discrete because G and $\widetilde{G}$ have the same structure near the identity since a small enough neighborhood of the identity has no topologically nontrivial paths. This is also why they have the same Lie algebra. It is worth emphasizing that G being a matrix group does *not* imply that $\widetilde{G}$ is a matrix group, and indeed we already mentioned this fails for $G = SL(2, \mathbb{R})$. This theorem is therefore one reason why general Lie groups are nicer than matrix Lie groups.

Combining these two theorems with the result of the previous section, we at last have the following result:

Theorem 5. ***Lie Group/Lie algebra correspondence****: If G_1 and G_2 are isomorphic Lie groups, then their Lie algebras $\mathfrak{g}_1$ and $\mathfrak{g}_2$ are also isomorphic. Moreover any abstract Lie algebra $\mathfrak{g}$ is isomorphic to the Lie algebra of a unique (up to isomorphism) connected simply-connected Lie group $\widetilde{G}$. Any other connected Lie group G whose Lie algebra is isomorphic to $\mathfrak{g}$ is itself Lie isomorphic to a quotient of $\widetilde{G}$ by a discrete central subgroup.*

Proof. As already mentioned, an isomorphism between G_1 and G_2 induces an isomorphism between $\mathfrak{g}_1$ and $\mathfrak{g}_2$. In the other direction, by theorem 2 and Ado's theorem $\mathfrak{g}$ is the Lie algebra of some connected matrix Lie group G. There could be more than one such G, but by theorem 4 each of these is a discrete central quotient of a universal covering group $\widetilde{G}$ and by theorem 3 these universal covers are all isomorphic. □

2.5 Representations of Lie groups

We have seen that many Lie groups can be realized as matrix groups. It is also interesting to consider a weaker sense in which matrices can be used to represent a Lie group. Indeed if G is a Lie group and V is a finite-dimensional vector space, a **representation of G on V** is a continuous[29] homomorphism $D : G \to Aut(V)$, where $Aut(V)$ is the set of invertible linear maps from V to itself, and homomorphism means that

$$D(g_1)D(g_2) = D(g_1 g_2). \tag{2.39}$$

If we pick a basis for V and work in components, then $Aut(V)$ is just $GL(N, \mathbb{C})$ for some N and the $D(g)$ are $N \times N$ matrices. The map D is *not* in general required to be injective - when it is then we say the representation is **faithful**. Clearly a Lie group G is a matrix group if and only if it has a faithful representation on some finite-dimensional V. A subspace $S \subset V$ is said to be **invariant** if $D(g)S = S$, and a representation is said to be **irreducible** if the only invariant subspaces are $S = V$ and $S = 0$. If V is a Hilbert space and the $D(g)$ are unitary then we say the representation is **unitary**.

Given an invariant subspace S for a unitary representation D, the orthogonal complement $S^\perp$ of S in V is also invariant. Indeed for any $s \in S, s' \in S^\perp, g \in G$ we have

$$(D(g)s', s) = (s', D(g^{-1})s) = 0 \tag{2.40}$$

since $D(g^{-1})s \in S$. This implies a very important fact about unitary representation theory:

Theorem 6. *Let D be a unitary representation of a Lie group G on a finite-dimensional vector space V. Then V is a direct sum of mutually-orthogonal invariant subspaces, on each of which D acts irreducibly.*

Proof. Let S_1 be a nonzero invariant subspace. Without loss of generality we can assume it has no smaller nonzero invariant subspace, in which case D acts irreducibly on S_1. If $S_1 = V$ then we are done, if not then consider a nonzero invariant subspace $S_2 \subset S_1^\perp$ which we can again take to have no smaller nonzero invariant subspace. D again acts irreducibly on S_2. If $S_2 = S_1^\perp$, then we are done. If not, then we consider a minimal nonzero invariant subspace $S_3 \subset (S_1 \oplus S_2)^\perp$ and check if $S_3 = (S_1 \oplus S_2)^\perp$, and so on. This process increases the dimensionality of $S_1 \oplus S_2 \oplus \ldots$ each time, so it must stop after a finite number of steps since V is finite-dimensional. □

[29] You might expect me to say "smooth" here, and I could, but it is a theorem that a continuous homomorphism between Lie groups is automatically smooth.

To say the result of this theorem in a more pedestrian way, there is an orthonormal basis for V in which the $D(g)$ matrices are all block diagonal with the same blocks and each block furnishes a unitary irreducible representation of G:

$$D(g) = \begin{pmatrix} D_1(g) & 0 & \dots \\ 0 & D_2(g) & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}. \tag{2.41}$$

You are hopefully familiar with this statement in the context of the addition of angular momentum in non-relativistic quantum mechanics.

In physics we are also interested in infinite-dimensional representations. For a general topological vector space V, possibly infinite-dimensional, a representation of G on V is defined to be a homomorphism from $D : G \to Aut(V)$ G such that $D(g)v$ is jointly continuous as a function of $g \in G$ and $v \in V$.[30] For example action of symmetries on Hilbert space in quantum mechanics gives an infinite-dimensional unitary representation of the symmetry group. For infinite-dimensional representations invariant subspaces are required to be closed subspaces of V, and the above theorem generalizes to say that any unitary representation of G is a "direct integral" of mutually-orthogonal irreducible representations.

Given any representation D of a Lie group G, we can always extract from it a representation of its Lie algebra $\mathfrak{g}$. We simply need to Taylor expand:

$$D(e^{i\epsilon\theta^a T_a}) = I + i\epsilon\theta^a \tau_a + O(\epsilon^2) \tag{2.42}$$

for some generator matrices τ_a. These are required to obey the Lie algebra

$$[\tau_a, \tau_b] = iC^c{}_{ab}\tau_c \tag{2.43}$$

for the same reason that the abstract generators T_a are. On the other hand we cannot always exponentiate an arbitrary set of matrices τ_a obeying (2.43) to get a valid representation of G: as discussed above we may only get a representation of its universal cover $\widetilde{G}$.

2.6 The adjoint representation

There is a particularly important representation that exists for any Lie group. It is easiest to describe for matrix groups, so we will start there. The idea is that we can construct a representation of G that acts on its own Lie algebra $\mathfrak{g}$ by conjugation. To establish this we need to show that if T is in the Lie algebra $\mathfrak{g}$ for some matrix Lie group G, then so is gTg^{-1} for any $g \in G$. For g near the identity we can establish using the BCH formula, but we want it to be true for any g so we need a better argument. Indeed for any $T \in \mathfrak{g}$ and $g \in G$ we can consider the quantity

$$ge^{i\epsilon T}g^{-1} = e^{i\epsilon gTg^{-1}}. \tag{2.44}$$

The left-hand side is clearly an element of G, so the right-hand side must be as well. Working to first order in ϵ, we thus see that gTg^{-1} must be a linear combination of the Lie algebra generators T_a and therefore must itself be an element of the Lie algebra. In other words we must have

$$gT_ag^{-1} = D^b{}_a(g)T_b \tag{2.45}$$

for some coefficients $D^b{}_a(g)$. You will show on the homework that this equation implies that

$$D^b{}_a(g_1g_2) = D^b{}_c(g_1)D^c{}_a(g_2), \tag{2.46}$$

so these matrices indeed provide a representation of the group G, which is called the **adjoint representation**. Its dimensionality is equal to the dimensionality of G as a manifold. Writing a general element of the Lie algebra as $T = \theta^a T_a$, we simply have

$$\theta^{a\prime} = D^a{}_b(g)\theta^b. \tag{2.47}$$

[30] It is not obvious a priori what is the right continuity or smoothness assumption to use. The choice here works for any topological group G (it doesn't need to be a Lie group), and moreover when G is a Lie group it implies that D is also smooth.

Note that the adjoint representation is not necessarily faithful, for example for $G = U(1)$ we simply have $D(g) = 1$ for all g. More generally $D(g) = 1$ whenever g is in the center of G (if G is connected this is if and only if).

The adjoint representation also has a geometric definition that works for any Lie group G. We first note that we can always define a right-multiplication map $R_h : g \mapsto gh$ on G. It is immediate that this commutes with L_h,

$$L_h \circ R_{h'} = R_{h'} \circ L_h, \tag{2.48}$$

since these both map g to hgh'. We may then define the **conjugation map**

$$C_h = L_h \circ R_{h^{-1}}, \tag{2.49}$$

which of course acts as

$$C_h(g) = hgh^{-1}. \tag{2.50}$$

In particular the identity is a fixed point of the conjugation map for any h, as is any other element of the center of G. We now define the adjoint map $Ad_h : \mathfrak{g} \to \mathfrak{g}$ by

$$Ad_h(X) = C_{h*}X. \tag{2.51}$$

You will show on the homework that this coincides with the matrix definition for matrix Lie groups.

The adjoint representation for G induces an adjoint representation of $\mathfrak{g}$. Working for simplicity in the matrix case, its generators τ_a^A can be determined by Taylor expanding (2.45) for g near the identity:

$$\left(\delta_b^c + i\epsilon\theta^a(\tau_a^A)^c{}_b + O(\epsilon^2)\right) T_c = T_b + i\theta^a[T_a, T_b] + O(\epsilon^2) = T_B - \theta^a C^c{}_{ab} T_c + O(\epsilon^2), \tag{2.52}$$

from which we have

$$(\tau_a^A)^c{}_b = iC^c{}_{ab}, \tag{2.53}$$

so the generators of the adjoint representation of the Lie algebra are the structure constants themselves! On the homework you will check that these indeed obey the Lie algebra (2.43).

2.7 Representations of compact groups

Representation theory is a rather involved subject, even in the finite-dimensional case. It is easier however for compact Lie groups, so we will discuss a few general results in that case. Many of these results follow from a deep fact about Lie groups, which is that they admit a left-invariant integration measure dg called the **Haar measure**. This exists for any locally-compact group, and it is unique up to an overall constant multiple. What left-invariant means here is that for any (integrable) function $f : G \to \mathbb{R}$ and any $h \in G$ we have

$$\int_G dg f(hg) = \int_G dg f(g). \tag{2.54}$$

In general the construction of the Haar measure requires some nontrivial analysis, but for Lie groups there is an easy construction.[31] Indeed say G is a Lie group of dimension n. To define an integration measure on G we need to choose a globally-defined n-form ω. We can then define the integral of a function $f : G \to \mathbb{R}$ to be

$$\int_G dg f(g) \equiv \int_G f\omega, \tag{2.55}$$

where on the right side we have the integral of an n-form over an n-manifold. The way that ω is constructed is straightforward: at the identity we choose an arbitrary n-form ω_e, and then we define

$$\omega(g) = L^*_{g^{-1}}\omega(e) \tag{2.56}$$

[31] Here "easy" means "easy if you know about the theory of integration of n-forms on n-manifolds". If you don't, then either look it up or just believe me.

where $L^*_{g^{-1}}$ indicates the pullback of a differential form by the left-multiplication map $L_{g^{-1}}$. By construction this choice of ω obeys

$$L^*_h \omega(g) = \omega(h^{-1}g) \tag{2.57}$$

for any $h \in G$. To see that this measure is left-invariant, we first note that

$$f(hg)\omega(g) = f(hg)L^*_h\omega(hg) = L^*_h(f\omega)(g). \tag{2.58}$$

We thus have

$$\int_G dg f(hg) = \int_G L_{h^*} f\omega = \int_G f\omega = \int_G dg f(g). \tag{2.59}$$

The second equality here follows from the fact that the integral of an n-form over a manifold is preserved under the pullback by any diffeomorphism (this is the geometric version of the change of variables formula). To see that the Haar measure constructed in this way is unique up to a constant multiple, we note that all n-forms at the identity are proportional and any left-invariant n-form must be determined elsewhere in G by (2.56).

When G is compact there are two more nice features of the Haar measure. The first is that the integral over the whole group is finite, so we can normalize it to one:[32]

$$\int_G dg = 1. \tag{2.60}$$

The second is that when G is compact the Haar measure is also right-invariant, meaning that for any integrable function $f : G \to \mathbb{R}$ we have

$$\int_G dg f(gh) = \int_G dg f(g). \tag{2.61}$$

The proof of this is that we can take the left-invariant volume form ω, pull it back by $R_{h^{-1}}$, and then integrate h over G (here we use compactness to ensure a finite answer). By construction this gives a volume form that is both left-invariant and right-invariant, and by the uniqueness of the Haar measure it must actually just be a multiple of the one we started with - so that one must have been right-invariant to begin with!

Now let's use the Haar measure to learn some fun things:

Theorem 7. *Let G be a compact Lie group and D a representation of G on a finite-dimensional vector space V. Then V admits an inner product for which D is unitary.*

Proof. Let $(\cdot,\cdot)_0$ be any inner product on V. We will define a new inner product for which D is unitary. Indeed for any $v, v' \in V$ we define

$$(v', v) = \int_G dg (D(g)v', D(g)v)_0, \tag{2.62}$$

which is finite since G is compact. $D(g)$ is unitary in this inner product since we have

$$(D(h)v', D(h)v) = \int_G dg (D(g)D(h)v', D(g)D(h)v)_0 = \int_G dg (D(gh)v', D(gh)v)_0 = \int_G dg (D(g)v', D(g)v)_0 = (v', v) \tag{2.63}$$

by the right-invariance of the Haar measure. We also need to check that it is positive definite: indeed

$$(v, v) = \int_G dg (D(g)v, D(g)v) \geq 0, \tag{2.64}$$

with vanishing only if $v = 0$. □

[32] The converse of this statement is also true: if the Haar measure of a Lie group is finite then the group is compact. The proof of this is that if G is not compact, there is a sequence of points $\{g_n\}$ that escapes every compact subset of G. Choosing a neighborhood U of the identity with compact closure, one then argues that this sequence can be "pruned" to generate an infinite sequence where the regions $g_n U$ are disjoint for all n. Since the Haar measure is left-invariant these regions all have the same volume, so the volume of their union is infinite and thus contradicts G having finite volume.

In other words all finite-dimensional representations of a compact Lie group are unitary![33] A similar argument establishes what is perhaps the most striking fact about the representation theory of compact Lie groups:

Theorem 8. *(Schur Orthogonality) Let G be a compact Lie group, let α label its inequivalent irreducible representations, and let d_α be the dimension of α. Denoting the representation matrices as $D^\alpha_{ij}(g)$, we have*

$$\int_G dg D^{\alpha'*}_{i'j'}(g) D^\alpha_{ij}(g) = \frac{1}{d_\alpha}\delta^{\alpha\alpha'}\delta_{ii'}\delta_{jj'}. \tag{2.65}$$

To prove this we first need to establish a famous lemma in representation theory:[34]

Lemma 2.1. *(Schur's lemma). Let G be a group, and D and D' be irreducible representations of G onto finite-dimensional vector spaces V and V' respectively. Let L be a linear* ***intertwining map*** *$L : V \to V'$ such that*

$$LD(g) = D'(g)L \tag{2.66}$$

for all $g \in G$. If D and D' are inequivalent then $L = 0$, while if $D = D'$ then L is proportional to the identity, $L = \lambda I$ for some $\lambda \in \mathbb{C}$.

Proof. We first note that the kernel and image of L are invariant subspaces of D and D' respectively. Indeed for the kernel if $Lv = 0$ then we have $LD(g)v = D'(g)Lv = 0$ for all $g \in G$. Similarly if $v' = Lv$ then $D'(g)v' = D'(g)Lv' = LD(g)v$. By irreducibility the kernel must be zero or V and the image must be zero or V'. Let's first suppose that the kernel is zero, so that L is injective. The image cannot be zero, so it must be V' and so L is also surjective. L is therefore invertible and we have

$$D'(g) = LD(g)L^{-1}, \tag{2.67}$$

so the two representations are equivalent. Thus if the two representations are not equivalent, the kernel of L must be equal to V or in other words we must have $L = 0$.

Now let's say that $D = D'$. Since L is a finite-dimensional matrix it must have an eigenvalue λ (this is a standard result in linear algebra), and we can define another intertwiner $\hat{L} = L - \lambda I$. The kernel of $\hat{L}$ is not zero since it annihilates the eigenvector of L with eigenvalue λ, so by irreducibility its kernel must be V. Therefore $\hat{L} = 0$, and thus $L = \lambda I$. □

Using this lemma we can now establish (2.65):

Proof. The idea is to define a map $L_{ii'} : V \to V'$ by

$$(L_{i'i})_{j'j} = \int_G dg D^{\alpha'*}_{i'j'}(g) D^\alpha_{ij}(g). \tag{2.68}$$

By the invariance of the Haar measure this is an intertwiner,

$$L_{i'i}D^\alpha(h) = D^{\alpha'}(h)L_{i'i}, \tag{2.69}$$

so by Schur's lemma it must vanish if α and α' are not equivalent. This establishes the $\delta_{\alpha\alpha'}$ in (2.65). If $\alpha = \alpha'$ then instead must have $L_{i'i} \propto I$, which establishes the $\delta_{jj'}$ in (2.65). Finally to get the $\delta_{ii'}$ we can instead define

$$(L_{jj'})_{ii'} \equiv (L_{i'i})_{j'j}\,, \tag{2.70}$$

[33] In fact this argument generalizes to the case where V is an infinite-dimensional Hilbert space, you just need to fill in some details about limits using the continuity in our definition of infinite-dimensional representations. This is less natural perhaps though, since the inner product you construct won't in general be the one you started with and the one you start with isn't arbitrary (as it was in the finite-dimensional case).

[34] This lemma also gives the uniqueness of intertwiners that we used back in in QFT I. Indeed say D and D' are equivalent representations with intertwiner L, $LD = D'L$. Equivalent means that $D'(g) = SD(g)S^{-1}$ for some invertible matrix S. We thus have $S^{-1}LD = DS^{-1}L$, so by Schur's lemma we have $S^{-1}L = \lambda I$, or in other words $L = \lambda S$. So L is indeed determined up to a constant multiple (S is also unique up to a constant multiple since if $SD(g)S^{-1} = \hat{S}D(g)\hat{S}^{-1}$ the lemma tells us that $\hat{S}^{-1}S \propto I$).

which again is an intertwiner by Haar invariance and thus must be proportional to $\delta_{ii'}$. The constant of proportionality $1/d_\alpha$ is determined by contracting i and i' and using the unitarity of $D^\alpha(g)$. □

The Schur orthogonality formula (2.65) has many remarkable consequences. One way to interpret it is as saying that the functions $\sqrt{d_\alpha} D^\alpha_{ij}(g)$ give an orthonormal set of functions labeled by α, i, j from the group G to the complex numbers. Another famous result called the **Peter-Weyl theorem** says that this set is complete, so $\sqrt{d_\alpha} D^\alpha_{ij}(g)$ is an orthonormal basis for $L^2(G)$! Using the Peter-Weyl theorem together with an induction argument similar to the one we gave for theorem 6, we can finally establish (see appendix A from my paper with Ooguri):

Theorem 9. *Every compact Lie group G has a faithful finite-dimensional (and unitary) representation.*

Thus for compact Lie groups we do not lose anything by restricting to matrix groups.

2.8 Semisimple algebras and groups

In group theory there is a special kind of group, called a simple group, from which all other groups can be built via a process called group extension. In general this process is complicated, but for connected compact Lie groups the decomposition is easy to state: every compact connected Lie group G is of the form

$$G = (T \times S_1 \times \ldots \times S_n)/F, \tag{2.71}$$

where T is an abelian torus $U(1)^m$, each S_i is a compact simply-connected simple Lie group, and F is a finite central subgroup. The definition of a simple Lie group is that it must be connected, non-abelian, and have no connected normal subgroups except for the identity and the full group.

We are actually more interested in the Lie algebra version of this statement, but to describe it we need two definitions: a Lie algebra $\mathfrak{g}$ is said to be **semisimple** if it has no nonzero abelian ideal, meaning an abelian Lie subalgebra $\mathfrak{y}$ whose commutator with anything else in $\mathfrak{g}$ stays in $\mathfrak{y}$. $\mathfrak{g}$ is said to be **compact semisimple** if it is semisimple and also the matrix

$$K_{ab} = \mathrm{Tr}\left(\tau^A_a \tau^A_b\right) \tag{2.72}$$

is positive definite, where τ^A_a are adjoint representation matrices (2.53). The matrix K_{ab} is called the **Killing form** on the Lie algebra.

Theorem 10. *Let $\mathfrak{g}$ be a Lie algebra. The following conditions are equivalent:*

1. *$\mathfrak{g}$ is the Lie algebra of a compact Lie group G.*
2. *$\mathfrak{g}$ as a vector space admits an inner product that is invariant under the action of the adjoint representation in the sense that*
$$\langle [X,Y], Z\rangle = \langle Y, [Z,X]\rangle \qquad \forall X, Y, Z \in \mathfrak{g}. \tag{2.73}$$
3. *$\mathfrak{g}$ has a direct sum decomposition as*
$$\mathfrak{g} = \mathfrak{z} \oplus \mathfrak{s}, \tag{2.74}$$
where $\mathfrak{z}$ is an abelian Lie algebra, $\mathfrak{s}$ is a compact semisimple Lie algebra, and $\mathfrak{z}$ and $\mathfrak{s}$ are mutually commuting.

Proof. The proof that 1 $\implies$ 2 works by choosing an arbitrary inner product $\langle\cdot,\cdot\rangle_0$ on the real vector space $\mathfrak{g}$ and then averaging in the Haar measure to improve it to an invariant one:

$$\langle X, Y\rangle \equiv \int dg \langle Ad_g X, Ad_g Y\rangle_0, \tag{2.75}$$

which obeys

$$\langle Ad_g X, Ad_g Y\rangle = \langle X, Y\rangle. \tag{2.76}$$

Expanding this for g near the identity gives (2.73).[35]

The proof that $2 \implies 3$ takes $\mathfrak{z}$ to be the center of $\mathfrak{g}$ and $\mathfrak{s}$ to be its orthogonal complement. To show that $\mathfrak{s}$ is semisimple, say that W were a nonzero element of an abelian ideal $\mathfrak{y}$ within $\mathfrak{s}$. Then for any $Z \in \mathfrak{g}$ we have

$$\langle [W,Z],[W,Z]\rangle = \langle Z,[W,[W,Z]]\rangle = 0, \tag{2.79}$$

with the second equality holding because $[W,Z]$ must be in $\mathfrak{y}$ since it is an ideal and the commutator with W must vanish since the ideal is abelian. The positive-definiteness of the inner product then shows that this implies that $[W,Z]=0$, or in other words that W is in the center of $\mathfrak{g}$. But this means it is in $\mathfrak{z}$, so it wasn't in $\mathfrak{s}=\mathfrak{z}^\perp$ in the first place. To see that $\mathfrak{s}$ is compact semisimple, we note that since the adjoint generators τ_a^A preserves the inner product on the Lie algebra they must exponentiate to orthogonal matrices. Therefore they must be in the Lie algebra $\mathfrak{so}(d_G)$ where d_G is the dimension of G. This is the set of imaginary anti-symmetric matrices, so the matrix K_{ab} is indeed positive-definite:

$$\mathrm{Tr}(A^2) = \sum_{ab} A_{ab}A_{ba} = -\sum_{ab} A_{ab}^2 \geq 0. \tag{2.80}$$

Finally we want to prove that $3 \implies 1$. We can easily exponentiate $\mathfrak{z}$ to get an abelian torus $U(1)^m$, so the challenge is to show that a compact semisimple Lie algebra $\mathfrak{s}$ is the Lie algebra of a compact Lie group. We have seen that $\mathfrak{s} \subset \mathfrak{so}(d_G)$, so by theorem 2 $\mathfrak{s}$ is the Lie algebra of a connected subgroup H of the compact group $SO(d_G)$. We need to show that this subgroup is closed, and thus compact. This is not obvious, a proof that isn't too bad is given in Knapp's "Lie groups beyond an introduction" (see proposition 4.27). The idea is to show that H is isomorphic to the set of linear transformations on $\mathfrak{g}$ that preserve the commutator, which is a closed subgroup of $GL(d_G,\mathbb{R})$. It is therefore also closed in $SO(d_G)$, and thus compact. □

To finish our demonstration of (2.71) at the Lie algebra level, we need to show that the semisimple Lie algebra $\mathfrak{s}$ decomposes into a direct sum of commuting simple Lie algebras $\mathfrak{s}_i$. A Lie algebra $\mathfrak{g}$ is **simple** if it is non-abelian and has no ideals except for 0 and $\mathfrak{g}$ (note that simple $\implies$ semisimple but is stronger, since non-abelian ideals are also forbidden (except for $\mathfrak{g}$)). A Lie algebra is called **compact simple** if it is compact semisimple and also simple. The decomposition proof is similar to that for theorem 6. Indeed let $\mathfrak{s}_1$ be a minimal nonzero ideal in $\mathfrak{s}$. If $\mathfrak{s}_1=\mathfrak{s}$ then we are done. If not then take its orthogonal complement using the invariant inner product on $\mathfrak{s}$. This orthogonal complement is also an ideal since for any $X \in \mathfrak{s}_1$, $Y \in \mathfrak{s}_1^\perp$, and $Z \in \mathfrak{s}$ we have

$$\langle X,[Y,Z]\rangle = \langle Y,[Z,X]\rangle = 0, \tag{2.81}$$

since $\mathfrak{s}_1$ is an ideal. We then choose a minimal ideal $\mathfrak{s}_2$ in $\mathfrak{s}_1^\perp$, check if it is equal to $\mathfrak{s}_1^\perp$, if not choose $\mathfrak{s}_3 \subset (\mathfrak{s}_1 \oplus \mathfrak{s}_2)^\perp$, and so on. None of the ideals produced in this way can be abelian since $\mathfrak{s}$ is semisimple.

We have now seen that every compact Lie algebra is a direct sum of abelian and compact simple Lie algebras. We recover the group version (2.71) as follows. First we can exponentiate or use theorem 2 to turn each of these Lie algebras into a Lie group, and by going to the universal cover we get a Lie group which is a product of $\mathbb{R}^m$ and some simply-connected simple Lie groups. A famous theorem due to Weyl says that the fundamental group of any compact Lie group with semisimple Lie algebra is finite, or in other words that the unique connected simply-connected Lie group with this Lie algebra is also compact. This almost gets us to (2.71), as it shows that every compact Lie group G is a discrete central quotient of $\mathbb{R}^N \times S_1 \times \ldots S_n$ with all S_i simple, simply-connected, and compact. Finally one can argue that this discrete central quotient can be split into an infinite part that converts $\mathbb{R}^m$ to $U(1)^m$ and then a finite part F, I will leave the details to you.

[35] If you are willing to use our proof that compact Lie groups are all all subgroups of $U(N)$ for some N, then there is a simpler description of this inner product: we simply take

$$\langle T,S\rangle = \mathrm{Tr}(TS), \tag{2.77}$$

which is indeed an inner product on the space of hermitian matrices. The invariance condition is immediate:

$$\mathrm{Tr}([X,Y]Z) = \mathrm{Tr}(XYZ - YXZ) = \mathrm{Tr}(Y[Z,X]). \tag{2.78}$$

2.9 Cartan classification

We have now seen that simply-connected compact simple Lie groups are the nontrivial building blocks for all compact Lie groups. To classify compact Lie groups, it is therefore enough to classify simply-connected compact simple Lie groups. A priori it is not clear how difficult this task will be. For example the classification of *finite* simple groups is famously complex, requiring $O(10^4)$ pages to establish the existence of 18 infinite families and an additional 26 sporadic groups, including the infamous "monster" group with $32! \times 10! \times (4!)^2 \times 2 \times 7 \times 13 \times 41 \times 47 \times 59 \times 71 \sim 8 \times 10^{53}$ elements (known to fans as the "friendly giant"). The situation with compact simple Lie groups turns out to be substantially better however, essentially because these are uniquely determined by their Lie algebras due to the Lie algebra-Lie group correspondence so it is enough to classify compact simple Lie algebras. The classification of these is given by the **Cartan catalog**, which lists all compact simple Lie algebras as follows:

- $\mathfrak{so}(N)$ for $N = 3, 4, \ldots$
- $\mathfrak{su}(N)$ for $N = 2, 3, \ldots$
- $\mathfrak{usp}(2N)$ for $N = 1, 2, \ldots$
- $\mathfrak{g}_2, \mathfrak{f}_4, \mathfrak{e}_6, \mathfrak{e}_7, \mathfrak{e}_8$.

Here $\mathfrak{so}(N)$ and $\mathfrak{su}(N)$ are the Lie algebras of $SO(N)$ and $SU(N)$ respectively, and $\mathfrak{usp}(2N)$ is the Lie algebra of a third infinite family of compact Lie groups called $USp(2N)$ and referred to as the unitary symplectic groups. $USp(2N)$ is defined as the set of $2N \times 2N$ unitary matrices that obey

$$U^T M U = M \tag{2.82}$$

for some non-degenerate antisymmetric matrix M that we can take without loss of generality to be given by

$$M = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}. \tag{2.83}$$

Finally there are the five exceptional compact simple Lie algebras, whose dimensions as real vector spaces are given respectively by $14, 52, 78, 133, 248$. All of them have appeared in some guise in physics, for example the Lie group E_8 appears in heterotic string theory and in grand unification models of particle physics. There are four "accidental equivalences" in the classification:

$$\begin{aligned} &\mathfrak{su}(2) \cong \mathfrak{so}(3) \cong \mathfrak{usp}(2) \\ &\mathfrak{so}(5) \cong \mathfrak{usp}(4) \\ &\mathfrak{su}(4) \cong \mathfrak{so}(6). \end{aligned} \tag{2.84}$$

Other than these the Lie algebras in the list are all distinct. The proof that this classification is complete uses something called Dynkin diagrams, you can read about them in Knapp's book.

2.10 Homework

1. What are the Lie algebras of $GL(N,\mathbb{C})$, $SL(2,\mathbb{C})$, $SU(N)$, $O(N)$, and $SO(N)$? What are their dimensionalities as real vector spaces? Check that they are all closed under taking a commutator and multiplying by i.

2. Check the third order term in the BCH formula (2.15).

3. Confirm that the Lie bracket is preserved by a diffeomorphism $f : M \to N$ as in equation (2.22). Hint: use (2.21) and (2.18), and make sure to take into account the transformation of the partial derivative.

4. Show that the definition (2.45) of the adjoint representation matrices implies the representation condition (2.46), and also show that the adjoint Lie algebra generators (2.53) obey the Lie algebra commutator (2.43).

5. Consider the representation of the spin group $SU(2)$ on the four-dimensional Hilbert space of two spins. Is this representation irreducible? If not, then what are the invariant subspaces and what are the irreducible representations that act on them? Hint: it is easier if you think about the representation of the Lie algebra first.

6. (extra credit) There is a clever alternative way to do geometric calculations in matrix Lie groups. The idea is to view G as a submanifold within the manifold $M_N(\mathbb{C})$ of $N \times N$ complex matrices. This is convenient because we can just use the components of the matrices as coordinates. So for example the left-multiplication map is simply
$$L_{g_1}(g_2)^{ij} = g_1^{ik} g_2^{kj}. \tag{2.85}$$
Using this representation, show that if $T = -iX$ is an element of the matrix Lie algebra then the left-invariant vector field corresponding to it is
$$X^{ij}(g) = g^{ik} X^{kj}. \tag{2.86}$$
Using this expression, give alternative arguments that the Lie bracket coincides with i times the matrix commutator and the matrix exponential map coincides with the geometric one. Also show that the pushforward of the conjugation map is
$$C_{h*}(X(g)) = hX(g)h^{-1}, \tag{2.87}$$
so the matrix and geometric definitions of the adjoint representation coincide. As you can see this approach is computationally easier than the one we used in the text based on the multiplication function $\widetilde{x}^a(x_1, x_2)$, but it is conceptually more subtle since we need to make sure that everything stays within the submanifold $G \subset M_N(\mathbb{C})$.

3 Gauge fields and Lagrangians

We now return to constructing non-abelian gauge theory. The idea is that we would like to have a local symmetry group G, called the **gauge group**, under which the matter fields transform in some unitary representation U:[36]

$$\phi'_n(x) = U_n{}^m\big(g(x)\big)\phi_m(x). \tag{3.1}$$

Without loss of generality we can take this representation to be irreducible, since we can decompose U into irreducible blocks and then view the components in different blocks as different fields. For now we will be agnostic about the nature of the gauge group G, assuming only that it is a Lie group so that we can take derivatives of $g(x)$.

3.1 Non-abelian gauge fields

The first problem we run into in trying to build a quantum field theory with the local symmetry (3.1) is that, as in Maxwell theory, we run into the problem that derivatives of the matter fields do not transform nicely under gauge transformations:

$$\partial_\mu\phi' = \partial_\mu U\phi + U\partial_\mu\phi, \tag{3.2}$$

where we have suppressed the representation indices and position dependence. We can try to fix this by defining a covariant derivative

$$D_\mu\phi_n(x) \equiv \partial_\mu\phi_n(x) - iA^\phi_{\mu,n}{}^m(x)\phi_m(x), \tag{3.3}$$

or more compactly

$$D_\mu\phi \equiv \partial_\mu\phi - iA^\phi_\mu\phi. \tag{3.4}$$

Here A^ϕ_μ is a matrix-valued one-form gauge field whose matrix dimensionality is the same as that of the matter representation U. Our goal is to choose the transformation law for A^ϕ_μ so that we have

$$D'_\mu\phi' = UD_\mu\phi. \tag{3.5}$$

Writing out both sides we see that we want

$$U\partial_\mu\phi + \partial_\mu U\phi - iA^{\phi\prime}_\mu U\phi = U\partial_\mu\phi - iUA^\phi_\mu\phi, \tag{3.6}$$

which will be true provided that we take

$$\begin{aligned} A^{\phi\prime}_\mu =& UA^\phi_\mu U^{-1} - i\partial_\mu UU^{-1} \\ =& U\left(A^\phi_\mu - iU^{-1}\partial_\mu U\right)U^{-1}. \end{aligned} \tag{3.7}$$

If we take $G = U(1)$ and $U = e^{i\Omega}$ then this reduces to the Maxwell gauge transformation

$$A'_\mu = A_\mu + \partial_\mu\Omega, \tag{3.8}$$

which is an encouraging sign.

On the other hand there is something inelegant about this discussion so far. We defined A^ϕ_μ as a matrix whose dimensionality is set by the matter representation U. Does this mean that each matter field transforming in an irreducible representation of G get its own separate gauge field A^ϕ_μ? What if there are no matter fields at all? And anyways how many matrix components of A^ϕ_μ are really independent? It would be nicer if we could accommodate all possible representations of the gauge group G with a single gauge field (this is what happened in Maxwell theory, where the same Maxwell field could couple to multiple matter fields with

[36] I write U instead of D here because I will need D_μ for the covariant derivative, and also because the representation in question is always unitary although we won't yet assume that here.

different charges). The way to address all these questions is to take A_μ^ϕ to live in the representation of the Lie algebra $\mathfrak{g}$ that is induced by the representation U of $\mathfrak{g}$:

$$A_\mu^\phi(x) = A_\mu^a(x)\tau_a, \tag{3.9}$$

where τ_a is the representation of the Lie algebra generator T_a in the U representation in the sense that

$$U(e^{i\epsilon\theta^a T_a}) = e^{i\epsilon\theta^a \tau_a} \tag{3.10}$$

for arbitrary θ^a and sufficiently small ϵ. We then take the A_μ^ϕ for different matter fields in arbitrary irreducible representations of G to be built out of the same gauge field coefficients $A_\mu^a(x)$, so it is really these coefficients that we should view as the gauge field.

As evidence for this proposal we can check that the gauge transformation (3.7) indeed acts within the linear span of the τ_a (i.e. it doesn't generate anything that isn't of the form (3.9)). To see this we need to show 1) the span is preserved by conjugation by U and 2) that the quantity $\partial_\mu U U^{-1}$ also lies within the span. The argument for 1) is the same that we used in defining the adjoint representation: we note that

$$U e^{i\epsilon\theta^a\tau_a} U^{-1} = e^{i\epsilon\theta^a U\tau_a U^{-1}}, \tag{3.11}$$

and then Taylor expand in ϵ. To see 2), for x near x_0 we write

$$U(x) = e^{i\epsilon\theta^a(x)\tau_a} U(x_0). \tag{3.12}$$

Differentiating with respect to x and then setting $x = x_0$ we have

$$\partial_\mu U U^{-1}(x_0) = i\epsilon\partial_\mu\theta^a(x_0)\tau_a, \tag{3.13}$$

which is indeed in the span of the τ_a. Specifying a gauge field configuration is thus the same as specifying a one-form whose components take values in the Lie algebra, which we can write as

$$A_\mu \equiv A_\mu^a T_a. \tag{3.14}$$

This is often called a **Lie algebra-valued one-form**. It is important to clearly distinguish A_μ and A_μ^ϕ: the former lives in $\mathfrak{g}$, while the latter lives in the representation of $\mathfrak{g}$ that is induced by the matter representation U. You should think of A_μ as the definition of the gauge field and A_μ^ϕ as a quantity that is built from it for use in matter covariant derivatives.

When G is a matrix group, the gauge transformation of A_μ is easy to describe:

$$A'_\mu = g\left(A_\mu - ig^{-1}\partial_\mu g\right)g^{-1}. \tag{3.15}$$

In particular if g is constant then A_μ transforms in the adjoint representation of G. We need to check that this transformation law implies the transformation (3.7) for A_μ^ϕ. We first note that

$$e^{i\epsilon\theta^a U(g)\tau_a U(g^{-1})} = U(g)e^{i\epsilon\theta^a\tau_a}U(g^{-1}) = U\left(ge^{i\epsilon\theta^a T_a}g^{-1}\right) = U\left(e^{i\epsilon\theta^a gT_a g^{-1}}\right), \tag{3.16}$$

so conjugation of the τ_a matrices by $U(g)$ gives the same adjoint transformation as conjugation of the T_a matrices by g. For the inhomegeneous term in the gauge transformations, near $x = x_0$ we can write

$$g(x) = g(x_0)e^{i\epsilon\theta^a(x)T_a}, \tag{3.17}$$

from which we find

$$g^{-1}\partial_\mu g(x_0) = i\epsilon\partial_\mu\theta^a(x_0)T_a. \tag{3.18}$$

Observing that

$$U(g(x)) = U(g(x_0))e^{i\epsilon\theta^a(x)\tau_a}, \tag{3.19}$$

we thus have

$$U^{-1}\partial_\mu U(x_0) = i\epsilon\partial_\mu\theta^a(x_0)\tau_a. \tag{3.20}$$

Thus the gauge transformation (3.15) for A_μ indeed implies the transformation (3.7) for A_μ^ϕ. The same result can be shown using geometric tools not assuming G is a matrix group, but I won't bother since we are about to see that it is only the matrix case which is interesting.

3.2 Field strength tensor

In Maxwell theory it was a good idea to define a gauge-invariant field strength tensor $F_{\mu\nu}$, in particular for use in constructing a gauge-invariant Lagrangian. We can similarly try to define a field strength for general gauge group G, but the more complicated transformation rule (3.15) means that we can't just use the $U(1)$ formula $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. The right non-abelian field strength is

$$F_{\mu\nu} \equiv \partial_\mu A_\nu - \partial_\nu A_\mu - i[A_\mu, A_\nu], \tag{3.21}$$

with the extra term leading to the nice gauge transformation

$$F'_{\mu\nu} = g F_{\mu\nu} g^{-1}, \tag{3.22}$$

as you will check on the homework. In other words the field strength transforms in the adjoint representation of G. It is sometimes convenient to also have an expression for the Lie algebra coefficients of $F_{\mu\nu}$:

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + C^a_{bc} A^b_\mu A^c_\nu. \tag{3.23}$$

It is worth emphasizing that $F_{\mu\nu}$ is gauge-invariant if and only if the adjoint representation of G is trivial. When G is connected this happens if and only if G is abelian, as it was in Maxwell theory. Otherwise to get a gauge-invariant operator local operator we need to consider something like

$$\mathrm{Tr}\left(F_{\mu\nu} F^{\mu\nu}\right). \tag{3.24}$$

3.3 Lagrangian

The local operator just mentioned looks like a good candidate for a term in the Lagrangian of a non-abelian gauge field, but to be open-minded let's try a more general quadratic action:

$$\mathcal{L} = -\frac{1}{2} g_{ab} F^a_{\mu\nu} F^{b\,\mu\nu}, \tag{3.25}$$

where g_{ab} is some symmetric matrix. There are two conditions we'd this matrix to obey:

(1) It should be gauge-invariant, meaning that

$$D^a{}_c(g) D^b{}_d(g) g_{ab} = g_{cd} \tag{3.26}$$

for arbitrary $g \in G$. Here $D^a{}_b(g)$ are the adjoint representation matrices. This is to ensure the Lagrangian is gauge-invariant.

(2) It should be positive, meaning that $g_{ab} X^a X^b > 0$ for all nonzero vectors X^a. This is to ensure that the Hamiltonian we define shortly is bounded from below.

Happily the existence of a matrix g_{ab} obeying these two conditions is precisely condition 2 of theorem 10 from the previous section, so we learn that the Lie algebra $\mathfrak{g}$ is a direct sum of an abelian Lie algebra $\mathfrak{z}$ and a semisimple Lie algebra $\mathfrak{s}$, and also that $\mathfrak{g}$ must be the Lie algebra of some compact Lie group $\hat{G}$. Somewhat annoyingly we cannot conclude from this that the gauge group G is itself compact, for example Maxwell theory makes perfect sense with a noncompact gauge group $G = \mathbb{R}$, whose Lie algebra coincides with that of the compact Lie group $\hat{G} = U(1)$. Moreover G could be disconnected, in which case the relationship between different connected components is not constrained by the Lie algebra. In general the most we can say is that the gauge group G must be of **compact type**, which just means that its Lie algebra coincides with that of some compact $\hat{G}$. If we restrict to connected G however, then the discussion after theorem 10 last time shows that G must be a discrete central quotient of the product of an abelian group $\mathbb{R}^n$ with some compact, simple, and simply-connected factors S_i. Thus any noncompactess is fundamentally either discrete or abelian in nature. It would still be a bother to keep track of this possibility however, and fortunately there is fairly

strong evidence that noncompact gauge groups of any kind are not compatible with quantum gravity (see my long paper with Ooguri). Moreover empirically the quantization of charge gives strong evidence that the Maxwell gauge group is in fact $U(1)$. From now on we will therefore simply assume that the gauge group G is compact. In particular by theorem 9 from the previous section, this means that G is always a matrix group.[37]

To get some intuition for the meaning of g_{ab}, let's first consider the Maxwell case $G = U(1)$. g_{ab} is then a one-dimensional matrix, otherwise known as a number g_{11}. Comparing (3.25) to the Maxwell Lagrangian you might be tempted to just set $g_{11} = 1/2$, but we need to be a bit careful about how we normalize the gauge field. We introduced the gauge field by way of the covariant derivative

$$D_\mu = \partial_\mu - iA^a_\mu \tau_a, \tag{3.27}$$

so its normalization is related to our choice of normalization for the Lie algebra basis T_a (once we have chosen this the normalization of τ_a in any representation is determined by (3.10)). For $G = U(1)$ the most obvious way to parametrize group elements is as

$$g = e^{i\theta}, \tag{3.28}$$

with $\theta \in [0, 2\pi)$. With this parameterization we have $T_1 = 1$, and the irreducible representations of G are

$$U_n(\theta) = e^{in\theta}, \tag{3.29}$$

with $n \in \mathbb{Z}$. The Lie algebra representations are thus

$$\tau_1^{\{n\}} = n, \tag{3.30}$$

where n is of course just telling us the electric charge of the representation in units of the fundamental charge e. But where is this fundamental charge in our discussion so far? After all this is what determines the fine structure constant α, which is the coupling constant of the theory. We have two ways of incorporating it. The first is what we implicitly did last semester, which is to instead normalize the Lie algebra generators so that

$$\tau_1^{\{n\}} = ne. \tag{3.31}$$

In the group this amounts to instead using a parameterization[38]

$$g = e^{ie\theta}, \tag{3.32}$$

where now $\theta \in [0, 2\pi/e]$. In this convention the factor of e is where you'd expect it to be in the covariant derivative,

$$D_\mu = \partial_\mu - ieA_\mu, \tag{3.33}$$

and we are free to set $g_{11} = 1/2$ to get Lagrangian

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu}. \tag{3.34}$$

From a logical point of view however it is arguably more natural to choose a group parametrization which is independent of the coupling constant, in which case the covariant derivative is simply

$$D_\mu = \partial_\mu - iA_\mu. \tag{3.35}$$

We would still like the theory to be equivalent to the previous one by a field redefinition. The only possibility is rescaling A_μ by a factor of e, which converts this covariant derivative to the previous one but also introduces

[37] If you are not happy with assuming compactness, I note that it is enough to assume that G has finitely many connected components. This condition, together with G being of compact type, are enough to establish that G is a matrix group.

[38] Sorry about the two meanings of e in this equation, downstairs it is 2.71828... while upstairs it is the proton charge.

a factor of e^2 in the kinetic term. To cancel this and get (3.34) there needed to be a factor of $1/e^2$ in the kinetic term, so when the covariant derivative is (3.35) the Maxwell Lagrangian is

$$\mathcal{L} = -\frac{1}{4e^2} F_{\mu\nu} F^{\mu\nu}. \tag{3.36}$$

Thus we see that in this convention we have $g_{11} = \frac{1}{2e^2}$, which is the main lesson of this discussion: if we fix our parametrization of the group in a coupling-independent way, and we want to preserve our expression (3.27) for the covariant derivative, then the gauge coupling constant goes into the matrix g_{ab}.

Let's now see how this works for more general compact G. We will first assume that $\mathfrak{g}$ is a simple Lie algebra, since the general case will be easy to deal with once we understand the simple case. Since G is compact it is a matrix group, and moreover it is unitary so we can take its generators T_a to be hermitian. We therefore have an immediate candidate for g_{ab},

$$g_{ab} \propto \mathrm{Tr}\left(T_a T_b\right). \tag{3.37}$$

We showed this obeys our two requirements of invariance and positivity in the previous section. We will now argue that any other choice for g_{ab} must be a scalar multiple of this one. As you can guess, the argument uses Schur's lemma. Using matrix notation, let's say that g and g' both obey the invariance condition,

$$\begin{aligned} D^T g D &= g \\ D^T g' D &= g', \end{aligned} \tag{3.38}$$

where D is any adjoint representation matrix. Using that g is non-degenerate we can multiply both of these equations on the left by g^{-1} and then combine them to find

$$g^{-1} g' D = D g^{-1} g', \tag{3.39}$$

so $g^{-1}g'$ is a matrix that commmutes with the adjoint representation. If we can show that the adjoint representation is irreducible, by Schur's lemma we therefore will have shown that

$$g' = \lambda g. \tag{3.40}$$

To show irreducibility, suppose that $\mathfrak{h} \subset \mathfrak{g}$ is an invariant subspace under the adjoint representation of G,

$$\hat{g}\mathfrak{h}\hat{g}^{-1} = \mathfrak{h} \qquad \forall \hat{g} \in G. \tag{3.41}$$

Working near the identity, this shows that the commutator of anything in $\mathfrak{h}$ with anything in $\mathfrak{g}$ must also be in $\mathfrak{h}$, or in other words that $\mathfrak{h}$ is an ideal of $\mathfrak{g}$. Since $\mathfrak{g}$ is simple this ideal must either be equal to zero or $\mathfrak{g}$, so indeed we have an irreducible representation. This completes the proof, so the most general invariant inner product on $\mathfrak{g}$ is of the form

$$g_{ab} = \frac{1}{g_{YM}^2} \mathrm{Tr}\left(T_a T_b\right), \tag{3.42}$$

where g_{YM} is an arbitrary parameter that is called the **gauge coupling**.[39] We at last can now write down the Lagrangian and gauge transformation for a non-abelian gauge field with simple gauge group G:

$$\begin{aligned} \mathcal{L} =& -\frac{1}{2g_{YM}^2} \mathrm{Tr}\left(F_{\mu\nu} F^{\mu\nu}\right) \\ F_{\mu\nu} =& \partial_\mu A_\nu - \partial_\nu A_\mu - i[A_\mu, A_\nu] \\ A'_\mu =& g\left(A_\mu - i g^{-1} \partial_\mu g\right) g^{-1} \\ D_\mu =& \partial_\mu - i A^a_\mu \tau_a. \end{aligned} \tag{3.43}$$

[39] I deeply apologize for the letter g appearing as a group element g, the matrix g_{ab}, and the gauge coupling g_{YM}. The blame for this is 100% on our elders, who clearly were lacking in basic common sense. Fortunately g_{ab} will not appear again after this section, and it will hopefully always be clear from the context whether we mean a group element or a gauge coupling (if not we will write g_{YM} for the coupling as in this section).

This Lagrangian was first written down by Yang and Mills in 1954, and non-abelian gauge theory is often called **Yang-Mills theory**.[40]

An essential point here is that this Lagrangian is *not* quadratic in the fields A_μ^a, due to the nonlinear term in the field strength (3.21). Unlike in the abelian case, a non-abelian gauge field interacts with itself! There is both a three-point and a four-point vertex, as we will see later when we discuss the Feynman rules. This nonlinearity also means that if we wish to rescale the gauge field to put g_{YM} into the covariant derivative instead in front of $\mathrm{Tr}\,(F^{\mu\nu}F_{\mu\nu})$, then it will also show up in the nonlinear term in the field strength and the gauge transformation (3.15). Thus the strength of matter interactions and the self-interactions of the gauge field are both controlled by the gauge coupling g_{YM}. With this choice of normalization we have

$$\begin{aligned}\mathcal{L} =& -\frac{1}{2}\mathrm{Tr}\,(F_{\mu\nu}F^{\mu\nu})\\ F_{\mu\nu} =& \partial_\mu A_\nu - \partial_\nu A_\mu - ig_{YM}[A_\mu, A_\nu]\\ A'_\mu =& g\left(A_\mu - ig_{YM}^{-1} g^{-1}\partial_\mu g\right) g^{-1}\\ D_\mu =& \partial_\mu - ig_{YM}A_\mu^a \tau_a.\end{aligned} \tag{3.44}$$

It is up to you whether you prefer convention (3.43) or the convention (3.44), both are common in the literature. (3.43) better respects the mathematical structure of the theory, while (3.44) makes the perturbative expansion in g_{YM} more manifest. As you might guess, we will mostly stick with (3.43).

Returning now to the more general case where the Lie algebra is a direct sum of abelian and simple factors, there is a separate gauge coupling for each such factor. The usual way to deal with this is to define a separate gauge field for each factor, and then give each of them its own kinetic term and gauge coupling in the action:

$$\mathcal{L} = -\sum_p \frac{1}{2g_p^2}\mathrm{Tr}_p\left(F_{\mu\nu}^{\{p\}}F^{\{p\},\mu\nu}\right), \tag{3.45}$$

where here p runs over the simple and $U(1)$ factors in $\mathfrak{g}$ and the trace happens within each factor. Nonperturbatively however we do have to be careful about the possibility of a discrete central quotient that mixes the different factors in G, which affects the allowed matter representations and also the possible gauge field configurations when the topology of spacetime is nontrivial. We will discuss both of these issues later.

The theory we have constructed makes sense for any compact gauge group G. In particular we have not required that G be connected. In fact G could be a finite group. In that case however the Lie algebra is trivial, so the theory as we have constructed it so far has zero degrees of freedom. Once we discuss how to define gauge fields on spacetime manifolds other than $\mathbb{R}^d$ however, we will see that a gauge field with finite G does in general have genuine degrees of freedom.

3.4 Including matter

Coupling our gauge fields to matter is straightforward since we have already introduced the covariant derivative D_μ. For example if we have a complex scalar field Φ_m transforming in an irreducible representation of a simple gauge group G with generators $(\tau_a)_{nm}$, then we can have a gauge theory with Lagrangian

$$\mathcal{L} = -\frac{1}{2g^2}\mathrm{Tr}\,(F_{\mu\nu}F^{\mu\nu}) - D_\mu\Phi^\dagger D^\mu\Phi - m\Phi^\dagger\Phi, \tag{3.46}$$

where we adopt the convenient convention that $\dagger$ takes the transpose of the representation index on ϕ. So for example

$$\Phi^\dagger\Phi = \Phi_m^*\Phi_m, \tag{3.47}$$

[40] An interesting anecdote about the genesis of Yang-Mills theory is the following. When it was presented by Yang at the Institute for Advanced Study he was criticized by Pauli, who asked where are the massless gauge bosons created by A_μ^a. Yang and Mills added a sentence to their paper acknowledging that they couldn't answer this question, and they submitted it anyways. Only much later was it understood that the answer to this question is that either 1) they are confined into hadrons by strong coupling or 2) they are Higgsed and massive. Sometimes you just have to put out the paper even when you don't have all the answers!

where $*$ indicates the Hilbert space adjoint. This Lagrangian is indeed invariant under

$$\begin{aligned} A'_\mu &= g\left(A_\mu - ig^{-1}\partial_\mu g\right)g^{-1} \\ \Phi' &= U(g)\Phi \end{aligned} \tag{3.48}$$

due to the covariant derivative transformation (3.5) and the unitarity of U. As in scalar QED, this theory has a three-point vertex with two scalars and one gauge boson, as well as a four point vertex with two scalars and two gauge bosons. Similarly we can include a spinor field Ψ_n transforming in some irrep of G by including a covariant derivative in the Dirac Lagrangian:

$$\mathcal{L} = -\frac{1}{2g^2}\mathrm{Tr}\left(F_{\mu\nu}F^{\mu\nu}\right) - i\overline{\Psi}\left(\not{D} + m\right)\Psi, \tag{3.49}$$

where we have again suppressed the representation index n and the $\Psi^\dagger$ in $\overline{\Psi}$ is understood to include a transpose that acts on this index (as well as the spinor index as usual).

3.5 Working in components

The Lagrangian (3.43) (or (3.44)) is valid for any choice of basis T_a for the Lie algebra. In calculations however it is sometimes convenient to pick a basis that diagonalizes the matrix $\mathrm{Tr}\left(T_a T_b\right)$, and the (somewhat mysterious) standard convention is to do this so that

$$\mathrm{Tr}\left(T_a T_b\right) = \frac{1}{2}\delta_{ab}. \tag{3.50}$$

In this basis we simply have

$$g_{ab} = \frac{1}{2g_{YM}^2}\delta_{ab}, \tag{3.51}$$

so we can also write the Yang-Mills Lagrangian as

$$\mathcal{L} = -\frac{1}{4g_{YM}^2}F^a_{\mu\nu}F^{a,\mu\nu}. \tag{3.52}$$

Indeed in this basis there is no reason to distinguish between raised and lowered adjoint indices. This has a nice consequence for the structure constants. In any basis, from the infinitesimal version of equation (3.26) we have

$$g_{cd}C^d{}_{ab} = -g_{ad}C^d{}_{cb}. \tag{3.53}$$

In a basis where (3.50) holds, this equation says that the structure constants are completely antisymmetric in all three indices (by definition they are antisymmetric under exchanging the second and third indices). So indeed it is more natural to write them as C_{abc}. Next time we will adopt an "intermediate" notation where

$$\mathrm{Tr}\left(T_a T_b\right) = \frac{1}{2}\hat{g}_{ab}, \tag{3.54}$$

where adjoint indices are raised and lowered using $\hat{g}_{ab}$. This allows us to write

$$C_{abc} = -C_{bac} \tag{3.55}$$

without committing to any particular basis.

3.6 Quantum chromodynamics

We are now at last in a position to write down the Lagrangian for **quantum chromodynamics**, or QCD for short, which is the modern theory of the strong nuclear force. This is a non-abelian gauge theory with gauge group $G = SU(3)$ and six spinor fields transforming in the three-dimensional defining representation of $SU(3)$, usually referred to as the **fundamental representation**. The gauge bosons are called **gluons** and the fermions are called **quarks**. The Lagrangian is

$$\mathcal{L} = -\frac{1}{2g^2}\mathrm{Tr}\left(F_{\mu\nu}F^{\mu\nu}\right) - i\sum_{i=1}^{6}\overline{\Psi}_i\left(\not{D} + m_i\right)\Psi_i, \tag{3.56}$$

where the i index here runs over the six flavors of quark that we discussed in the first section: up, down, charm, strange, top, and bottom. As promised in the first section the action is invariant under a global $U(1)^6$ **flavor symmetry** that rotates the phase of each quark field independently, so the number of ups, downs, etc must be conserved in any QCD scattering process (again with the caveat that antiquarks count with opposite sign). In particular this includes the baryon number symmetry that rotates all of the quark phases together. It is also invariant under $\mathcal{C}$, $\mathcal{R}$, and $\mathcal{T}$ symmetries by essentially the same calculations you did in the homework earlier this semester. Since the up and down quark masses are approximately equal compared to the other quark masses, there is also an approximate global symmetry called **isospin symmetry** that mixes the up and down quark fields with an $SU(2)$ transformation.

The QCD Lagrangian may look simple, but the dynamics it gives rise to are quite complex. In fact we will spend quite a bit of the remaining semester understanding some of the remarkable phenomena it leads to. For now we will just make one remaining comment, which is that there is actually another term we could include in the Lagrangian if we are willing to violate $\mathcal{T}$ symmetry:

$$\Delta\mathcal{L} = \frac{\theta}{32\pi^2}\epsilon^{\mu\nu\alpha\beta}\mathrm{Tr}\left(F_{\mu\nu}F_{\alpha\beta}\right). \tag{3.57}$$

Here θ is a dimensionless parameter that is actually periodic with periodicity 2π, so it is called the θ-angle. This term is actually a total derivative, as you will check on the homework, so it doesn't affect the equations of motion. It does still have interesting non-perturbative effects that we will study later.

3.7 Maximally supersymmetric gauge theory

Another important gauge theory is the maximally supersymmetric gauge theory which is dual to quantum gravity in $AdS_5 \times \mathbb{S}^5$. The Lagrangian of this theory is

$$\mathcal{L} = -\frac{1}{g^2}\mathrm{Tr}\left(\frac{1}{2}F_{\mu\nu}F^{\mu\nu} + D_\mu\Phi^I D^\mu\Phi^I + i\overline{\Psi}^A\gamma^\mu D_\mu\Psi_A - \overline{\Psi}^A\left(\left(\Sigma^I\right)_A{}^B P_L + \left(\tilde{\Sigma}^I\right)_A{}^B P_R\right)[\Phi^I,\Psi_B] - \frac{1}{2}[\Phi^I,\Phi^J][\Phi^I,\Phi^J]\right), \tag{3.58}$$

where $I = 1,2,\ldots,6$, $A = 1,2,3,4$, Φ^I are six real scalar fields transforming in the adjoint representation of G, and Ψ_A are four Majorana fermions also transforming in the adjoint representation. $P_L = (1+\gamma)/2$ and $P_R = (1-\gamma)/2$ are the projections onto left and right Weyl spinors, and the quantities $\left(\Sigma^I\right)_A{}^B$ and $\left(\tilde{\Sigma}^I\right)_A{}^B$ are constructed from from the ten-dimensional γ-matrices, which arise from the easiest construction of this Lagrangian as a dimensional reduction of the maximally supersymmetric Yang-Mills theory in $9+1$ spacetime dimensions. The first three terms here are the usual gauge, scalar, and Majorana spinor kinetic terms, while the last two are Yukawa and ϕ^4 type interactions. We will not need the details of this theory in this class, the point I want to make is that we now have all we need to write it down.

3.8 Homework

1. Show that the non-abelian field strength (3.21) has the gauge transformation (3.22).

2. Find bases for the Lie algebras of $SU(2)$ and $SU(3)$ that obey the normalization condition (3.50). Also show that the Lie algebra of $SL(2,\mathbb{R})$ does *not* admit such a normalization.

3. Starting from the Lagrangian
$$\mathcal{L} = -\frac{1}{4g^2} F^a_{\mu\nu} F^{a,\mu\nu} + J^{a,\mu} A^a_\mu, \tag{3.59}$$
show that the equation of motion is
$$D_\mu F^{a,\nu\mu} = g^2 J^{a,\nu} \tag{3.60}$$
where D_μ is the covariant derivative in the adjoint representation.

4. Show that the Lagrangian term (3.57) is a total derivative, meaning it is the divergence of a vector field. Hint: the vector field is a linear combination of $\epsilon^{\mu\nu\alpha\beta}\mathrm{Tr}\,(A_\nu F_{\alpha\beta})$ and $\epsilon^{\mu\nu\alpha\beta}\mathrm{Tr}\,(A_\nu A_\alpha A_\beta)$.

4 Hamiltonian formulation and quantization of Yang-Mills theory

4.1 Lagrangian preliminaries

We've now constructed the Lagrangian for Yang-Mills theory with any compact gauge group G. We can write it as

$$\mathcal{L} = -\frac{1}{2g^2}\mathrm{Tr}\left(F_{\mu\nu}F^{\mu\nu}\right) + \mathcal{L}_M\left(\Phi, D\Phi\right), \tag{4.1}$$

where Φ_n are some matter fields that transform in a (possibly reducible) representation of G with generators τ_a. We have dropped the subscript "YM" on the gauge coupling g since we do not anticipate confusion with elements of the gauge group. Calculations are a bit easier if we work in component fields, which we will do by introducing a notation

$$\mathrm{Tr}\left(T_a T_b\right) = \frac{1}{2}\hat{g}_{ab} \tag{4.2}$$

and then using $\hat{g}_{ab}$ to raise and lower adjoint indices $a, b, \ldots$. As mentioned last time we could go further and choose a basis where $\hat{g}_{ab} = \delta_{ab}$, in which case we would not need to distinguish between raised and lowered indices, but I prefer not to do this since otherwise it is not clear which quantities depend on this choice and which do not.[41] It is also a useful check that valid index contractions always involve one upper and one lower index. We can thus write the action as

$$\mathcal{L} = -\frac{1}{4g^2}F^a_{\mu\nu}F^{\mu\nu}_a + \mathcal{L}_M\left(\Phi, D\Phi\right). \tag{4.3}$$

In the previous homework you showed that the equation of motion for the gauge field is

$$D_\mu F^{\nu\mu}_a = g^2 J^\nu_a, \tag{4.4}$$

where

$$\begin{aligned} J^\mu_a &\equiv \frac{\partial \mathcal{L}_M}{\partial A^a_\mu(x)} \\ &= \frac{\partial \mathcal{L}_M}{\partial D_\nu \Phi_n}\frac{\partial D_\nu \Phi_n}{\partial A^a_\mu} \\ &= -i\frac{\partial \mathcal{L}_M}{\partial D_\mu \Phi_n}(\tau_a)_n{}^m \Phi_m. \end{aligned} \tag{4.5}$$

For example if the matter is a spinor field Ψ transforming in some representation with generators τ_a, we have

$$J^\mu_a = -\overline{\Psi}\gamma^\mu \tau_a \Psi. \tag{4.6}$$

You will show in the homework that the gauge invariance of $\mathcal{L}_M$ requires the current to be covariantly conserved:

$$D_\mu J^\mu_a \equiv \partial_\mu J^\mu_a - C^c{}_{ba} A^b_\mu J^\mu_c = 0 \tag{4.7}$$

Using the invariance condition

$$C_{abc} = \hat{g}_{ad}C^d{}_{bc} = -\hat{g}_{bd}C^d{}_{ac} = -C_{bac} \tag{4.8}$$

we can also write this with the adjoint index up,

$$D_\mu J^{a,\mu} \equiv \partial_\mu J^{a,\mu} + C^a{}_{bc} A^b_\mu J^{c,\mu} = 0, \tag{4.9}$$

which shows the consistency with our adjoint generator formula

$$\left(\tau^A_b\right)^a{}_c = iC^a{}_{bc}. \tag{4.10}$$

[41] Note that A^a_μ and $F^a_{\mu\nu}$ are defined with the adjoint index up, while J^μ_a is defined with the adjoint index down. So if they appear with the index in the other position, as in (4.4), then $\hat{g}$ or its inverse has been used.

If we define a matrix-valued current $J^\mu \equiv J^{a\mu}T_a$, then we can write the equation of motion and current conservation equations more elegantly as

$$\begin{aligned} D_\mu F^{\nu\mu} &= \partial_\mu F^{\nu\mu} - i[A_\mu, F^{\nu\mu}] = g^2 J^\mu \\ D_\mu J^\mu &= \partial_\mu J^\mu - i[A_\mu, J^\mu] = 0. \end{aligned} \tag{4.11}$$

4.2 Boundary conditions and gauge transformations

In computing the variation of the action

$$S = \int d^d x \mathcal{L}, \tag{4.12}$$

a boundary term is generated of the form

$$\delta S \supset -\frac{1}{g^2}\int_\Gamma d^{d-1}x\, n_\mu F_a^{\mu\nu}\delta A_\nu^a. \tag{4.13}$$

Here Γ is the spatial boundary, consisting for example of a spatial sphere of large radius together with time. In order for the action to be stationary up to future/past boundary terms we therefore need to choose boundary conditions so that this term vanishes. The simplest choice is to require that

$$A_\nu^a t^\nu\big|_\Gamma = 0, \tag{4.14}$$

where t^ν is any vector that is tangent to Γ. In more geometric language, the pullback of the one-form A^a to Γ should vanish. To preserve these boundary conditions we should also restrict gauge transformations $g(x)$ to those which become constant on Γ. As in Maxwell theory, we will soon see that any gauge transformation that approaches the identity on Γ needs to be viewed as redundancies of the theory. Otherwise the initial value formulation of the theory is not well-posed: specifying A_μ^a and $\dot{A}_\mu^a$ on a time slice is not sufficient to determine them elsewhere.

4.3 Hamiltonian formulation

Let's now construct the Hamiltonian. The canonical momentum for the gauge field is

$$\Pi_a^\mu \equiv \frac{\partial \mathcal{L}}{\partial \dot{A}_\mu^a} = \frac{1}{g^2}F_a^{\mu 0}, \tag{4.15}$$

and using our expression for $F_{\mu\nu}^a$ we have

$$\begin{aligned} \dot{A}_\mu^a &= F_{0\mu}^a + \partial_\mu A_0^a - C^a{}_{bc}A_0^b A_\mu^c \\ &= g^2\Pi_\mu^a + D_\mu A_0^a. \end{aligned} \tag{4.16}$$

We note in particular that

$$\Pi_a^0 = 0, \tag{4.17}$$

which is thus a constraint just as it was in Maxwell theory. We therefore can write the Hamiltonian density as

$$\begin{aligned} \mathcal{H} &= \Pi_a^0\dot{A}_0^a + \Pi_a^i\dot{A}_i^a + \frac{1}{4g^2}F_a^{\mu\nu}F_{\mu\nu}^a + \mathcal{H}_M \\ &= \Pi_a^0\dot{A}_0^a + \Pi_a^i\left(g^2\Pi_i^a + D_i A_0^a\right) - \frac{g^2}{2}\Pi_a^i\Pi_i^a + \frac{1}{4g^2}F_a^{ij}F_{ij}^a + \mathcal{H}_M \\ &= \frac{g^2}{2}\Pi_a^i\Pi_i^a + \frac{1}{4g^2}F_a^{ij}F_{ij}^a + \mathcal{H}_M - A_0^a D_i\Pi_a^i + \Pi_a^0\dot{A}_0^a + \partial_i\left(\Pi_a^i A_0^a\right). \end{aligned} \tag{4.18}$$

Here $\mathcal{H}_M$ is the Hamiltonian derived from the matter Lagrangian. The last term is a total derivative that vanishes when we integrate to get the Hamiltonian and impose the boundary conditions (4.14), and the second to last term vanishes when we impose the Π^0_a constraint (4.17). We would like to interpret the fourth term as being part of a Gauss constraint

$$D_i \Pi^i_a + J^0_a = 0, \tag{4.19}$$

which is the $\nu = 0$ component of the equation of motion (4.4), but to see this we need to deal with the fact that $\mathcal{H}_M$ depends on A_0 through the covariant derivative. More explicitly we have

$$\mathcal{H}_M = \Pi^m \dot{\Phi}_m - \mathcal{L}_M, \tag{4.20}$$

with

$$\Pi^m \equiv \frac{\partial \mathcal{L}_M}{\partial \dot{\Phi}_m} = \frac{\partial \mathcal{L}_M}{D_0 \Phi_m}. \tag{4.21}$$

We can solve this equation for $D_0\Phi_m$, which thus is a function of Φ_m, Π^m, and $D_i\Phi_m$ but *not* of A^a_0. Therefore the only A_0 dependence of $\mathcal{H}_M$ comes through $\dot{\Phi}_m$, so we have

$$\begin{aligned} \mathcal{H}_M &= \Pi^m \left(D_0 \Phi_m + i A^a_0 (\tau_a)_m{}^n \Phi_n \right) - \mathcal{L}_M \\ &= -A^a_0 J^0_a + \hat{\mathcal{H}}_M, \end{aligned} \tag{4.22}$$

where we have used (4.5) and defined

$$\hat{\mathcal{H}}_M = \Pi^m D_0 \Phi_m - \mathcal{L}_M, \tag{4.23}$$

which depends only on Φ_m, $D_i\Phi_m$, and Π^m. For example for a spinor field coupled to the gauge field we have

$$\hat{\mathcal{H}}_M = i\overline{\Psi}\gamma^i D_i \Psi. \tag{4.24}$$

Returning to the full Hamiltonian, we thus have

$$H = \int d^{d-1}x \Big[\frac{g^2}{2} \Pi^i_a \Pi^a_i + \frac{1}{4g^2} F^{ij}_a F^a_{ij} + \hat{\mathcal{H}}_M - A^a_0 \left(D_i \Pi^i_a + J^0_a \right) + \Pi^0_a \dot{A}^a_0 \Big]. \tag{4.25}$$

Note that all dependence on A^a_0 and $\dot{A}^a_0$ is multiplying the constraints. We can view these constraints as generators of the gauge transformations which vanish at spatial infinity, just as in Maxwell theory, and as in Maxwell theory we have two options for dealing with them:

- **Gauge-fixing:** impose some additional condition on the gauge field to remove the gauge redundancy, after which we can view A^a_μ as physical.
- **Invariant quotient:** Begin with a "big" Hilbert space $\mathcal{H}_{big}$ of wave functionals of all components of A^a_μ and Φ_m, and then define the "physical" Hilbert space $\mathcal{H}_{phys}$ as the subspace of $\mathcal{H}_{big}$ annihilated by the constraints (4.17) and (4.19).

The first approach is more common in QFT textbooks, but it is rather confusing since it destroys manifest locality and often also manifest Lorentz invariance. Thus in our discussion of Maxwell theory we treated the second approach as fundamental and derived the first within it. We will follow the same logic here. Thus we define the physical Hilbert space to be

$$\mathcal{H}_{phys} \equiv \Big\{ |\psi\rangle \in \mathcal{H}_{big} \,\Big|\, \Pi^0_a(\vec{x})|\psi\rangle = \Big(D_i \Pi^i_a(\vec{x}) + J^0_a(\vec{x}) \Big) |\psi\rangle = 0 \Big\}, \tag{4.26}$$

and acting on $\mathcal{H}_{phys}$ the Hamiltonian is simply

$$H = \int d^{d-1}x \Big[\frac{g^2}{2} \Pi^i_a \Pi^a_i + \frac{1}{4g^2} F^{ij}_a F^a_{ij} + \hat{\mathcal{H}}_M \Big]. \tag{4.27}$$

I emphasize that on $\mathcal{H}_{big}$, A^a_μ and Π^μ_a are well-defined operators obeying the usual commutation relations

$$[A^a_\mu(\vec{x}), \Pi^\nu_b(\vec{y})] = i\delta^\nu_\mu \delta^a_b \delta^{d-1}(\vec{x}-\vec{y}). \tag{4.28}$$

The two constraints commute with each other since the Gauss constraint (expressed in terms of Φ_m and Π^m) does not involve A^a_0, and with some work one can show that the different components of the Gauss constraint have mutual commutators that are again proportional to the constraints (this is what it means for them to be first-class, and it is a consequence of the fact that the constraints are the generators of gauge transformations so they form an infinite-dimensional Lie algebra). Thus it is consistent to set them all to zero. Imposing the $\Pi^0_a = 0$ constraint is easy: we only consider wave functionals that are independent of A^a_0. It is worth emphasizing that the Hamiltonian is only well-defined once we impose this constraint: otherwise it depends on the quantity $\dot{A}^a_0$ that has no definition as an operator on $\mathcal{H}_{big}$.[42]

4.4 Gauge-invariant operators

Since the constraints are the generators of the gauge transformations which vanish at spatial infinity, the physical Hilbert space $\mathcal{H}_{phys}$ is often called the gauge-invariant Hilbert space. This is not quite correct, since as we saw in Maxwell theory it is *not* invariant under gauge transformations that approach a nonzero constant at infinity, but the terminology is standard so we will use it.[43] Similarly we say that an operator is gauge-invariant if it is invariant under all gauge transformations that vanish at spatial infinity. These are operators that are well-defined on $\mathcal{H}_{phys}$, and thus the only candidates for physical observables.

The simplest gauge-invariant operators are local operators such as $\Phi^\dagger\Phi, \mathrm{Tr}\,(F_{\alpha\beta}F_{\gamma\delta}), \overline{\Psi}D_\mu\Psi$, etc. In Maxwell theory however we found we could also define interesting non-local gauge-invariant operators by using the **Wilson line**

$$W_q(C) \equiv e^{iq\int_C A\cdot dx}, \tag{4.29}$$

where C is a curve from a point x_i to a point x_f in spacetime and $q = ne$ is a possible charge. In Maxwell theory under a gauge transformation

$$A'_\mu = A_\mu + \partial_\mu\Omega \tag{4.30}$$

the Wilson line has gauge transformation

$$W_q(C) = e^{iq(\Omega(x_f)-\Omega(x_i))}W_q(C), \tag{4.31}$$

so if Φ is a field of charge q then the operator

$$\Phi^\dagger(x_f)W_q(C)\Phi(x_i) \tag{4.32}$$

is gauge-invariant. We also defined gauge-invariant operators

$$\widetilde{\Phi}_C(x) = W_q(C)\Phi(x) \tag{4.33}$$

with C a curve going from x to spatial infinity. $\widetilde{\Phi}_C(x)$ is gauge-invariant because the gauge transformations we view as redundancies vanish at infinity. Finally if $x_i = x_f$ then $W_q(C)$ is itself gauge-invariant, and in this case it is referred to as a **Wilson loop**.

Returning now to general compact G, we would like to again define a Wilson line for use in constructing nonlocal gauge-invariant operators. The definition is more complicated than before however, since A is now a matrix and the matrices at different points on the path C do not need to commute. This is similar to the problem of defining the time-evolution operator for a time-dependent Hamiltonian in quantum mechanics,

[42] We could try to define it on $\mathcal{H}_{big}$ by omitting this term, but then it would commute with A^a_0 so it wouldn't generate the right equations of motion.

[43] In QCD states which transform nontrivially under constant gauge transformations actually carry infinite energy due to confinement, so the Hilbert space of finite-energy states is indeed gauge-invariant without this caveat. More generally this depends on the phase of the gauge theory, as we will discuss later.

and the solution is the same. The idea is that for any representation α of G with Lie algebra generators τ_a, we have

$$W_\alpha(C) = Pe^{i\int_C A^a\tau_a\cdot dx}, \tag{4.34}$$

where P indicates the **path-ordered exponential**. There are two ways of defining this. One way is that if we parameterize C as $x^\mu(s)$ with

$$\begin{aligned} x^\mu(0) &= x_i^\mu \\ x^\mu(1) &= x_f^\mu, \end{aligned} \tag{4.35}$$

and we define a unitary matrix $U(s)$ with dimensionality equal to that of α as the solution of the differential equation

$$\frac{dU}{ds} = iA_\mu^a\tau_a\frac{dx^\mu}{ds}U \tag{4.36}$$

with initial condition $U(0) = I$, then

$$W_\alpha(C) \equiv U(1). \tag{4.37}$$

The other way is to discretize C into a sequence of points $x_0, x_1, \ldots, x_N$, with $x_0 = x_i$ and $x_N = x_f$, and then define

$$W_\alpha(C) \equiv \lim_{N\to\infty}\left(I + iA_\mu^a(x_{N-1})\tau_a\Delta x_{N-1}^\mu\right)\ldots\left(I + iA_\mu^a(x_0)\tau_a\Delta x_0^\mu\right) \tag{4.38}$$

where

$$\Delta x_i^\mu = x_{i+1}^\mu - x_i^\mu. \tag{4.39}$$

The gauge transformation of the Wilson line is more easily determined from the second definition. Indeed to first order in Δx_i^μ we have

$$\begin{aligned} I + iA_\mu^a(x_i)'\tau_a\Delta x_i^\mu &= I + iU(x_i)\left(A_\mu^a\tau_a - iU^{-1}(x_i)\partial_\mu U(x_i)\right)U(x_i)^{-1}\Delta x_i^\mu \\ &= U(x_i)\Big(I + iA_\mu^a(x_i)\tau_a\Delta x_i^\mu + U(x_i)^{-1}\partial_\mu U(x_i)\Delta x_i^\mu\Big)U(x_i)^{-1} \\ &= U(x_i)\Big((I + U(x_i)^{-1}\partial_\mu U(x_i)\Delta x_i^\mu\Big)\Big(I + iA_\mu^a(x_i)\tau_a\Delta x_i^\mu\Big)U(x_i)^{-1} \\ &= U(x_{i+1})\Big(I + iA_\mu^a(x_i)\tau_a\Delta x_i^\mu\Big)U(x_i)^{-1} \end{aligned} \tag{4.40}$$

Here $U(x)$ is the representation of the gauge transformation $g(x)$ in the representation α. In the product representation (4.38) of the Wilson line these U matrices cancel between the different factors, except at the beginning and the end of the product. Therefore we have the gauge transformation

$$W_\alpha(C)' = U(x_f)W_\alpha(C)U(x_i)^{-1}. \tag{4.41}$$

In particular if Φ is a field that transforms in representation α of G, then the operator

$$\Phi^\dagger(x_f)W_\alpha(C)\Phi(x_i) \tag{4.42}$$

is gauge invariant. Moreover if $x_i = x_f$ then we have the gauge-invariant Wilson loop

$$\widetilde{W}_\alpha(C) \equiv \mathrm{Tr}\left(W_\alpha(C)\right). \tag{4.43}$$

Unlike in the abelian case, a trace is necessary to get a gauge-invariant operator. When we discuss lattice gauge theory later in the semester we will see that in some sense the Wilson line is a more natural fundamental object than the gauge field A_μ. We already see that it has a simpler gauge transformation, and we can also use it directly to define the covariant derivative:

$$D_\mu\Phi \equiv \lim_{\epsilon\to 0}\frac{W_\alpha(C)^\dagger\Phi(x+\epsilon e_\mu) - \Phi(x)}{\epsilon}, \tag{4.44}$$

where e_μ is a unit coordinate displacement in the μ direction and C is the infinitesimal curve

$$x^\nu(s) = x^\nu + \epsilon s \delta^\nu_\mu. \tag{4.45}$$

From this point of view $W_\alpha(C)$ can be viewed as telling us how to "parallel transport" a field transforming in representation α along a curve C so that we can compare its values at different points in spacetime, so what the covariant derivative does is parallel transport $\Phi(x+\epsilon e_\mu)$ back to x before it compares it with $\Phi(x)$.

4.5 Gauge-fixed operators in the gauge-invariant formalism

The alternative approach to dealing with gauge redundancy that we mentioned above is gauge-fixing. This means we impose some additional condition on the fields of the theory that uniquely picks one representative of each gauge equivalence class. In Maxwell theory we discussed several possible gauge-fixing conditions, with the two we discussed the most being **Coulomb gauge**

$$\partial_i A_i = 0 \tag{4.46}$$

and **axial gauge**

$$A_1 = 0. \tag{4.47}$$

Both of these conditions break manifest Lorentz invariance, and axial gauge further breaks rotational invariance. For general compact G we can similarly try to fix the gauge, but it becomes harder to find a simple gauge-fixing condition that really picks one representative of each gauge equivalence class. In particular the non-abelian Coulomb gauge condition

$$\partial_i A^a_i = 0 \tag{4.48}$$

does not work due to what is called the **Gribov ambiguity**, which means that for some choices of A^i_a there are nontrivial gauge transformations that preserve this condition. In Minkowski space however the axial gauge condition

$$A^a_1 = 0 \tag{4.49}$$

still essentially works, as we will now argue. There are two things we need to show: that any gauge field configuration can be put in axial gauge by a valid gauge transformation, and that once this is achieved no further gauge transformations are allowed. For the former, given a gauge field configuration $A_\mu(x)$, we want to find a gauge transformation $g(x)$ such that

$$ig^{-1}\partial_1 g = A_1, \tag{4.50}$$

which we can rewrite as

$$\partial_1 g^{-1} = iA_1 g^{-1}. \tag{4.51}$$

We have just learned how to solve this equation, the solution is

$$g(t, x^1, x^\perp)^{-1} = Pe^{i\int_{-\infty}^{x^1} A_1(t, x^{1\prime}, x^\perp) dx^{1\prime}}. \tag{4.52}$$

Moreover this g is unique since it is the solution of an ODE with boundary conditions specified, so there is no Gribov ambiguity. There are two points we need to be careful about however: first of all the integral could potentially be divergent, and we also need to check that it defines a gauge transformation that becomes trivial at spatial infinity since those are the only ones we quotient by. The convergence is easily dealt with by requiring the gauge field to go to zero at infinity: for A_1 this is true in every direction except for the x^1 direction by our boundary conditions, while in the x^1 direction we can get it by restricting to finite-energy configurations. The gauge transformation g vanishes in any spatial direction except for the x^1 direction for the same reason, and the integral also vanishes by construction when $x^1 = -\infty$. So the only direction where it might not vanish is $x^1 = +\infty$. And indeed in general it might not: there is a gauge-invariant degree of freedom given by the Wilson line from $x^1 = -\infty$ to $x^1 = +\infty$, and we cannot remove it by a gauge

transformation. The best way to take this into account is to treat this Wilson line as a single extra degree of freedom in the theory, together with a gauge field obeying (4.49). The extra degree of freedom can only be measured by moving a charged particle from $x^1 = -\infty$ to $x^1 = \infty$, so it has no effect on experiments done in a finite region.

The way this gauge-fixing procedure is implemented in our gauge-invariant formalism is that the gauge field in axial gauge is actually a gauge-invariant operator on the physical Hilbert space:

$$\widetilde{A}_\mu \equiv g\left(A_\mu - ig^{-1}\partial_\mu g\right)g^{-1}, \tag{4.53}$$

with $g(x)$ given by (4.52). This is a highly non-local operator due to the path-ordered exponential in (4.52), just as we found for the gauge field in Coulomb gauge in Maxwell theory. This non-locality is hidden if we try to view $\widetilde{A}_\mu$ as the fundamental field, as is done in most textbook treatments of gauge-fixing, but it reappears if we compute commutators of $\widetilde{A}_\mu$ with other gauge-invariant operators such as $\mathrm{Tr}\,(F_{\mu\nu}F^{\mu\nu})$ so in my view it is better to have it there explicitly in the operator. If you do not like it, the best recourse is to avoid gauge-fixing altogether by working with the gauge-invariant formalism.

4.6 Homework

1. Write down the condition that a matter action $\mathcal{L}(\Phi, D\Phi)$ depending on some matter fields Φ_n and their covariant derivatives $D_\mu \Phi_n$ be invariant under infinitesimal gauge transformations

$$\Phi'_n(x) = \Phi_n(x) + i\epsilon\theta^a(x)\,(\tau_a)_n{}^m\,\Phi_m(x). \tag{4.54}$$

   Using this condition, and also the equations of motion obtained by varying the matter action with respect to Φ_n, show the covariant conservation (4.7) of the matter current (4.5). Also show that the left-hand side of the equation of motion (4.4) is identically covariantly conserved.

2. Consider the matter Lagrangian

$$\mathcal{L}_M = -\left(D_\mu \Phi\right)^\dagger D_\mu \Phi - V(\Phi^\dagger \Phi). \tag{4.55}$$

   for a complex gauge field Φ_m transforming in some representation of the gauge group G. Compute the gauge current J^μ_a for this action, and also show that the matter Hamiltonian (4.23) is independent of A^a_0 when written in terms of Φ_m and Π^m.

3. So far we have discussed general gauge transformations of A_μ. Defining $g(x) = e^{i\epsilon^a(x)T_a}$, write out the infinitesimal form of a gauge transformation of A_μ to first order in ϵ^a. Also write the transformation of A^a_μ.

4. Compute the commutator of $A^a_\mu(\vec{x})$ and $\Phi_n(\vec{x})$ with $-\int d^{d-1}x \epsilon^a(\vec{x})\left(D_i \Pi^i_a + J^0_a\right)$, with ϵ^a vanishing at spatial infinity, and compare to the result of the previous problem to convince yourself that the Gauss constraint indeed generates gauge transformations.

5. Consider a Wilson loop around a small square-shaped loop of side length ϵ, which you can take to lie in the $x^1 x^2$ plane oriented with the axes. Working to second order in ϵ, express it in terms of gauge-invariant local operators. Hint: this is easiest if you Taylor-expand the gauge fields around point in the center of the loop, which you might as well take to be the origin of the $x^1 x^2$ plane. Also at this order it is enough to approximate the contribution from each edge as $(1 + i\epsilon A)$ with A evaluated at the midpoint of the edge and pointing along it, you are free to assume this but if you would like to derive it in a systematic way you can use the Magnus expansion

$$P e^{\int_0^1 ds X(s)} = \exp\left[\int_0^1 ds X(s) + \frac{1}{2}\int_0^1 ds_1 \int_0^{s_1} ds_2 [X(s_1), X(s_2)] + \ldots\right] \tag{4.56}$$

   and Taylor expand X about $s = 1/2$ so that you can evaluate the integrals.

5 Path integral formulation of Yang-Mills theory

Our primary tool for doing concrete calculations in interacting field theories is the path integral formulation, and that is true for Yang-Mills theory as well. As in Maxwell theory, our strategy will be to derive the path integral from the gauge-invariant Hamiltonian formulation of the theory. This allows us to avoid the problems with defining a non-perturbative gauge-fixing that we encountered in the last section, although a completely rigorous formulation will have to wait until we go to the lattice in a few sections.

5.1 Gauge-invariant path integral

We'll begin by recalling that in ordinary quantum mechanics the path integral is derived by repeatedly using the formula

$$\langle q'|e^{-i\epsilon H(Q,P)}|q\rangle \approx \int \frac{dp}{2\pi} e^{ip(q'-q)-i\epsilon H(q',p)} \tag{5.1}$$

to split up a transition amplitude into infinitesimal pieces. Here H is ordered with Q to the left and P to the right. The approximation is at small ϵ. In gauge theory however the analogous quantity does not make sense, since states of definite A live in $\mathcal{H}_{big}$ but the Hamiltonian is only defined on $\mathcal{H}_{phys}$. In Maxwell theory we dealt with this in two steps. We first introduced an "intermediate" Hilbert space $\mathcal{H}_{int}$ of wave functionals that are independent of A_0, and then on $\mathcal{H}_{int}$ we introduced a projection P_{GI} onto $\mathcal{H}_{phys}$ that imposes the Gauss constraint and then constructed an integral representation of

$$\langle \vec{a}', \phi'|e^{-i\epsilon H} P_{GI}|\vec{a}, \phi\rangle. \tag{5.2}$$

We will follow the same strategy here for general compact gauge group G.

5.1.1 Group theory preliminaries

To begin we first introduce some new terminology. So far we have been using the term "gauge group" for the group G that the matter fields at a given spacetime point x transform in representations under. G is a finite dimensional Lie group. There are two other notions of gauge group that will be important in this section, so to avoid confusion we will define them now. The first of these is the set of maps from space $\mathbb{R}^{d-1}$ to G that become the identity at spatial infinity. We will refer to this as the **canonical gauge group**, and denote it $\mathcal{G}_c$. $\mathcal{G}_c$ is the group whose infininitesimal generators are the Gauss constraints, and it acts unitarily on $\mathcal{H}_{int}$. The physical Hilbert space $\mathcal{H}_{phys}$ is the set of states that are invariant under $U(g)$ for all $g \in \mathcal{G}_c$. The other notion of gauge group that we will need is the set of maps from the full spacetime $\mathbb{R}^d$ to G that become the identity at spatial infinity. We will refer to this as the **spacetime gauge group**, and denote it by $\mathcal{G}$. Classically $\mathcal{G}$ is the set of what we usually refer to as gauge redundancies. Both $\mathcal{G}_c$ and $\mathcal{G}$ are infinite-dimensional Lie groups, and both are compact since a product of compact spaces is compact even if it is an infinite product (this is called Tychonoff's theorem).[44] In particular this means that we can write the projection onto gauge-invariant states as

$$P_{GI} \equiv \int_{\mathcal{G}_c} dg U(g). \tag{5.3}$$

This clearly acts as one on gauge-invariant states, and acting on any state $|\psi\rangle$ it produces a gauge-invariant state since

$$U(h)\int_{\mathcal{G}_c} dg U(g)|\psi\rangle = \int_{\mathcal{G}_c} dg U(hg)|\psi\rangle = \int_{\mathcal{G}_c} dg U(g)|\psi\rangle. \tag{5.4}$$

[44] This statement would not have been true had we had restricted the elements of $\mathcal{G}_c$ and $\mathcal{G}$ to be smooth functions. This is the usual problem we have in path integral arguments, where a pointwise measure is dominated by strongly discontinuous functions. It is better justified on the lattice, as you will see in a few sections.

In constructing the path integral however, a more natural representation of P_{GI} is this one:

$$P_{GI} = \frac{\int \mathcal{D}a_0 e^{i\int d^{d-1}x a_0^a \left(D_i \Pi_a^i + J_a^0\right)}}{\int \mathcal{D}a_0}. \tag{5.5}$$

Here $a_0^a(\vec{x})$ is a function that will become the time component of the gauge field, and the measure is the cartesian measure

$$\mathcal{D}a_0 = \prod_{a,\vec{x}} da_0^a(\vec{x}). \tag{5.6}$$

Deriving this formula is somewhat subtle, the simplest argument I could find is based on using something called the Weyl integration formula to rewrite the integral in terms of a maximal abelian subalgebra and then using the fact that in the abelian case the integral over the Lie algebra gives a delta function of the charge. I will instead show something a bit weaker that is still sufficient for our construction of the path integral, which is that if we raise the right-hand side to the power N then it approaches P_{GI} in the limit that $N \to \infty$. Provisionally denoting the right-hand side of (5.5) as $\hat{P}_{GI}$, we want to show that

$$\lim_{N\to\infty} \hat{P}_{GI}^N = P_{GI}. \tag{5.7}$$

We'll first show that for any $g \in \mathcal{G}_c$ we have

$$U(g)\hat{P}_{GI}U(g^{-1}) = \hat{P}_{GI}. \tag{5.8}$$

This is because the constraints $D_i\Pi_a^i + J_a^0$ are the generators of $\mathcal{G}_c$, so conjugation by U acts on them in the adjoint representation:

$$U(g)\int \mathcal{D}a_0 e^{i\int d^{d-1}x a_0^a \left(D_i \Pi_a^i + J_a^0\right)} U(g^{-1}) = \int \mathcal{D}a_0 e^{i\int d^{d-1}x D^a{}_b(g) a_0^b \left(D_i \Pi_a^i + J_a^0\right)}. \tag{5.9}$$

We can undo this adjoint transformation by a change of coordinates

$$a_0^{a\prime}(\vec{x}) = D^a{}_b(g) a_0^b \tag{5.10}$$

in the a_0 integral, at most at the cost of a Jacobian factor in the measure. And in fact there is no such factor, since we have

$$D^a{}_b(g) D^c{}_d(g) \hat{g}_{ac} = \hat{g}_{bd} \tag{5.11}$$

and thus $Det(D) = \pm 1$. The change of variables formula uses the absolute value of the determinant, so there is no Jacobian factor. Therefore $\hat{P}_{GI}$ commutes with $U(g)$. Thus by Schur's lemma $\hat{P}_{GI}$ acts as a constant λ on each irreducible representation in the decomposition of $\mathcal{H}_{int}$ into irreps of $\mathcal{G}_c$. Clearly $\lambda = 1$ on $\mathcal{H}_{phys}$, so to establish (5.5) we need to show that $\lambda = 0$ for each nontrivial irrep. This is the step where the Weyl integration formula is needed. To avoid it we will instead just show that acting on nontrivial irreps we have $|\lambda| < 1$. This is enough to show (5.7). The idea is that that if $|i\rangle$ and $|j\rangle$ are states in the same irrep, then

$$\langle j|\hat{P}_{GI}|i\rangle = \lambda \delta_{ij} \tag{5.12}$$

is really just averaging the representation matrix $D_{ij}(g)$ over some set of group elements g. Since the D_{ij} are unitary (remember that $\mathcal{G}_c$ is compact), their components must have absolute value less than or equal to one. If the representation is nontrivial then some of these components have absolute value less than one, in which case the average must be strictly less than one.

5.1.2 Deriving the path integral

Let's now see how we can use (5.5) (or (5.7)) to derive the gauge theory path integral. Other than (5.5) the steps are the same as we have already done several times, so we will be fairly terse. For an infinitesimal time

evolution it goes like this:

$$\begin{aligned}\langle \vec{a}', \phi' | e^{-i\epsilon H t} P_{GI} | \vec{a}, \phi \rangle &= \frac{1}{\int \mathcal{D}a_0} \int \mathcal{D}\pi \mathcal{D}a_0 \langle \vec{a}' \phi' | e^{-i\epsilon H t} | \pi \rangle \langle \pi | e^{i\epsilon \int d^{d-1}x a_0^a (D_i \Pi_a^i + J_a^0)} | \vec{a}, \phi \rangle \\ &\approx \frac{1}{\int \mathcal{D}a_0} \int \mathcal{D}\pi \mathcal{D}a_0 \exp\left[i\epsilon \int d^{d-1}x \left(\frac{a_i^{a\prime} - a_i^a}{\epsilon} \pi_a^i + \frac{\phi_m' - \phi_m}{\epsilon} \pi^m - \mathcal{H} + a_0^a (D_i \pi_a^i + j_a^0) \right) \right] \\ &\propto \frac{1}{\int \mathcal{D}a_0} \int \mathcal{D}a_0 e^{i\epsilon \int d^{d-1} \mathcal{L}}, \end{aligned} \tag{5.13}$$

where

$$\mathcal{H} = \frac{g^2}{2} \pi_a^i \pi_i^a + \frac{1}{4g^2} f_{ij}^a f_a^{ij} + \hat{\mathcal{H}}_M \tag{5.14}$$

is the canonical Hamiltonian density we derived last time, $\mathcal{D}\pi$ denotes an integral over both π_a^i and π^m, and in the last step we have performed the Gaussian integral over the π to replace them by their saddle point values

$$\begin{aligned} \pi_a^i &= \frac{1}{g^2} f_a^{i0} \\ \pi^m &= \frac{\partial \mathcal{L}_M}{\partial \dot{\phi}^m} \end{aligned} \tag{5.15}$$

at the cost of a field-independent determinant factor that we have dropped. In the last step we have the Lorentz-invariant and gauge-invariant Lagrangian density

$$\mathcal{L} = -\frac{1}{4g^2} f_{\mu\nu}^a f_a^{\mu\nu} + \mathcal{L}_M \tag{5.16}$$

that we started with, so all the non-covariant parts of the Hamiltonian analysis have disappeared! Throughout lower-case fields indicate integration variables or field eigenvalues, while upper case fields indicate operators. The first equality is (5.5) together with inserting a complete set of states, and the approximation is using small ϵ linearize the exponentials so that the operators (with appropriate ordering) can be replaced by their eigenvalues, after which they are re-exponentiated. Once we have (5.13), we can then proceed to vacuum expectations of gauge-invariant operators as before:

$$\langle \Omega | T \mathcal{O}_1(A, \Phi) \dots \mathcal{O}_n(A, \Phi) | \Omega \rangle = \frac{\int \mathcal{D}a \mathcal{D}\phi \mathcal{O}_1(a, \phi) \dots \mathcal{O}_n(a, \phi) e^{iS}}{\int \mathcal{D}a \mathcal{D}\phi e^{iS}}, \tag{5.17}$$

where as usual the action is evaluated using an $i\epsilon$ prescription to project onto the ground state. There were multiple heuristic steps in this derivation (to say the least!), leading to the usual issues of regularization and renormalization that we have discussed at length in the previous semesters.

Before moving on to discuss the gauge-fixed version of (5.17), I will mention why (5.7) is enough to derive it. What the derivation we just gave really did is supply a factor of $\hat{P}_{GI}$ at each time step. Since the Hamiltonian is gauge-invariant it commutes with $\hat{P}_{GI}$, so when we evolve for a finite time we can move all of those factors of $\hat{P}_{GI}$ through the time evolution operators and combine them. In the continuum limit of many time steps this thus becomes P_{GI} by (5.7).

5.2 Fadeev-Popov gauge-fixing

In QED we saw that the expression (5.17) was in some sense "too pure" for practical calculations. This was because the numerator and the denominator are both divergent due to the invariance of the integrand under gauge transformations in the spacetime gauge group $\mathcal{G}$ (this is in addition to the various UV divergences which we will deal with later using renormalization). In fact the factor of $\frac{1}{\int \mathcal{D}a_0}$ in (5.13) was precisely there to cancel this divergence, but this factor canceled in (5.17). This doesn't mean that (5.17) is wrong,

and indeed its Euclidean version is the starting point for lattice gauge theory. In perturbative calculations however this divergence is quite inconvenient, and indeed in QED we saw that it led to a photon kinetic term that could not be inverted to find a well-defined propagator.

The standard approach to this problem is to gauge-fix the path integral (5.17) using a non-abelian version of the Fadeev-Popov (FP) procedure we used to set up a gauge-fixed path integral for QED last semester. In this section we will describe this procedure *assuming* that we have a good set of gauge-fixing conditions

$$B^{a,x}[A,\Phi] = 0 \tag{5.18}$$

that pick a unique representative of each equivalence class of field configurations under $\mathcal{G}$. In fact we will assume a bit more, which is that for any function $f^a(x)$ we can find a unique gauge transformation $g \in \mathcal{G}$ such that

$$B^{a,x}[A_g,\Phi_g] = f^a(x), \tag{5.19}$$

where A_g and Φ_g are the gauge transformations of A and Φ by g. It must be acknowledged however that non-perturbative issues such as the Gribov ambiguity for Coulomb gauge or the residual Wilson line for axial gauge make either of these things hard to do in practice. Fortunately for us however, in perturbation theory it is enough to pick a gauge-fixing that works near $A = 0, \Phi = 0$ and that is much easier to do.

We will present the FP procedure in a rather general way. We begin with a path integral

$$Z_{\mathcal{O}} \equiv \int \mathcal{D}\phi \mathcal{O}[\phi] e^{iS[\phi]} \tag{5.20}$$

where both the operator $\mathcal{O}[\phi]$ and the path integral measure $\mathcal{D}\phi e^{iS[\phi]}$ are invariant under a gauge transformation

$$\phi' = \phi_g \tag{5.21}$$

with $g \in \mathcal{G}$ a spacetime gauge transformation coming from a compact gauge group G. Here I have combined the gauge and matter fields into a single field ϕ (in general at the cost of having ϕ transform in a reducible representation of whatever symmetries are present). In this argument we could combine $\mathcal{O}$ and e^{iS} into a single gauge-invariant functional, but I will leave them explicit to remind you why we are doing this. The first step of the FP procedure is to define

$$Z_\xi \equiv \int \mathcal{D}f e^{-\frac{i}{2g^2\xi}\int d^dx f^a(x) f_a(x)} \tag{5.22}$$

with $\xi > 0$, in terms of which we have

$$Z_{\mathcal{O}} = \frac{1}{Z_\xi}\int \mathcal{D}\phi \mathcal{D}f \mathcal{O}[\phi] e^{iS[\phi] - \frac{i}{2g^2\xi}\int d^dx f^a(x) f_a(x)}. \tag{5.23}$$

The factor of g^2 in the definition of Z_ξ is there so that our ξ matches the standard one in the literature. So far this is trivial. The key step is that we now change variables in the integral over the $f^\alpha(x)$ in a field-dependent way, by solving the equation (5.19) for g (which we assumed can be uniquely done). We thus have

$$Z_{\mathcal{O}} = \frac{1}{Z_\xi}\int \mathcal{D}\phi \mathcal{D}g \det\left(\frac{\delta B[\phi_g]}{\delta g}\right) \mathcal{O}[\phi] e^{iS[\phi] - \frac{i}{2g^2\xi}\int d^dx B^{a,x}[\phi_g] B^x_a[\phi_g]}, \tag{5.24}$$

where the determinant arises from the change of variables matrix

$$\frac{\delta B^{a,x}[\phi_g]}{\delta g^{b,y}} \tag{5.25}$$

with $g^{a,x}$ being coordinates on $\mathcal{G}$ and the measure $\mathcal{D}g$ is the cartesian measure on $g^{a,x}$.[45] Using the gauge invariance of $\mathcal{O}[\phi]$ and $\mathcal{D}\phi e^{iS[\phi]}$ we can rewrite this as

$$Z_{\mathcal{O}} = \frac{1}{Z_\xi}\int \mathcal{D}\phi_g \mathcal{D}g \det\left(\frac{\delta B[\phi_g]}{\delta g}\right) \mathcal{O}[\phi_g] e^{iS[\phi_g] - \frac{i}{2g^2\xi}\int d^dx B^{a,x}[\phi_g] B^x_a[\phi_g]}. \tag{5.26}$$

[45] This measure on $\mathcal{G}$ is not so natural but don't worry, we will soon convert it to the Haar measure!

So far this is the same as the manipulation we did for QED, with the only new ingredient being that the determinant factor is now field-dependent so we cannot just discard it. The next step in QED was to change integration variables from ϕ_g to ϕ, which removes the g dependence of all terms except for the determinant. The new complication in the non-abelian case is what to do about this determinant. For this purpose it is useful to note that if we denote by $(hg)^{ax}$ the coordinates of $hg \in \mathcal{G}$, then by the chain rule we have[46]

$$\frac{\delta B^{ax}[\phi_{hg}]}{\delta h^{by}} = \frac{\delta B^{ax}[\phi_{hg}]}{\delta (hg)^{cy}} \frac{\partial (hg)^{cy}}{\partial h^{by}}, \tag{5.27}$$

so taking the determinant and setting $h = e$ we have

$$\det\left(\frac{\delta B[\phi_g]}{\delta g}\right) = \det\left(\frac{\delta B[\phi_{hg}]}{\delta h}\Big|_{h=e}\right) \Big/ \det\left(\frac{\partial (hg)}{\partial h}\Big|_{h=e}\right). \tag{5.28}$$

The quantity

$$\frac{\mathcal{D}g}{\det\left(\frac{\partial (hg)}{\partial h}\Big|_{h=e}\right)} \tag{5.29}$$

is actually proportional to the Haar measure on $\mathcal{G}$, as you will check in the homework. Moreover the quantity

$$\det\left(\frac{\delta B[\phi_{hg}]}{\delta h}\Big|_{h=e}\right) \tag{5.30}$$

depends on ϕ only in the combination ϕ_g. Thus we are free to now change variables $\phi_g \to \phi$, after which we have

$$Z_{\mathcal{O}} = \frac{\mathrm{Vol}(\mathcal{G})}{Z_\xi} \int \mathcal{D}\phi\, \Delta_{FP}[\phi]\, \mathcal{O}[\phi]\, e^{iS[\phi] + i\int d^d x \mathcal{L}_{gf}}, \tag{5.31}$$

where

$$\Delta_{FP}[\phi] \equiv \det\left(\frac{\delta B[\phi_h]}{\delta h}\Big|_{h=e}\right) \tag{5.32}$$

is called the **Fadeev-Popov determinant** and

$$\mathcal{L}_{gf}(x) = -\frac{1}{2g^2\xi} B^{a,x}[\phi] B_{a,x}[\phi] \tag{5.33}$$

is called the **gauge-fixing Lagrangian**. $\mathrm{Vol}(\mathcal{G})$ is the volume of $\mathcal{G}$ in the measure (5.29), so we have succeeded in removing the divergence due to integration over gauge transformations. In the end the result is fairly simple: up to an overall constant that cancels in (5.17), the gauge-fixed path integral differs from the gauge-invariant one only by the presence of the FP determinant $\Delta_{FP}[\phi]$ and the gauge-fixing Lagrangian.

5.3 Ghosts

Equation (5.31) is the starting point for most perturbative calculations in gauge theory. In order to make it practical however we need to pick a more useful representation of the FP determinant $\Delta_{FP}[\phi]$. Here it is useful to remember our general Gaussian integration formulae, which give us a way of representing determinants as path integrals. For bosonic path integrals however what we get is an inverse power of the determinant, so to get a positive power (in particular one) we need to use a Grassman integral. The identity we will use is that

$$\int d\xi_1 d\psi^1 \dots d\xi_M d\psi^M e^{-\xi^T A \psi} = \det(A), \tag{5.34}$$

[46] In the second factor we have ordinary partial derivatives since $hg(y)$ only depends on $h(y)$. We could alternatively write a functional derivative and then have h^{cz} instead of h^{cy} and integrate over z, but the functional derivative produces a $\delta^d(y-z)$ which is then removed by integrating over z so we end in the same place.

where ψ^n and ξ_n are real Grassmann variables and A is any $M \times M$ matrix. Applying this formula to the FP determinant gives

$$\Delta_{FP} \propto \int \mathcal{D}b\mathcal{D}c \exp\left[i \int d^d x d^d y b_a(x) \frac{\delta B^{ax}}{\delta h^{by}}\Big|_{h=e} c^b(y)\right], \tag{5.35}$$

where b_a is like $i\xi$ and c^a is like ψ in (5.34). The $\propto$ indicates that we dropped a field-independent phase arising from this i and also from ambiguity in how exactly we order the fermion measure. This phase always cancels when we compute a ratio of path integrals in order to get normalized correlation functions or S-matrix elements. The fields b_a and c^a are called **Faddeev-Popov ghosts** (sometimes c^a are called ghosts and b_a antighosts). They have the rather counterintuitive property of being anticommuting fields that transform as Lorentz scalars, which seems to contradict the spin-statistics theorem. We will discuss their physical interpretation later in the section.

In the literature b_a is often called $\overline{c}_a$ or c_a^*, which suggests that b_a and c_a are complex conjugates, but this is not actually true: instead the reality assignments for the ghost fields are

$$\begin{aligned} c^{a*} &= c^a \\ b_a^* &= -b_a. \end{aligned} \tag{5.36}$$

One way to see this is from the comparison to ψ and ξ above, and we will also motivate it in a moment from the reality of the Lagrangian. It is quite tempting to take b_a to be real by dropping the factor of i in relating it to ξ, but keeping this i is a rather standard convention since it makes the ghost propagator match the usual free scalar propagator as we will see next section.

Speaking of the Lagrangian, we would of course like to view the quantity in the exponent of (5.35) as a contribution to the Lagrangian together with $\mathcal{L}_{gf}$ but it does not look local. This however is a purely notational problem: if the $B^{a,x}$ are local functionals of the fields, then the functional derivative with respect to h^{by} always introduces a δ-function that does one of the spacetime integrals. We can illustrate this by making the concrete choice

$$B^{a,x}[A] = \partial^\mu A^a_\mu, \tag{5.37}$$

which is the non-abelian version of the Lorenz gauge condition that we used in quantum electrodynamics. Parametrizing $h \in \mathcal{G}$ near the identity as

$$h = e^{i \int d^d x \epsilon^a(x) \tau_{ax}} \tag{5.38}$$

and using the expression

$$\delta_\epsilon A^a_\mu = D_\mu \epsilon^a \tag{5.39}$$

for an infinitesimal gauge transformation that you derived on the previous homework, we have

$$\frac{\delta B^{ax}}{\delta \epsilon^b(y)}\Big|_{\epsilon=0} = \partial^\mu D_\mu \delta^a_b \delta^d(x-y), \tag{5.40}$$

with the derivatives acting on x, and thus

$$\Delta_{FP} \propto \int \mathcal{D}b\mathcal{D}c\, e^{i \int d^d x \mathcal{L}_{FP}}, \tag{5.41}$$

with

$$\mathcal{L}_{FP} = -\partial^\mu b_a D_\mu c^a. \tag{5.42}$$

Here we took the liberty of integrating by parts to make a Lagrangian with no second derivatives. This Lagrangian is indeed real given our reality assignments (5.36), provided that we are careful to remember that the complex conjugate exchanges the order of Grassman variables.

5.4 What do ghosts mean?

At this point you may be rather concerned about the physical meaning of the Fadeev-Popov ghost fields. Why is it ok that they violate the spin-statistics relation? Do they really correspond to physical particles? If so then where were they in the gauge-invariant presentation of the theory? In particular I emphasize that the ghost action $\mathcal{L}_{FP}$ includes a cubic interaction term where a b and a c ghost interact with a gauge boson. Does this mean that we can physically produce b and c excitations out of gauge bosons? Even worse, if your memory goes back to the beginning of the previous semester you may recall that in order to make distinct anticommuting fields be hermitian we need to define an inner product which is not positive semi-definite. Does that mean that unitarity is somehow violated?

The answers to all these questions become clear once we remember what we are really doing. We started with a Hilbert space $\mathcal{H}_{big}$ with positive inner product, and we projected to the gauge-invariant subspace $\mathcal{H}_{phys}$ that inherits this positive inner product. We are computing the expectation values of gauge-invariant operators built out of the gauge and matter fields, which we can think of as operators on $\mathcal{H}_{phys}$. They obey the usual spin-statistics relations, as they must. The ghost fields are best thought of as purely calculational devices in the path integral, which are there to help us fix the gauge, and we should not include them in the operators that we are computing expectation values of. From a perturbative perspective, they are allowed to run in loops but not to appear as external lines.

You may nonetheless wonder if we can assign some kind of Hilbert space interpretation to the ghost fields. Indeed we can, but this requires extending the $\mathcal{H}_{phys}$ to a larger vector space, different from $\mathcal{H}_{big}$, whose inner product has indefinite norm. Our proof of the spin-statistics theorem used the positivity of the Hilbert space inner product, so there is no contradiction that the conclusions of the theorem do not hold on this larger vector space. This idea leads to what is called **BRST quantization**, which is a formalism for giving the gauge-fixed path integral an operator interpretation and then constructing the physical Hilbert space by a kind of quotient. We won't develop this formalism in detail, as we already have a satisfactory Hilbert space formalism based on gauge-invariant quantization, but we will say a little about it. Before doing so however, we first need to introduce the related notion of BRST symmetry.

5.5 BRST symmetry

Let's now have a look at our general gauge-fixed Lagrangian:

$$\mathcal{L}_{total} = \mathcal{L}_{GI} - \frac{1}{2g^2\xi} B^{a,x} B_{a,x} + b_a \int d^d y \frac{\delta B^{ax}}{\delta h^{by}}\Big|_{h=e} c^b(y). \tag{5.43}$$

Here $\mathcal{L}_{GI}$ is the gauge-invariant Lagrangian we started with. The new terms are *not* gauge invariant, which after all was the whole point of doing gauge-fixing. Nonetheless it was realized by Becchi, Rouet, Stora, and Tyutin that it still possesses a residual form of gauge symmetry called **BRST symmetry**. This is most clearly demonstrated if we introducing an auxiliary (bosonic) scalar field n_a as

$$\widetilde{\mathcal{L}}_{total} = \mathcal{L}_{GI} + \frac{g^2\xi}{2} n_a n^a + n_a B^{a,x} + b_a \int d^d y \frac{\delta B^{ax}}{\delta h^{by}}\Big|_{h=e} c^b(y). \tag{5.44}$$

Doing the Gaussian integral over n_a replaces it by its saddle point value

$$n^a = -\frac{1}{g^2\xi} B^{ax}, \tag{5.45}$$

which turns this action back into (5.43). The BRST transformation is an infinitesimal transformation labeled by a Grassman parameter we will call θ, with the rough idea being that it is an inifitesimal gauge transformation with gauge-transformation parameter

$$\epsilon^a = \theta c^a. \tag{5.46}$$

Acting on all of the fields the BRST transformation is

$$\begin{aligned}
\delta_\theta \phi_n &= i\theta c^a \left(\tau_a\right)_n{}^m \phi_m \\
\delta_\theta a_\mu^a &= \theta D_\mu c^a \\
\delta_\theta c^a &= -\frac{1}{2}\theta C^a{}_{bc} c^b c^c \\
\delta_\theta b_a &= -\theta n_a \\
\delta_\theta n_a &= 0.
\end{aligned} \tag{5.47}$$

One can confirm by direct calculation that the action $\widetilde{\mathcal{L}}_{total}$ is invariant under this transformation. This is obviously true for $\mathcal{L}_{GI}$, since it depends only on the physical fields ϕ_n and a_μ^a and for these the BRST transformation is just a gauge transformation; the gauge invariance of $\mathcal{L}_{GI}$ is right there in the name. The term involving $n_a n^a$ is also obviously invariant. It is not hard to argue that when the transformation acts on $B^{a,x}$ in the third term it cancels the term where the transformation acts on b_a in the fourth term; this follows from

$$\delta_\theta B^{a,x} = \int d^d y \frac{\delta B^{a,x}}{\delta h^{by}}\Big|_{h=e} \theta c^b(y), \tag{5.48}$$

which is true since $B^{a,x}$ only depends on the gauge and matter fields (not the ghosts or the auxiliary field), and so its transformation is just an infinitesimal gauge transformation with gauge parameter (5.46). The tricky thing is to argue that the quantity $\int d^d y \frac{\delta B^{a,x}}{\delta h^{b,y}}\Big|_{h=e} c^b(y)$ is invariant. In the homework you will confirm this for the special case of Lorenz gauge, but to show it more generally it is worthwhile to first take a detour and realize that the BRST transformation squares to zero in the sense that

$$\delta_{\theta_1}\delta_{\theta_2} = 0 \tag{5.49}$$

acting on all the fields. This is obvious acting on b_a and n_a, but for the other fields we need to check it. I will do it for ϕ_n, and leave a_μ^a and c^a to you in the homework. For ϕ_n we have

$$\begin{aligned}
\delta_{\theta_1}\delta_{\theta_2}\phi &= i\theta_2 \tau_a \left(\delta_{\theta_1} c^a \phi + c^a \delta_{\theta_1}\phi\right) && (5.50)\\
&= i\theta_2\theta_1 \left(-\frac{1}{2} C^a{}_{bc} c^b c^c \tau_a \phi - i c^a c^b \tau_a \tau_b \phi\right) && (5.51)\\
&= -i\theta_2\theta_1 \left(-\frac{1}{2} C^a{}_{bc} c^b c^c \tau_a \phi + \frac{1}{2} C^c{}_{ab} c^a c^b \tau_c \phi\right) && (5.52)\\
&= 0, && (5.53)
\end{aligned}$$

where in the third equality we used the antisymmetry of $c^a c^b$ to replace $\tau_a\tau_b$ by $\frac{1}{2}[\tau_a, \tau_b]$ and then used the Lie algebra. (5.49) also holds on products of fields, for example

$$\delta_{\theta_1}\delta_{\theta_2}(\phi_1\phi_2) = \delta_{\theta_1}\phi_1\delta_{\theta_2}\phi_2 + \delta_{\theta_2}\phi_1\delta_{\theta_1}\phi_2 = 0, \tag{5.54}$$

with the second equality holding because θ_1 and θ_2 are anticommuting and we need to move them past each other to exchange them.

We can now use (5.49) to finish showing the invariance of $\widetilde{\mathcal{L}}_{total}$ under the BRST transformation (5.47). The key point is to recognize that we have

$$\widetilde{\mathcal{L}}_{total} = \mathcal{L}_{GI} + \frac{d}{d\theta}\delta_\theta \left(-b_a B^{a,x} - \frac{g^2\xi}{2} b_a n^a\right), \tag{5.55}$$

which follows from

$$\delta_\theta \left(b_a n^a\right) = -\theta n_a n^a \tag{5.56}$$

and (5.48). Thus if we compute $\delta_{\theta'}$ of the right-hand side of (5.55) then $\mathcal{L}_{GI}$ is invariant by assumption and the rest is also invariant since $\delta_{\theta'}\delta_\theta = 0$.

5.6 Applications of BRST symmetry

There are three primary applications of BRST symmetry:

(1) In renormalizing Yang-Mills theory we need to introduce counterterms as usual, and for the theory to be self-consistent these counterterms need to be gauge-invariant. On the other hand the ghost and gauge-fixing terms in the gauge-fixed path integral are *not* gauge invariant, which raises the concerning possibility that the gauge-fixed path integral could generate divergences which require a gauge non-invariant counterterm to cancel. The standard proof that this does not happen uses BRST symmetry to constrain the forms of possible divergences, making sure that only those that can be absorbed by gauge-invariant counterterms can be generated. You can see this argument for example in Weinberg's book. From my point of view however this is something of a distraction: in the gauge-invariant path integral (5.17) there are no violations of gauge symmetry, and if we choose a gauge-invariant regulator such as lattice gauge theory then no gauge-variant counterterms can be generated.

(2) We declared that ghost fields do not lead to physical particles, but in the gauge-fixed path integral this is not obvious since we can easily write down diagrams that look like they could produce ghosts from physical initial states. The self-consistency of the gauge-invariant path integral makes it clear that these diagrams do not actually contribute to physical amplitudes, but the standard approach to confirming this to all orders in perturbation theory uses BRST symmetry.

(3) We can use BRST symmetry as an alternative starting point to the quantization of gauge theories. The idea is to introduce a larger vector space $\mathcal{H}_{BRST}$ on which the ghost fields act as anticommuting linear operators, and which is equipped with an "inner product" that includes states of negative and zero norm. We then introduce a fermionic BRST operator Q_B on $\mathcal{H}_{BRST}$ such that

$$i[\theta Q_B, \mathcal{O}] = \delta_\theta \mathcal{O} \tag{5.57}$$

for any operator $\mathcal{O}$. By construction the BRST operator obeys

$$Q_B^2 = 0. \tag{5.58}$$

Looking at the action (5.55), we see that the physical part of the action commutes with Q_B while the gauge-variant part is a commutator of Q_B with something else. This motivates us to define a similar condition at the level of states: physical states in $\mathcal{H}_{BRST}$ are those which are annihilated by Q_B, but with the caveat that two such states which differ by a state which is Q_B acting on another state should be identified. Thus in this approach we define the physical Hilbert space to be the kernel of Q_B quotiented by its image:

$$\mathcal{H}_{phys} \equiv \Big\{ |\psi\rangle \in \mathcal{H}_{BRST} \Big| Q_B|\psi\rangle = 0, \qquad |\psi\rangle \sim |\psi\rangle + Q_B|\chi\rangle \Big\}. \tag{5.59}$$

A mathematician would describe this as saying that physical states live in the **cohomology** of the nilpotent operator Q_B. In this language states in the kernel of Q_B are called **closed** and states in its image are called **exact**. In Weinberg or Srednicki you can read about how this leads to the same $\mathcal{H}_{phys}$ for a free Maxwell field that we constructed using gauge-invariant quantization, and this has been proven to all orders in perturbation theory for general Yang-Mills theories. I am not sure however if it really works non-perturbatively without further ingredients, since the BRST approach is ultimately based on gauge-fixing and in the non-abelian case this suffers from global problems such as the Gribov ambiguity.

Given the complaints I made about each of these applications, you might ask why I told you about BRST symmetry at all. This is fair, but it is quite popular in many corners of theoretical physics so it is good for you to learn something about it. One reason why its adherents like it is that they can formally realize manifest Lorentz invariance on $\mathcal{H}_{BRST}$, while in our gauge-invariant construction Lorentz invariance is only

manifest once we go to the path integral. My view however is that I would rather have unitarity be manifest and Lorentz invariance be something we need to check than vice versa. After all any non-perturbative UV regularization will break Lorentz invariance, but we can preserve gauge-invariance and unitarity exactly on the lattice as we will soon see.

5.7 Homework

1. What are the gauge-fixing and ghost actions in axial gauge? Argue that in the limit $\xi \to 0$ the ghosts decouple from the gauge field and thus can be ignored.

2. Check the Gaussian integral (5.34) for $M = 2$ by expanding the exponential and computing the Grassman integral.

3. Confirm the BRST invariance of the action (5.44) under the BRST transformation (5.47) for the special case of Lorenz gauge, with gauge-fixing condition (5.37). Do this by direct calculation, i.e. not using (5.49).

4. Show that $\delta_{\theta_1}\delta_{\theta_2}$ vanishes acting on the gauge field a^a_μ and the ghost c^a. Hint: you will need to use the Jacobi identity for the structure constants.

5. Extra credit: show that the measure (5.29) is indeed proportional to the Haar measure. Hint: it easier to show that it is right-invariant, and then you can use compactness to infer left-invariance. Alternatively, you can show directly that it is the pullback by $R_{g^{-1}}$ of a volume form at the identity, which again shows it is right-invariant.

6 Perturbative Yang-Mills theory

We have now set up a gauge-fixed path integral for evaluating correlation functions of gauge-invariant operators in Yang-Mills theory:

$$\langle\Omega|T\mathcal{O}_1(A,\Phi)\dots\mathcal{O}_n(A,\Phi)|\Omega\rangle = \frac{\int \mathcal{D}a\mathcal{D}\phi\mathcal{D}b\mathcal{D}c\,\mathcal{O}_1(a,\phi)\dots\mathcal{O}_n(a,\phi)e^{i\int d^dx\mathcal{L}_{tot}(a,\phi,b,c)}}{\int \mathcal{D}a\mathcal{D}\phi\mathcal{D}b\mathcal{D}c\,e^{i\int d^dx\mathcal{L}_{tot}(a,\phi,b,c)}}, \tag{6.1}$$

with

$$\mathcal{L}_{tot} = -\frac{1}{4g^2}f^a_{\mu\nu}f_a^{\mu\nu} + \mathcal{L}_M(\phi,D\phi) - \frac{1}{2g^2\xi}\partial^\mu a^a_\mu\partial_\nu a^\nu_a - \partial^\mu b_a D_\mu c^a. \tag{6.2}$$

In this section we will learn how to do perturbative calculations in this theory assuming that the gauge coupling g is small. To be concrete we will take the matter field to be a spinor transforming in some representation α of G with generators τ_a, so the matter Lagrangian is

$$\mathcal{L}_M = -i\overline{\psi}\left(\not{D}+m\right)\psi. \tag{6.3}$$

In this section we will adopt the alternative gauge field normalization

$$\hat{a}_\mu \equiv \frac{1}{g}a_\mu, \tag{6.4}$$

since this is the normalization that is most intuitive for perturbative calculations. Writing everything out, the action is

$$\begin{aligned}\mathcal{L}_{tot} = &-\frac{1}{4}\left(\partial_\mu\hat{a}^a_\nu - \partial_\nu\hat{a}^a_\mu + gC^a{}_{bc}\hat{a}^b_\mu\hat{a}^c_\nu\right)\left(\partial^\mu\hat{a}^\nu_a - \partial^\nu\hat{a}^\mu_a + gC_a{}^{de}\hat{a}^\mu_d\hat{a}^\nu_e\right) - \frac{1}{2\xi}\partial^\mu\hat{a}^a_\mu\partial_\nu\hat{a}^\nu_a \\ &- i\overline{\psi}\left(\not{\partial} + m - ig\tau_a\gamma^\mu\hat{a}^a_\mu\right)\psi - \partial^\mu b_a\left(\partial_\mu c^a + gC^a{}_{bc}\hat{a}^b_\mu c^c\right).\end{aligned} \tag{6.5}$$

If we set $g=0$ this is a free action of $\dim(G)$ Maxwell fields, $\dim(\alpha)$ spinor fields, and $2\dim(G)$ ghost fields with an off-diagonal kinetic term. The Feynman propagators for these fields are obtained by inverting these kinetic terms:

$$\begin{aligned}S_n{}^m(x-y) &\equiv \langle\psi_n(x)\overline{\psi}^m(y)\rangle = \delta^m_n\int\frac{d^dp}{(2\pi)^d}\frac{i(\not{p}+im)}{p^2+m^2-i\epsilon}e^{ip\cdot(x-y)} \\ \Delta^{ab}_{\mu\nu}(x-y) &\equiv \langle\hat{a}^a_\mu(x)\hat{a}^b_\nu(y)\rangle = \hat{g}^{ab}\int\frac{d^dp}{(2\pi)^d}\left(\frac{-i\eta_{\mu\nu}}{p^2-i\epsilon} + i(1-\xi)\frac{p_\mu p_\nu}{(p^2-i\epsilon)^2}\right)e^{ip\cdot(x-y)} \\ \Theta^a{}_b(x-y) &\equiv \langle c^a(x)b_b(y)\rangle = \int\frac{d^dp}{(2\pi)^d}\frac{-i\delta^a_b}{p^2-i\epsilon}e^{ip\cdot(x-y)}.\end{aligned} \tag{6.6}$$

Here the notation $\langle\cdot\rangle$ means evaluating the gauge-fixed path integral (6.1) with the indicated quantity inserted. I have suppressed the spinor indices on the fermion propagator, so the indicated indices are the representation indices only. It is worth emphasizing however that these insertions are *not* gauge-invariant, so these propagators should not be interpreted as expectation values of gauge-invariant local operators on $\mathcal{H}_{phys}$. This is clear for example from the ξ-dependence of $\Delta^{ab}_{\mu\nu}$. They should be viewed as intermediate quantities for use in the perturbative calculation of expectation values of gauge-invariant operators. As in Maxwell theory we will work in the Feynman gauge $\xi=1$, which eliminates the second term in the gauge field propagator. The graphical representations of these propagators in momentum space are shown in figure 6.

In order to do perturbative calculations we also need the interaction vertices. These can be read off from equation (6.5). For example each insertion of the spinor interaction (including an i because we have $i\mathcal{L}_{tot}$ in the path integral) brings down a factor of

$$-ig\overline{\psi}\tau_a\gamma^\mu\psi\hat{a}^a_\mu \tag{6.7}$$

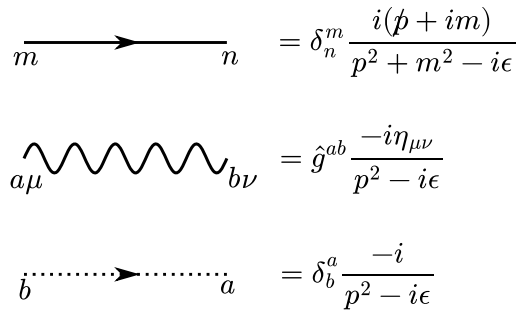

Figure 6: Momentum-space propagators for spinors, gauge fields, and ghosts.

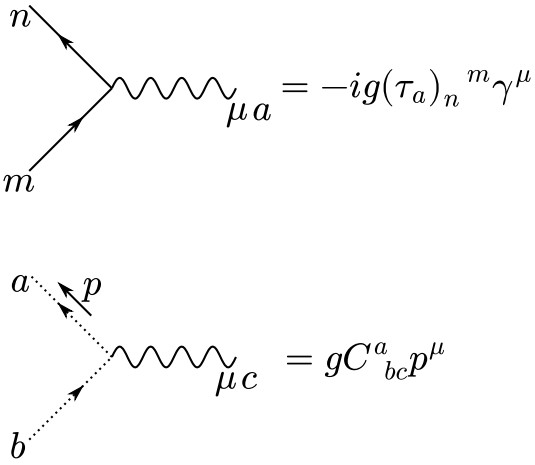

Figure 7: Interaction vertices for spinors and ghosts with a gauge field.

in the path integral. The fields get attached to propagators, with the ψ being the end of a propagator and the $\overline{\psi}$ being the beginning of one. In our usual rule that we multiply propagators along a fermion line from right to left as matrices, this gives an interaction vertex of

$$-ig\tau_a\gamma^\mu. \tag{6.8}$$

Each ghost interaction insertions brings a factor of

$$-igC^a{}_{cb}\partial^\mu b_a \hat{a}^c_\mu c^b. \tag{6.9}$$

To turn this into a momentum-space interaction vertex, we need to remember that in momentum space a derivative in an interaction corresponds to a factor of ip^μ if the momentum is ingoing and $-ip^\mu$ if the momentum is outgoing, with p being the momentum of the propagator which is attached to the field with the derivative. In the ghost propagator we should think of b as the beginning and c as the end (you can compare to $\overline{\psi}$ and ψ), so if p is the momentum of the outgoing b leg the interaction provides a factor of

$$-igC^a{}_{cb} \cdot (-ip^\mu) = gC^a{}_{bc} p^\mu. \tag{6.10}$$

We show the pictures for these vertices in figure 7.[47]

[47] In the ghost vertex if you like you can cyclically permute the adjoint indices so that a is the gauge field, b is the b-ghost, and c is the c-ghost. This then gives the structure constant $C^b{}_{ca}$, you can decide which is easier to remember.

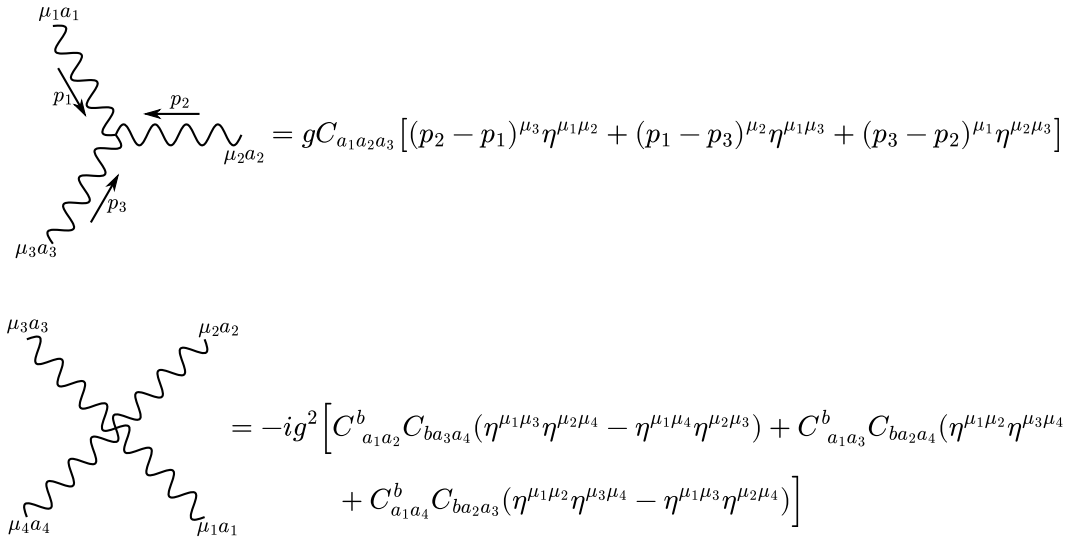

Figure 8: Three-point and four-point interaction vertices for a gauge field.

We also need to consider the gauge field interaction vertices. There are two cross terms in (6.5) that give a three-point vertex for the gauge field, and each time we bring down their sum in the path integral we get a factor of

$$-\frac{ig}{2}C_{abc}\left(\eta^{\mu\alpha}\eta^{\nu\beta} - \eta^{\mu\beta}\eta^{\nu\alpha}\right)\partial_\mu \hat{a}^a_\nu \hat{a}^b_\alpha \hat{a}^c_\beta. \tag{6.11}$$

The path integral attaches this vertex to three gauge-field propagators, but there are six ways of doing this attachment since each propagator has a choice of which of the three gauge fields to use at its endpoint. The interaction term is invariant under $(\alpha, b) \leftrightarrow (\beta, c)$, so to get the full contribution we can multiply by two and sum over cyclic permutations of (ν, a), (α, b), and (β, c). Such permutations do not change the structure constant. Adopting a convention where we refer to the external labels as $(p_1, \mu_1, a_1)\ldots(p_3, \mu_3, a_3)$ and take all momenta to be ingoing, we have the interaction vertex

$$gC_{a_1a_2a_3}\left[p_1^{\mu_2}\eta^{\mu_1\mu_3} - p_1^{\mu_3}\eta^{\mu_1\mu_2} + p_2^{\mu_3}\eta^{\mu_2\mu_1} - p_2^{\mu_1}\eta^{\mu_2\mu_3} + p_3^{\mu_1}\eta^{\mu_3\mu_2} - p_3^{\mu_2}\eta^{\mu_3\mu_1}\right], \tag{6.12}$$

which we can condense a bit into

$$gC_{a_1a_2a_3}\big[(p_2-p_1)^{\mu_3}\eta^{\mu_1\mu_2} + (p_1-p_3)^{\mu_2}\eta^{\mu_1\mu_3} + (p_3-p_2)^{\mu_1}\eta^{\mu_2\mu_3}\big]. \tag{6.13}$$

Finally we need to consider the four-point gauge field vertex, which comes from insertions of

$$-\frac{ig^2}{4}C^a_{\ bc}C_{ade}\eta^{\mu\alpha}\eta^{\nu\beta}\hat{a}^b_\mu \hat{a}^c_\nu \hat{a}^d_\alpha \hat{a}^e_\beta. \tag{6.14}$$

There are 24 ways for the four incoming propagator lines to attach to this vertex, but since the vertex is invariant under $(\mu\nu, bc) \leftrightarrow (\alpha\beta, de)$ and $(\mu, b) \leftrightarrow (\alpha, c)$ we can fix the position of one of them so there are only six distinct ways of attaching. Denoting the external labels as $(\mu_1, a_1)\ldots(\mu_4, a_4)$, we can sum over these six permutations as

$$1234 + 1243 + 1324 + 1342 + 1423 + 1432 \tag{6.15}$$

and then multiply by four. This gives the vertex factor

$$-ig^2\Big[C^b_{\ a_1a_2}C_{ba_3a_4}(\eta^{\mu_1\mu_3}\eta^{\mu_2\mu_4} - \eta^{\mu_1\mu_4}\eta^{\mu_2\mu_3}) + C^b_{\ a_1a_3}C_{ba_2a_4}(\eta^{\mu_1\mu_2}\eta^{\mu_3\mu_4} - \eta^{\mu_1\mu_4}\eta^{\mu_2\mu_3})$$
$$+ C^b_{a_1a_4}C_{ba_2a_3}(\eta^{\mu_1\mu_2}\eta^{\mu_3\mu_4} - \eta^{\mu_1\mu_3}\eta^{\mu_2\mu_4})\Big]. \tag{6.16}$$

These interaction vertices are shown in figure 8.

6.1 External leg factors

The Feynman rules we just described are enough to compute correlation functions of gauge-invariant operators in Yang-Mills theory. In the previous semesters we also introduced Feynman rules for connected scattering amplitudes, for use in computing differential cross sections and decay rates. This required us to introduce additional factors for external legs, which we derived non-perturbatively using the LSZ formula. In Yang-Mills theory however this derivation is complicated by the existence of confinement: when the coupling becomes strong at long distances, the asymptotic scattering states of the theory can have no simple relationship to the fields appearing in the Lagrangian. This possibility is realized in QCD, as we will learn starting next time. Nonetheless it can still be useful to define a "perturbative scattering amplitude" $\widetilde{\mathcal{M}}_c^{pert}$, where we ignore the existence of confinement and treat the fields as if they did indeed correspond to asymptotic scattering states. Roughly speaking this can be a good idea because in theories like QCD the fields are weakly coupled at short distances, so we can think of the perturbative scattering amplitude as describing the "high-energy part" of a hadronic collision. It then needs to be sewn onto some kind of non-perturbative input to explain how the quarks and gluons from the perturbative scattering process are extracted from and converted into hadrons. We will say a bit more about these operations work once we finish setting up the rules and use them to compute a few simple processes. In the meantime the idea is to simply define the perturbative scattering amplitude using the same leg factors we had in quantum electrodynamics:

- Ingoing fermion with momentum p and spin $\sigma \to \sqrt{Z_2} u(p,\sigma)$
- Outgoing fermion with momentum p and spin $\sigma \to \sqrt{Z_2} \overline{u}(p,\sigma)$
- Ingoing antifermion with momentum p and spin $\sigma \to -\sqrt{Z_2} \overline{v}(p,\sigma)$
- Outgoing antifermion with momentum p and spin $\sigma \to -\sqrt{Z_2} v(p,\sigma)$
- Ingoing gauge boson with momentum p and helicity $\sigma \to \sqrt{Z_3} e_\mu(p,\sigma)$
- Outgoing gauge boson with momentum p and helicity $\sigma \to \sqrt{Z_3} e^*_\mu(p,\sigma)$

Typically the minus signs in the antifermion factors are ignored since they always cancel when we square the amplitude, and we will do this too. The wave function renormalization constants here cannot be defined to be on-shell residues as they were in massive theories, due to the strong coupling at long distances, so we need to define them in some less physical way by choosing an arbitrary normalization for the two-point functions of the spinor and gauge fields, for example by specifying their values at some particular momentum. Fortunately we will only use these external factors at tree level however, so we will not have to worry about this.

For some purposes it is also useful to define external leg factors for ghosts. At tree level these can simply be taken to one, while at loop level their renormalization is related to that of the gauge field by gauge-invariance.

6.2 Tree-level quark annihilation and production

One particularly simple amplitude we can consider is the annihilation of a quark and its antiparticle to produce a quark of different flavor and its antiparticle, which we will call $q\overline{q} \to q'\overline{q}'$. For example we could be considering the production of the top quark at a hadron collider like the LHC at CERN or the Tevatron at Fermilab, with the initial q and $\overline{q}$ being constituents of the hadrons that are being collided.This process has a single Feynman diagram that contributes, shown in figure 9. Using the Feynman rules we just derived, the connected covariant scattering amplitude is

$$\begin{aligned} i\widetilde{M}_c &= (-ig)^2 \overline{v}(p_2,\sigma_2)\gamma^\mu u(p_1,\sigma_1) \frac{-i\eta_{\mu\nu}}{(p_1+p_2)^2 - i\epsilon} \overline{u}(p_1',\sigma_1')\gamma^\nu v(p_2',\sigma_2') \, (\tau_a)_{n_2}{}^{n_1} \, (\tau_b)_{n_1'}{}^{n_2'} \, \hat{g}^{ab} \\ &= ig^2 \frac{\overline{v}_2 \gamma^\mu u_1 \cdot \overline{u}_{1'} \gamma_\mu v_{2'}}{(p_1+p_2)^2} \, (\tau_a)_{n_2}{}^{n_1} \, (\tau^a)_{n_1'}{}^{n_2'} , \end{aligned} \tag{6.17}$$

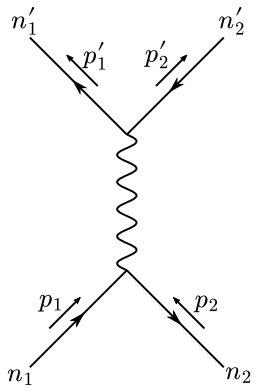

Figure 9: Tree-level $q\overline{q} \to q'\overline{q}'$ scattering.

where in the second line we adopted an abbreviated notation for u and v as in our discussion of spinors last semester and also set $\epsilon = 0$ since we do not need to perform any loop integrals. As in QED we multiply γ-matrices from right to left as we move along a fermion line, and the same is true for the color matrices τ_a. Indeed this amplitude is exactly the same as the amplitude we had in QED for $e^+e^- \to \mu^+\mu^-$, except for the two factors of τ_a that keep track of the colors of the external quarks. When we square the amplitude and sum over spins we will therefore get the same amplitude we had there times what is usually called a "color factor". Moreover we usually average over initial colors and sum over final colors for the same reasons we do for spins, in which case the only difference with the QED amplitude is the presence of the color factor

$$\left| \left(\tau_a\right)_{n_2}{}^{n_1} \left(\tau^a\right)_{n_1'}{}^{n_2'} \right|^2 = \mathrm{Tr}\left(\tau_a\tau_b\right) \mathrm{Tr}\left(\tau^a\tau^b\right). \tag{6.18}$$

We will learn how to compute this kind of quantity later in the section, in the meantime looking back at our QED results we have

$$\sum_{\sigma,\sigma',n,n'} |\widetilde{M}_c|^2 = \mathrm{Tr}\left(\tau_a\tau_b\right) \mathrm{Tr}\left(\tau^a\tau^b\right) \frac{32g^4}{(p_1+p_2)^2} \Big((p_1\cdot p_1')(p_2\cdot p_2') + (p_1\cdot p_2')(p_2\cdot p_1') - m_{q'}^2(p_1\cdot p_2) - m_q(p_1'\cdot p_2') + 2m_q^2 m_{q'}^2 \Big). \tag{6.19}$$

I will leave it to your imagination to convert this into a differential cross section in the center of mass frame, or you can look it up in the lecture notes for QFT II. In particular in the limit that $m_q \ll m_{q'}$ which is relevant for top-quark production, the total spin/color averaged cross section is

$$\sigma_{ave} = \frac{\mathrm{Tr}\left(\tau_a\tau_b\right) \mathrm{Tr}\left(\tau^a\tau^b\right)}{9} \frac{\alpha_s^2\pi}{4\omega^2} \sqrt{1-\frac{m_{q'}}{\omega^2}} \left(1+\frac{m_{q'}^2}{2\omega^2}\right), \tag{6.20}$$

where ω is the energy of each particle in the center of mass frame (they all have to be the same by energy and momentum conservation) and we have defined a strong analogue of the fine structure constant:

$$\alpha_s \equiv \frac{g^2}{4\pi}. \tag{6.21}$$

Measuring this process gives one way of experimentally determining the gauge coupling g. Strictly speaking however this coupling is scale-dependent due to renormalization group effects, as we will see next time, so the usual convention is to report it at the scale of the Z-boson mass:

$$\alpha_s(m_Z) \approx .118. \tag{6.22}$$

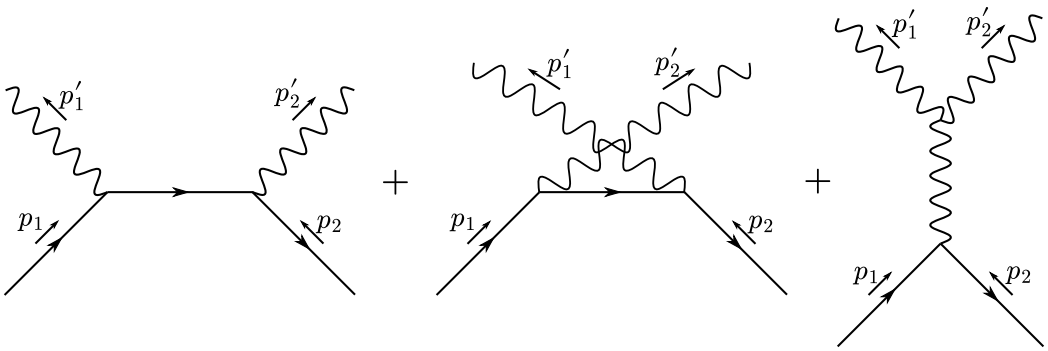

Figure 10: Tree-level contributions to $q\overline{q} \to gg$ scattering.

This is quite a bit larger than the fine structure constant $\alpha \approx .007$, but it is still small enough that perturbative QCD gives a decent description of the process up to the non-perturbative hadronization effects mentioned above.

Indeed this is as good a place as any to say a bit more about how those hadronization effects are dealt with in practice. The basic idea is that each hadron has a nonperturbative **parton distribution function** that tells us how likely we are to find each type of quark and gluon within it, with such and such energy and spin. These functions cannot be computed in perturbation theory, but they can either be fit to experiment or (in some cases) computed numerically using lattice QCD. Similarly each outgoing quark or gluon has a non-perturbative **fragmentation function**, which tells how likely its outgoing jet is to contain each type of hadron given its initial energy, color, and spin. In high-energy hadron scattering the physics which determines these non-perturbative functions happens over much longer length scales than the length scales that are relevant for the hard scattering. So what people do in practice is integrate perturbative QCD scattering amplitudes of the type just computed against parton distribution functions for the initial quarks/gluons and fragmentation functions for the final quarks/gluons, leading to an "inclusive" amplitude for scattering hadrons into jets. Clearly much more could be said about this, but since this is not a class in collider QCD we will leave it here.

6.3 Gluon polarization and ghosts

Another tree-level process we can consider in QCD is the scattering of a quark and antiquark to two gluons, $q\overline{q} \to gg$. The diagrams for this are shown in figure 10. The first two diagrams are the same as we had in QED for Compton scattering (up to being turned on the side), while the third is new and involves the three-gluon vertex. To make things a little nicer we will take the quark masses to zero, in which case the values of these diagrams are

$$\begin{aligned}
i\widetilde{M}_c^{\{1\}} &= (-ig)^2 e^*_{\mu_1} e^*_{\mu_2} \frac{i\overline{v}_2 \gamma^{\mu_2}(\not{p}_1 - \not{p}_1')\gamma^{\mu_1} u_1}{(p_1 - p_1')^2} (\tau_{a_2}\tau_{a_1})_{n_2}{}^{n_1} \\
i\widetilde{M}_c^{\{2\}} &= (-ig)^2 e^*_{\mu_1} e^*_{\mu_2} \frac{i\overline{v}_2 \gamma^{\mu_1}(\not{p}_1 - \not{p}_2')\gamma^{\mu_2} u_1}{(p_1 - p_2')^2} (\tau_{a_1}\tau_{a_2})_{n_2}{}^{n_1} \\
i\widetilde{M}_c^{\{3\}} &= (-ig)^2 e^*_{\mu_1} e^*_{\mu_2} \frac{\overline{v}_2 \gamma_\nu u_1}{(p_1 + p_2)^2} (\tau_b)_{n_2}{}^{n_1} C^b_{a_1 a_2} \big((p_1' - p_2')^\nu \eta^{\mu_1\mu_2} - (p_2' + 2p_1')^{\mu_2} \eta^{\nu\mu_1} + (p_1' + 2p_2')^{\mu_1} \eta^{\nu\mu_2} \big)
\end{aligned} \tag{6.23}$$

where in our formula (6.13) for the three-gluon vertex we have used the ingoing momentum assignments $p_1 \to -p_1'$, $p_2 \to -p_2'$, and $p_3 \to p_1 + p_2$. If you remember our calculation of the Compton scattering cross section last semester, even in the abelian case squaring this amplitude and summing over spins is no picnic.

In QCD it is worse, since now the square has nine terms instead of four. We will not attempt to compute it in detail (see Peskin and Schroeder problem 17.3a if you are feeling brave), and will instead just content ourselves with using this amplitude to illustrate something interesting about the role of ghosts in gauge-fixed amplitudes.[48]

We begin by recalling that in quantum electrodynamics we had a clever trick for evaluating the sum over external photon polarizations. The true polarization sum is

$$\sum_\sigma e_\mu(p,\sigma)e^*_\nu(p,\sigma) = \eta_{\mu\nu} + \ell_\mu p_\nu + p_\mu \ell_\nu, \tag{6.24}$$

where ℓ^μ is a gauge-dependent null vector called the rigging vector that obeys

$$\begin{aligned} \ell \cdot p &= -1 \\ \ell \cdot e(p,\sigma) &= 0. \end{aligned} \tag{6.25}$$

The extra two terms are needed to make sure we only sum over two physical polarization states. What we noticed however was that due to the conservation of the electromagnetic current J^μ, any QED scattering amplitude has a property called the Ward-Takehashi identity. This says that replacing any external polarization vector $e^\mu(p,\sigma)$ by p^μ causes the amplitude to vanish. Therefore in the polarization sum (6.24) we could drop the second two terms, since contracted with the rest of the amplitude they vanish. Unfortunately the situation is not as simple in the non-abelian case, ultimately because the matter current obeys the covariant conservation law

$$D_\mu J^\mu_a = 0 \tag{6.26}$$

instead of the more simple $\partial_\mu J^\mu$ that we had in the abelian case. We will now illustrate this using our tree-level $q\overline{q} \to gg$ amplitude. Indeed let's replace $e_{\mu_1}(p_1',\sigma)$ by p_1'. This gives

$$\begin{aligned} i\widetilde{M}_c^{\{1\}} &\to -ig^2 \frac{\overline{v}_2 \not{\epsilon}^*_{2'}(\not{p}_1 - \not{p}_1')\not{p}_1' u_1}{(p_1 - p_1')^2} \tau_{a_2}\tau_{a_1} \\ &= ig^2 \overline{v}_2 \not{\epsilon}^*_{2'} u_1 \tau_{a_2}\tau_{a_1}, \end{aligned} \tag{6.27}$$

where in going from the first to the second line we have used the massless Dirac equation $\not{p}_1 u_1 = 0$ to replace $\not{p}_1'$ by $\not{p}_1' - \not{p}_1$, after which we used $\not{v}\not{v} = v \cdot v$ to cancel the denominator. Similarly for the second diagram we have

$$\begin{aligned} i\widetilde{M}_c^{\{2\}} &\to -ig^2 \frac{\overline{v}_2 \not{p}_1'(\not{p}_1 - \not{p}_2')\not{\epsilon}^*_{2'} u_1}{(p_1 - p_2')^2} \tau_{a_1}\tau_{a_2} \\ &= -ig^2 \overline{v}_2 \not{\epsilon}^*_{2'} u_1 \tau_{a_1}\tau_{a_2}, \end{aligned} \tag{6.28}$$

where in going from the first to the second line we again used the Dirac equation $\overline{v}_2 \not{p}_2 = 0$ to replace $\not{p}_1'$ by $\not{p}_1' - \not{p}_2 = \not{p}_1 - \not{p}_2'$. Thus the sum of the two diagrams gives

$$i\widetilde{M}_c^{\{1\}} + i\widetilde{M}_c^{\{2\}} \to ig^2 \overline{v}_2 \not{\epsilon}^*_{2'} u_1 [\tau_{a_2}, \tau_{a_1}] = g^2 \overline{v}_2 \not{\epsilon}^*_{2'} u_1 C^b{}_{a_1 a_2} \tau_b. \tag{6.29}$$

In QED this vanishes due to the vanishing of the structure constants, confirming the Ward-Takehashi identity, but in QCD it doesn't so we need to say more. Fortunately we also have the third diagram in figure 10, and in an encouraging sign it has the same color structure of $C^b{}_{a_1 a_2}\tau_b$ so some cancellation is possible. In the third diagram we have

$$\begin{aligned} i\widetilde{M}_c^{\{3\}} &\to -g^2 C^b{}_{a_1 a_2}\tau_b \frac{1}{(p_1+p_2)^2} \overline{v}_2 \Big[e^*_{2'} \cdot p_1'(\not{p}_1' - \not{p}_2') - e^*_{2'} \cdot (p_2' + 2p_1')\not{p}_1' + (p_1+p_2)^2 \not{\epsilon}^*_{2'} \Big] u_1 \\ &= -g^2 C^b{}_{a_1 a_2}\tau_b \frac{1}{(p_1+p_2)^2} \overline{v}_2 \Big[(p_1+p_2)^2 \not{\epsilon}^*_{2'} + (e^*_{2'} \cdot p_2')\not{p}_2' \Big] u_1, \end{aligned} \tag{6.30}$$

[48] One thing that one quickly learns in this business is that using the methods we have introduced so far, one has to work quite hard to compute amplitudes whose final forms end up being surprisingly simple. This suggests that better methods should be available, and indeed many have been discovered. If you are going to compute these things for a living, you will need to learn these methods. One nice on-ramp is Peskin's 2011 review "Simplifying multi-jet QCD computation".

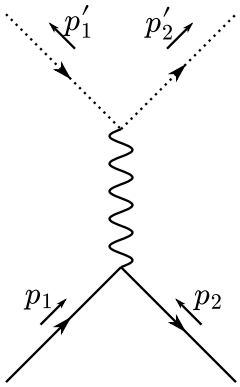

Figure 11: The tree-level contribution to $q\overline{q} \to bc$ "scattering". The purpose of this diagram is only to cancel unphysical gluon polarizations should we choose to include them in the sum over final state polarizations by replacing the sum by $\eta_{\mu\nu}$.

where in going from the first line to the second we used momentum conservation to eliminate p_1' and also used the Dirac equation to set $\overline{v_2}(\not{p}_1 + \not{p}_2)u_1 = 0$. The first term in (6.30) precisely cancels the contribution (6.29) from the first two diagrams. The second term also vanishes provided that we choose $e_{2'}$ to be transversely polarized,

$$e_{2'} \cdot p_2' = 0, \tag{6.31}$$

as indeed we should do for a physical amplitude. On the other hand we see that if we do NOT take $e_{2'}$ to be transversely polarized, then the amplitude is not zero. This is a purely non-abelian effect, that would not have been allowed by the Ward-Takahashi identity.

Because of this failure of the Ward identity, we cannot ignore the rigging vector terms in the spin sum equation (6.24). There are two things we can do about this. The most obvious, which these days is what is actually often done, is to simply compute the amplitudes at fixed gluon helicity and only sum over this at the end. See Peskin and Schroeder problem 17.3 for this approach. An alternative approach which also works is to drop the rigging vector terms in (6.24) but then include an additional contribution to the squared amplitude where we include the possibility of producing a ghost pair in the final state. The diagram for this is shown in figure 11. This may seem like a wacky thing to do, but in fact the ghosts are doing precisely what they are designed to do: canceling the unphysical polarizations of the gluons. We can see this explicitly by evaluating the above diagram:

$$i\widetilde{M}_c^{\{ghost\}} = -g^2 \tau^b C^{a_2}{}_{a_1 b} \frac{\overline{v}_2 \not{p}_2' u_1}{(p_1 + p_2)^2}. \tag{6.32}$$

Choosing $e_{2'}$ to be the rigging vector ℓ in (6.30) and lowering the ghost index a_2 to match what we get for an external gluon, we see that the second term in (6.30) is precisely the same as this ghost contribution (remember that $\ell \cdot p_2' = -1$). Moreover when we square the ghost amplitude and sum over final states we get a minus sign due to the peculiarity of the ghost inner product, so this contribution precisely cancels the unphysical gluon polarizations in the sum.[49] Thus we can either stick to physical gluon polarizations or include unphysical gluon polarizations and also ghosts, and either way we get the same answer! What happened in this example can be shown to hold to all orders in perturbation theory using BRST symmetry, which generalizes the Ward-Takehashi identity to something called the Slavnov-Taylor identity.

[49]The way this sign arises is that using the ghost inner product the completeness relation is $I = |0\rangle\langle 1| + |1\rangle\langle 0|$, so the ghosts in the final state are exchanged when we compute the amplitude squared by inserting a complete set of states. See QFT II for some more discussion of this inner product, or the 1979 paper of Kugo and Ojima for a complete discussion. In Peskin and Schroeder they cheat to get this sign by sticking the amplitude inside of a loop diagram, in which case it comes from the factor of -1 for fermion loops.

6.4 Color factors

So far we have not discussed how to evaluate color factors. We already saw the quantity $\mathrm{Tr}\left(\tau_a\tau_b\right)\mathrm{Tr}\left(\tau^a\tau^b\right)$ in our discussion of $q\overline{q}\to q'\overline{q}'$, and for example if we square the first term in our expression for $q\overline{q}\to gg$ and sum over gluon adjoint indices we encounter $\mathrm{Tr}\left(\tau_a\tau_b\tau^a\tau^b\right)$. Such quantities can be computed using representation theory, as we will now explain. The first thing to note is that for any representation α of a simple Lie algebra $\mathfrak{g}$ we have

$$\mathrm{Tr}\left(\tau_a\tau_b\right) = C_1(\alpha)\hat{g}_{ab} \tag{6.33}$$

for some number $C_1(\alpha)$. This follows from Schur's lemma, since this quantity defines an adjoint-invariant metric on the Lie algebra. Morever we can observe that

$$[\tau_a\tau^a,\tau_b] = [\tau_a,\tau_b]\tau^a + \tau_a[\tau^a,\tau_b] = iC^c{}_{ab}\tau_c\tau^a + iC^c{}_{ab}\tau^a\tau_c = 0 \tag{6.34}$$

by the antisymmetry of C_{cab} in the first two indices, so again by Schur's lemma we must have

$$\tau_a\tau^a = C_2(\alpha) \tag{6.35}$$

for some number $C_2(\alpha)$ that is called the **quadratic Casimir invariant** of the representation. For example if α_j is the spin-j representation of $SU(2)$, then

$$C_2(\alpha_j) = J^2 = j(j+1). \tag{6.36}$$

These two numbers are related by contracting the adjoint indices in (6.33), giving

$$d_\alpha C_2(\alpha) = d_G C_1(\alpha), \tag{6.37}$$

where d_G is the dimension the gauge group G as a manifold and d_α is the matrix dimension of the representation α. Using these we immediately have

$$\mathrm{Tr}\left(\tau_a\tau_b\right)\mathrm{Tr}\left(\tau^a\tau^b\right) = C(\alpha)^2 d_G, \tag{6.38}$$

and with a bit more work we have

$$\begin{aligned}
\tau_a\tau_b\tau^a &= [\tau_a,\tau_b]\tau^a + \tau_b\tau_a\tau^a \\
&= iC^c{}_{ab}\tau_c\tau^a + C_2(\alpha)\tau_b \\
&= \frac{1}{2}iC^{ca}{}_b[\tau_c,\tau_a] + C_2(\alpha)\tau_b \\
&= \frac{1}{2}iC^{ca}{}_b iC^d{}_{ca}\tau_d + C_2(\alpha)\tau_b \\
&= -\frac{1}{2}(iC^c{}_{ba})(iC^a{}_{dc})\tau^d + C_2(\alpha)\tau_b \\
&= -\frac{1}{2}\mathrm{Tr}\left(\tau_b^A\tau_d^A\right)\tau^d + C_2(\alpha)\tau_b \\
&= \left(C_2(\alpha) - \frac{1}{2}C_1(A)\right)\tau_b,
\end{aligned} \tag{6.39}$$

where A indicates the adjoint representation. Thus in the sum over gluon states we have

$$\begin{aligned}
\mathrm{Tr}\left(\tau_a\tau_b\tau^a\tau^b\right) &= \left(C_2(\alpha) - \frac{1}{2}C_1(A)\right)\mathrm{Tr}\left(\tau_b\tau^b\right) \\
&= \left(C_2(\alpha) - \frac{1}{2}C_1(A)\right)C_2(\alpha)d_\alpha.
\end{aligned} \tag{6.40}$$

For future reference we will now compute $C_1(\alpha)$ and $C_2(\alpha)$ for the trivial ($\alpha = 1$), fundamental ($\alpha = F$), and adjoint ($\alpha = A$) representations of $SU(N)$.[50] In the trivial representation we of course have

$$C_1(1) = C_2(1) = 0, \tag{6.41}$$

since the Lie algebra generators vanish. In the fundamental representation we by definition have

$$C_1(F) = \frac{1}{2}, \tag{6.42}$$

so

$$C_2(F) = \frac{d_G}{d_\alpha} C(\alpha) = \frac{N^2 - 1}{2N}. \tag{6.43}$$

In particular when $N = 2$ this gives $3/4$, which agrees with the total angular momentum $j(j+1)$ with $j = \frac{1}{2}$. Computing these for the adjoint representation is a bit trickier, since a direct computation requires us to know the structure constants. We can avoid this however by being clever about representation theory. The trick is to observe that the tensor product of a fundamental and antifundamental representation of $SU(N)$ has a direct sum decomposition

$$F \otimes \overline{F} = 1 \oplus A. \tag{6.44}$$

To see this we note that this product representation acts on the complex vector space $M_n(\mathbb{C})$ of $N \times N$ complex matrices as

$$M' = UMU^\dagger \qquad U \in SU(N). \tag{6.45}$$

We can choose a basis for $M_n(\mathbb{C})$ consisting of the identity matrix together with some basis of traceless hermitian matrices. The former is invariant, which gives the trivial representation, while the latter is just the Lie algebra of $SU(N)$, and thus transforms in the adjoint representation. We can check that the dimensionalities add correctly,

$$N^2 = 1 + (N^2 - 1). \tag{6.46}$$

For any representations α_1, α_2 the generators of the tensor product representation are $\tau_a^1 \otimes I + I \otimes \tau_a^2$, so we have

$$\begin{aligned} \mathrm{Tr}\left((\tau_a^1 \otimes I + I \otimes \tau_a^2)(\tau_b^1 \otimes I + I \otimes \tau_b^2)\right) &= d_{\alpha_2} \mathrm{Tr}\left(\tau_a^1 \tau_b^1\right) + d_{\alpha_1} \mathrm{Tr}\left(\tau_a^2 \tau_b^2\right) \\ &= \left(d_{\alpha_1} C(\alpha_2) + d_{\alpha_2} C(\alpha_1)\right) \hat{g}_{ab}. \end{aligned} \tag{6.47}$$

Here we used that τ_a^1, τ_a^2 are traceless since $\mathfrak{g}$ is simple. We thus have

$$C_1(\alpha_1 \otimes \alpha_2) = d_{\alpha_1} C_1(\alpha_2) + d_{\alpha_2} C_1(\alpha_1), \tag{6.48}$$

so in particular for $SU(N)$ we have

$$C_1(F \otimes \overline{F}) = N. \tag{6.49}$$

We can then observe that by the decomposition (6.44) we also have

$$C_1(F \otimes \overline{F}) = C_1(1) + C_1(A), \tag{6.50}$$

and thus

$$C_1(A) = N. \tag{6.51}$$

Finally from (6.37) we also have

$$C_2(A) = N. \tag{6.52}$$

For $N = 2$ this agrees with the total angular momentum of a spin one particle.

[50] Note that A and F in this paragraph have nothing to do with the gauge field and field strength!

6.5 Homework

1. Derive the propagators (6.6) by evaluating the Gaussian path integral (6.1) at $g = 0$. If you need help I recommend consulting section 5.3 of QFT I and section 6.5 of QFT II.

2. Derive the two interaction vertices for a complex scalar field ϕ coupled to a Yang-Mills field in a representation of G with generators $(\tau_a)_n{}^m$, and also write down the propagator for ϕ.

3. Draw the tree-level diagrams for producing $t\bar{t}$ from two incoming gluons, $gg \to t\bar{t}$. Using the Feynman rules write out the values for each of these diagrams (but you do not need to square the amplitude or sum/average over spins/colors).

4. Draw the one-loop diagrams that renormalize the gauge-field three-point and four-point vertices in Yang-Mills theory coupled to a spinor field. You do not need to evaluate them.

5. Compute the tree-level matrix element squared for $qq' \to qq'$ scattering, summed over initial and final spins and colors. For the QED part you can consult QFT II. If you are feeling enthusiastic you can convert it to a differential cross-section in the center of mass frame.

7 Yang-Mills theory at one loop

We will now study the one-loop behavior of Yang-Mills theory. Our goal is to compute the β-function for the gauge coupling g, which will tell us that with sufficiently few matter fields the theory flows to strong coupling at long distances. In order to show this however we first need a definition of the gauge coupling g. We can of course just use the parameter appearing in the action, which we will now denote g_0 and call the bare coupling, but due to the UV divergences of field theory the relationship between g_0 and what is actually measured is subtle. In the field theories we studied so far we dealt with this problem by defining a renormalized coupling constant using some particularly simple scattering amplitude such as $\phi\phi \to \phi\phi$ in ϕ^4 theory or Coulomb scattering in QED. This option is not available to us in non-abelian gauge theory because of confinement, so we need to instead define g using some kind of correlation function at a momentum scale which is high enough that the theory is still weakly coupled. We thus need to say which correlation function we use, and in particular we need to deal with the fact that the UV divergences also cause the bare fields a and ψ themselves to pick up infinite renormalization factors in their correlation functions. The complete calculation involves seven one-loop Feynman diagrams, so we'll start with some high-level discussion of where we are going.

7.1 Defining renormalization parameters

We begin by writing the bare Lagrangian of Yang-Mills theory prior to gauge fixing:

$$\mathcal{L} = -\frac{1}{4}\hat{f}^a_{\mu\nu}\hat{f}^{\mu\nu}_a - i\overline{\psi}\left(\not{D} + m_0\right)\psi, \tag{7.1}$$

with

$$f^a_{\mu\nu} = \partial_\mu \hat{a}^a_\nu - \partial_\nu \hat{a}^a_\mu + g_0 C^a{}_{bc}\hat{a}^b_\mu \hat{a}^c_\nu \tag{7.2}$$

and

$$D_\mu \psi = \partial_\mu \psi - i g_0 \hat{a}^a_\mu \tau_a \psi. \tag{7.3}$$

The parameters of this theory are the bare fermion mass m_0 and the bare coupling constant g_0. Our goal is to define renormalized versions m and g of these parameters, and also renormalized field operators

$$\begin{aligned} \widetilde{a}^a_\mu &= Z_3^{-1/2}\hat{a}^a_\mu \\ \widetilde{\psi} &= Z_2^{-1/2}\psi, \end{aligned} \tag{7.4}$$

such that the correlation functions of $\widetilde{a}_\mu$ and $\widetilde{\psi}$ are independent of the cutoff when expressed in terms of g and m. In QED we defined m as the location of a pole in the two point function $\langle\psi\overline{\psi}\rangle$ and Z_2 the residue of that pole, and we similarly defined Z_3 as the residue of the $p^2 = 0$ pole in $\langle\hat{a}_\mu\hat{a}_\nu\rangle$. We then defined the renormalized coupling e using the low-energy scattering of an electron off of a background electromagnetic field. These definitions do not work in QCD due to confinement, so we instead need to just choose Z_2, Z_3, and m in some arbitrary way to cancel the divergences. Roughly speaking we expect the bare correlators to have divergences as

$$\begin{aligned} \langle\psi\overline{\psi}\rangle &= Z_2 \times \text{finite} \\ \langle\hat{a}_\mu\hat{a}_\nu\rangle &= Z_3 \times \text{finite} \\ \langle\hat{a}_\mu\psi\overline{\psi}\rangle &= Z_2^2 Z_3 g_0 Z_1^{-1} \times \text{finite}, \end{aligned} \tag{7.5}$$

where in the three point function the factors of Z_2^2 and Z_3 come from quantum corrections to the external propagators, g_0 is there because the correlator vanishes if $g_0 = 0$, and Z_1^{-1} captures any intrinsic divergence in the three-point one-particle-irreducible vertex. For the renormalized three-point function we thus have

$$\langle\widetilde{a}_\mu\widetilde{\psi}\widetilde{\overline{\psi}}\rangle = \frac{Z_2 Z_3^{1/2}}{Z_1} g_0 \times \text{finite}, \tag{7.6}$$

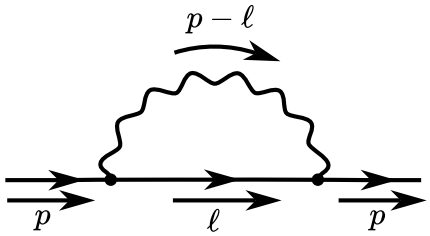

Figure 12: The one-loop contribution to the fermion self-energy.

so we can get a finite renormalized three-point function by defining a finite renormalized coupling constant

$$g \equiv \frac{Z_2\sqrt{Z_3}}{Z_1} g_0. \tag{7.7}$$

To use this formula we of course need to give more precise definitions of Z_1, Z_2, and Z_3 that remove the ambiguity of multiplying each of them by a finite factor. We will do this below when we compute them, and we will then simply adopt (7.7) as our definition of g. To compute the β-function we therefore need to extract the cutoff dependence of Z_1, Z_2, and Z_3.

There is one additional parameter that shows up in our gauge-fixed Lagrangian, which is the gauge-fixing parameter ξ. Nothing physical depends on this parameter, but since we are using gauge-variant quantities such as $\widetilde{a}_\mu$ and ψ to define our renormalization scheme we need to allow for the possibility of a renormalization of ξ. We will therefore refer to the bare parameter as ξ_0, and we will soon see that to preserve our gauge fixing condition at higher loops we need to do an additional renormalization

$$\xi_0 = Z_3 \xi. \tag{7.8}$$

We did not need to discuss this renormalization in QED since we only aimed for a finite S-matrix, which is gauge-invariant.

7.2 Fermion self-energy

We begin by studying Z_2 and m. These are extracted from the fermion two-point function, which we will parametrize in momentum space as

$$\langle \psi(p)\overline{\psi}(p')\rangle = (2\pi)^d \delta^d(p+p') \frac{i}{\not{p} - i(m_0 + \Sigma(\not{p}))} \tag{7.9}$$

just as we did in quantum electrodynamics. $\Sigma(\not{p})$ is called the self-energy of the electron, and by Lorentz invariance it must have the form

$$\Sigma(\not{p}) = A(p^2) + i\not{p} B(p^2). \tag{7.10}$$

In perturbation theory $\Sigma(\not{p})$ is the sum of one-particle irreducible (1PI) diagrams with an external ψ and an external $\overline{\psi}$ but with no external propagator, just as it was in QED. At one-loop this receives a contribution from only one diagram, shown in figure 12. The diagram evaluates to

$$\begin{aligned}\Sigma(\not{p}) &= (-ig_0^d)^2 \tau_a \tau^a \int \frac{d^d\ell}{(2\pi)^d} \frac{\gamma^\mu(\not{\ell} + im_0)\gamma_\mu}{(\ell^2 + m_0^2 - i\epsilon)((\ell-p)^2 - i\epsilon)} \\ &= -(g_0^d)^2 C_2(\alpha) \int \frac{d^d\ell}{(2\pi)^d} \frac{\gamma^\mu(\not{\ell} + im_0)\gamma_\mu}{(\ell^2 + m_0^2 - i\epsilon)((\ell-p)^2 - i\epsilon)},\end{aligned} \tag{7.11}$$

where in anticipation of dimensional regularization we've defined

$$g_{0,d} = \mu^{\frac{4-d}{2}} g_0 \tag{7.12}$$

where μ is some arbitrary mass scale to make sure the coupling has the right dimensions. We already computed this loop integral in our study of QED, with the only new ingredient being the color factor $C_2(\alpha)$. We thus can just use our QED result: setting

$$d = 4 - 2\epsilon, \tag{7.13}$$

we have

$$\Sigma(\not{p}) = \frac{g^2}{8\pi^2} C_2(\alpha) \int_0^1 dx \left[(2m + ix\not{p}) \left(\frac{1}{\epsilon} + \log(4\pi) - \gamma - 1 + \log\left(\frac{\mu^2}{x(1-x)p^2 + (1-x)m^2} \right) \right) + m \right]. \tag{7.14}$$

Here we have replaced g_0 by g and m_0 by m since we are working only to second order in g_0. Introducing

$$\delta Z_2 = Z_2 - 1 \tag{7.15}$$

and

$$\delta m = m - m_0, \tag{7.16}$$

we see that in order for the two-point function of the renormalized field $\widetilde{\psi}$ to be finite we need

$$Z_2 \left(\not{p} - i(m_0 + \Sigma) \right) = \not{p} - im + \delta Z_2(\not{p} - im) - i(\Sigma - \delta m) + \dots \tag{7.17}$$

to be finite.

To proceed further we need to adopt a renormalization scheme that precisely defines Z_2 and δm. The simplest rule is called **minimal subtraction**, or MS for short, in which we define $\delta m/m$ and δZ_2 to consist only of a sum of positive powers of $1/\epsilon$, with coefficients that depend only on the coupling g and not any masses. In other words we choose δm and δZ_2 to subtract only the divergent parts of $\Sigma(\not{p})$. In fact the standard convention in the QCD community is not quite this one however, instead it is something called **modified minimal subtraction**, abbreviated $\overline{\text{MS}}$. This like MS, except that we do the redefinition

$$\mu^2 = \frac{e^\gamma \widetilde{\mu}^2}{4\pi}, \tag{7.18}$$

before choosing δm and δZ_2 to subtract the divergent parts of $\Sigma(\not{p})$. The motivation for this modification is that it removes some annoying factors generated by dimensional regularization, for example we now have

$$\Sigma(\not{p}) = \frac{g^2}{8\pi^2} C_2(\alpha) \int_0^1 dx \left[(2m + ix\not{p}) \left(\frac{1}{\epsilon} - 1 + \log\left(\frac{\widetilde{\mu}^2}{x(1-x)p^2 + (1-x)m^2} \right) \right) + m \right]. \tag{7.19}$$

δm and δZ_2 are still required to be sums of positive powers of $\frac{1}{\epsilon}$ with mass-independent coefficients, with the subleading powers differing in general from those for MS.[51] The divergent part of Σ is

$$\begin{aligned} \Sigma^{div}(\not{p}) &= \frac{g^2}{8\pi^2} C_2(\alpha) \int_0^1 dx \left(2m + ix\not{p} \right) \frac{1}{\epsilon} \\ &= \frac{g^2}{8\pi^2} C_2(\alpha) \left(2m + \frac{i}{2}\not{p} \right) \frac{1}{\epsilon}. \end{aligned} \tag{7.20}$$

Looking at the coefficient of $\not{p}$ in (7.17), we apparently need

$$\delta Z_2 = -\frac{g^2}{16\pi^2} C_2(\alpha) \frac{1}{\epsilon} \tag{7.21}$$

to cancel the divergence. To cancel the divergence which does not have a $\not{p}$, we then need to have

$$\delta m = m \left(\frac{g^2}{4\pi^2} C_2(\alpha) \frac{1}{\epsilon} + \delta Z_2 \right) = \frac{3g^2}{16\pi^2} C_2(\alpha) m \frac{1}{\epsilon}. \tag{7.22}$$

We have thus determined two out of our four renormalization constants! Unfortunately the other two will require more work.

[51] At one loop there are no subleading powers, so δm and δZ_2 coincide between MS and $\overline{\text{MS}}$. The finite renormalized two-point function however looks a bit different since in $\overline{\text{MS}}$ the $\log(4\pi) - \gamma$ has been absorbed into $\widetilde{\mu}$.

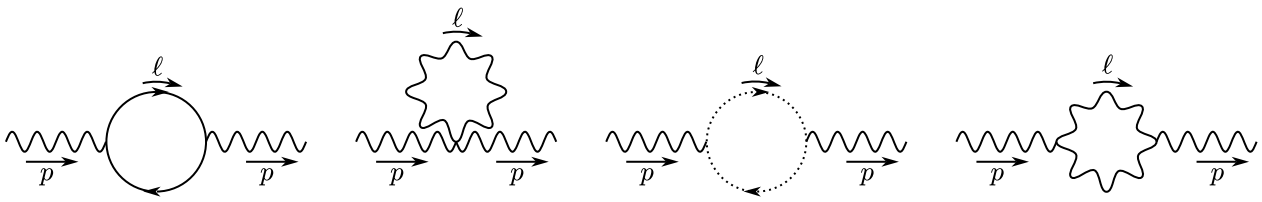

Figure 13: One-loop contributions to the gauge-field self-energy.

7.3 Gauge field self-energy

Next on our list is to compute Z_3. This is defined using the exact gauge-field propagator, which as in QED we can take to have the form

$$\langle \hat{a}^a_\mu(p)\hat{a}^b_\nu(p')\rangle = (2\pi)^d\delta^d(p+p')\hat{g}^{ab}\left(\hat{\Delta}^{-1}-\Pi\right)^{-1}_{\mu\nu}, \tag{7.23}$$

where

$$(\hat{\Delta}^{-1})^{\mu\nu} = i\left(p^2\eta^{\mu\nu}-\left(1-\frac{1}{\xi_0}\right)p^\mu p^\nu\right) \tag{7.24}$$

is the matrix we invert to get the bare momentum-space propagator $\hat{\Delta}$ and $\Pi^{\mu\nu}(p)$ is the self-energy. $\Pi^{\mu\nu}$ is computed by the sum of all 1PI diagrams with two external gauge fields, no external propagators, and a factor of $\hat{g}_{ab}$ removed. In order to have a finite renormalized correlator, we need

$$Z_3\left(p^2\eta^{\mu\nu}-\left(1-\frac{1}{\xi_0}\right)p^\mu p^\nu + i\Pi^{\mu\nu}\right) \tag{7.25}$$

to be finite. Writing[52]

$$\Pi^{\mu\nu} = -i(p^2\eta^{\mu\nu}-p^\mu p^\nu)\Pi, \tag{7.26}$$

this means that we need

$$Z_3\left((1+\Pi)p^2\eta^{\mu\nu}-\left(1-\frac{1}{\xi_0}+\Pi\right)p^\mu p^\nu\right) = Z_3(1+\Pi)\left(p^2\eta^{\mu\nu}-\left(1-\frac{1}{\xi_0(1+\Pi)}\right)p^\mu p^\nu\right) \tag{7.27}$$

to be finite. In $\overline{\text{MS}}$ we do this by choosing

$$Z_3^{-1} = 1+\Pi^{div}, \tag{7.28}$$

where Π^{div} is the divergent part of Π expressed as a polynomial in $\frac{1}{\epsilon}$ with coefficients that are independent of p and m. We can ensure both terms in (7.27) are finite by further choosing

$$\xi_0 = Z_3\xi \tag{7.29}$$

as quoted above, which gives

$$\xi_0(1+\Pi) = \xi\frac{1+\Pi^{div}+\Pi^{finite}}{1+\Pi^{div}} = \xi. \tag{7.30}$$

In particular we can choose $\xi = 1$ to get Feynman gauge to all orders in perturbation theory.

[52]In QED we derived this form using current conservation. That argument does not work in the non-abelian case, but this equation still follows from BRST invariance. We won't give the argument, but we will see that it is true at one loop by explicit calculation.

We now compute $\Pi^{\mu\nu}$ at one loop. There are four diagrams that contribute, shown in figure 13. We computed the first diagram in QED up to a color factor $C_1(\alpha)$ coming from $\mathrm{Tr}\,(\tau_a\tau_b) = C_1(\alpha)\hat{g}_{ab}$, using our QED result we have

$$\begin{aligned}\Pi^{\mu\nu}(p) &\supset -(-ig_{0,d})^2 C_1(\alpha)\int \frac{d^d\ell}{(2\pi)^d}\frac{\mathrm{Tr}\left(i(\not{\ell}+im)\gamma^\mu i(\not{\ell}-\not{p}+im)\gamma^\nu\right)}{(\ell^2+m^2-i\epsilon)((\ell-p)^2+m^2-i\epsilon)} \\ &= -i(p^2\eta^{\mu\nu}-p^\mu p^\nu)C_1(\alpha)\frac{g^2}{2\pi^2}\int_0^1 dx x(1-x)\left[\frac{1}{\epsilon}+\log\left(\frac{\widetilde{\mu}^2}{x(1-x)p^2+m^2}\right)\right]. \end{aligned} \tag{7.31}$$

Thus the divergent part of Π is

$$\Pi^{div} \supset C_1(\alpha)\frac{g^2}{12\pi^2}\frac{1}{\epsilon}. \tag{7.32}$$

Writing

$$Z_3 = 1+\delta Z_3, \tag{7.33}$$

we thus have

$$\delta Z_3 \supset -C_1(\alpha)\frac{g^2}{12\pi^2}\frac{1}{\epsilon}. \tag{7.34}$$

This is the result which gave us the β-function in QED.

To determine the rest of δZ_3 we need to compute the other three diagrams in figure 13. For the second diagram the color factor is

$$C^c{}_{ad}C_{cb}{}^d = -C_1(A)\hat{g}_{ab}, \tag{7.35}$$

so the contribution to the vacuum polarization is

$$\Pi^{\mu\nu} \supset i(g_{0,d})^2\cdot 2(d-1)C_1(A)\eta^{\mu\nu}\int\frac{d^d\ell}{(2\pi)^d}\frac{-i}{\ell^2-i\epsilon}. \tag{7.36}$$

This actually vanishes in dimensional regularization. To see this, recall our key loop integration formula:

$$\int_0^\infty d\ell\frac{\ell^{a-1}}{(\ell^2+D)^b} = \frac{D^{a/2-b}}{2}\frac{\Gamma(a/2)\Gamma(b-a/2)}{\Gamma(b)}, \tag{7.37}$$

Here we are interested in the case $a=d$, $b=1$, $\sigma=0$. Since $a-2b=d-2$, and in dimensional regularization we should take $D\to 0$ before taking $d\to 4$, we get a vanishing answer. For future reference we note now that two useful instances of this formula are

$$\begin{aligned}\int\frac{d^d\ell}{(2\pi)^d}\frac{1}{(\ell^2+D-i\epsilon)^2} &= \frac{i\Gamma\left(2-\frac{d}{2}\right)}{2^d\pi^{\frac{d}{2}}}D^{\frac{d-4}{2}} \\ \int\frac{d^d\ell}{(2\pi)^d}\frac{\ell^2}{(\ell^2+D-i\epsilon)^2} &= \frac{i\Gamma\left(2-\frac{d}{2}\right)}{2^d\pi^{\frac{d}{2}}\left(\frac{2}{d}-1\right)}D^{\frac{d-2}{2}}. \end{aligned} \tag{7.38}$$

To evaluate these we perform the Wick rotation $\ell^0 = i\ell^0_E$, the angular integral to get a factor of

$$\Omega_{d-1} = \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)}, \tag{7.39}$$

and then the radial integral using (7.37). In the second case we also use the Γ-function identity

$$\Gamma(x+1) = x\Gamma(x). \tag{7.40}$$

Turning now to the ghost loop, labeling the adjoint index of the upper ghost propagator as c and the lower one as d, the color factor is

$$C^c{}_{da}C^d{}_{cb} = -C_1(A)\hat{g}_{ab}. \tag{7.41}$$

The contribution to the self-energy is thus

$$\Pi^{\mu\nu} \supset -(g_{0,d})^2 C_1(A) \int \frac{d^d\ell}{(2\pi)^d} \frac{\ell^\mu (\ell-p)^\nu}{(\ell^2 - i\epsilon)((\ell-p)^2 - i\epsilon)}, \tag{7.42}$$

where the explanation of the overall sign is that we get a minus sign from the fermion loop, a minus sign from the color factor, and a minus sign from $(-i)^2$ in the product of ghost propagators. We did not encounter this loop integral in QED, so we will actually need to compute it. As usual we combine the denominators using the Feynman parameter identity

$$\frac{1}{AB} = \int_0^1 dx \frac{1}{(xA + (1-x)B)^2}, \tag{7.43}$$

which gives

$$\begin{aligned}
\Pi^{\mu\nu} &\supset -(g_{0,d})^2 C_1(A) \int_0^1 dx \int \frac{d^d\ell}{(2\pi)^d} \frac{\ell^\mu (\ell-p)^\nu}{\Big((\ell-(1-x)p)^2 + x(1-x)p^2 - i\epsilon\Big)^2} \\
&= -(g_{0,d})^2 C_1(A) \int_0^1 dx \int \frac{d^d\ell}{(2\pi)^d} \frac{(\ell+(1-x)p)^\mu (\ell-xp)^\nu}{\left(\ell^2 + x(1-x)p^2 - i\epsilon\right)^2} \\
&= -(g_{0,d})^2 C_1(A) \int_0^1 dx \int \frac{d^d\ell}{(2\pi)^d} \frac{\frac{1}{d}\ell^2\eta^{\mu\nu} - x(1-x)p^\mu p^\nu}{\left(\ell^2 + x(1-x)p^2 - i\epsilon\right)^2}
\end{aligned} \tag{7.44}$$

In the first line we completed the square in the denominator, in going from the first to the second we did a linear shift of the integration variable to simplify the denominator, in going from the second to the third we eliminated the terms which are linear in ℓ in the numerator since these integrate to zero and also made the replacement $\ell^\mu \ell^\nu \to \frac{1}{d}\eta^{\mu\nu}\ell^2$ using Lorentz invariance. We can evaluate the ℓ integrals using (7.38), giving

$$\Pi^{\mu\nu} \supset ig^2 C_1(A) \frac{\Gamma\left(2 - \frac{d}{2}\right)}{2^d \pi^{d/2}} \int_0^1 dx\, x(1-x) \left(\frac{x(1-x)p^2}{\mu^2}\right)^{\frac{d-4}{2}} \left(\frac{1}{d-2} p^2 \eta^{\mu\nu} + p^\mu p^\nu\right). \tag{7.45}$$

Finally setting $d = 4 - 2\epsilon$ and expanding in ϵ we get

$$\Pi^{\mu\nu} \supset \frac{ig^2 C_1(A)}{32\pi^2} \int_0^1 dx\, x(1-x) \left[\left(\frac{1}{\epsilon} + 1 + \log\left(\frac{\widetilde{\mu}^2}{x(1-x)p^2}\right)\right) p^2 \eta^{\mu\nu} + 2\left(\frac{1}{\epsilon} + \log\left(\frac{\widetilde{\mu}^2}{x(1-x)p^2}\right)\right) p^\mu p^\nu\right]. \tag{7.46}$$

Thus the contribution to the divergent part of $\Pi^{\mu\nu}$ from this diagram is

$$\Pi^{\mu\nu} \supset \frac{ig^2 C_1(A)}{32\pi^2 \epsilon} \left(\frac{1}{6} p^2 \eta^{\mu\nu} + \frac{1}{3} p^\mu p^\nu\right). \tag{7.47}$$

This is not proportional to $p^2\eta^{\mu\nu} - p^\mu p^\nu$, but we will see that it becomes so when combined with the final diagram.

For the final diagram, labeling the adjoint index for the top propagator as c and the bottom one as d, the color factor is

$$C_{adc} C_b{}^{cd} = -C_1(A) \hat{g}_{ab}, \tag{7.48}$$

so the contribution to the self-energy is

$$\Pi^{\mu\nu} \supset \frac{1}{2}(g_{0,d})^2 C_1(A) \int \frac{d^d\ell}{(2\pi)^d} \frac{N^{\mu\nu}}{(\ell^2 - i\epsilon)((\ell-p)^2 - i\epsilon)} \tag{7.49}$$

with

$$N^{\mu\nu} \equiv \Big((\ell-2p)^\alpha \eta^{\mu\beta} + (\ell+p)^\beta \eta^{\mu\alpha} - (2\ell-p)^\mu \eta^{\alpha\beta}\Big)\Big((\ell-2p)_\alpha \delta^\nu_\beta + (\ell+p)_\beta \delta^\nu_\alpha - (2\ell-p)^\nu \eta_{\alpha\beta}\Big). \tag{7.50}$$

For the overall sign a minus sign from the color factor cancels a minus sign from the two gauge-field propagators, and the factor of 1/2 is a symmetry factor. Introducing Feynman parameters as usual we have

$$\Pi^{\mu\nu} \supset \frac{1}{2}(g_{0,d})^2 C_1(A) \int_0^1 dx \int \frac{d^d\ell}{(2\pi)^d} \frac{N^{\mu\nu}}{\left(\ell^2 + x(1-x)p^2 - i\epsilon\right)^2}, \tag{7.51}$$

where we shifted the integration variable as $\ell \to \ell + (1-x)p$ so the numerator is now

$$\begin{aligned} N^{\mu\nu} =& (4d-6)\ell^\mu\ell^\nu + \left(2\ell^2 + \left((2-x)^2 + (1+x)^2\right)p^2\right) \\ &+ \Big(d(1-2x)^2 - 2(1+x)(2-x) - 2(1-2x)(2-x) + 2(1+x)(1-2x)\Big)p^\mu p^\nu \\ =& \left(6\left(1-\frac{1}{d}\right)\ell^2 + (5-2x(1-x))p^2\right) + \big(d(1-2x)^2 - 6(x^2-x+1)\big)p^\mu p^\nu. \end{aligned} \tag{7.52}$$

In the first line we dropped terms that are linear in ℓ since these integrate to zero, and in the second line we used that we can replace $\ell^\mu\ell^\nu \to \frac{1}{d}\eta^{\mu\nu}$ by Lorentz invariance and we also combined some terms. We can now evaluate the loop integral using (7.38), which gives

$$\begin{aligned} \Pi^{\mu\nu} \supset \frac{i}{2} g^2 C_1(A) \frac{\Gamma\left(2-\frac{d}{2}\right)}{2^d \pi^{\frac{d}{2}}} \int_0^1 dx \left(\frac{x(1-x)p^2}{\mu^2}\right)^{\frac{d-4}{2}} \Big[&\left(5 - \left(\frac{6}{d-2} + 8\right)x(1-x)\right)p^2\eta^{\mu\nu} \\ &+ \big(d(1-2x)^2 - 6(x^2-x+1)\big)\,p^\mu p^\nu\Big]. \end{aligned} \tag{7.53}$$

It is not hard to expand this to zeroth order in ϵ, but we only need the divergent piece, which is given by

$$\Pi^{\mu\nu} \supset \frac{ig^2 C_1(A)}{32\pi^2\epsilon}\left(\frac{19}{6}p^2\eta^{\mu\nu} - \frac{11}{3}p^\mu p^\nu\right). \tag{7.54}$$

Combining now the contributions to $\Pi^{\mu\nu}$ from the ghost and gauge field loops, we find the divergent part

$$\Pi^{\mu\nu} \supset \frac{5ig^2 C_1(A)}{48\pi^2\epsilon}\left(p^2\eta^{\mu\nu} - p^\mu p^\nu\right). \tag{7.55}$$

This has the form (7.26) as promised, so we can extract

$$\Pi^{div} \supset -\frac{5g^2 C_1(A)}{48\pi^2\epsilon}, \tag{7.56}$$

and thus

$$\delta Z_3 \supset \frac{5g^2 C_1(A)}{48\pi^2\epsilon}. \tag{7.57}$$

Our full result for Z_3 at one-loop is thus

$$\delta Z_3 = \frac{g^2}{12\pi^2\epsilon}\left(\frac{5}{4}C_1(A) - C_1(\alpha)\right). \tag{7.58}$$

7.4 Vertex renormalization

We now turn to computing Z_1. There are two diagrams, shown in figure 14. Referring to the full 1PI result as $V_3^{\mu,a}$, the first diagram contributes

$$V_3^{\mu,a} \supset -i(-ig_{0,d})^3 \tau_b\tau_a\tau_b \int \frac{d^d\ell}{(2\pi)^d} \frac{\gamma^\nu i(\not{p}' - \not{p} + \not{\ell} + im)\gamma^\mu i(\not{\ell} + im)\gamma_\nu}{(\ell^2 + m^2 - i\epsilon)((\ell + p' - p)^2 + m^2 - i\epsilon)((\ell - p)^2 - i\epsilon)}. \tag{7.59}$$

Except for the color factor

$$\tau_b\tau_a\tau^b = \left(C_2(\alpha) - \frac{1}{2}C_1(A)\right)\tau_a \tag{7.60}$$

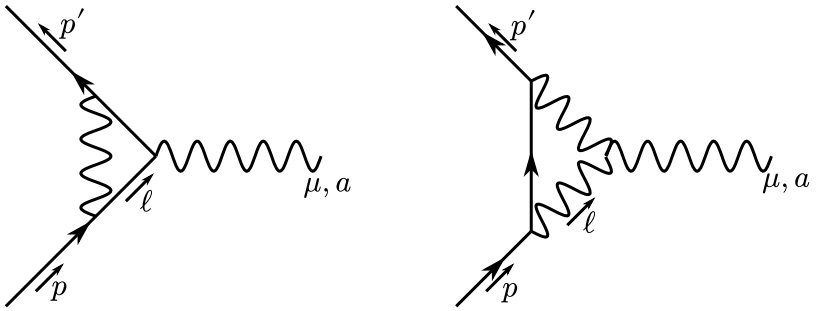

Figure 14: One-loop contributions to the renormalization of the fermion-gauge field vertex.

this is the same loop integral that we studied in QED when we computed $g-2$, and with a little work we can extract from that calculation the divergence we need to find the contribution to Z_1. On the other hand that calculation made several assumptions that are not so natural here: it put the external momenta p and p' on shell, and sandwiched the amplitude between $\overline{u}$ and u in order to extract the coupling of the QED current to a background electric field. We also used Pauli-Villars instead of dimensional regularization. Rather than working out the details of this extraction, we will instead just proceed from scratch. We can save a lot of work however by deciding at the beginning that we will only keep contributions that can contribute to the divergent part of the integral. In particular this means we can set $m = 0$ and $p = p'$, which amounts to working in the approximation that

$$|p^2| \gg |(p-p')^2|, m^2. \tag{7.61}$$

We cannot however set $p = 0$, as that would lead to an infrared divergence.[53]

Proceeding with the calculation, in this approximation we have

$$V_3^{\mu,a} \supset -g_{0,d}^3 \left(C_2(\alpha) - \frac{1}{2} C_1(A) \right) \tau_a \int \frac{d^d\ell}{(2\pi)^d} \frac{\gamma^\nu \not{\ell} \gamma^\mu \not{\ell} \gamma_\nu}{(\ell^2 - i\epsilon)^2((\ell-p)^2 - i\epsilon)}. \tag{7.62}$$

We can combine the dominators using

$$\frac{1}{ABC} = 2 \int_0^1 dx \int_0^1 dy \int_0^1 dz \delta(x+y+z-1) \frac{1}{(xA+yB+zC)^3}, \tag{7.63}$$

which gives

$$\begin{aligned} V_3^{\mu,a} &\supset -2g_{0,d}^3 \left(C_2(\alpha) - \frac{1}{2} C_1(A) \right) \tau_a \int_0^1 dz (1-z) \int \frac{d^d\ell}{(2\pi)^d} \frac{\gamma^\nu \not{\ell} \gamma^\mu \not{\ell} \gamma_\nu}{((\ell - zp)^2 + z(1-z)p^2 - i\epsilon)^3} \\ &= -2g_{0,d}^3 \left(C_2(\alpha) - \frac{1}{2} C_1(A) \right) \tau_a \int_0^1 dz (1-z) \int \frac{d^d\ell}{(2\pi)^d} \frac{\gamma^\nu (\not{\ell} + z\not{p}) \gamma^\mu (\not{\ell} + z\not{p}) \gamma_\nu}{(\ell^2 + z(1-z)p^2 - i\epsilon)^3}. \end{aligned} \tag{7.64}$$

In the first line we used that

$$\int_0^1 dx \int_0^1 dy \int_0^1 dz \delta(x+y+z-1) f(z) = \int_0^1 dz \int_0^{1-z} dx f(z) = \int_0^1 dz (1-z) f(z), \tag{7.65}$$

while in the second we shifted the integration variable. We can drop the terms that are linear in ℓ in the numerator as usual, and actually only the quadratic term can potentially lead to a logarithmic divergence.

[53] You might ask why we did not make such early-stage approximations in the previous calculations for this section. The reason is that for the two-point functions there are potentially both quadratic and logarithmic divergences, so extracting the logarithmic divergence is more subtle. For $V_3^{\mu,a}$ the logarithmic divergence is the leading one, so it is easier to see what terms can contribute to it.

We can further simplify the numerator using our γ-matrix identities

$$\begin{aligned}\gamma^\nu\gamma^\mu\gamma_\nu &= -(d-2)\gamma^\mu \\ \gamma^\nu\gamma^\alpha\gamma^\mu\gamma^\beta\gamma_\nu &= -2\gamma^\beta\gamma^\mu\gamma^\alpha - (d-4)\gamma^\alpha\gamma^\mu\gamma^\beta\end{aligned} \tag{7.66}$$

(see the problems for section 11 of QFT II), and also the replacement $\ell^\alpha\ell^\beta \to \frac{1}{d}\ell^2\eta^{\alpha\beta}$, giving

$$V_3^{\mu,a} \supset -2g_{0,d}^3\left(C_2(\alpha) - \frac{1}{2}C_1(A)\right)\tau_a\gamma^\mu\frac{(d-2)^2}{d}\int_0^1 dz(1-z)\int\frac{d^d\ell}{(2\pi)^d}\frac{\ell^2}{\left(\ell^2 + z(1-z)p^2 - i\epsilon\right)^3}. \tag{7.67}$$

We can evaluate the loop integral using

$$\int\frac{d^d\ell}{(2\pi)^d}\frac{\ell^2}{\left(\ell^2 + D - i\epsilon\right)^3} = i\frac{d}{2^{d+2}\pi^{\frac{d}{2}}}D^{\frac{d-4}{2}}\Gamma\left(2 - \frac{d}{2}\right), \tag{7.68}$$

so

$$V_3^{\mu,a} \supset -ig_{0,d}g^2\left(C_2(\alpha) - \frac{1}{2}C_1(A)\right)\tau_a\gamma^\mu\frac{(d-2)^2\Gamma\left(2-\frac{d}{2}\right)}{2^{d+1}\pi^{\frac{d}{2}}}\int_0^1 dz(1-z)\left(\frac{z(1-z)p^2}{\mu^2}\right)^{\frac{d-4}{2}}. \tag{7.69}$$

The divergence comes from

$$\Gamma(2 - \frac{d}{2}) = \Gamma(\frac{1}{\epsilon}) = \frac{1}{\epsilon} + \ldots, \tag{7.70}$$

so everywhere else we can set $d = 4$ to get the divergent part

$$V_3^{\mu,a} \supset -ig_{0,d}\gamma^\mu\tau_a\left(C_2(\alpha) - \frac{1}{2}C_1(A)\right)\frac{g^2}{16\pi^2\epsilon}. \tag{7.71}$$

To determine the contribution to Z_1, we note that we are requiring that

$$Z_1 V_3^{\mu,a} \tag{7.72}$$

be finite. Including the tree-level and one-loop contributions to $V_3^{\mu,a}$ (the tree-level contribution is just $-ig_{0,d}\gamma^\mu\tau_a$) and also writing

$$Z_1 = 1 + \delta Z_1, \tag{7.73}$$

we thus need

$$\delta Z_1 \supset -\frac{g^2}{16\pi^2\epsilon}\left(C_2(\alpha) - \frac{1}{2}C_1(A)\right). \tag{7.74}$$

Turning now to the second diagram, its contribution to $V_3^{\mu,a}$ is

$$\begin{aligned}V_3^{\mu,a} &\supset (-ig_{0,d})^3 C_{abc}\tau^b\tau^c\int\frac{d^d\ell}{(2\pi)^d}\frac{\gamma_\alpha\left(\not{p} - \not{\ell} + im\right)\gamma_\beta}{(\ell^2 - i\epsilon)((\ell+q)^2 - i\epsilon)((\ell-p)^2 + m^2 - i\epsilon)}\Big[-(2q+\ell)^\beta\eta^{\mu\alpha} + (q-\ell)^\alpha\eta^{\mu\beta} + (q+2\ell)^\mu\eta^{\alpha\beta}\Big] \\ &= -\frac{1}{2}C_1(A)\tau_a(g_{0,d})^3\int\frac{d^d\ell}{(2\pi)^d}\frac{-\gamma^\mu(\not{p} - \not{\ell} + im)(2\not{q} + \not{\ell}) + (\not{q} - \not{\ell})(\not{p} - \not{\ell} + im)\gamma^\mu + \gamma_\nu(\not{p} - \not{\ell} + im)\gamma^\nu(q+2\ell)^\mu}{(\ell^2 - i\epsilon)((\ell+q)^2 - i\epsilon)((\ell-p)^2 + m^2 - i\epsilon)},\end{aligned} \tag{7.75}$$

where

$$q = p' - p \tag{7.76}$$

is the **momentum transfer**. As for the previous diagram we are only interested in the divergent part, so we can set $q = 0$ and $m = 0$ to get

$$V_3^{\mu,a} \supset -\frac{1}{2}C_1(A)\tau_a(g_{0,d})^3\int\frac{d^d\ell}{(2\pi)^d}\frac{\gamma^\mu(\not{\ell} - \not{p})\not{\ell} + \not{\ell}(\not{\ell} - \not{p})\gamma^\mu - 2\gamma_\nu(\not{\ell} - \not{p})\gamma^\nu\ell^\mu}{(\ell^2 - i\epsilon)^2((\ell-p)^2 - i\epsilon)}. \tag{7.77}$$

Using again the Feynman parameter identity (7.63), we have

$$\begin{aligned}
V_3^{\mu,a} &\supset -C_1(A)\tau_a(g_{0,d})^3 \int_0^1 dz(1-z) \int \frac{d^d\ell}{(2\pi)^d} \frac{\gamma^\mu(\ell\!\!\!/ - p\!\!\!/)\ell\!\!\!/ + \ell\!\!\!/(\ell\!\!\!/ - p\!\!\!/)\gamma^\mu - 2\gamma_\nu(\ell\!\!\!/ - p\!\!\!/)\gamma^\nu \ell^\mu}{\left((\ell - zp)^2 + z(1-z)p^2 - i\epsilon\right)^3} \\
&\supset -C_1(A)\tau_a(g_{0,d})^3 \int_0^1 dz(1-z) \int \frac{d^d\ell}{(2\pi)^d} \frac{2\ell^2\gamma^\mu + 2(d-2)\ell\!\!\!/\ell^\mu}{\left(\ell^2 + z(1-z)p^2 - i\epsilon\right)^3} \\
&= -C_1(A)\gamma^\mu\tau_a(g_{0,d})^3 \int_0^1 dz(1-z)\frac{4(d-1)}{d} \int \frac{d^d\ell}{(2\pi)^d} \frac{\ell^2}{\left(\ell^2 + z(1-z)p^2 - i\epsilon\right)^3} \qquad (7.78) \\
&= -ig_{0,d}\gamma^\mu\tau_a C_1(A) g^2 \frac{(d-1)\Gamma\left(2 - \frac{d}{2}\right)}{2^d \pi^{\frac{d}{2}}} \int_0^1 dz(1-z) \left(\frac{z(1-z)p^2}{\mu}\right)^{\frac{d-4}{2}}, \qquad (7.79)
\end{aligned}$$

where in going from the first line to the second we did a linear shift of inegration variable, keeping only quadratic terms in ℓ in the numerator, and also used a γ-matrix identity (7.66). In going from the second to the third line we used the replacement $\ell^\mu\ell^\alpha \to \frac{1}{d}\eta^{\mu\alpha}$, and in going from the third to the fourth we used the loop integral (7.68). The divergent part is thus

$$V_3^{\mu,a} \supset -ig_0\gamma^\mu\tau_a \cdot \frac{3C_1(A)g^2}{32\pi^2\epsilon}, \qquad (7.80)$$

which gives a contribution

$$\delta Z_1 \supset -\frac{3C_1(A)g^2}{32\pi^2\epsilon} \qquad (7.81)$$

to Z_1. Our full one-loop result for the vertex renormalization factor is thus

$$\delta Z_1 = -\frac{g^2}{16\pi^2\epsilon}\left(C_2(\alpha) + C_1(A)\right). \qquad (7.82)$$

7.5 Calculation of the β function

We have now completed our calculation of the renormalization of Yang-Mills theory coupled to a spinor field. The renormalization constants in the $\overline{\text{MS}}$ scheme are

$$\begin{aligned}
m &= m_0\left(1 + \frac{3g^2}{16\pi^2\epsilon}C_2(\alpha) + \dots\right) \\
Z_1 &= 1 - \frac{g^2}{16\pi^2\epsilon}\left(C_2(\alpha) + C_1(A)\right) + \dots \\
Z_2 &= 1 - \frac{g^2}{16\pi^2\epsilon}C_2(\alpha) + \dots \\
Z_3 &= 1 + \frac{g^2}{16\pi^2\epsilon}\left(\frac{5}{3}C_1(A) - \frac{4}{3}n_F C_1(\alpha)\right) + \dots,
\end{aligned} \qquad (7.83)$$

where in all cases the neglected terms are of order g^4. Here we have slightly generalized the above calculation to allow for n_F fermions transforming in the representation α of G: the only diagram which is modified is the first one in figure 13, which gets multiplied by n_F.[54] Using our definition (7.7) of the renormalized coupling constant we thus have

$$\begin{aligned}
g &= g_0\left(1 + \delta Z_2 - \delta Z_1 + \frac{1}{2}\delta Z_3 + \dots\right) \\
&= g_0\left(1 + \frac{g^2}{32\pi^2\epsilon}\left(\frac{11}{3}C_1(A) - \frac{4}{3}n_F C_1(\alpha)\right) + \dots\right).
\end{aligned} \qquad (7.84)$$

[54] If any of the fermions obey a Majorana condition, then as you might guess they count as 1/2 in n_F.

This formula is the main result of this section, but to extract some physics from it we need think about the physical meaning of the quantity ϵ.

In QED we proposed a relationship

$$\log \Lambda^2 = \frac{1}{\epsilon} + \log \widetilde{\mu}^2 + O(1) \tag{7.85}$$

between the parameters of dimensional regularization and some more physical cutoff scheme with an explicit energy cutoff Λ. We motivated this by comparing the results of dimensional regularization to those of a Pauli-Villars regulator. We can also motivate it more directly by the structure of loop integrals: any logarithmic divergence at one loop has the qualitative form

$$\int_0^\Lambda d\ell \frac{\ell^{2n-1}}{(\ell^2+D)^n} = \log\left(\frac{\Lambda}{\sqrt{D}}\right) + O(1), \tag{7.86}$$

where D is some quantity with energy dimension two that is built out of external momenta and masses. If we assume that these all have ratios to some fixed renormalization scale $\widetilde{\mu}$ that are $O(1)$, then in the right-hand side we can replace $\sqrt{D}$ in the logarithm by $\widetilde{\mu}$ while maintaining the validity of this expression. In dimensional regularization we instead assign a value to this integral of

$$\mu^{2\epsilon} \int_0^\infty d\ell \frac{\ell^{2(n-\epsilon)-1}}{(\ell^2+D)^n} = \frac{1}{2}\left(\frac{\mu^2}{D}\right)^\epsilon \frac{\Gamma(n-\epsilon)\Gamma(\epsilon)}{\Gamma(n)} = \frac{1}{2\epsilon} + O(1). \tag{7.87}$$

Comparing (7.86) to (7.87) and replacing $\sqrt{D} \to \widetilde{\mu}$ in the logarithm, we see that at one loop we can indeed use the replacement (7.85) to convert dimensional regularization results for the logarithmically divergent part of some quantity to results from some more physical cutoff scheme such as a lattice (the differences between the physical schemes show up only in the $O(1)$ parts).

With this understanding, we can express the bare coupling as a function of the renormalized coupling as

$$g_0 = g\left(1 - \frac{g^2}{16\pi^2}\left(\frac{11}{3}C_1(A) - \frac{4}{3}n_F C_1(\alpha)\right)\log\left(\frac{\Lambda}{\widetilde{\mu}}\right) + \dots\right). \tag{7.88}$$

From the Wilsonian point of view what this equation is telling us is how to vary the bare coupling g_0 as we change Λ to preserve the IR physics. Usually this is captured by the renormalization group β-function

$$\Lambda \frac{dg_0}{d\Lambda} \equiv \beta(g_0) = -\frac{g_0^3}{16\pi^2}\left(\frac{11}{3}C_1(A) - \frac{4}{3}n_F C_1(\alpha)\right) + \dots, \tag{7.89}$$

which is the famous formula that won Gross, Politzer, and Wilzcek the 2004 Nobel Prize in Physics. In particular if

$$\frac{11}{3}C_1(A) > \frac{4}{3}n_F C_1(\alpha), \tag{7.90}$$

then the bare coupling constant decreases as Λ increases compared to the scale $\widetilde{\mu}$. This is called **asymptotic freedom**, since the theory becomes free at very high energy.

We can also think about this physics by writing the renormalized coupling as a function of the bare coupling as

$$g = g_0\left(1 + \frac{g_0^2}{16\pi^2}\left(\frac{11}{3}C_1(A) - \frac{4}{3}n_F C_1(\alpha)\right)\log\left(\frac{\Lambda}{\widetilde{\mu}}\right) + \dots\right). \tag{7.91}$$

This tells us that when (7.90) holds, if we fix the bare coupling g_0 and cutoff Λ, then the renormalized coupling g grows as we lower the scale $\widetilde{\mu}$ at which we do experiments. From this point of view it is more natural to define the β-function as

$$\widetilde{\mu}\frac{dg}{d\widetilde{\mu}} \equiv \beta(g) = -\frac{g^3}{16\pi^2}\left(\frac{11}{3}C_1(A) - \frac{4}{3}n_F C_1(\alpha)\right) + \dots. \tag{7.92}$$

This function looks the same either way because to determine g from g_0 or g_0 from g we solve the same renormalization group differential equation, merely exchanging the initial and final conditions. Whichever way we think about it, for QCD we have

$$C_1(A) = 3 \tag{7.93}$$

and

$$C_1(\alpha) = 1/2, \tag{7.94}$$

so (7.90) holds provided that

$$n_F < 33/2. \tag{7.95}$$

This is indeed true in QCD since it has only six quark flavors (and in fact we will see shortly that it effectively has fewer at low energies), so QCD is asymptotically free. We can write the energy scale at which g becomes order one in terms of the bare parameters as

$$\widetilde{\mu}_{strong} \sim \Lambda e^{-\frac{16\pi^2}{g_0^2}\frac{1}{\frac{11}{3}C_1(A)-\frac{4}{3}n_F C_1(\alpha)}}, \tag{7.96}$$

in QCD this energy scale is conventionally called Λ_{QCD}. This mechanism for generating an exponentially low energy scale in terms of the bare coupling and cutoff scale is sometimes called **dimensional transmutation**.

7.6 Asymptotic freedom in QCD

Why did Gross, Politzer, and Wilczek get the Nobel prize for showing that QCD is asymptotically free? The reason is that it gives a plausible resolution of the biggest puzzle about the strong interactions: the quantum numbers of hadrons are consistent with them being bound states of quarks, but no quarks have ever been seen in isolation. In fact by the late 1960s this puzzle was even sharper, since scattering experiments at SLAC of high-energy (say $E \gtrsim 10$GeV) electrons off of a fixed proton target, usually called **deep inelastic scattering**, were showing a strong deflection of the electrons as if they scattering electromagnetically off of individual quarks within the proton (the relevant term here is "Bjorken scaling", although I won't try to explain what was scaling with what). Asymptotic freedom fit the bill for this perfectly: at high energies (or short distances/times) the quarks in the proton are weakly-interacting constituents, often called **partons** in this context, but at lower energies they must be strongly interacting to explain why the proton has such a large binding energy and quarks are not seen in isolation. It was the discovery of asymptotic freedom that convinced most high-energy physicists that QCD is likely the right theory of the strong interactions.

7.7 More on the β function

7.7.1 Renormalization of other vertices

Having worked so hard to compute the β function, we should try to learn a little more about it. The first point to make is that we chose to define the renormalized coupling using the $\langle\widetilde{a}\widetilde{\psi}\widetilde{\overline{\psi}}\rangle$ three-point function. You might ask why we did not $\langle\widetilde{a}\widetilde{a}\widetilde{a}\rangle$ or $\langle\widetilde{a}\widetilde{a}\widetilde{a}\widetilde{a}\rangle$, and indeed we could have. Would this have given a different result? In other words can we be sure that the renormalization that made $\langle\widetilde{a}\widetilde{\psi}\widetilde{\overline{\psi}}\rangle$ finite also made these other renormalized correlators finite? In the bare action the vertices for these are all related by gauge invariance, which is encouraging, but we can we sure that these relations carry over at higher loops? To address these questions we can repeat our qualitative discussion of the divergence of the fermion vertex. Namely the bare three-point function fo the gauge field should have the form

$$\langle\hat{a}\hat{a}\hat{a}\rangle = \frac{Z_3^3}{Z_{1,3g}} g_0 \times \text{finite}, \tag{7.97}$$

where the factors of Z_3 come from the external exact propagators and $Z_{1,3g}$ captures whatever divergence is generated by loop corrections to the 1PI three-point gauge field vertex. As in our discussion of Z_1 above,

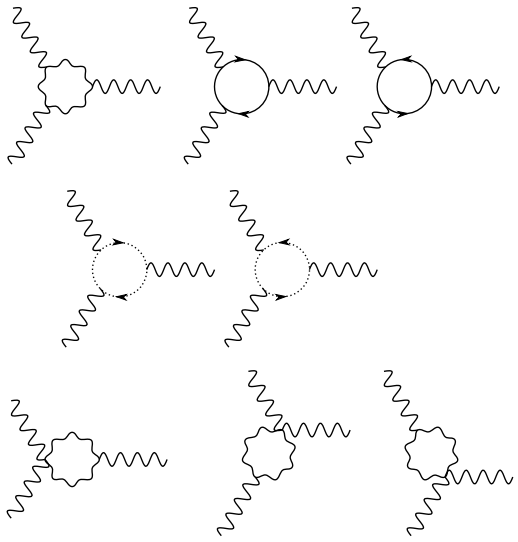

Figure 15: One-loop contributions to the renormalization of the gauge field three-point vertex. Note that there are only really four distinct diagrams to compute, but then we need to sum over permutations of how they are attached to the external legs.

we determine $Z_{1,3g}$ in perturbation theory by requiring that its product with the 1PI three-point gauge field vertex is finite. The renormalized three-point function is

$$\langle \widetilde{a}\widetilde{a}\widetilde{a}\rangle = \frac{Z_3^{3/2}}{Z_{1,3g}} g_0 \times \text{finite} = \frac{Z_3 Z_1}{Z_{1,3g} Z_2} g \times \text{finite}, \tag{7.98}$$

where in the second equality we used our definition (7.7) of the renormalized coupling. Thus to get a finite renormalized three-point function, we need

$$Z_{1,3g} = \frac{Z_3 Z_1}{Z_2}, \tag{7.99}$$

which at one loop is equivalent to requiring that

$$\delta Z_{1,3g} = \delta Z_1 + \delta Z_3 - \delta Z_2. \tag{7.100}$$

The Feynman diagrams for computing $\delta Z_{1,3g}$ at one loop are shown in figure 15, in the homework you will evaluate them to show that (7.100) indeed holds.

We can also do a similar analysis for the four-point gauge field vertex, the bare correlator obeys

$$\langle \hat{a}\hat{a}\hat{a}\hat{a}\rangle = \frac{Z_3^4}{Z_{1,4g}} g_0^2 \times \text{finite} \tag{7.101}$$

and the renormalized correlator is

$$\langle \widetilde{a}\widetilde{a}\widetilde{a}\widetilde{a}\rangle = \frac{Z_3^2}{Z_{1,4g}} g_0^2 \times \text{finite} = \frac{Z_3 Z_1^2}{Z_{1,4g} Z_2^2} g^2 \times \text{finite}, \tag{7.102}$$

the finiteness of which implies

$$Z_{1,4g} = \frac{Z_3 Z_1^2}{Z_2^2}, \tag{7.103}$$

and thus

$$\delta Z_{1,4g} = \delta Z_3 + 2\delta Z_1 - 2\delta Z_2. \tag{7.104}$$

I won't make you check this one, but it indeed is also true.

7.7.2 Integrating out massive fields

Our expression (7.92) for evolution of the renormalized coupling constant as a function of the renormalization scale $\widetilde{\mu}$ has a somewhat counterintuitive property: it includes all spinor fields transforming in representation α of the gauge group without any regard for how their mass m compares to the renormalization scale $\widetilde{\mu}$. For example say there were a 7th quark with a mass of 10,000 GeV: could we actually infer its existence from the running of the coupling constant measured at low energies? Clearly the answer should be no, and that means that there is something artificial about our $\overline{\text{MS}}$ prescription for defining the renormalized coupling.

The source of the problem is that in $\overline{\text{MS}}$ we defined Z_1, Z_2, Z_3 by removing only a polynomial in $\frac{1}{\epsilon}$ with mass-independent coefficients: this is enough to make the renormalized correlators finite, as we saw explicitly at one loop, but in general it is *not* enough to make them $O(1)$. It will do this when all of the dimensionful parameters such as p^2, m^2, $\widetilde{\mu}^2$ are of similar size, but if there is a large hierarchy between them then this can lead to large factors such as $\log \frac{\widetilde{\mu}^2}{m^2}$. A truly physical renormalization scheme should be one that defines the renormalized correlators to be $O(1)$ when the external momenta are of order $\widetilde{\mu}$, and $\overline{\text{MS}}$ does not do this. To illustrate this lets return to our expression (7.31) for the contribution of a fermion loop to the vacuum polarization $\Pi^{\mu\nu}$ (this is where the fermion contribution to the β function came from). The simplest scheme that ensures a renormalized two-point function which is $O(1)$ is to define Z_3 so that it sets $\Pi^{\mu\nu} = 0$ when $p^2 = \widetilde{\mu}^2$, which it does by subtracting $\Pi_{\mu\nu}$ evaluated at this momentum. In other words we want

$$\begin{aligned}\delta Z_3 \supset & -C_1(\alpha)\frac{g^2}{2\pi^2}\int_0^1 dx x(1-x)\left[\frac{1}{\epsilon} + \log\left(\frac{\widetilde{\mu}^2}{x(1-x)\widetilde{\mu}^2 + m^2}\right)\right] \\ = & -C_1(\alpha)\frac{g^2}{2\pi^2}\int_0^1 dx x(1-x)\log\left(\frac{\Lambda^2}{x(1-x)\widetilde{\mu}^2 + m^2}\right). \end{aligned} \tag{7.105}$$

Using this in our expression (7.84) for the renormalized coupling, we see that it now has the form

$$g = g_0\left(1 + \frac{g_0^2}{16\pi}\frac{11}{3}C_1(A)\log\left(\frac{\Lambda}{\widetilde{\mu}}\right) - C_1(\alpha)\frac{g_0^2}{2\pi^2}\int_0^1 dx x(1-x)\log\left(\frac{\Lambda^2}{x(1-x)\widetilde{\mu}^2 + m^2}\right)\right). \tag{7.106}$$

When $m \ll \mu$ this gives the same formula we had before, but when $m \gg \mu$ then the $\widetilde{\mu}$-dependence of the second term is suppressed by a factor of $\frac{\widetilde{\mu}^2}{m^2}$, as we should expect in a physical scheme where heavy fields should indeed decouple.

7.7.3 Defining the β function without Λ

Our derivation of the β function in $\overline{\text{MS}}$ so far used the somewhat heuristic replacement (7.85). This is valid at one-loop, but not at higher loops where the divergence structure is more complicated. This is not a problem in more physical schemes, which directly produce Λ-dependence rather than ϵ-dependence, but it would still be nice to have a definition in $\overline{\text{MS}}$ that works at higher loops. The idea is the following. In $\overline{\text{MS}}$ we by definition have

$$g_{0,d} = \mu^\epsilon Z(g,\epsilon)g, \tag{7.107}$$

where

$$Z \equiv \frac{Z_1}{Z_2\sqrt{Z_3}}. \tag{7.108}$$

The idea is that at fixed g_0, as we change $\widetilde{\mu}$ we should also change $g(\widetilde{\mu})$ to preserve this equation, so differentiating both sides of (7.107) with respect to $\widetilde{\mu}$ we have

$$\widetilde{\mu}\frac{dg}{d\widetilde{\mu}} = -\frac{\epsilon g Z}{Z + g\frac{\partial Z}{\partial g}}. \tag{7.109}$$

We thus can define the β function as

$$\beta(g) \equiv -\lim_{\epsilon\to 0}\frac{\epsilon g Z}{Z + g\frac{\partial Z}{\partial g}}. \tag{7.110}$$

It is not obvious from this definition that the limit exists, after all Z has poles as a function of ϵ, but the consistency of the renormalization group equation (which says that the change in g with respect to $\widetilde{\mu}$ can only depend on the couplings at the scale $\widetilde{\mu}$) requires it to. At one loop we have

$$Z = 1 - \frac{g^2}{32\pi^2\epsilon}\left(\frac{11}{3}C_1(A) - \frac{4}{3}n_F C_1(\alpha)\right) + \ldots, \tag{7.111}$$

so we have

$$\begin{aligned}
\beta &= -\lim_{\epsilon\to 0}\epsilon g \frac{\left(1 - \frac{g^2}{32\pi^2\epsilon}\left(\frac{11}{3}C_1(A) - \frac{4}{3}n_F C_1(\alpha)\right) + \ldots\right)}{\left(1 - \frac{3g^2}{32\pi^2\epsilon}\left(\frac{11}{3}C_1(A) - \frac{4}{3}n_F C_1(\alpha)\right) + \ldots\right)} \\
&= -\lim_{\epsilon\to 0}\epsilon g\left(1 + \frac{g^2}{16\pi^2\epsilon}\left(\frac{11}{3}C_1(A) - \frac{4}{3}n_F C_1(\alpha)\right) + \ldots\right) \\
&= -\frac{g^3}{16\pi^2}\left(\frac{11}{3}C_1(A) - \frac{4}{3}n_F C_1(\alpha)\right) + \ldots,
\end{aligned} \tag{7.112}$$

which coincides with (7.92) above. This calculation has been extended to higher loops, with five loops being the most recent to be computed. The two loop result looks like this:

$$\beta(g) = -\frac{g^3}{16\pi^2}\left(\frac{11}{3}C_1(A) - \frac{4}{3}n_F C_1(\alpha)\right) - \frac{g^5}{(16\pi^2)^2}\left(\frac{34}{3}C_2(A)^2 - \frac{20}{3}C_2(A)C_1(\alpha)n_F - 4C_2(\alpha)C_1(\alpha)n_F\right) + \ldots. \tag{7.113}$$

Such higher loop results can be important in precision matching of QCD to experiment, for example when we want to compare scattering results at energies other than m_Z to $\alpha_s(M_Z)$.

7.8 Homework

1. Compute the contribution to the β function from a complex scalar field Φ transforming in some representation α of G. Hint: you can still use the formula (7.7) with the same Z_1 and Z_2 that we computed, the only thing that changes is Z_3.[55] There are two diagrams that contribute, and you can take the scalar to be massless which you should find sets one of the diagrams to zero.

2. Check the counterterm relation (7.100) by computing the Feynman diagrams in figure 15. Hints: you only need to compute the divergent parts of the diagrams, use our calculation of Z_1 above as inspiration. Also beware that the bottom three have a nontrivial symmetry factor! I have fond memories of doing this problem as a graduate student myself, make sure you leave plenty of time for it.

[55] The logic here is that we are still using the fermion coupling to the gauge field to define the gauge coupling even though it is the scalar contribution we are interested in. If you find this confusing then consider we are computing the β-function for a theory with one scalar and n_F fermions, after which we can take $n_F \to 0$ if desired to get the answer with no fermions. If you still don't like that then you can renormalize the scalar vertex instead, but then you need to compute more diagrams.

8 Lattice gauge theory

We have now seen that, due to asymptotic freedom, QCD is perturbative at high energies and becomes strongly-coupled at low energies. This raises the hope that it can give a unified theory of both high-energy nuclear phenomena such as Bjorken scaling and low-energy nuclear phenomena such as confinement.[56] To be sure of this however, we need to find a way to understand what actually happens in the low-energy strong-coupling regime. By far the most successful approach to this problem is **lattice gauge theory**, which will be the topic of this section. Lattice gauge theory is important for two reasons:

- It gives us a way to *think* about gauge theory non-perturbatively, telling us what is "really going on" without need for gauge-fixing or weak coupling approximations.
- In some cases it can be used to do precise numerical calculations in the strong-coupling regime, leading for example to the first-principles calculations of the spectrum of light hadrons I showed in the first section.

In this class we are more concerned with understanding than precision computation, so we will focus more on the first of these advantages, but we will also say a bit about how lattice gauge theory calculations are actually done in practice.

8.1 Hamiltonian lattice gauge theory

8.1.1 What is a gauge field?

We will begin with the Hamiltonian approach to lattice gauge theory. Naively you might guess that the way to do this is to choose a spatial lattice N, and then for each $\vec{x} \in N$ define a vector-valued gauge field operator $\vec{A}(\vec{x})$. Doing things this way however makes it difficult to preserve exact gauge-invariance on the lattice. To find something better, we need to think about what we really need the gauge field to do. In the continuum we motivated the introduction of the gauge field by using it to define a covariant derivative,[57]

$$D_\mu \Phi_n = \partial_\mu \Phi_n - i A^a_\mu (\tau_a)_{nm} \Phi_m. \tag{8.1}$$

Let's consider the same quantity for a scalar field $\Phi(\vec{x})$ define on a cubic spatial lattice with lattice spacing a. The partial derivative in the continuum is replaced by the finite difference

$$\partial_i \Phi_n(\vec{x}) \to \frac{\Phi_n(\vec{x} + \vec{\delta}_i) - \Phi_n(\vec{x})}{a}, \tag{8.2}$$

where $\vec{\delta}_i$ is a displacement by one lattice unit in the i direction. We would like to introduce a gauge symmetry with gauge group G, under which Φ transforms in some irreducible representation α_ϕ of G:

$$\Phi'_n(\vec{x}) = D^{\{\alpha_\phi\}}_{nm}(g_{\vec{x}}) \Phi_m. \tag{8.3}$$

Here $D^{\{\alpha_\phi\}}_{nm}(g)$ is the representation matrix for g in irreducible representation α_ϕ, and $g_{\vec{x}}$ is the gauge transformation at the spatial point $\vec{x}$. Specifying $g_{\vec{x}}$ for all $\vec{x} \in N$ is the same as picking an element of what we called the canonical gauge group $\mathcal{G}_c$ in the continuum. From this point of view, the problem with the finite-difference (8.2) is that the two terms transform with different elements of G, $g_{\vec{x}}$ and $g_{\vec{x}+\vec{\delta}_i}$, so the gauge transformation of the difference is rather ugly. To get something with a nicer gauge transformation,

[56] A "true" nuclear physicist would view anything involving quarks and gluons as "high energy", as there is also the even lower energy question of how hadrons are bound together into nuclei. We won't say too much about this question in this class: in principle we should use the Yukawa theory to describe nuclei as bound states of pions and nucleons, but this is not so practical when we consider large nuclei. We will return to pion physics in a few sections.

[57] In this section we will write all color indices down to avoid unpleasant notational issues.

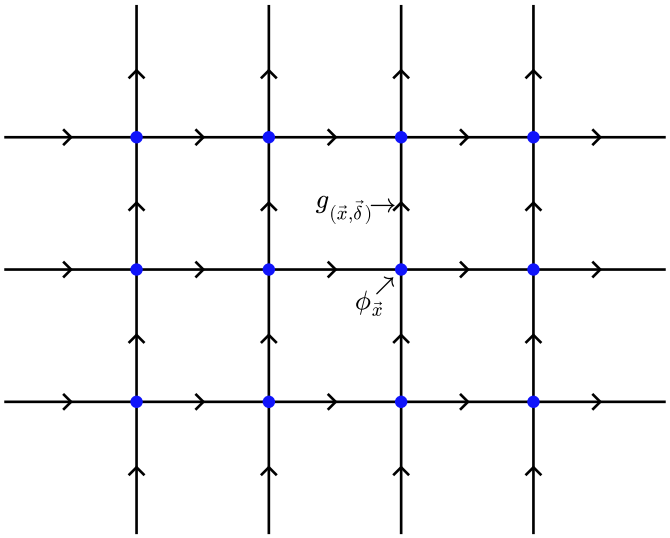

Figure 16: A spatial lattice, with sites colored blue and (oriented) links marked with arrows. In Hamiltonian lattice gauge theory we assign matter degrees of freedom to each site and a group element to each link for the gauge field.

we need a way to convert a gauge transformation at $\vec{x}$ to a gauge transformation at $\vec{x}+\vec{\delta}_i$. Fortunately in the continuum we had just such an object: the Wilson line

$$W_{\alpha,nm}(C) = (Pe^{i\int_C A^a_\mu \tau_a dx^\mu})_{nm}, \tag{8.4}$$

which we saw has the gauge transformation

$$W_\alpha(C)' = D^{\{\alpha\}}(g(x_f))W_\alpha(C)D^{\{\alpha\}}(g(x_i)^{-1}) \tag{8.5}$$

where x_i and x_f are the initial and final endpoints of the curve C. Using this we can construct a finite difference between Φ at two neighboring lattice points as

$$D_i\Phi \to \frac{W_\alpha(C)^\dagger\Phi(\vec{x}+\vec{\delta}_i)-\Phi(\vec{x})}{a}, \tag{8.6}$$

where we can take C to be a straight line from $\vec{x}$ to $\vec{x}+\vec{\delta}_i$. Taking the limit of this quantity as $a\to 0$ recovers the covariant derivative (8.1). Thus to define a covariant derivative on the lattice, what we need is a rule that assigns a Wilson line to each **link** of the lattice. The link needs to carry an orientation so that we know which way the Wilson line goes, so when we use the term "link" we will always mean it to include an orientation. We will parametrize links as $(\vec{x},\vec{\delta})$, since to specify a link we can give its starting vertex and a lattice displacement vector for its direction. See figure 16 for a spatial lattice together with its links. In fact even this is not quite what we *really* need however, since we would like to be able to differentiate fields in arbitrary representations of G without introducing an independent Wilson line for each. So what we *really* need on each link is simply a group element $g_{(\vec{x},\vec{\delta})}\in G$ whose representation $D^{\{\alpha\}}(g_{(\vec{x},\vec{\delta})})$ gives the Wilson line for this link in representation α. This is our main lesson: *in lattice gauge theory, a gauge field is an assignment of a group element $g_{(\vec{x},\vec{\delta})}$ to each (oriented) link of the lattice N.*

Before continuing I want to emphasize a potentially confusing point: in lattice gauge theory a gauge field is a set of group elements $g_{(\vec{x},\vec{\delta})}$ assigned to links, while a gauge transformation is a set of group elements $g_{\vec{x}}$ assigned to sites. They are NOT the same thing. From (8.5) the gauge transformation of the gauge field is

$$g'_{(\vec{x},\vec{\delta})} = g_{\vec{x}+\vec{\delta}}\, g_{(\vec{x},\vec{\delta})}\, g_{\vec{x}}^{-1}. \tag{8.7}$$

We should also now say more carefully what we mean by a spatial lattice N. At a minimum we should mean a set of points $\vec{x} \in \mathbb{R}^{d-1}$, but we will also take N to include a set of links $\vec{x}, \vec{\delta}$, with the understanding that each link appears with exactly one orientation. In other words if $(\vec{x}, \vec{\delta})$ is included in N then its opposite $\vec{x} + \vec{\delta}, -\vec{\delta}$ is not included. For a cubic lattice in $\mathbb{R}^{d-1}$ it may seem "obvious" which links should be included, for example we can just take the neighboring links that point in the direction of increasing $x^1, x^2, \ldots$. Sometimes it is useful to consider other lattices however, for example triangular, hexagonal, etc, where the choice of which links to include and how to orient them is less obvious. It is also sometimes useful to consider other spatial topologies such as $\mathbb{S}^{d-1}$, and then again it is not so clear how the sites should be connected. Thus it is safest to explicitly declare which links are included. We will see in a little while that it also turns out to be a good idea to explicitly say which oriented faces are included in the lattice: we will defer discussion of this to a bit later in the section. In mathematics there is actually a name for a lattice defined as a collection of oriented points, links, faces, volumes, etc: it is called a **CW complex**. I won't tell you the definition here, but you can look it up if you are curious.[58]

8.1.2 Hilbert space and operators

In the continuum we started with a big Hilbert space $\mathcal{H}_{big}$ of wave functionals of a_μ and ϕ, and we constructed the physical Hilbert space by requiring it to be annihilated by the constraints. We will follow the same approach in Hamiltonian lattice gauge theory, but to be more efficient we will immediately impose the $\Pi^0_a = 0$ constraint to go to the intermediate Hilbert space $\mathcal{H}_{int}$ whose wave functionals are independent of A^a_0. Thus we will only assign group elements to oriented spatial links, as indeed we were already doing above. In the absence of matter fields a basis for $\mathcal{H}_{int}$ consists of states $|g\rangle$ labeled by a choice of group element $g_{(\vec{x},\vec{\delta})}$ for each link. This is the tensor product

$$\mathcal{H}_{int} = \bigotimes_{(\vec{x},\vec{\delta})\in N} \mathcal{H}_{(\vec{x},\vec{\delta})} \tag{8.8}$$

of Hilbert spaces $\mathcal{H}_{(\vec{x},\vec{\delta})}$ for each link in N. Each of these local Hilbert spaces has a basis of states of the form[59]

$$|g\rangle_{(\vec{x},\vec{\delta})}. \tag{8.9}$$

Wave functions in this Hilbert space are elements of $L^2(G)$, i.e. states

$$|\psi\rangle = \int dg\, \psi(g) |g\rangle_{(\vec{x},\vec{\delta})} \tag{8.10}$$

such that

$$\langle\psi|\psi\rangle = \int dg |\psi(g)|^2 < \infty \tag{8.11}$$

where dg is the Haar measure on G. Note that this means that

$$\langle g'|g\rangle = \delta(g, g'), \tag{8.12}$$

with the δ-function defined so that

$$\int dg' \delta(g, g') f(g') = f(g) \tag{8.13}$$

for any reasonable function $f : \mathbb{G} \to \mathbb{C}$.

[58] For the lattice gauge theory we construct in this section it is enough to stop at the faces, but for "higher form" gauge fields we need the higher-dimensional volumes as well.

[59] I emphasize that we do NOT assign separate tensor factors for the opposite orientations of the link, only one of the orientations is included in N. This is because the Wilson line with opposite orientation is not an independent degree of freedom; we will see in a moment how to define it.

There are three natural sets of operators we can define on $\mathcal{H}_{(\vec{x},\vec{\delta})}$:

$$\begin{aligned} W_{\alpha,nm}(\vec{x},\vec{\delta})|g\rangle_{(\vec{x},\vec{\delta})} &= D^{\{\alpha\}}_{nm}(g)|g\rangle_{(\vec{x},\vec{\delta})} \\ L_h(\vec{x},\vec{\delta})|g\rangle_{(\vec{x},\vec{\delta})} &= |hg\rangle_{(\vec{x},\vec{\delta})} \\ R_h(\vec{x},\vec{\delta})|g\rangle_{(\vec{x},\vec{\delta})} &= |gh\rangle_{(\vec{x},\vec{\delta})}. \end{aligned} \tag{8.14}$$

We will refer to these as the Wilson lines, the left-multiplications, and the right-multiplications. You can think of the Wilson lines as "position" operators and the left/right multiplications as "momentum" operators for a particle moving on G. You will show on the homework that their algebra (dropping the link labels) is:

$$\begin{aligned} L_h L_{h'} &= L_{hh'} \\ R_h R_{h'} &= R_{h'h} \\ L_h R_{h'} &= R_{h'} L_h \\ W_{\alpha,nm} L_h &= D^{\{\alpha\}}_{n,n'}(h) L_h W_{\alpha,n',m} \\ W_{\alpha,nm} R_h &= R_h W_{\alpha,nm'} D^{\{\alpha\}}_{m',m}(h). \end{aligned} \tag{8.15}$$

You will also show that adjoints of the left/right multiplication operators are given by

$$\begin{aligned} L_h^\dagger &= L_{h^{-1}} = L_h^{-1} \\ R_h^\dagger &= R_{h^{-1}} = R_h^{-1}, \end{aligned} \tag{8.16}$$

so in particular they are unitary. The adjoint of the Wilson line also has a nice interpretation: it is a Wilson line whose orientation is opposite to that of the link:

$$W_\alpha^\dagger(\vec{x},\vec{\delta}) \equiv W_\alpha(\vec{x}+\vec{\delta},-\vec{\delta}), \tag{8.17}$$

where the $\dagger$ is understood to be the Hilbert space adjoint together with a transpose on the color indices nm, i.e.

$$W^\dagger_{\alpha,nm}(\vec{x},\vec{\delta}) = W_{\alpha,mn}(\vec{x},\vec{\delta})^* \tag{8.18}$$

where $*$ means the Hilbert space adjoint.

8.1.3 Gauge transformations and the physical Hilbert space

To get from $\mathcal{H}_{int}$ to the physical Hilbert space we need to restrict to gauge-invariant states. We thus need to write down unitary operators on $\mathcal{H}_{int}$ that implement the gauge transformations. A gauge transformation at $\vec{x}$ implements a gauge transformation on each Wilson line living on a link that is attached to $\vec{x}$, including both the outgoing links $(\vec{x},\vec{\delta})$ and the ingoing links $(\vec{x}-\vec{\delta},\vec{\delta})$. It also needs to act with a unitary operator U^{matt} on the matter fields living at $\vec{x}$. The full expression for the unitary operator in $\mathcal{H}_{int}$ implementing a gauge transformation by g at $\vec{x}$ is

$$U_{\vec{x}}(g) = \prod_{\vec{\delta}\,\text{ingoing}} L_g(\vec{x}-\vec{\delta},\vec{\delta}) \prod_{\vec{\delta}\,\text{outgoing}} R_{g^{-1}}(\vec{x},\vec{\delta})\, U^{matt}_{\vec{x}}(g). \tag{8.19}$$

We can check that this gives the correct gauge transformation using our Wilson line algebra (8.15):

$$W_\alpha(\vec{x},\vec{\delta})' = U_{\vec{x}+\vec{\delta}}(g_{\vec{x}+\vec{\delta}})^\dagger U_{\vec{x}}(g_{\vec{x}})^\dagger W_\alpha(\vec{x},\vec{\delta}) U_{\vec{x}}(g_{\vec{x}}) U_{\vec{x}+\vec{\delta}}(g_{\vec{x}+\vec{\delta}}) = D^{\{\alpha\}}(g_{\vec{x}+\vec{\delta}}) W_\alpha(\vec{x},\vec{\delta}) D^{\{\alpha\}}(g_{\vec{x}}^{-1}), \tag{8.20}$$

and it of course also implements the matter gauge transformation

$$U_{\vec{x}}(g)^\dagger \Phi(\vec{x}) U_{\vec{x}}(g) = D^{\{\alpha_\phi\}}(g)\Phi(\vec{x}). \tag{8.21}$$

We then define the physical Hilbert space as the set of states invariant under all $U_{\vec{x}}(g)$:[60]

$$\mathcal{H}_{phys} \equiv \Big\{ |\psi\rangle \in \mathcal{H}_{int} \Big| U_{\vec{x}}(g)|\psi\rangle = |\psi\rangle \qquad \forall \vec{x}, g \Big\}. \tag{8.22}$$

We note that the operators $U_{\vec{x}}(g)$ together give a unitary representation on $\mathcal{H}_{int}$ of the lattice version of the canonical gauge group $\mathcal{G}_c$. We also observe that they commute for different sites,

$$U_{\vec{x}}(g)U_{\vec{x}'}(g') = U_{\vec{x}'}(g')U_{\vec{x}}(g) \qquad (\vec{x} \neq \vec{x}'), \tag{8.23}$$

which is a consequence of the commutativity of L_g and R_g.

8.1.4 What is the Hamiltonian?

Having constructed the gauge-invariant Hilbert space, we now want to construct the Hamiltonian. In the continuum we had

$$H = \int d^{d-1}x \left(\frac{g^2}{2} \Pi_a^i \Pi_i^a + \frac{1}{4g^2} F_{ij}^a F_a^{ij} + \hat{\mathcal{H}}_M \right), \tag{8.24}$$

so our goal is to figure out what replaces the first two terms on the lattice. We can guess lattice analogs of A_i^a and Π_a^i by defining

$$\begin{aligned} W_\alpha(\vec{x}, \vec{\delta}_i) &\equiv e^{iaA^a(\vec{x},\vec{\delta}_i)\tau_a} \\ L_{e^{i\theta^a T_a}}(\vec{x}, \vec{\delta}_i) &\equiv e^{-i\theta^a a^{d-2}\Pi_a(\vec{x},\vec{\delta}_i)}. \end{aligned} \tag{8.25}$$

The first of these is the obvious discretization of the continuum Wilson line. To motivate the second, we can observe that from (8.15) we must have

$$e^{ia^{d-2}\theta^a \Pi_a} e^{iaA^b\tau_b} e^{-ia^{d-2}\theta^a\Pi_a} = e^{i\theta^a\tau_a} e^{iaA^a\tau_a}. \tag{8.26}$$

Expanding both terms to leading order in the exponent we have

$$ia\left(A^b + ia^{d-2}\theta^a[\Pi_a, A^b]\tau_b + \ldots\right) = i\left(aA^b\tau_b + \theta^b\tau_b + \ldots\right), \tag{8.27}$$

where "..." indicates terms that are higher order in θ and/or a. Thus at leading order in a we need to have

$$[\Pi_a, A^b] = -\frac{i}{a^{d-1}}\delta_a^b, \tag{8.28}$$

which is precisely the lattice version of the canonical commutation relation (the factor of $1/a^{d-1}$ is necessary to produce a spatial δ-function in the continuum). One way to interpret the factor of a^{d-2} in the definition of Π_a is that it is there so that we can interpret L_h as

$$L_h = e^{i\theta \times \text{electric flux}}, \tag{8.29}$$

where the electric flux is through a lattice hypercube that is punctured by the link. In any case we can therefore write a lattice version of the electric flux term in the Hamiltonian as

$$H \supset \sum_{(\vec{x},\vec{\delta})\in N} a^{d-1}\frac{g^2}{2}\Pi_a(\vec{x},\vec{\delta})\Pi^a(\vec{x},\vec{\delta}). \tag{8.30}$$

[60] In the continuum we required this only for gauge transformations which become trivial at spatial infinity. On the lattice we can implement this automatically using boundary conditions where any spatial boundary is as a set of outgoing links that do not end on vertices, see figure 16 for an illustration (not drawing the vertices means that we do not include matter fields at the ends of these links, nor do we impose a gauge constraint there). Essentially this is Dirichlet boundary conditions for the gauge field, since there are no gauge fields connecting the absent vertices.

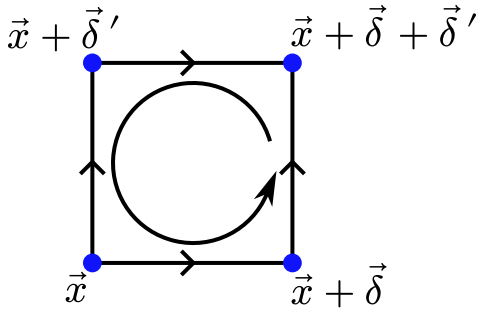

Figure 17: A plaquette in a cubic lattice. The circular arrow indicates the orientation of the Wilson loop.

Using Schur orthogonality we can give this electric flux term a nice group-theoretic interpretation. The idea is that we can think of the action of L_g on $L^2(G)$ as a big reducible representation of G, usually called the **regular representation**. Schur orthogonality tells us how to decompose this into irreducible representations: defining an orthonormal basis

$$|\alpha, nm\rangle \equiv \int dg \sqrt{d_\alpha} D^{\{\alpha\}}_{nm}(g)|g\rangle \tag{8.31}$$

for $L^2(G)$, we have

$$L_g|\alpha, nm\rangle = \sum_\ell D^{\{\alpha\}}_{n\ell}(g^{-1})|\alpha, \ell m\rangle. \tag{8.32}$$

Thus for each α the states $|\alpha, nm\rangle$ transform in the inverse transpose of α, or equivalently its conjugate. From (8.25) we see that $-a^{d-2}\Pi_a$ are precisely the generators of the regular representation of G, so the action of $\Pi_a\Pi^a$ within each block is proportional to the quadratic Casimir invariant:[61]

$$\Pi_a\Pi^a|\alpha, nm\rangle = a^{-2(d-2)}C_2(\alpha)|\alpha, nm\rangle. \tag{8.33}$$

Therefore we can think of this term in the Hamiltonian as punishing states of the link according to their quadratic Casimir in the $|\alpha, nm\rangle$ basis, with the trivial representation $\alpha = 1$ being punished the least! This is the non-abelian lattice version of the idea that this term wants to minimize the magnitude of the electric field.

Turning now to the magnetic term in the Hamiltonian, we need a way to think about the field strength in terms of the Wilson line. You showed on the homework a few sections ago that for a small Wilson loop on a curve C bounding a surface D to first order in the area of D we get the exponential of the trace of the flux of F through D. To get the trace of F^2 we need to revisit this problem to second order in the area, which we will now do. The idea is that we will consider a square Wilson loop that is one lattice unit on each side. Such a minimal square loop is called a **plaquette**. We will denote plaquettes in our cubic lattice as $(\vec{x}, \vec{\delta}, \vec{\delta}')$ with $\vec{\delta}\cdot\vec{\delta}' = 0$, as shown in figure 17. This labeling is redundant since we do not care which site is the beginning of the loop, so we identify

$$(\vec{x}, \vec{\delta}, \vec{\delta}') \sim (\vec{x}+\vec{\delta}, \vec{\delta}', -\vec{\delta}) \sim (\vec{x}+\vec{\delta}+\vec{\delta}', -\vec{\delta}, -\vec{\delta}') \sim (\vec{x}+\vec{\delta}', -\vec{\delta}', \vec{\delta}). \tag{8.34}$$

We will further extend our definition of the lattice N to include the list of all the plaquettes in the lattice, again with the convention that if $(\vec{x}, \vec{\delta}, \vec{\delta}')$ is included then its opposite orientation $(\vec{x}, \vec{\delta}', \vec{\delta})$ is not. The Wilson loop for the plaquette is

$$W_\alpha(\vec{x}, \vec{\delta}, \vec{\delta}') \equiv \mathrm{Tr}\left(W_\alpha(\vec{x}+\vec{\delta}', -\vec{\delta}')W_\alpha(\vec{x}+\vec{\delta}+\vec{\delta}', -\vec{\delta})W_\alpha(\vec{x}+\vec{\delta}, \vec{\delta}')W_\alpha(\vec{x}, \vec{\delta})\right), \tag{8.35}$$

[61] Here we used that a representation α and its conjugate always have the same quadratic Casimir. Also note that we can interpret this formula as an exact solution of the quantum mechanics of a particle moving on the group manifold G.

which respects the identifiction (8.34) due to the cyclicity of the trace. This Wilson loop is often also called the plaquette, to distinguish it from the geometric loop itself we will call it the **Wilson plaquette**.

To understand the continuum limit, we write the Wilson plaquette in the xy plane in terms of the gauge field as

$$W_\alpha(\vec{x}, a\hat{x}, a\hat{y}) = \mathrm{Tr}\left(e^{-ia\left(A_y - \frac{a}{2}\partial_x A_y\right)} e^{-ia\left(A_x + \frac{a}{2}\partial_y A_x\right)} e^{ia\left(A_y + \frac{a}{2}\partial_x A_y\right)} e^{ia\left(A_x - \frac{a}{2}\partial_y A_x\right)}\right). \tag{8.36}$$

To preserve more symmetry we've chosen to Taylor expand the gauge field about the center point of the plaquette; this makes the Wilson plaquette an even function of a since sending $a \to -a$ gives a cyclic permutation inside the trace. Using the BCH formula, we have

$$\log\left(e^{X_1} e^{X_2} e^{X_3} e^{X_4}\right) = \sum_i X_i + \frac{1}{2}\sum_{i<j}[X_i, X_j] + \ldots, \tag{8.37}$$

which gives

$$W_\alpha(\vec{x}, a\hat{x}, a\hat{y}) = \mathrm{Tr}\left(e^{ia^2 F_{xy} + \cdots}\right), \tag{8.38}$$

where "..." indicates commutator terms that are at most of order a^3. Expanding the exponential we thus have

$$W_\alpha(\vec{x}, a\hat{x}, a\hat{y}) = \mathrm{Tr}\left(1 + ia^2 F_{xy} - \frac{a^4}{2} F_{xy}^2 + O(a^6)\right). \tag{8.39}$$

To justify the size of the error term, we note that in the linear term from expanding the exponential there are no higher order terms since the trace of a commutator is zero. In the quadratic term there in principal could have been a term of order a^5 from multiplying F_{xy} by a commutator, but this term must vanish since the Wilson plaquette is an even function of a. If we choose α to be the (possibly reducible) defining representation of G, then we can recogtnize the third term in the expansion as being precisely proportional to the magnetic term in the Hamiltonian that we are after. To get rid of the second term we can take the real part:

$$\Re(W_\alpha(\vec{x}, a\hat{x}, a\hat{y})) = \frac{W_\alpha(\vec{x}, a\hat{x}, a\hat{y}) + W_\alpha(\vec{x}, a\hat{x}, a\hat{y})^\dagger}{2} = \mathrm{Tr}\left(1 - \frac{a^4}{2} F_{xy}^2 + O(a^6)\right). \tag{8.40}$$

We can think of the real part as averaging over the two possible orientations of the plaquette. The leading constant term we can view as a renormalization of the cosmological constant, which in field theory has no dynamical effects. We thus are led to the **Kogut-Susskind** Hamiltonian of lattice gauge theory:[62]

$$H = \frac{g^2}{2} a^{d-1} \sum_{(\vec{x},\vec{\delta}) \in N} \Pi_a(\vec{x}, \vec{\delta}) \Pi^a(\vec{x}, \vec{\delta}) - \frac{1}{g^2} a^{d-5} \sum_{(\vec{x},\vec{\delta},\vec{\delta}') \in N} \Re\left(W_\alpha(\vec{x}, \vec{\delta}, \vec{\delta}')\right). \tag{8.41}$$

If we allow the sum over plaquettes to include both orientations then we can replace the $\Re$ by a factor of $1/2$. As a reminder, here α is the defining representation of G.[63]

Finally we should say a little about the matter Hamiltonian. We will only consider a complex scalar field Φ to avoid the fermion doubling problem that we discussed in QFT II. The continuum Hamiltonian density is

$$\hat{\mathcal{H}}_M = \Pi^\dagger \Pi + (D_i \Phi)^\dagger D_i \Phi + V(|\Phi|). \tag{8.42}$$

On the lattice we have in $\mathcal{H}_{int}$ in addition to the link Hilbert spaces a Hilbert space $\mathcal{H}_{\vec{x}}$ at each site, with a set of basis states $|\phi\rangle$ that are eigenenstates of Φ:

$$\Phi(\vec{x})|\phi\rangle = \phi_{\vec{x}}|\phi\rangle. \tag{8.43}$$

[62] Although we have used a cubic lattice to motivate this definition, it works for any lattice as long as we sum over all links in the first term and all plaquettes in the second term.

[63] For general compact G there is an arbitrary choice here of which faithful representation we view as defining; we already had to make that choice in the continuum Lagrangian when we wrote it in terms of a matrix-valued gauge field. This choice does not really change the theory however: the only place the defining representation appears is in the definition of $\hat{g}_{ab}$, so changing our choice of defining representation only rescales the gauge coupling.

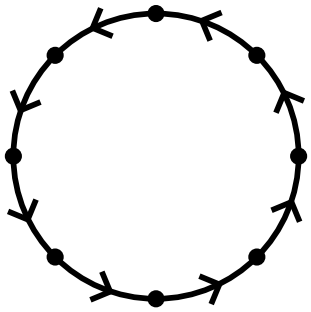

Figure 18: A circular spatial lattice in $1+1$ dimensions.

To match the continuum we define the canonical momentum as

$$\Pi(\vec{x}) \equiv -i \frac{1}{a^{d-1}} \frac{\partial}{\partial \phi(\vec{x})}, \tag{8.44}$$

so the Hamiltonian contains a term

$$H \supset \sum_{\vec{x} \in N} \Pi^\dagger(\vec{x}) \Pi(\vec{x}) \tag{8.45}$$

The square of the discrete covariant derivative along a link $(\vec{x}, \vec{\delta})$ is

$$\frac{\left(\Phi^\dagger(\vec{x}+\vec{\delta}) W_{\alpha_\phi}(\vec{x},\vec{\delta}) - \Phi^\dagger(\vec{x})\right)\left(W_{\alpha_\phi}(\vec{x},\vec{\delta})^\dagger \Phi(\vec{x}+\vec{\delta}) - \Phi(\vec{x})\right)}{a^2}. \tag{8.46}$$

The terms which do not involve W_α in the square are just a renormalization of the mass of the scalar, so we can write the lattice scalar Hamiltonian like this:

$$H_M = a^{d-1} \sum_{\vec{x} \in N} \left(\Pi(\vec{x})^\dagger \Pi(\vec{x}) + V(|\Phi|)\right) - a^{d-3} \sum_{(\vec{x},\vec{\delta}) \in N} \left(\Phi^\dagger(\vec{x}+\vec{\delta}) W_{\alpha_\phi}(\vec{x},\vec{\delta}) \Phi(\vec{x}) + \text{h.c.}\right). \tag{8.47}$$

If we allow the sum over links to include both orientations then we can drop the "+h.c.".

8.1.5 A notational aside

So far we have been keeping around powers of the lattice spacing a. In this business it of course standard to work in units where $a = 1$: we've postponed doing this so that build some intuition for the continuum limit, but from now on we will join the party.

8.1.6 $1+1$ dimensions

The simplest nontrivial example of Hamiltonian lattice gauge theory is pure gauge theory on a spatial $\mathbb{S}^1$. The lattice is shown in figure 18. Labeling sites by $i = 1, \ldots N$ and links by the site they are outgoing from, the Hamiltonian is simply

$$H = \frac{g^2}{2} \sum_{i=1}^{N} \Pi_a(i) \Pi^a(i). \tag{8.48}$$

It will be instructive to first consider the case $N = 1$, where there is one link attached twice to the same site. $\mathcal{H}_{int}$ is just equal to $L^2(G)$, and $\mathcal{H}_{phys}$ is the subspace of wave functions obeying

$$\psi(hgh^{-1}) = \psi(g) \tag{8.49}$$

for all $g, h \in G$. The set $Cl(g)$ of elements in G which are related to g by conjugation by some $h \in G$ is called the **conjugacy class** of g, and a function in $L^2(G)$ which is constant on conjugacy classes is called a **class function** on G. Thus we see that the set of gauge invariant states for our one-link theory is the set of class functions. For any irreducible representation α of G there is a natural class function

$$\chi_\alpha(g) = \mathrm{Tr}\left(D^{\{\alpha\}}(g)\right) \tag{8.50}$$

which is called the **character** of the representation. We then have the following theorem:

Theorem 11. *Let G be a compact Lie group. The set of characters $\chi_\alpha(g)$ form an orthonormal basis for the set of class functions on G.*

Proof. The orthonormality

$$\int dg \chi_\alpha(g)^* \chi_{\alpha'}(g) = \delta_{\alpha\alpha'} \tag{8.51}$$

is an immediate consequence of Schur orthogonality, so what we need to show is that the set of $\chi_\alpha(g)$ is complete for class functions. We first note that for any function $f \in L^2(G)$, we can construct a class function by acting with the linear operator

$$Pf(g) = \int dh f(hgh^{-1}). \tag{8.52}$$

Moreover this operation squares to one by the invariance of the Haar measure, and thus gives a projection onto the set of class functions. From Schur orthogonality and the Peter-Weyl theorem we know that the representation matrices $D^{\{\alpha\}}_{nm}(g)$ give an orthonormal basis for G. We will show that acting on these with P just gives the characters. Indeed consider the quantity

$$PD^{\{\alpha\}}_{nm}(g) = \int dh D^{\{\alpha\}}_{nm}(hgh^{-1}). \tag{8.53}$$

By invariance of the Haar measure we have

$$D^{\{\alpha\}}_{nn'}(h') PD^{\{\alpha\}}_{nm}(g) = PD^{\{\alpha\}}_{nm}(g) D^{\{\alpha\}}_{nn'}(h'), \tag{8.54}$$

so by Schur's lemma we must have

$$PD^{\{\alpha\}}_{nm}(g) = \lambda(g)\delta_{nm}. \tag{8.55}$$

Taking the trace of both sides, we see that we have

$$\lambda(g) = \frac{1}{d_\alpha}\chi_\alpha(g), \tag{8.56}$$

so we have shown that

$$PD^{\{\alpha\}}_{nm}(g) = \frac{1}{d_\alpha}\delta_{nm}\chi_\alpha(g). \tag{8.57}$$

Since any class function can be expanded in terms of the $D^{\{\alpha\}}_{nm}(g)$, and since acting with P on a class function does not change it, it must be that it can also be expanded in the characters $\chi_\alpha(g)$. □

Thus the characters $\chi_\alpha(g)$ give a basis for the set of gauge-invariant states of the link:

$$|\alpha\rangle \equiv \int dg \chi_\alpha(g)|g\rangle. \tag{8.58}$$

In fact these are energy eigenstates, since we showed that acting on states of definite α the electric flux term is just proportional to the quadratic Casimir of α:

$$H|\alpha\rangle = \frac{g^2}{2}C_2(\alpha)|\alpha\rangle. \tag{8.59}$$

Turning now to the general case of N sites connected by N links, you will show on the homework that what the gauge-invariance condition at each link tells us about the wave function $\psi(g_1, \ldots, g_N)$ is that we must have

$$\psi(g_1, \ldots, hg_i, g_{i+1}h^{-1}, \ldots g_N) = \psi(g_1, \ldots, g_N) \tag{8.60}$$

for all $h \in G$ and $i = 1, \ldots, N$ (identifying g_{N+1} with g_1). You will also show that to satisfy this condition we need the wave function to have the form

$$\psi(g_1, \ldots, g_N) = f\left(g_N g_{N-1} \ldots g_1\right), \tag{8.61}$$

with the function $f : G \to \mathbb{C}$ being a class function. We may then simply port over our discussion of the one-link Hilbert space to conclude that there is again an orthonormal basis for the gauge-invariant states given by

$$|\alpha\rangle \equiv \int dg_1 \ldots dg_N \chi_\alpha(g_N \ldots g_1)|g_1 \ldots g_N\rangle. \tag{8.62}$$

There is a very simple way to describe these states: they are obtained by acting on the trivial representation state

$$|1\rangle = \int dg_1 \ldots dg_N |g_1 \ldots g_N\rangle \tag{8.63}$$

with a Wilson loop going once around the circle:

$$|\alpha\rangle = W_\alpha(\mathbb{S}^1)|1\rangle. \tag{8.64}$$

As you may have guessed these states are eigenstates of the Hamiltonian: to see this we can use the algebra (8.15) to see that for each link we have

$$[\Pi_a, W_\alpha] = \tau_a W_\alpha, \tag{8.65}$$

and thus

$$\begin{aligned}
[\Pi_a \Pi^a, W_\alpha] &= \Pi_a [\Pi^a, W_\alpha] + [\Pi_a, W_\alpha]\Pi^a \\
&= \tau^a [\Pi_a, W_\alpha] + 2\tau_a W_\alpha \Pi^a \\
&= C_2(\alpha) W_\alpha + 2\tau_a W_\alpha \Pi^a.
\end{aligned} \tag{8.66}$$

Since the ground state $|1\rangle$ is annihilated by all Π^a since it is invariant under acting with L_h, each link contributes a factor of $\frac{g}{2} C_2(\alpha)$ to the energy:

$$H|\alpha\rangle = \frac{g^2}{2} N C_2(\alpha)|\alpha\rangle. \tag{8.67}$$

This completes the solution of Yang-Mills theory with compact gauge group G in $1+1$ dimensions: the eigenstates of the Hamiltonian are states of definite electric flux wrapping the circle,[64] with the fluxes labeled by irreducible representations of G, and the energy density in the flux is

$$\sigma = \frac{g^2}{2} C_2(\alpha). \tag{8.68}$$

8.2 Confinement in the strong-coupling expansion

Having set up the machinery of Hamiltonian lattice gauge theory, we will now use it to do some physics. The Hamiltonian is

$$H = \frac{g^2}{2} \sum_{(\vec{x}, \vec{\delta}) \in N} \Pi^a \Pi_a - \frac{1}{g^2} \sum_{(\vec{x}, \vec{\delta}, \vec{\delta}') \in N} \Re\left(W_\alpha(\vec{x}, \vec{\delta}, \vec{\delta}')\right) + H_M. \tag{8.69}$$

[64] Somewhat annoyingly the direction of the electric flux is opposite to the direction of the Wilson line, due to (8.65) and the relative sign between E_a and Π_a. This is because it is really $W^\dagger$ that creates the electric flux in the direction of the line, just as it is $\overline{\psi}$ which creates the electron.

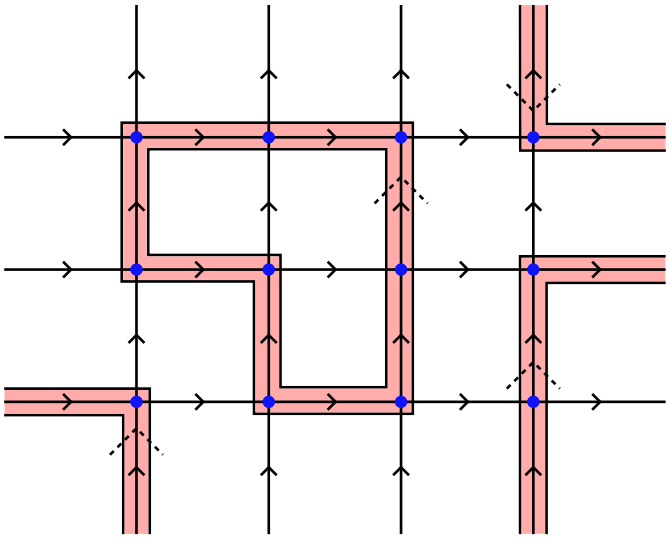

Figure 19: A loop network eigenstate of the strong coupling Hamiltonian. Sitting on top of the vacuum is a network of disjoint loops of electric flux, each labeled by an irreducible representation of G, and each loop contributes an energy that is proportional to its length and quadratic Casimir. The direction of flux is indicated by the dashed arrows, it is opposite to the direction of the Wilson line.

In the continuum we studied this theory at weak coupling, $g \ll 1$. In that limit the electric flux term is not important compared to the magnetic flux term, so roughly speaking the gauge field likes to minimize the magnetic flux. At length scales

$$a \ll \ell \ll a e^{\frac{16\pi^2}{g^2}\frac{1}{\frac{11}{3}C_1(A)-\frac{4}{3}n_F C_1(\alpha)}}, \tag{8.70}$$

the theory is described by weakly-interacting gauge bosons and matter particles in the continuum.[65] For bigger length scales the theory becomes strongly-coupled. To understand what happens then we will now consider the opposite expansion, the strong-coupling expansion

$$g \gg 1. \tag{8.71}$$

In this limit the electric flux term is dominant, so the ground state is just the tensor product of the trivial representation state on every link:

$$|\Omega\rangle = \bigotimes_{(\vec{x},\vec{\delta})\in N} |1\rangle_{(\vec{x},\vec{\delta})}. \tag{8.72}$$

Moreover by using our Wilson line commutator (8.66), we see that acting on this state with any Wilson loop $W_\alpha(L)$ that does not use the same link more than once we have

$$HW_\alpha(L)|\Omega\rangle = \frac{g^2}{2} d(L) C_2(\alpha) W_\alpha(L)|\Omega\rangle. \tag{8.73}$$

Furthermore if we act on the vacuum with several disjoint Wilson loops, each not using the same link more than once, then we have

$$HW_{\alpha_1}(L_1)\dots W_{\alpha_n}(L_n)|\Omega\rangle = \frac{g^2}{2}\Big(C_2(\alpha_1)d(L_1) + \dots + C_2(\alpha_n)d(L_N)\Big) W_{\alpha_1}(L_1)\dots W_{\alpha_n}(L_n)|\Omega\rangle. \tag{8.74}$$

[65] Here we are assuming that the matter content is such that the β function is negative as in QCD, and indeed in this formula we assumed the matter was just n_F spinors in representation α. For more general matter one just changes the second factor in the exponent as appropriate.

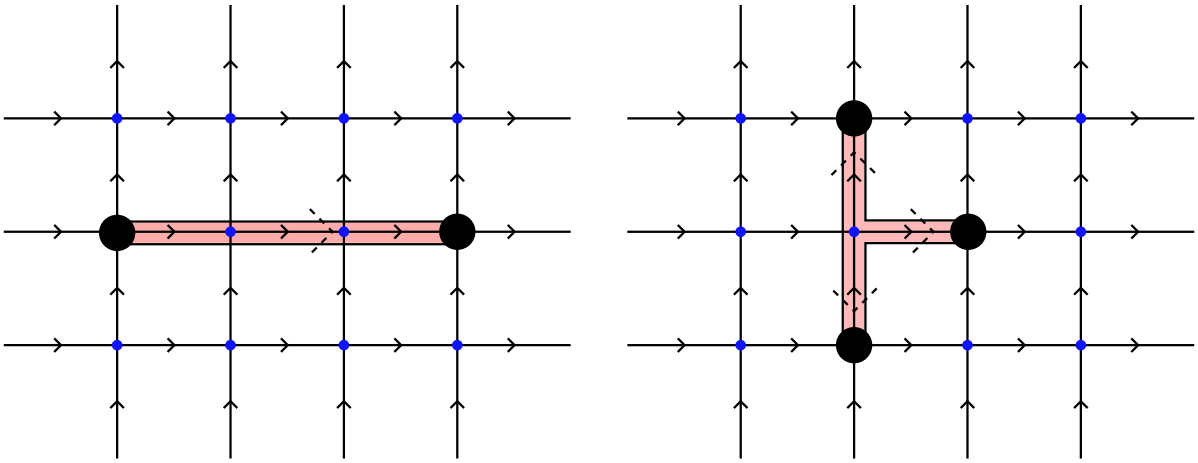

Figure 20: Meson and baryon eigenstates in the strong-coupling approximation.

In this way we can generate a large family of exact energy eigenstates in the $g \gg 1$ limit, see figure 19 for an illustration.[66]

Now let's see how this picture give rise to confinement. Say we want to create a charged scalar/antiscalar pair by acting on the strong-coupling vacuum with

$$\Phi(\vec{x}_2)W_\alpha^\dagger(C)\Phi^\dagger(\vec{x}_1), \tag{8.75}$$

where C is a path from $\vec{x}_1$ to $\vec{x}_2$. There are many options for the path C, but if we want to create the least energy in the strong-coupling approximation then we should take C to be the path from $\vec{x}_1$ to $\vec{x}_2$ that traverses the fewest links. This is our big result: the potential energy between a scalar/antiscalar pair grows linearly with distance:

$$V(\vec{x}_1, \vec{x}_2) = \frac{g^2}{2} C_2(\alpha) d(\vec{x}_1, \vec{x}_2). \tag{8.76}$$

This is precisely the picture we motivated back in the first section for the confinement of quarks: the electric flux is collimated into a thin tube between a quark/antiquark pair, resulting in an energy that grows linearly with separation! See figure 20 for an illustration.

If we take $G = SU(3)$ then we can also realize baryons in the strong-coupling approximation. Indeed consider acting on the strong-coupling vacuum with

$$\epsilon_{m_1 m_2 m_3} W^\dagger_{\alpha, m_1 n_1}(C_{41}) W^\dagger_{\alpha, m_2 n_1}(C_{42}) W^\dagger_{\alpha, m_3 n_1}(C_{43}) \Phi^\dagger_{n_1}(\vec{x}_1) \Phi^\dagger_{n_2}(\vec{x}_2) \Phi^\dagger_{n_3}(\vec{x}_3), \tag{8.77}$$

where C_{4i} is a path from x_i with $i = 1, 2, 3$ to a point x_4. The minimal energy configuration is a tree-like one where we choose the points x_4 and the paths C_{4i} to minimize the total number of links appearing in the three paths. Again there is a linear energy cost in trying to increase the size of the baryon. See figure 20 for an illustration.

It is interesting to consider the role of the mass of the matter particle in this discussion. If the potential energy in the pair is large compared to the particle mass, then there is a lower-energy eigenstate where we split the flux tube with a scalar/antiscalar pair. To avoid this we need

$$g^2 d \ll m. \tag{8.78}$$

[66] Unlike in the $1+1$ case this this set of eigenstates is not complete: to get a full eigenbasis we need to allow for acting with Wilson lines that fuse together using intertwiners, which leads to something called spin network states. We will meet an example of this momentarily when we talk about baryons.

In order to have a sizable distance range where pair production is suppressed we need

$$1 \ll d \ll \frac{m}{g^2}, \tag{8.79}$$

where I'll remind you we are working in units where $a = 1$. Moreover for the strong coupling expansion to be valid we want $g^2 \gg 1$, so to have a distance range where both the strong-coupling approximation is valid and pair production is suppressed we need

$$1 \ll g^2 \ll m. \tag{8.80}$$

In particular this means the mass of the particle must be large in lattice units. The theory is still confining when this is not the case, but it becomes hard to directly measure the linear potential since the disconnected contribution becomes dominant. In QCD this breaking of the flux tube is the essential mechanism for the hadronization of the outgoing partons in a hard scattering process.

8.2.1 What did we learn?

We have just seen that the confinement of quarks is an immediate and beautiful consequence of the strong coupling expansion in lattice gauge theory. But what does this really tell us? After all the argument works just as well if G is abelian, so $U(1)$ lattice gauge theory is also confining in the strong-coupling expansion! The difference however is that the question we are really interested in is *not* what happens when the theory is strongly-coupled at the lattice scale: instead we want to assume the theory is *weakly* coupled at the lattice scale, and then ask what happens at distances which are much larger. In the abelian case the answer to this question is clear: the β-function has a positive sign, so the coupling just gets weaker and weaker as we flow to the long distance. The abelian theory never enters the regime of validity of the strong-coupling expansion, so the fact that it would predict confinement is irrelevant. The novelty in QCD is that even if we start with a weak gauge coupling at the lattice scale, as we flow to long distance the coupling gets stronger and stronger due to asymptotic freedom. Once the coupling becomes $O(1)$ we can no longer compute what happens analytically, but we can *hope* that it continues to increase, eventually entering the regime of validity of the strong coupling expansion where confinement is manifest. And indeed all the available evidence from numerical lattice gauge theory, and also of course from experiment, is that this is indeed what happens! The situation can be summarized like this: the lattice strong coupling expansion gives a plausible explanation for how confinement is realized in real QCD, and numerics and experiment show that this explanation is correct. To *prove* this without using numerics or experiment, we would need to control the renormalization group at $O(1)$ coupling with enough precision to show that it indeed enters the strong-coupled confining phase at long distance. If you can do this there is a million dollar prize waiting for you, but in my view there isn't really anything else it could plausibly do so I am willing to believe it and move on.

8.3 Euclidean path integral formulation

So far we have been discussion Hamiltonian lattice gauge theory. This is the most intuitive place to think about the dynamics of gauge theory, but it is not well-suited for numerical computations (this may change if we get sufficiently powerful quantum computers). The approach which has been the most powerful for concrete calculations is the lattice version of the Euclidean path integral. In the continuum this has the form

$$\langle \Omega | f[A, \Phi] | \Omega \rangle = \frac{\int \mathcal{D}a \mathcal{D}\phi f[a, \phi] e^{-S_E[a,\phi]}}{\int \mathcal{D}a \mathcal{D}\phi e^{-S_E[a,\phi]}}, \tag{8.81}$$

where

$$S_E = \int d^d x \left(\frac{1}{2g^2} \mathrm{Tr}\,(F_{\mu\nu} F^{\mu\nu}) + \mathcal{L}_{M,E} \right) \tag{8.82}$$

is the Euclidean action and $\mathcal{L}_{M,E}$ is the Euclidean matter Lagrangian density. For example for a scalar field we have

$$\mathcal{L}_{M,E} = (D_\mu \phi)^\dagger D^\mu \phi + V(|\phi|). \tag{8.83}$$

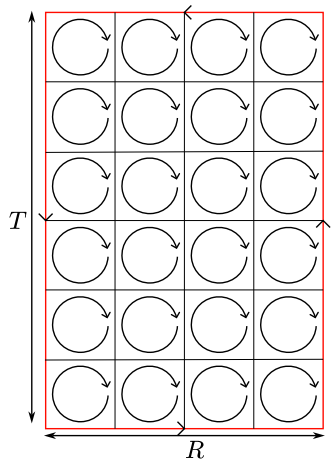

Figure 21: Tiling the interior of a rectangular Wilson loop with plaquettes to give a nonvanishing contribution in the Haar integral.

In Euclidean signature spacetime indices are raised and lowered with the Euclidean metric $\delta_{\mu\nu}$. The gauge-invariant operator f on the left-hand side is built out of time-ordered Euclidean Heisenberg fields with time evolution

$$\mathcal{O}(\tau) = e^{\tau H}\mathcal{O}(0)e^{-\tau H}. \tag{8.84}$$

In Euclidean lattice gauge theory we use a *spacetime lattice*, which includes links in the time direction as well as in space. So for example in figure 16 we can re-interpret the lattice as extending in both space and time. The gauge field is an assignment of a group element to each spacetime link, and the matter field is an assignment of a field value to each spacetime site. Since we already learned how to discretize the magnetic flux term in the Hamiltonian, we can use the same approach here for the spacetime action. This gives us the **Wilson action** of lattice gauge theory:

$$S_E = -\frac{1}{g^2}\sum_{(x,\delta,\delta')\in N} \Re\left(W_\alpha(x,\delta,\delta')\right), \tag{8.85}$$

or

$$S_E = -\frac{1}{g^2}\sum_{(x,\delta,\delta')\in N} \Re\left(W_\alpha(x,\delta,\delta')\right) - \sum_{(x,\delta)\in N}\left(\phi^\dagger_{x+\delta}W_{\alpha_\phi}(x,\delta))\phi_x + \text{c.c}\right) + \sum_{x\in N} V(|\phi|) \tag{8.86}$$

if we include scalar matter. To compute the path integral we then "merely" have to integrate e^{-S_E} times an observable over ϕ_x and $g_{(x,\delta)}$ for all x and (x,δ) in N:

$$\langle\Omega|f[\Phi,W_\alpha]|\Omega\rangle = \frac{\int dgd\phi f[\phi,W_\alpha(g)]e^{-S_E}}{\int dgd\phi e^{-S_E}}. \tag{8.87}$$

8.3.1 Strong coupling again

We will first use the Euclidean approach to reproduce our result that lattice gauge theory is confining in the strong-coupling limit. Since we no longer have direct access to the Hamiltonian, we need to do it in a somewhat less obvious way. The idea is to look at the expectation value in the vacuum of a large rectangular Wilson loop $W_\alpha(L)$, where L has length T in a direction we will view as time and R in a direction we will view as space. What we will show is that in the absence of matter fields, at large T we have

$$\langle\Omega|W_\alpha(L)|\Omega\rangle \propto e^{-\sigma TR}, \tag{8.88}$$

which is called **area-law** behavior for the Wilson loop. The interpretation of this behavior is that the expectation value of this Wilson line is proportional to

$$e^{-TV(R)} \tag{8.89}$$

where V is the potential between an infinitely-massive quark/antiquark pair with spatial separation R. This is because the temporal parts of the Wilson line precisely give the worldline action for the quark and the antiquark, so at large T this calculation is telling us the ground state energy for the gauge field in the presence of these fixed background particles. Therefore area law-behavior for the Wilson line is the same as having a linear attraction between quarks and antiquarks, just as we found more directly in the Hamiltonian approach.

To demonstrate the area law it is convenient to specialize to $G = SU(N)$, with $N \geq 3$. From Schur orthogonality we then have the integrals

$$\begin{aligned} \int dU\, U_{nm} &= 0 \\ \int dU\, U_{nm} U_{n'm'} &= 0 \\ \int dU\, U_{nm} U^*_{n'm'} &= \frac{1}{N} \delta_{nn'} \delta_{mm'}. \end{aligned} \tag{8.90}$$

The first of these we can think of as the integral against the trivial representation, while the second is the integral against the conjugate representation, so both vanish by Schur orthogonality between inequivalent irreps (what is different for $SU(2)$ is that the fundamental and antifundamental representations are equivalent so the second integral doesn't have to vanish). Since $1/g^2$ is small, the idea is then that we can evaluate the expectation value of $W_\alpha(L)$ by bringing down as few powers of the Wilson action as are needed in order to get a nonzero integral over U. To start with we can bring down a plaquette with a U^* for each U in L. But these plaquettes have other Us, so we need to bring down more plaquettes to provide a U^* for each of these. Continuing in this way we build up a surface D bounded by L that is tiled by the plaquettes that we brought down (see figure 21). We bring down the fewest plaquettes by taking this to be a minimal area surface, so the dependence on the parameters of the loop is

$$\langle \Omega | W_\alpha(L) | \Omega \rangle \propto \left(\frac{\sharp}{g^2} \right)^{TR}, \tag{8.91}$$

where $\sharp$ is some $O(1)$ number we haven't tried to compute. We can rewrite this as

$$\langle \Omega | W_\alpha(L) | \Omega \rangle \propto e^{-\sigma RT}, \tag{8.92}$$

where σ is energy density of the flux tube, also sometimes called the string tension, and is given by

$$\sigma = \log(g^2) + O(1). \tag{8.93}$$

This completes the demonstration of the area law. Notice that our expression for σ in terms of the bare coupling is not the same as we found in the Hamiltonian approach, but that is ok since they are different regularization schemes.

8.3.2 Monte Carlo evaluation

Finally we will say a little about how Euclidean path integrals are evaluated numerically outside of the strong-coupling regime. In principal we could try to just do the multiple integral (8.87), but direct evaluation of high-dimensional integrals is computationally very expensive due to the exponentially growing volume of the

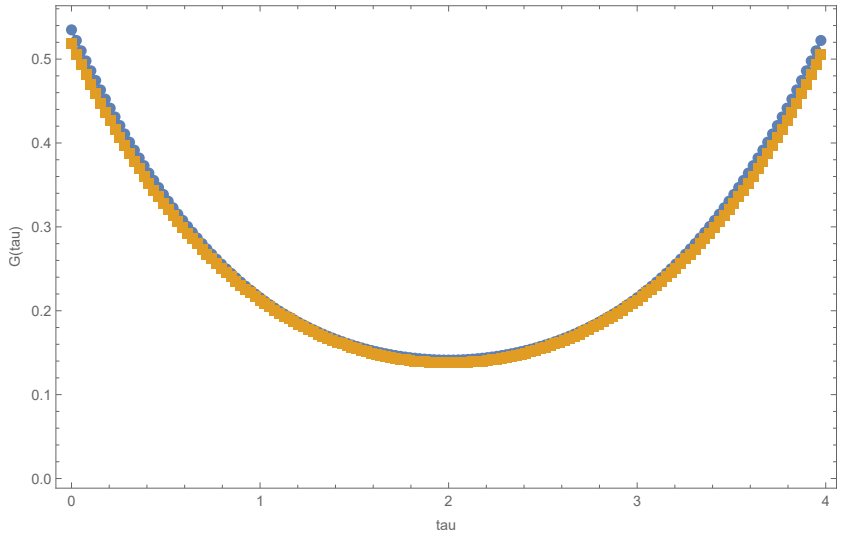

Figure 22: The Euclidean thermal correlation function for the harmonic oscillator with $\omega = 1$ and $\beta = 4$. The orange line is the exact answer, while the blue line is the result of running the Metropolis algorithm. The error is about 2%. This took about 30 minutes on my laptop, although my code was in mathematica and not optimized at all.

region of integration. A better approach is to use an essential feature of Euclidean action (8.86) which is that it is real. We can therefore view the quantity

$$p[g,\phi] \equiv \frac{e^{-S_E[g,\phi]}}{\int \mathcal{D}g'\mathcal{D}\phi' e^{-S_E[g',\phi']}} \tag{8.94}$$

as a probability distribution for Euclidean gauge/matter field configurations, and the integral (8.87) as computing the expectation value of a random variable f. What this means is that we do not actually need to sample the integrand for *all* field configurations g and ϕ, we just need to create a big enough set of samples the averaging f over the sample set is close to computing true average. Of course this is easier said than done, and in particular the term "big enough" is doing a lot of work here. How big is big enough? There is no definite answer to this problem that works in all cases, but two useful rules of thumb are that we want a big enough set that adding more samples does not substantially change the average, and we also want it to be the case that if we run the algorithm again the result that we get is not substantially different. This approach is called the **Monte Carlo method** for numerical integration.

The key question in Monte Carlo is how we generate our samples to average over. As you can imagine this is a very refined art, with many special optimizations for specific problems. Here I will only describe one of the oldest and simplest ways of generating samples, the **Metropolis algorithm**. The idea of the algorithm is quite straightforward: we define an evolution on the set of field configurations which goes through site/link by site/link and randomly proposes an "update" to the configuration at that site/link: the update is then accepted with probability one if it decreases S_E and with probability $e^{-\Delta S_E}$ if it increases S_E. Repeating this process many times generates a long string of samples, which we can then use as our ensemble for estimating the expectation value of f. The Metropolis algorithm is the canonical example of what is called a **Markov Chain/Monte Carlo** algorithm, or **MC/MC** for short.

To illustrate the Metropolis algorithm, we can consider the thermal two point function in the harmonic oscillator:

$$G(\tau) \equiv \frac{\mathrm{Tr}\left(e^{-\beta H}X(\tau)X(0)\right)}{\mathrm{Tr}\left(e^{-\beta H}\right)} = \frac{\int \mathcal{D}x x(\tau)x(0)e^{-S_E}}{\int \mathcal{D}x e^{-S_E}}, \tag{8.95}$$

with

$$H = \frac{P^2}{2} + \frac{\omega^2}{2} X^2 \tag{8.96}$$

$$S_E = \frac{1}{2}\dot{x}^2 + \frac{\omega^2}{2} X^2. \tag{8.97}$$

Since it is the harmonic oscillator we know the exact answer, it is given by

$$G(\tau) = \frac{\cosh\left[\omega\left(\tau - \beta/2\right)\right)]}{2\omega \sinh\left[\omega\beta/2\right]} \tag{8.98}$$

with $\tau \in (0, \beta)$. To use the Metropolis algorithm we discretize Euclidean time using a lattice

$$\tau_n = n\frac{\beta}{N}, \tag{8.99}$$

in terms which we integrate over field configurations x_n obeying the boundary condition

$$x_0 = x_N. \tag{8.100}$$

The discretized Euclidean action is

$$S_E = \sum_{n=0}^{N-1} \left(\frac{(x_{n+1} - x_n)^2}{2a} + \frac{a\omega^2}{2} x_n^2 \right). \tag{8.101}$$

The job of Metropolis is to create a sequence $x_n^{\{a\}}$ of configurations using the following rule. Say we have the configuration $x_n^{\{a\}}$ and we want to create $x_n^{\{a+1\}}$. For each n (we can go through them in order) we propose the change

$$x_n^{\{a+1\}} = x_n^{\{a\}} + \Delta x, \tag{8.102}$$

where Δx is chosen uniformly randomly in an interval $\Delta x \in (-\epsilon, \epsilon)$. If this change decreases the Euclidean action we accept it, while if it does not then we accept it with probability $e^{-\Delta S_E}$. Each time we go through the list of N positions is called a "sweep", and after many sweeps we generate a long chain of x_n configurations. We then average our observable, $x(\tau)x(0)$ in this case, over the chain (with a few minor refinements I'll mention in a moment). In figure 22 you can see a comparison of the result of this algorithm to the exact answer, computed with 156 lattice sites, $400,000$ sweeps, and initialized in the $x_n = 0$ configuration. In the homework you will make your own version of this plot.

We will close with a few general comments on issues that come up in using this algorithm:

- One obvious potential problem is initial bias: we have to begin the chain somewhere, for a while it remembers these initial conditions and that biases the samples. The usual solution to this is simply to discard a long enough initial part of the chain that this bias is forgotten. This step in the algorithm is usually called "burn-in", for the run that produced figure 22 I ignored the first 120,000 sweeps in computing the average.

- Another obvious problem is that each samples which are near each other in the chain will be fairly strongly correlated, since it takes a fair number of sweeps to "re-randomize". This means that we cannot view the samples as independent, and this means that we have to run the algorithm for longer than we would if they were truly independent. This problem is usually called autocorrelation. One way to mitigate it is to "thin" the chain by only using say every 20th sample in computing the average of the observable. This does help with the autocorrelation, but actually it isn't really necessary because even if we use all the samples the "size" of the blocks with high autocorrelation will be roughly similar throughout the chain and so we can just compute the average over everything and not get into much trouble.

- The running time of the algorithm depends quite strongly on the range ϵ over which we choose Δx. If ϵ is too large then very few proposed changes will be accepted, leading to a long convergence time, while if ϵ is too small then many changes will be accepted but they will barely change the samples, again leading to a long convergence time. The usual rule of thumb is that you should choose ϵ so that $30 - 70\%$ of the proposed changes are accepted. In my code I used $\epsilon = .3$.

- Another problem is that since the proposed updates only change x_n one lattice site at a time, the chain can get stuck for an exponentially long time in a metastable minimum that requires many changes to get out. For example in applying this algorithm to the Ising model at low temperature with a small external magnetic field, if we start with the "wrong" magnetization it is almost impossible to get out. In order to fix this we need to improve the algorithm to occasionally try "larger" updates just in case we are stuck.

- As a matter of practice it is usually better to use your configurations x_n as you generate them, i.e. for each configuration compute its contribution to the average of the observable right away, and then discard it. Otherwise you would need to keep all your configurations in memory until the end and then compute one giant average. Also you can get better convergence when computing a correlation function like $G(\tau)$ if you average over the locations of the two points keeping their distance fixed.

- Many of the theories we are interested in, such as QCD, have fermions. The fermionic path integral is over Grassmann variables, and it cannot be given a probabilistic interpretation. This seems like a fatal problem for the Monte Carlo approach, but in fact there is a way out: we simply integrate out the fermions, which usually have a Gaussian action, to get a complicated determinant. We can then absorb this determinant as a nonlocal part of the action if it is positive, or else try treating it as part of the observable. Unfortunately the non-locality makes it more expensive to compute updates of the chain. This problem is on top of the fermion doubling problem that we met with lattice fermions last semester.

- Not all field theories, even without fermions, have actions that are real in Euclidean signature. Moreover we are not only interested in Euclidean field theory: we would like a way to compute correlation functions in Lorentzian signature where we actually live! Either of these situations leads to an integrand which is not real and positive, which is usually called the **sign problem**. So far the sign problem in general is an insurmountable obstacle to numerical calculations in field theory, although in some special examples there are tricks that can be used to make progress.

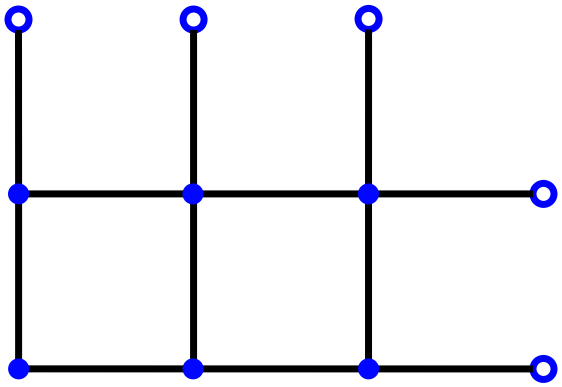

Figure 23: A cubic lattice on a torus, with $N_x = 3$ and $N_y = 2$. The hollow circles indicate gluings to the site on the opposite side of the lattice, so this lattice has 6 sites and 12 links.

8.4 Homework

1. Derive the lattice algebra (8.15) and (8.16) from the definitions (8.14).

2. Show that gauge-invariance of the state of $1+1$ Hamiltonian lattice gauge theory is indeed equivalent to (8.60). Also show that this requires the wave function to have the form (8.61).

3. In this problem you will solve the simplest $2+1$ dimensional gauge theory, the $\mathbb{Z}_2$ gauge theory. On each link we have a two-dimensional Hilbert space, and it is conventional to define $Z = W_\alpha$ and $X = L_{-1} = R_{-1}$, where α is the defining (and only nontrivial) representation of $\mathbb{Z}_2$. Check that these operators indeed act on the group space $|\pm 1\rangle$ as the Pauli Z and X operators. The Hamiltonian is

$$H = -\frac{g^2}{2} \sum_{(\vec{x},\vec{\delta})\in N} X(\vec{x},\vec{\delta}) - \frac{2}{g^2} \sum_{(\vec{x},\vec{\delta},\vec{\delta}')\in N} Z(\vec{\delta}+\vec{\delta}',-\vec{\delta}')Z(\vec{x}+\vec{\delta}+\vec{\delta}',-\vec{\delta})Z(\vec{x}+\vec{\delta},\vec{\delta}')Z(\vec{x},\vec{\delta}), \quad (8.103)$$

   where the second term is the usual magnetic flux term and the first term is a discrete analogue of the electric flux term. At large g we want to have $X = 1$ on each link to minimize the energy, which is a discrete version of the tensor product strong-coupling vacuum that we found above. It is unique on any spatial lattice.

   In this problem however we will study the theory in the weak-coupling limit, where it is the plaquette term we need to minimize. We will work on a cubic lattice with the topology of a spatial torus, as shown in figure 23. Thus to find the ground state we want to find the set of states where

$$Z(\vec{\delta}+\hat{y},-\hat{y}')Z(\vec{x}+\hat{x}+\hat{y},-\hat{x})Z(\vec{x}+\hat{x},\hat{y})Z(\vec{x},\hat{x}) = 1 \quad (8.104)$$

   for all plaquettes $(\vec{x},\hat{x},\hat{y})$, together with the gauge-invariant constraint

$$X(\vec{x},\hat{x})X(\vec{x}-\hat{x},\hat{x})X(\vec{x},\hat{y})X(\vec{x}-\hat{y},\hat{y}) = 1 \quad (8.105)$$

   for all sites. We are interested in identifying the dimensionality of the subspace that satisfies these two constraints.

   (i) Assuming the lattice has $N_x \times N_y$ sites, how many sites, links, and plaquettes are there? If we assume that all of the plaquette and site constraints are independent, what would we conclude about the number of ground states?

(ii) In fact the constraints are not all independent. How many ways are there to come up with a nontrivial product of constraints which is equal to the identity? What does this say about the dimensionality of the ground state subspace?

(iii) Identify some gauge-invariant operators that we can simultaneously diagonalize to uniquely label the ground states. Can you construct some gauge-invariant operators that move you between these ground states?

(iv) What must happen to the theory as we vary g from weak to strong coupling?

In the continuum limit all states of this theory except for the ground subspace go off to infinite energy, resulting in a nontrivial theory with zero Hamiltonian: the continuum $\mathbb{Z}_2$ gauge theory. This is the simplest nontrivial example of a **topological field theory**, which means a field theory whose expectation values are independent of the spacetime metric.[67]

4. Using the code of your choice, write a program to compute the thermal correlator (8.95) in the harmonic oscillator using the Metropolis algorithm, and compare your result to the exact expression (8.98). You can use the parameters I mentioned in the text, but of course you should also play around to see how the converges works. I don't mind if you consult with an LLM, but please try to write the code yourself first.

[67] At some point quantum information people started calling the $\mathbb{Z}_2$ lattice gauge theory on the torus the "toric code", essentially because this ground state subspace is fairly "well-protected" against errors which act on it with local operators.

9 Chiral symmetry breaking and the effective action for pions and nucleons

In this section we will discuss the classic example of spontaneous breaking of a global symmetry in particle physics, which leads to the realization that the relative lightness of pions and kaons compared to other hadrons is explained by their being Goldstone bosons for an approximate global symmetry of QCD that becomes exact in the limit of vanishing quark masses. This also allows us to write down an effective action for the interactions of pions and nucleons which is called the **chiral Lagrangian**.

9.1 Flavor symmetry in massless QCD

To get started we can consider a simplified version of QCD where the only matter is N_F massless quarks q_i. We will primarily be interested in this model with $N_F = 2$ or $N_F = 3$, with the former being a model that treats the u and d quarks as massless and ignores the rest and the latter doing the same for u, d, and s. The Lagrangian of this model is

$$\mathcal{L} = -\frac{1}{2g^2}\mathrm{Tr}\,(f_{\mu\nu}f^{\mu\nu}) - i\overline{q}^i \not{D} q_i, \tag{9.1}$$

where D_μ is the covariant derivative in the fundamental representation of $SU(3)$ and $i = 1, 2, \ldots N_F$. What we are interested in in this section are the global symmetries of this action. There is an immediate $U(N_F)$ symmetry that rotates the q_i into each other, but this is not actually the full internal symmetry of this Lagrangian. The reason is that it does not mix the left- and right-handed Weyl components of the q_i, so we can actually do independent $U(N_F)$ transformations on these. Therefore the full flavor symmetry group is

$$G_F \equiv U(N_F)_L \times U(N_F)_R, \tag{9.2}$$

which acts on the quark fields as

$$q_i' = U_{L,ij}P_L q_j + U_{R,ij}P_R q_j, \tag{9.3}$$

with

$$\begin{aligned} P_L &= \frac{1+\gamma}{2} \\ P_R &= \frac{1-\gamma}{2} \end{aligned} \tag{9.4}$$

being the projections onto the left-handed and right-handed components of a Dirac spinor.

There are several subgroups of G_F that are worth discussing specifically. First of all there is the $U(1)_B$ **baryon number** symmetry

$$q_i' = e^{i\theta} q_i, \tag{9.5}$$

which is a symmetry of QCD even when the quark masses are restored. It is generated by the baryon number current

$$j_B^\mu = -\frac{1}{3}\overline{q}^i \gamma^\mu q_i, \tag{9.6}$$

where the factor of 3 is included so that proton has baryon number one. Another important subgroup is **isospin** symmetry $SU(N_F)_I$, which is the diagonal subgroup of G_F where $U_L = U_R = U$ and $\det(U) = 1$:

$$q_i' = U_{ij} q_j. \tag{9.7}$$

The Noether currents for isospin symmetry are

$$j_{I,a}^\mu = -\overline{q}^i (T_a)_i{}^j \gamma^\mu q_j, \tag{9.8}$$

where $(T_a)_i{}^j$ are the generators of $SU(N_F)_I$. These currents are Lorentz vectors. There is also an **axial symmetry** subgroup $U(1)_A$ that acts as

$$q_i' = e^{i\theta\gamma} q_i. \tag{9.9}$$

Its Noether current is

$$j_A^\mu = \overline{q}\gamma\gamma^\mu q. \tag{9.10}$$

These three subgroups account for $N_F^2 + 1$ of the generators of G_F: the remaining $N_F^2 - 1$ generators correspond to the pseudovector currents[68]

$$j_{P,a}^\mu = \overline{q}^i (T_a)_i{}^j \gamma\gamma^\mu q_j. \tag{9.11}$$

Isospin symmetry and baryon number symmetry would persist if we gave the quarks nonzero but equal masses, while the rest of the flavor symmetry group would be explicitly broken.

9.2 Flavor symmetry of the hadron spectrum

We now consider to what extent these flavor symmetries are realized by the observed spectrum of light hadrons. Baryon number symmetry is clearly present, for example in determining which hadron is a meson and which is a baryon. Isospin also turns out to be a good approximate symmetry of the hadron spectrum, as we will now discuss.

9.2.1 Isospin

We'll begin with $N_F = 2$, in which case there are only u and d quarks. Referring back to our table of hadrons, the three pions $\pi^\pm, \pi^0$ with masses $m_\pi \approx 140$MeV are a natural candidate for the three-dimensional adjoint (vector) representation of $SU(2)$. The three ρ mesons, with mass $m_\rho \approx 775$MeV, are another plausible adjoint, and the ω, with mass $m_\omega \approx 783$MeV, is a plausible singlet. There is also a linear combination that we'll call η_0 of the η and η' which persists in the limit that we remove the strange quark, and this is also an isospin singlet.[69] Moving on to baryons, the proton and neutron form a plausible doublet of $SU(2)$ with $m_n \approx 940$MeV, while the four Δ baryons give a plausible spin 3/2 representation with $m_\Delta \approx 1232$MeV. This exhausts our list of hadrons that do not involve the strange quark, which we thus see is indeed consistent with $SU(2)$ isospin symmetry.

We can also understand these isospin representations directly using the quark fields: the three pions correspond to the local operators

$$\overline{q}T_a\gamma q, \tag{9.12}$$

which is indeed an isospin adjoint, the three ρ mesons correspond to the local operators

$$\overline{q}T_a\gamma^\mu q, \tag{9.13}$$

which again is an adjoint, and the ω corresponds to

$$\overline{q}\gamma^\mu q. \tag{9.14}$$

There is also the pseudoscalar operator

$$\overline{q}\gamma q, \tag{9.15}$$

that creates the η_0.[70]

We turn now to the case of $N_F = 3$, where we restore the strange quark. There are two different senses in which we could study isopspin in this situation. The first is to treat the strange quark as massless and

[68] The axial symmetry current is also a pseudovector, but we will see in a moment that it is a good idea to treat it separately.

[69] Since the η_0 is a mixture of η and η', each of which involve the strange quark, and which have rather different masses, we can not take its mass approximately from experiment. Lattice simulations however suggest that in a world without a strange quark its mass would be around $800 - 1000$MeV.

[70] You might ask how I knew where to put γ in these formulas, for example why isn't there a meson created by $\overline{q}q$? There is actually, but it is very broad since it has the same quantum numbers as a pair of pions so it decays almost immediately. Similarly $\overline{q}T_a q$ gives us a broad particle that quickly decays into a pion and an η_0 (at least if this is kinematically allowed, without knowing the precise masses of this particle and the η_0 we cannot be sure and I couldn't find them in the lattice literature; what is certainly true is that if we put back the strange quark then this decay happens promptly to $\pi\eta$).

discuss the full $SU(3)$ isospin symmetry. This however will be broken fairly badly since $m_s \gg m_u, m_d$. More accurate is to continue to use the $SU(2)$ isospin we discussed above, taking it to act trivially on the strange quark. Thus the approximately degenerate multiplets of $SU(3)$ decompose into more precisely degenerate multiplets of $SU(2)$. For example we can guess that the three pions together with the four kaons $K^{\pm}$, K^0_S, K^0_L, and also the η, together transform in the eight-dimensional adjoint representation of $SU(3)$. This is sometimes called an **octet**, and we can decompose it into $SU(2)$ representations as

$$8 = 3 \oplus 2 \oplus 2 \oplus 1, \tag{9.16}$$

where the 3 is the pions, the pair of 2s is the kaons (they are degenerate with each other due to charge conjugation symmetry), and the 1 is the η.[71] The η' is a singlet under $SU(3)$ and thus also $SU(2)$. This $SU(3)$ adjoint is a pseudoscalar under Lorentz symmetry. There is another $SU(3)$ adjoint octet that is a Lorentz vector, consisting of the three ρ mesons, the four K^* mesons, and a linear combination of the ϕ and ω mesons. The remaining linear combination of ϕ and ω is an $SU(3)$ singlet.[72] The operators for these two octets are again (9.12) and (9.13), with q now being a three-component object.

We can similarly consider how the spectrum of light baryons fit into approximate $SU(3)$ isospin representations. The spin 1/2 proton, neutron, Λ, Σ, and Ξ baryons all plausibly fit into yet another adjoint octet of $SU(3)$. In the decomposition (9.16) of this octet into $SU(2)$ representations, the proton and neutron are an $SU(2)$ fundamental, as are the Ξ baryons, the Σ baryons are an $SU(2)$ adjoint, and the Λ is an $SU(2)$ singlet. The remaining spin 3/2 baryons in our table, the Δ, Σ^*, Ξ^* and Ω baryons, fit into a ten-dimensional $SU(3)$ representation called the **decuplet** that you will study in the homework. We have thus used isospin symmetry to explain the approximate mass degeneracies of all of the light hadrons we discussed back in the first section. The main lesson to remember is that the hadron masses are much closer within an $SU(2)$ subrepresentation than they are in a full $SU(3)$ representation, reflecting the fact that $SU(2)$ isospin is a much better approximate symmetry than $SU(3)$ isospin.

9.2.2 Chiral symmetry breaking

At this point you may be concerned. We accounted for all of the approximate degeneracies in the light hadron spectrum using isospin symmetry, so what is left for the rest of our flavor symmetry group G_F to explain? We can make this problem more severe by noting that the pseudovector charges

$$Q_{P,a} = \int d^{d-1}x j^0_{P,a} \tag{9.17}$$

anticommute with spatial reflection,

$$Q_{P,a} U_{\mathcal{R}} = -U_{\mathcal{R}} Q_{P,a}. \tag{9.18}$$

This means that for each pseudoscalar hadron such as a pion, by acting on it with $Q_{P,a}$ we should expect to find a scalar hadron with approximately the same mass. There are no such hadrons in our observed table, so apparently the pseudovector charges do not generate symmetries of the hadron spectrum. This becomes even more clear if we consider the observed scattering amplitudes and decay rates of hadron: they do not obey any flavor selection rules beyond those implied by isospin and baryon number. Fortunately we have an alternative: the pseudovector directions in G_F could be spontaneously broken. In that case acting with $Q_{P,a}$ on a hadron would not need to result in another type of hadron, instead it would create a two-particle state consisting of the hadron we started with together with a pseudoscalar Goldstone boson. We can thus immediately test this proposal: where are these pseudoscalar Goldstone bosons? We first consider the case $N_F = 2$, in which case there are three pseudovector currents so we should expect three pseudoscalar Goldstone bosons. These of course are just the three pions, $\pi_{\pm}$ and π_0, and they indeed transform in

[71] One useful trick in doing these decompositions is to note that hadrons with different numbers of strange quarks must be in different $SU(2)$ representations.

[72] The reason that the mass eigenstates are quite different from the flavor eigenstates for the ϕ and ω is that they have different numbers of strange quarks, so the $SU(3)$ breaking is quite bad. This does not happen for the η and η' because the former is a Goldstone boson and the latter is heavy due to an anomaly, as we will soon see.

an adjoint of the unbroken $SU(2)$ isospin symmetry just as the pseudovector currents do. Moreover this explains perhaps the most striking feature of our table of hadrons, which is that the pions are much lighter than the other hadrons - this is because they are approximate Goldstone bosons! In particular their mass differences are of order $\sim 5\text{MeV} \sim m_u, m_d$, just as we would expect from isospin breaking by quark masses (electromagnetism, which we have not included in our discussion, also gives splitting of the same order). Turning to $N_F = 3$, there are now eight pseudovector currents so we can view all of the pseudoscalar meson octet as Goldstone bosons. This explains why the kaons and the η are substantially lighter than the rest of the hadrons. It also explains why they are not as light as the pions; again this is because $SU(3)$ isospin is a worse symmetry than $SU(2)$ isospin due to the large strange quark mass.

The proposed phenomenon that flavor symmetry is spontaneously broken to isospin and baryon number symmetry is called **chiral symmetry breaking**.[73] At the level of the Lie algebra, if we ignore for a moment the axial symmetry then the symmetry breaking pattern is

$$\mathfrak{su}(N_F)_L \otimes \mathfrak{su}(N_F)_R \otimes \mathfrak{u}(1)_B \to \mathfrak{su}(N_F)_I \otimes \mathfrak{u}(1)_B. \tag{9.19}$$

As we have seen, this rather simple proposal is able to qualitatively explain most of the features of the spectrum of light hadrons.

9.2.3 The fate of axial symmetry

There is one remaining generator of G_F, the generator of the axial symmetry $U(1)_A$. Its current is also a pseudovector, so the associated charge again anticommutes with parity. Thus if it were an unbroken global symmetry, it would lead to an unobserved doubling of the hadron spectrum. Based on our experience with chiral symmetry breaking we might therefore guess that axial symmetry is spontaneously broken, but this also does not work: there is no plausible Goldstone boson in the hadron spectrum.[74] So what happened to this symmetry? In the early days of QCD this was called the "$U(1)$ problem", and it was a major defect in the quark model of hadrons. Eventually it was understood by 't Hooft that axial symmetry is actually not a symmetry at all, it is explicitly violated by non-perturbative effects called instantons. We will return to instantons at the end of the semester.

9.3 Which symmetries can spontaneously break?

There is one glaring issue with our discussion so far: we have not shown that flavor symmetry is actually broken as (9.19) in QCD. In particular there is no classical potential for an order parameter like what we had in previous examples of spontaneous symmetry breaking, so if chiral symmetry breaking happens it must be due to strong non-perturbative quantum effects. Of course the consistency of lattice calculations with the observed hadron spectrum is strong numerical evidence that it does happen, but we would also like some theoretical understanding of how and why. As in the case of confinement it is possible to study this problem in the strong-coupling approximation on the lattice, but in this case we run directly into the fermion doubling problem. In particular there is no lattice realization of fermions which preserves the full flavor symmetry, and this makes it difficult to use strong-coupling methods to study its possible breaking away from the continuum limit.[75] The most we can say on general grounds is that the order parameter is rather constrained: we want to have a Lorentz scalar which is charged under the pseudovector flavor generators but not under isospin or baryon number (or charge conjugation). The simplest possibility is to have the expectation value

$$\langle \overline{q}^i P_L q_j \rangle = iv\delta^i_j, \tag{9.20}$$

[73] I do not like this name, since it makes it sound like the pseudovector currents generate a symmetry called "chiral symmetry" which is spontaneously broken, but they do not form a Lie algebra since their associated charges are not closed under commutators. We are stuck with it however.

[74] The least implausible candidate is the η', but its mass is similar to the rest of the hadrons so it isn't light.

[75] In 1+1 dimensions it is possible to preserve some discrete analogue of flavor symmetry on the lattice, and then one can indeed check that it is spontaneously broken. This however is tied up with an anomaly that is not present in $3+1$ dimensions, so it is not clear to what extent this is a good model for what happens in higher dimensions.

where by dimensional analysis we can guess that we should have

$$v \sim \Lambda_{QCD}^3. \tag{9.21}$$

This expectation value is sometimes called the **chiral condensate**. The factor of i is included for later convenience, we can change it with a field redefinition.

So far it has not been possible to give a compelling theoretical explanation for a nonzero chiral condensate in $3+1$ dimensions, but there are some plausibility arguments that are worth mentioning. First of all there is a fairly simple argument due to Vafa and Witten that isospin and baryon number symmetry are *not* spontaneously broken. Thus chiral symmetry breaking really just says that in massless QCD all of the spontaneous symmetry breaking which is allowed on general grounds does indeed happen, which doesn't sound so unreasonable to me. There is also an argument based on anomalies that *something* interesting needs to happen at long distances in massless QCD, and chiral symmetry breaking is the simplest something that can satisfy this argument. We will discuss that argument further once we have studied anomalies, for now we will just present the Vafa-Witten argument since it has several enlightening features.

The idea of Vafa and Witten is to study the Euclidean expectation value of some product of fermion and anti-fermion operators, dressed by Wilson lines to make them gauge-invariant:

$$\langle \Psi_1 \dots \Psi_n \overline{\Psi}_1 \dots \Psi_n W \rangle = \frac{\int \mathcal{D}\psi \mathcal{D}\overline{\psi} \mathcal{D}a, \psi_1 \dots \psi_n \overline{\psi}_1 \dots \overline{\psi}_n W e^{-S_E}}{\int \mathcal{D}\psi \mathcal{D}\overline{\psi} \mathcal{D}a \, e^{-S_E}}. \tag{9.22}$$

Here I have suppressed the positions and flavor, color, and spinor indices of the fermions, and I have rather heuristically written the Wilson lines as W. In fact to make the argument more precise we will assume that the locations of the fermions (as well as the endpoints of the dressing Wilson lines) are integrated against smooth test function spinors of compact support whose spinor indices are contracted with those of Ψ or $\overline{\Psi}$. In a more explicit notation we would write

$$\psi_i = \int d^d x_i f_i^{\alpha_i}(x_i)^* \psi_{\alpha_i, k_i, n_i}(x_i), \tag{9.23}$$

where α_i is a spinor index, k_i is a flavor index, and n_i is a color index. The color indices are all contracted with color indices on W, so the only free indices in this correlation function are the flavor indices. Since we are in Euclidean signature the fermion action is

$$S_E \supset \int d^d x \overline{\psi} \left(\not{D} + M \right) \psi, \tag{9.24}$$

where M is a mass matrix that is diagonal in the spinor and color indices but can be nontrivial for the flavor indices. In Euclidean signature we have

$$\gamma_E^0 = i\gamma^0, \tag{9.25}$$

and ψ and $\overline{\psi}$ are independent Grassmann variables (as operators however we have $\overline{\Psi} = \Psi^\dagger \gamma_E^0$). We will assume for now that the eigenvalues of M are all positive, but for example they could all be equal in which case isospin would remain a good symmetry. What we will argue is that in this case all flavor symmetries that commute with M are unbroken.

The first thing to do is integrate out the fermions using the Gaussian action (9.24), which gives

$$\langle \Psi_1 \dots \Psi_n \overline{\Psi}_1 \dots \Psi_n W \rangle = \frac{1}{Z} \sum_{\pi \in S_n} (-1)^{p(\pi)} \int \mathcal{D}a \, (\not{D}+M)^{-1}_{1,\pi(1)} \dots (\not{D}+M)^{-1}_{n,\pi(n)} W \, \det(\not{D}+M) e^{-S_E^g}. \tag{9.26}$$

Here

$$S_E^g = \frac{1}{2g^2} \int d^d x \mathrm{Tr} \left(f_{\mu\nu} f^{\mu\nu} \right) \tag{9.27}$$

is the Euclidean gauge field action, and the notation $(\not{D} + M)^{-1}_{i,j}$ is shorthand for the Euclidean propagator from antifermion j to fermion i in the presence of the background gauge field a. The denominator Z is

$$Z = \int \mathcal{D}a \, \det(\not{D} + M) e^{-S_E^g}, \tag{9.28}$$

and $p(\pi)$ indicates the parity of a permutation π of n objects. The quantity on the right-hand side of (9.26) is manifestly invariant under acting on the flavor indices of the propagators with any flavor symmetry transformation that commutes with M, since these are automatically symmetries of the propagator since $\not{D}$ is diagonal in the flavor indices. This however does not really show that the these symmetries are not spontaneously broken, instead it merely shows that the Euclidean path integral projects onto a symmetric state within the vacuum subspace. To really show that these symmetries are not spontaneously broken, we need to show that a generic small deformation δM of the mass matrix leads to a small deformation of the correlation function. This is because in a situation where the flavor symmetries were spontaneously broken, such a small perturbation would drastically change the expectation value from that in the symmetric state to that in a pointer state.

The way that Vafa and Witten showed that the right-hand side of (9.26) is stable under a symmetry breaking perturbation δM is by giving a bound on the propagators which is independent of the gauge-field configuration and also the spatial volume, and then arguing that $\det(\not{D} + M)$ is positive. Since the Wilson lines W are unitary they are also bounded, and this means that we are computing the expectation value of a bounded quantity against a positive probability measure. The average thus obeys the same bound, and so in particular there can be no divergence as we take $\delta M \to 0$. The key ingredient to showing both of these things is realizing that the Euclidean Dirac operator $\not{D}$ is antihermitian, which you will show on the homework. This means that given an eigenspinor

$$\not{D}\psi = i\lambda\psi \tag{9.29}$$

with $\lambda \neq 0$ we always have another eigenspinor $\gamma\psi$ obeying

$$\not{D}\gamma\psi = -\gamma\not{D}\psi = -i\lambda\gamma\psi, \tag{9.30}$$

so the nonzero eigenvalues are always paired. Thus their contribution to the determinant is

$$(M + i\lambda)(M - i\lambda) = M^2 + \lambda^2, \tag{9.31}$$

which is positive. There can also be zero eigenvalues which do not have paired eigenspinors, but for these we are just computing the determinant of M which is again positive by assumption. To get a bound on the propagator, we first adopt a flavor basis where M is diagonal. We may then write

$$\left(\not{D} + M\right)^{-1}_{ij} = \pm \int d^d x d^d y f_i^\dagger(x) \int_0^\infty ds \left(e^{\mp(\not{D}+m_i)s}\right)_{xy} f_j(y), \tag{9.32}$$

where $\pm$ indicates the sign of m_i, and we have used that $\not{D}$ is antihermitian to ensure the convergence of the s integral. x and y are the position labels on the differential operator $e^{\mp(\not{D}+m_i)s}$, and f_i and f_j are the smooth spinors of compact support that we have smeared ψ and $\overline{\psi}$ against. We take these to be normalized as

$$\int d^d x f^\dagger(x) f(x) = 1, \tag{9.33}$$

where we have contracted the unwritten spinor indices. Noting that $e^{s\not{D}}$ is unitary, by the Cauchy-Schwarz inequality for operators on spinors we have

$$|\left(\not{D} + M\right)^{-1}_{ij}| \leq \int_0^\infty ds e^{-|m_i|s} = \frac{1}{|m_i|} \leq \frac{1}{m_0}, \tag{9.34}$$

where m_0 is the absolute value of the smallest (in absolute value) eigenvalue of M. This bound is indeed smooth under small deformations of M, provided that we begin with M positive.

What we have shown so far is that any symmetry of massless QCD that persists when we turn on some positive mass matrix M cannot be spontaneously broken once we do so. In other words any candidate vacuum that breaks these symmetries must have strictly higher energy. This part of the argument is quite solid. Now we ask what can happen as we tune M to zero. Unless there is a coincidence, we should not expect these symmetry-breaking vacua to become degenerate with the unbroken vacuum in this limit. Indeed the only reason for a degeneracy to appear as $M \to 0$ is if a new symmetry emerges. Of course there *are* new symmetries that emerge when $M = 0$, and these are precisely the pseudovector symmetries which we expect to be spontaneously broken! The point however is that if these new vacua are related to the one where the $M > 0$ symmetries are not broken by a new symmetry, then (possibly up to a relabeling) the $M > 0$ symmetries should also not be broken in these new vacua. In order to have them be broken, we would need an *additional* coincidental degeneracy beyond that provided by chiral symmetry breaking.

9.4 Effective actions for general Goldstone bosons

We now turn to the question of how to describe the interactions of Goldstone bosons. In general say that we have a compact internal global symmetry group G broken to a closed subgroup H by some expectation value

$$v_i = \langle v|O_i|v\rangle, \tag{9.35}$$

where O_i is a basis for the local operators in the theory. Writing the action of the symmetry group on the local operators as

$$U^\dagger(g)O_iU(g) = D_i{}^j(g)O_j, \tag{9.36}$$

where D by definition is a faithful representation of G, the unbroken subgroup H is defined by

$$H = \{h \in G|D(h)v = v\}. \tag{9.37}$$

$|v\rangle$ here are the set of degenerate vacua. In this situation Goldstone's theorem tells us that there are $\dim(G) - \dim(H)$ massless Goldstone bosons. At low energies we can hope to build an effective field theory describing the physics of these Goldstone bosons, together with any other light degrees of freedom which happen to be around. To do this we postulate the existence of some Goldstone boson fields, determine their symmetry transformation properties, and write down an effective action containing all possible terms built out of these fields that respect the symmetries. Higher derivative terms will be suppressed by powers of the energy cutoff on this effective field theory.

The first step in this algorithm is to describe the Goldstone boson fields. From Goldstone's theorem these should be $\dim(G) - \dim(H)$ independent scalar fields, transforming in the adjoint representation of H since that is what the broken currents do. An elegant way to implement these two requirements is to take the Goldstone fields to be valued in the quotient vector space $\mathfrak{g}/\mathfrak{h}$. In other words we take the field at each point to be

$$T = \xi^a T_a, \tag{9.38}$$

where T_a are the generators of $\mathfrak{g}$, and we identify

$$T \sim T + X \tag{9.39}$$

for any $X \in \mathfrak{h}$. The unbroken symmetry H acts on these Goldstone boson fields as

$$T' = h^{-1}Th, \tag{9.40}$$

which respects the quotient (9.39) since $\mathfrak{h}$ is closed under conjugation by elements of H. To describe this quotient more concretely, we can use the adjoint-invariant inner product on $\mathfrak{g}$ to pick an orthonormal set of broken generators $\hat{T}_a$, each of which is orthogonal to everything in $\mathfrak{h}$, and then we can simply take

$$T = \xi^a \hat{T}_a \tag{9.41}$$

with no further quotient. The adjoint action of H keeps us within this subspace since for any $X \in \mathfrak{h}$ we have

$$\langle h^{-1}\hat{T}_a h, X\rangle = \langle \hat{T}_a, hXh^{-1}\rangle = 0. \tag{9.42}$$

There are two problems however with this simple approach. The first problem is that the effective action we write down should be invariant under all of G, not just H, and the G transformation of T is complicated. This is because conjugation by a general element of G takes us out of the subspace spanned by the $\hat{T}_a$, so we need to remove by hand the part of $g^{-1}Tg$ that lives in $\mathfrak{h}$. The other problem is that changing the constant modes of ξ^a should correspond to changing the symmetry-breaking vacuum $|v\rangle$ that we are expanding around, but the set of distinct vacua is not parametrized by $\mathfrak{g}/\mathfrak{h}$. Instead it is described by the quotient space G/H, meaning the group G quotiented by the equivalence relation

$$g \sim gh \tag{9.43}$$

for all $h \in H$. To see this, we note that the expectation value has the symmetry transformation

$$v_i' = \langle v|U(g)^\dagger O_i U(g)|v\rangle = D_i{}^j(g)v_j. \tag{9.44}$$

Assuming that there are no "accidental" degeneracies that do not arise from symmetry, every possible expectation value can be reached from any particular one by acting with $D(g)$ for some $g \in G$.[76] Acting with an element of H however does not change the expectation value, so the set of distinct vacua are labeled by points in G/H.

The solution to the second problem is easy to guess: we should take our Goldstone boson fields to be valued in G/H instead of $\mathfrak{g}/\mathfrak{h}$. This actually solves the first problem as well. We can abstractly write elements in this space as $[g]$, meaning the equivalence class that contains the group element G. These equivalence classes are usually called **cosets** of G by H. The action of G on the coset field follows from our vacuum transformation law (9.44), it is simply

$$U(g')^\dagger [g(x)] U(g') = [g'g(x)]. \tag{9.45}$$

Near the identity class $[e]$ we can parametrize the classes as

$$g(x) = e^{i\xi^a(x)\hat{T}_a}, \tag{9.46}$$

which gives the $\mathfrak{g}/\mathfrak{h}$ parametrization we described above. The complicated G transformation of ξ^a arises because we have

$$g' e^{i\xi^a(x)\hat{T}_a} = e^{i\xi^{a'}(x)\hat{T}_a} h(\xi, g') \tag{9.47}$$

for $h(\xi, g')$ some complicated function of ξ and g' that is discarded when we pass to the coset.

For general G and H the construction of the effective action proceeds in the following way. We first introduce a G-valued field $g(x)$. We would like to construct an action which assigns the same value to $g(x)$ and $g(x)h(x)$ for any $h(x)$, since this is an action which is well-defined on cosets. For this purposes it is convenient to introduce the Maurer-Cartan one-form

$$\omega_\mu = g^{-1}\partial_\mu g, \tag{9.48}$$

which we showed before is valued in the Lie algebra $\mathfrak{g}$. ω is automatically invariant under the left-multiplication G transformation (9.45). Under a change[77]

$$g'(x) = g(x)h(x)^{-1} \tag{9.49}$$

[76] The only theories I know with robust degenerate vacua not related by symmetry are supersymmetric. Even in these cases however the conclusion here is still correct: the Goldstone bosons propagate on G/H. There are just additional massless scalars, called moduli, whose zero modes also change which vacuum we are in.

[77] The inverse on h here is included for later convenience, as you will see in a moment. Also beware that g' now means a change of representative rather than a symmetry transformation.

of the coset representative, we have

$$\omega' = h\left(g^{-1}\partial_\mu g - h^{-1}\partial_\mu h\right)h^{-1}. \tag{9.50}$$

This is useful because we can decompose

$$g^{-1}\partial_\mu g = E_\mu + F_\mu, \tag{9.51}$$

with $E_\mu \in \mathfrak{h}$ and $F_\mu \in \mathfrak{h}^\perp$, in terms of which the change of coset representative gives

$$\begin{aligned} E'_\mu &= h(E_\mu - h^{-1}\partial_\mu h)h^{-1} \\ F'_\mu &= hF_\mu h^{-1}. \end{aligned} \tag{9.52}$$

In other words F transforms in some linear representation of H that is induced by the adjoint representation of G,[78] and E transforms like a gauge field for H! Using these transformations it is simple to write down effective actions that are G-invariant and also independent of our representative choice in G/H. For example a simple two-derivative Lagrangian for our Goldstone boson fields is

$$\mathcal{L} = f_\pi^{d-2}\mathrm{Tr}\left(F_\mu F^\mu\right), \tag{9.53}$$

where f_π is an energy scale we have included to make the units correct. In the homework you will show that near the identity you can write this term as

$$\mathcal{L} = -\frac{1}{2}f_\pi^{d-2}\hat{g}_{ab}\partial_\mu\xi^a\partial^\mu\xi^b + O(\xi^3), \tag{9.54}$$

with

$$\mathrm{Tr}\left(\hat{T}_a\hat{T}_b\right) = \frac{1}{2}\hat{g}_{ab} \tag{9.55}$$

being our usual metric on the Lie algebra (here restricted to the broken generators), so this term is a kinetic term for the Goldstone bosons together with higher order interaction terms that are needed to make the action G-invariant. In order for these fields to be canonically normalized we should define

$$\xi^a \equiv \frac{\pi^a}{f_\pi}, \tag{9.56}$$

so these interaction terms are suppressed by increasing powers of the energy scale f_π.

We can also include other light matter fields that are charged under G-symmetry. For example say that we have a field Φ that transforms in some representation

$$\Phi' = D(g')\Phi \tag{9.57}$$

of G. We can define a new "dressed" field

$$\widetilde{\Phi} \equiv D(g^{-1}(x))\Phi(x) \tag{9.58}$$

which is automatically invariant under G transformations. Unfortunately it now has a nontrivial dependence on our coset representative:

$$\widetilde{\Phi}'(x) = D(h(x))\Phi(x). \tag{9.59}$$

We are very good however at constructing actions which are invariant under such transformations: anytime we see a derivative of $\widetilde{\Phi}$ we need to turn it into a covariant deriative using a gauge field - and indeed we have a gauge field available, E_μ! We thus need to construct our Lagrangian using

$$D_\mu\widetilde{\Phi} \equiv \partial_\mu\widetilde{\Phi} - iE_\mu^a\tau_a\widetilde{\Phi}, \tag{9.60}$$

[78] Note that it is *not* the adjoint representation of H, since F lives in $\mathfrak{h}^\perp$ instead of $\mathfrak{h}$.

together with $\widetilde{\Phi}$ and F_μ, in such a way that it is invariant under the "gauge transformations" (9.52) and(9.59). Any Lagrangian so constructed will be automatically invariant under G transformations and also independent of our coset representative! In particular if Φ is a complex scalar field then we have the simple two-derivative action

$$\mathcal{L} = f_\pi^{d-2}\mathrm{Tr}\,(F_\mu F^\mu) - (D^\mu\widetilde{\Phi})^\dagger D_\mu\widetilde{\Phi} - V(|\widetilde{\Phi}|). \tag{9.61}$$

Once we rescale ξ to canonically normalize it, the Goldstone boson interactions in this Lagrangian are again all suppressed by powers of f_π.

9.5 The chiral Lagrangian

There is clearly more that we could say about the general case of the breaking of G to H, but in the special case of chiral symmetry breaking there is a trick that makes the problem substantially easier and it would be remiss of me to go further without making use of this. The trick is that the coset space G/H is actually itself equivalent to the group manifold $SU(N_f)$, so we can simply use a field $U(x)$ valued in $SU(N_f)$ as the Goldstone boson field without any further mention of quotients. To see this, we need to be a bit more careful about the global structure of the flavor symmetry group than we have been so far. There are two subtleties we need to be careful about. First of all $U(1)_A$ axial symmetry is violated by the anomaly we will discuss soon except for a discrete subgroup $\mathbb{Z}_{2N_F}$. This means that the Lie algebra of flavor symmetry is actually

$$\mathfrak{su}(N_F)_L \oplus \mathfrak{su}(N_f)_R \oplus \mathfrak{u}(1)_B. \tag{9.62}$$

This does not however mean that the flavor symmetry group is the product of these groups, and in fact it isn't. The reason is that there are elements of $SU(N_F)_L \times SU(N_F)_R$ which are shared with $U(1)_B$, and also with the discrete remnant $\mathbb{Z}_{2N_F}$ of $U(1)_A$. For example $U_L = U_R = e^{2\pi i/N_F}$ has determinant one and thus is in $SU(N_F)_L \times SU(N_F)_R$, but it is also in $U(1)_B$. The actual flavor symmetry is thus given by

$$G_F = \frac{SU(N_F)_L \times SU(N_F)_R \times U(1)_B \times \mathbb{Z}_{2N_F}}{K}, \tag{9.63}$$

where K is a discrete central subgroup that identifies these shared elements that we won't need to describe explicitly. Similarly the true remaining symmetry after chiral symmetry breaking is

$$H = \frac{SU(N_F)_I \times U(1)_B \times \mathbb{Z}_{2N_F}}{K}, \tag{9.64}$$

with K being the same discrete central subgroup in both cases. What we are really after however is the quotient space G/H, and since the discrete identifications are the same in G and H, and in particular $K \subset H$, we simply have

$$G/H = \frac{SU(N_F)_L \times SU(N_F)_R \times U(1)_B \times \mathbb{Z}_{2N_F}}{SU(N_F)_I \times U(1)_B \times \mathbb{Z}_{2N_F}} = \frac{SU(N_F)_L \times SU(N_F)_R}{SU(N_F)_I}. \tag{9.65}$$

In the second equality we used that $U(1)_B \times \mathbb{Z}_{2N_F}$ is in the center of the numerator and denominator, and thus can be canceled. This last quotient is easily identified as being equal to $SU(N_F)$ as a manifold: the correspondence is

$$[(U_L, U_R)] \leftrightarrow U_L U_R^{-1}. \tag{9.66}$$

Indeed under right multiplication by an element $U_I$of $SU(N_F)_I$ we have

$$(U_L, U_R) \to (U_L U_I, U_R U_I), \tag{9.67}$$

which does not change $U_L U_R^{-1}$, and going the other way we can simply map $U \to (U, 1)$. Thus we can take our pion field to simply be $U(x)$ valued in $SU(N_F)$. Moreover the action of G_F on U is simply

$$U'(x) = LUR^\dagger, \tag{9.68}$$

with $L \in SU(N_F)_L$ and $R \in SU(N_F)_R$ (this is a valid representation of G_F since the $U(1)_B \times \mathbb{Z}_{2N_f}$ part commutes with U, as does the central subgroup K). The leading invariant term in the effective Lagrangian for the Goldstone modes can thus be written as

$$\mathcal{L} = -f_\pi^2 \mathrm{Tr}\left(\partial_\mu U^\dagger \partial^\mu U\right), \tag{9.69}$$

which is called the **chiral Lagrangian**. If we parametrize near the identity as

$$U = e^{i\xi^a T_a}, \tag{9.70}$$

with now T_a being all the generators of $SU(N_F)$, then this becomes

$$\mathcal{L} = -f_\pi^2 \left(\frac{1}{2}\partial_\mu \xi^a \partial^\mu \xi_a + \left(\frac{1}{4}\partial_\mu(\xi^a\xi^b)\partial^\mu(\xi^c\xi^d) - \frac{1}{3}\partial_\mu \xi^a \partial^\mu(\xi^b\xi^c\xi^d)\right)\mathrm{Tr}\left(T_a T_b T_c T_d\right) + \ldots\right) \tag{9.71}$$

In particular for $N_F = 2$ matching these interactions to experiment gives[79]

$$f_\pi \approx 46\mathrm{MeV}. \tag{9.72}$$

In fact the easiest way to measure f_π is via the electroweak decays of pions, so it is usually called the **pion decay constant**. There are also higher derivative terms we can add, such as $\mathrm{Tr}\left(\partial_\mu U^\dagger \partial^\mu U \partial_\nu U^\dagger \partial^\nu U\right)$. There are no more two-derivative terms, as you can check for yourself.

You may be somewhat concerned here that f_π as we have defined it is too low of a scale to justify validity of this effective field theory as low-energy expansion. After all the pion masses are of order 150MeV which is substantially higher than f_π! The point however is that what we are really expanding in when we compute loops is

$$\frac{p^2}{32\pi^2 f_\pi^2} \approx \left(\frac{\mathrm{energy}}{820\mathrm{MeV}}\right)^2, \tag{9.73}$$

which is of order .1 for low-energy pions. The $\frac{1}{16\pi^2}$ is the usual loop factor from $\frac{4\pi^2}{(2\pi)^4}$ and the additional $1/2$ comes from our normalization convention for the T_a.

Finally we can consider how to include the couplings of pions to nucleons. One approach is to use the general coset construction of the previous section to define a dressed nucleon field $\widetilde{N}$, but then we lose the clarity of writing things in terms of U. The representation of G_F on nucleons depends strongly on N_F, so we will specialize to $N_f = 2$ for the rest of this subsection, in which case the nucleons are in the fundamental representation of $SU(2)_I$. We thus have two nucleon spinors N_i transforming under $SU(2)_L \times SU(2)_R$ in the same way as the quarks:

$$N' = LP_L N + RP_R N. \tag{9.74}$$

The leading action coupling N to U is

$$\mathcal{L} = -f_\pi^2 \mathrm{Tr}\left(\partial_\mu U^\dagger \partial^\mu U\right) - i\overline{N}\partial\!\!\!/ N - im_N \overline{N}\left(U^\dagger P_L + U P_R\right) N - \frac{i}{2}(1 - g_A)\overline{N}\gamma^\mu\left(U\partial_\mu U^\dagger P_L + U^\dagger \partial_\mu U P_R\right) N, \tag{9.75}$$

where the relative sign in the second term comes from hermiticity and the relative sign in the third term comes from imposing parity symmetry. g_A is called the **axial vector coupling**, we already met it back in our first section where we saw it can be measured using the lifetime of the neutron. Empirically

$$g_A \approx 1.3. \tag{9.76}$$

Expanding out this Lagrangian to leading nontrivial order in ξ gives an interaction

$$\mathcal{L} \supset \frac{g_A}{2}\partial_\mu \xi^a \overline{N}\gamma^\mu \gamma T_a N. \tag{9.77}$$

[79] Unfortunately there are many different conventions for how to define f_π. So far I have seen values which are mine multipled by 2, $2\sqrt{2}$, and 4. Mine introduces the fewest factors of 2 into the Lagrangian.

Integrating by parts and using the Dirac equation $\not{\partial} N = -m_N N$, and also rescaling ξ, we find a Yukawa type interaction

$$\mathcal{L} \supset -\frac{g_A m_N}{2f_\pi} \pi^a \overline{N} T_a \gamma N, \tag{9.78}$$

so the pion/nucleon Yukawa coupling is given by[80]

$$g_{\pi \overline{N} N} = \frac{g_A m_N}{2f_\pi} \approx 13.3, \tag{9.79}$$

which is called the **Goldberger-Treiman relation**. Experimentally it holds to about 5% accuracy, and historically it was a key indication that pions are likely Goldstone bosons.

9.6 Goldstone masses

We can use this formalism to get some estimates for the pion, kaon, and eta masses in terms of QCD parameters. To do this we need to consider how the chiral Lagrangian gets modified in the presence of a quark mass term

$$\mathcal{L} \supset -i\overline{q} \left(M P_L + M^\dagger P_R \right) q, \tag{9.80}$$

where we have allowed M to be an arbitrary $N_f \times N_F$ matrix on flavor indices (the fact that we have $M^\dagger$ in the second term follows from hermiticity). We observe that an $SU(N_F)_L \times SU(N_F)_R$ transformation (L, R) on the quarks can be canceled by a transformation

$$M' = RML^\dagger \tag{9.81}$$

of the mass matrix, so the partition function is invariant under this transformation of the mass matrix. This must therefore also be true for the effects of M on the low-energy effective action, which means that the leading effect of the quark masses on the chiral Lagrangian is a linear term

$$\mathcal{L} \supset B \mathrm{Tr} \left(UM + M^\dagger U^\dagger \right), \tag{9.82}$$

since this is hermitian and invariant under (9.68) and (9.81). We match this M dependence between the UV and IR theories by equating the contribution to the partition function that is linear in M in the two theories:

$$\langle \overline{q} \left(M P_L + M^\dagger P_R \right) q \rangle = iB \langle \mathrm{Tr} \left(UM + M^\dagger U^\dagger \right) \rangle. \tag{9.83}$$

We can compute the left-hand side using the chiral condensate (9.20), and on the right-hand side the vacuum is $\langle U \rangle = 1$ since we are expanding about $\xi^a = 0$. Thus we have

$$iv \mathrm{Tr} \left(M + M^\dagger \right) = iB \mathrm{Tr} \left(M + M^\dagger \right), \tag{9.84}$$

and thus

$$B = v. \tag{9.85}$$

We can Taylor expand the term (9.82) to second order in ξ to get a mass term, setting $M = M^\dagger$ we have

$$\mathcal{L} \supset -v \xi^a \xi^b \mathrm{Tr} \left(T_a T_b M \right). \tag{9.86}$$

Considering first the case $N_f = 2$, and taking

$$M = \begin{pmatrix} m_u & 0 \\ 0 & m_d \end{pmatrix}, \tag{9.87}$$

[80] You might worry that this Yukawa coupling is too large to justify perturbation theory in the usual Yukawa sense we discussed last semester. Indeed it is, but our real expansion parameter is (9.73). The tree-level Yukawa calculation still gives a good explanation of the nucleon potential however, as the momentum transfer is quite small.

we can use the $SU(2)$ fact that

$$\{T_a, T_b\} = \frac{1}{2}\delta_{ab} \tag{9.88}$$

to get

$$\mathcal{L} \supset -\frac{1}{4} v \xi^a \xi_a (m_u + m_d). \tag{9.89}$$

Finally setting $\xi^a = \pi^a / f_\pi$ to go to canonical normalization, we get

$$m_\pi^2 = \frac{v(m_u + m_d)}{2 f_\pi^2}, \tag{9.90}$$

which is called the **Gell-Mann/Oakes/Renner relation**. Using the observed values of m_π, m_u, m_d, and f_π we get

$$v \approx (225\text{MeV})^3 . \tag{9.91}$$

Finally we can do a similar analysis for $N_f = 3$: choosing generators

$$\begin{aligned}
T_1 &= \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & -1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad T_2 = \begin{pmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad T_3 = \begin{pmatrix} 0 & -i/2 & 0 \\ i/2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
T_4 &= \begin{pmatrix} 0 & 0 & 1/2 \\ 0 & 0 & 0 \\ 1/2 & 0 & 0 \end{pmatrix} \qquad T_5 = \begin{pmatrix} 0 & 0 & -i/2 \\ 0 & 0 & 0 \\ i/2 & 0 & 0 \end{pmatrix} \qquad T_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \\
T_7 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i/2 \\ 0 & i/2 & 0 \end{pmatrix} \qquad T_8 = \begin{pmatrix} 1/\sqrt{12} & 0 & 0 \\ 0 & 1/\sqrt{12} & 0 \\ 0 & 0 & -1/\sqrt{3} \end{pmatrix}
\end{aligned} \tag{9.92}$$

and mass matrix

$$M = \begin{pmatrix} m_u & 0 & 0 \\ 0 & m_d & 0 \\ 0 & 0 & m_s \end{pmatrix}, \tag{9.93}$$

and recognizing T_1 as π_0, T_2, T_3 as $\pi_\pm$, T_3, T_4 as $K_\pm$, T_5, T_6 as K_0 and $\overline{K}_0$, and T_8 as η, we have the predictions[81]

$$\begin{aligned}
m_\pi^2 &= \frac{v(m_u + m_d)}{2 f_\pi^2} \approx (138\text{MeV})^2 \\
m_{K_\pm}^2 &= \frac{v(m_s + m_u)}{2 f_\pi^2} \approx (512\text{MeV})^2 \\
m_{K_0}^2 &= \frac{v(m_s + m_d)}{2 f_\pi^2} \approx (518\text{MeV})^2 \\
m_\eta^2 &= \frac{v(4m_s + m_u + m_d)}{6 f \pi^2} \approx (589\text{MeV})^2.
\end{aligned} \tag{9.94}$$

Comparing to our table from the first section these are actually pretty good, although we shouldn't get credit for the pion mass since we used the experimental value it to determine v and we neglected electromagnetism (as well as the mixing between π_0 and η and also η and η'). Historically these formulas were of course used in the opposite direction, to learn the quark masses from the observed hadron spectrum.

[81] There is also a mass mixing between the π_0 and the η, which is of order $v(m_u - m_d)/f_\pi^2$. The true mass eigenstates are thus superpositions of these

9.7 Wess-Zumino-Witten term

There is actually another term in the chiral Lagrangian which I have not explained, which is very important in understanding the relationship between pions and anomalies. It also connects to some beautiful topics in algebraic topology. I will explain it at the end of the semester once we have enough mathematical machinery.

9.8 Homework

1. In this problem we will understand the baryon decuplet.

   (a) Argue that the tensor product of three $SU(3)$ fundamentals decomposes into $SU(3)$ irreps as

   $$3 \otimes 3 \otimes 3 = 10 \oplus 8 \oplus 8 \oplus 1, \tag{9.95}$$

   where the two 8s are adjoints and the 10 is a symmetric three-index tensor. Hint: a general three index tensor T_{ijk} transforms in the product representation, and you can decompose it into irreps by contracting with ϵ^{ijk} in various ways.

   (b) Decompose the 10 of $SU(3)$ into $SU(2)$ irreducibles. Do these fit the degeneracy pattern of the spin 3/2 baryons?

2. Argue that $i\not{D}$ is hermitian in Euclidean signature.

3. Consider the case of Goldstone bosons where $U(1)_L \times U(1)_R$ is spontaneously broken to the diagonal $U(1)$ subgroup $(e^{i\theta}, e^{i\theta})$. Writing a general element of G as $(e^{i\theta_L}, e^{i\theta_R})$, what are the quantities E_μ and F_μ? How does G symmetry act on θ_L and θ_R? Let Φ be a field that transforms with charges (q_L, q_R) under G. What is $D_\mu\widetilde{\Phi}$? Write out the quantities $\mathrm{Tr}\,(F_\mu F^\mu)$ and $D_\mu\widetilde{\Phi}^\dagger D^\mu\Phi$ in terms of θ_L, θ_R, and Φ and check that they are invariant under G symmetry.

4. Confirm the Goldstone mass calculations (9.94).

10 Electroweak theory and the standard model

Having finally finished with QCD, it is time for us to to understand the weak force responsible for the β-decay of the neutron. I'll remind you of several features of this theory that we inferred back in the first section:

- The weak force is mediated by three massive vector bosons, the $W_\pm$ and Z. The $W_\pm$ are charged under electromagnetism, while the Z is neutral.
- Interaction with the $W_\pm$ must be able to turn one quark flavor into another, and also electrons/muons/taus into neutrinos.
- The neutrinos must be very light or massless, and they should not interact with electromagnetism or the strong force
- $\mathcal{C}$, $\mathcal{R}$, and $\mathcal{T}$ symmetry should all be violated, with the $\mathcal{R}$ and $\mathcal{C}$ violation being the strongest since the $W_\pm$ and Z bosons only couple to left-handed quarks and leptons.

In this section we will construct a quantum field theory, the **standard model of particle physics**, that meets all of these requirements and also incorporates QCD.

10.1 Gauge sector

We will begin with the gauge sector. In addition to the $SU(3)$ gauge fields for QCD, we need four more gauge generators to account for the $W_\pm$, the Z, and the photon. At least two of them must be non-abelian in order for the $W_\pm$ to be able to change flavor, so the only possible option for the gauge group is

$$(SU(3) \times SU(2) \times U(1)_Y)/K. \tag{10.1}$$

Here K is a discrete central subgroup that we will ignore for now (eventually we will argue that $K = \mathbb{Z}_6$ is the most natural choice), and $U(1)_Y$ is conventionally normalized so that its charge is quantized in units of $1/6$. Since the $W_\pm$ and Z are all heavy we clearly are going to want to Higgs some of this gauge group. The most naive guess is that we should just Higgs $SU(2)$ to nothing, leaving $U(1)_Y$ to be the electromagnetic gauge group, but this possibility is excluded since the $W_\pm$ carry electric charge. Therefore we need to Higgs in such a way that the unbroken $U(1)_E$ subgroup for electromagnetism is a mixture of the fundamental $SU(2)$ and $U(1)$:

$$SU(2) \times U(1)_Y \to U(1)_E. \tag{10.2}$$

The charge associated to $U(1)_Y$ is called **hypercharge**, while the charge associated to $U(1)_E$ is of course just ordinary electric charge. The Higgs field ϕ that causes this breaking must thus be charged under both $SU(2)$ and $U(1)_Y$. The simplest guess is to take ϕ to transform in the fundamental of $SU(2)$, and we will take its hypercharge to be $-1/2$ for reasons that will be clear in a moment. The $SU(3)$ of QCD should be unbroken, so ϕ must be a color singlet. These gauge charge assignments are usually summarized by saying the Higgs representation is $(1, 2, -\frac{1}{2})$. We can thus write the Lagrangian for the gauge and Higgs sector of the standard model as[82]

$$\mathcal{L}_{gauge} = -\frac{1}{2g_3^2}\mathrm{Tr}\,(G_{\mu\nu}G^{\mu\nu}) - \frac{1}{2g_2^2}\mathrm{Tr}\,(F_{\mu\nu}F^{\mu\nu}) - \frac{1}{4g_1^2}B_{\mu\nu}B^{\mu\nu} - (D_\mu\phi)^\dagger D^\mu\phi - \frac{\lambda}{4}\left(\phi^\dagger\phi - \frac{1}{2}v^2\right)^2. \tag{10.3}$$

Here

$$G_{\mu\nu} = \partial_\mu G_\mu - \partial_\nu G_\mu - i[G_\mu, G_\nu] \tag{10.4}$$

is the field strength for the $SU(3)$ gauge field G_μ of QCD,

$$F_{\mu\nu} = \partial_\mu A_\mu - \partial_\nu A_\mu - i[A_\mu, A_\nu] \tag{10.5}$$

[82] In this section we will follow standard practice and capitalize the gauge fields while writing matter fields in lower case. This is not consistent with our usual rule of writing classical fields in lower case and field operators in upper case, sorry.

is the field strength for the $SU(2)$ gauge field A_μ, and

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \tag{10.6}$$

is the hypercharge field strength. The covariant derivative on ϕ is

$$D_\mu \phi = \partial_\mu \phi - iT_a A^a_\mu \phi + \frac{i}{2} B_\mu \phi, \tag{10.7}$$

where $T_a = \sigma_a/2$ are the generators of $SU(2)$.

To identify the physical degrees of freedom in the gauge sector we can go to unitarity gauge. We leave a general discussion of the definition of this gauge to the homework, here only noting that it amounts to choosing ϕ to have the form

$$\phi(x) \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} v + h(x) \\ 0 \end{pmatrix} \tag{10.8}$$

with $v + h > 0$ and $\langle h(x) \rangle = 0$. This does not completely fix the $SU(2) \times U(1)_Y$ gauge symmetry however, since we can have

$$e^{i\theta^3 T_3} e^{-i\omega/2} \langle \phi \rangle = \begin{pmatrix} e^{i\theta^3/2} & 0 \\ 0 & e^{-i\theta^3/2} \end{pmatrix} \begin{pmatrix} e^{-i\omega/2} & 0 \\ 0 & e^{-i\omega/2} \end{pmatrix} \begin{pmatrix} v/\sqrt{2} \\ 0 \end{pmatrix} = \begin{pmatrix} v/\sqrt{2} \\ 0 \end{pmatrix} \tag{10.9}$$

if

$$\theta^3 = \omega \mod 4\pi. \tag{10.10}$$

The periodicity of θ here is 4π and the periodicity of ω is 12π (the latter is because we choose the minimal hypercharge to be $1/6$), so this describes a $U(1)$ subgroup of $SU(2) \times U(1)_Y$ that wraps $SU(2)$ three times for each time it wraps $U(1)_Y$. We can take the generator of this subgroup to be[83]

$$Q = T_3 + Y, \tag{10.11}$$

where Y is the hypercharge generator. We will see in a moment that Q is electric charge. The transformation of the gauge field under this unbroken subgroup is given by

$$A^{a\prime}_\mu T_a + B'_\mu = e^{i\Omega T_3} A^a_\mu T_a e^{-i\Omega T_3} + B_\mu + (1 + T_3)\partial_\mu \Omega, \tag{10.12}$$

which motivates us to introduce a change of variables on the gauge fields,

$$\begin{aligned} W_\mu &\equiv \frac{1}{\sqrt{2g_2^2}} \left(A^1_\mu - iA^2_\mu \right) \\ A_\mu &\equiv \frac{1}{\sqrt{g_1^2 + g_2^2}} \left(\frac{g_1}{g_2} A^3_\mu + \frac{g_2}{g_1} B_\mu \right) \\ Z_\mu &\equiv \frac{1}{\sqrt{g_1^2 + g_2^2}} \left(A^3_\mu - B_\mu \right). \end{aligned} \tag{10.13}$$

The reason for the factors of g_1/g_2 and g_2/g_1 in the expression for A_μ, as well as the overall factors in front, is that we want the transformation from A^3_μ and B_μ to A_μ and Z_μ to be an orthogonal transformation when expressed in terms of the canonically normalized fields $\hat{A}^a_\mu = A^a_\mu/g_2$ and $\hat{B}_\mu = B_\mu/g_1$, since this will ensure that the new fields are also canonically normalized. Using these new fields we have the gauge transformations

$$\begin{aligned} W'_\mu &= e^{i\Omega} W_\mu \\ Z'_\mu &= Z_\mu \\ A'_\mu &= A_\mu + \frac{\sqrt{g_1^2 + g_2^2}}{g_1 g_2} \partial_\mu \Omega. \end{aligned} \tag{10.14}$$

[83] This formula is the reason for the somewhat bizarre normalization of hypercharge; if we set the minimal hypercharge to one then we would have $6Y$ here. I actually think that would be better, since it would make the appeal of the $\mathbb{Z}_6$ quotient in the gauge group more obvious, but it is too widely accepted to change it.

Thus we can identify A_μ as the electromagnetic gauge field, and the $W^\pm$ that are annihilated/created by the complex field W_μ indeed have charge ± 1. Since A_μ is canonically normalized we can identify the electromagnetic gauge coupling from its gauge transformation as

$$e = \frac{g_1 g_2}{\sqrt{g_1^2 + g_2^2}}. \tag{10.15}$$

The mass term for these gauge bosons is

$$\begin{aligned}\mathcal{L} &\supset -\frac{v^2}{8}\begin{pmatrix}1 & 0\end{pmatrix}\begin{pmatrix}A_\mu^3 - B_\mu & A_\mu^1 - iA_\mu^2 \\ A_\mu^1 + iA_\mu^2 & -A_\mu^3 - B_\mu\end{pmatrix}\begin{pmatrix}A^{\mu,3} - B^\mu & A^{\mu,1} - iA^{\mu,2} \\ A^{\mu,1} + iA^{\mu,2} & -A^{\mu,3} - B^\mu\end{pmatrix}\begin{pmatrix}1 \\ 0\end{pmatrix} \\ &= -m_W^2 W_\mu^\dagger W^\mu - \frac{1}{2}m_Z^2 Z_\mu Z^\mu,\end{aligned} \tag{10.16}$$

with

$$\begin{aligned}m_W &= \frac{v}{2}g_2 \\ m_Z &= \frac{v}{2}\sqrt{g_1^2 + g_2^2}.\end{aligned} \tag{10.17}$$

Thus we see that indeed the W and Z are massive, while the photon remains massless as required by its unbroken $U(1)_E$ gauge invariance. These low-energy parameters are typically expressed in terms the $U(1)_E$ gauge coupling e and the Weinberg mixing angle

$$\begin{aligned}g_1 &= \sqrt{g_1^2 + g_2^2}\sin\theta_W \\ g_2 &= \sqrt{g_1^2 + g_2^2}\cos\theta_W,\end{aligned} \tag{10.18}$$

which is also the angle in the orthogonal transformation between $(\hat{A}_\mu^3, \hat{B}_\mu)$ and (Z_μ, A_μ). In terms of these we have

$$\begin{aligned}e &= g_2 \sin\theta_W \\ m_W &= \frac{ve}{2\sin\theta_W} \\ m_Z &= \frac{m_W}{\cos\theta_W}.\end{aligned} \tag{10.19}$$

Observationally we have $m_W \approx 80.4\,\text{GeV}$, $m_Z \approx 91.2\,\text{GeV}$, and $e \approx .303$, from which we see that[84]

$$\begin{aligned}v &\approx 250\,\text{GeV} \\ \sin\theta_W &\approx .472 \\ g_2 &\approx .642.\end{aligned} \tag{10.20}$$

In degrees the Weinberg angle is about $28°$, so the photon is more $\hat{B}_\mu$ than it is $\hat{A}_\mu^3$. We can also look at the Higgs potential, this is given by

$$V(h) = \frac{\lambda v^2}{4}h^2 + \frac{\lambda v}{4}h^3 + \frac{\lambda}{16}h^4. \tag{10.21}$$

In particular the Higgs mass is given by

$$m_H = \sqrt{\frac{\lambda v^2}{2}}. \tag{10.22}$$

[84]These results are scheme-dependent in detail, for example we should really use e at m_Z instead of m_e, which is more like .313, and in $\overline{\text{MS}}$ the Weinberg angle is defined so that $\sin\theta_W \approx .481$. With these modifications, we instead get $v \approx 247\,\text{GeV}$ and $g_2 \approx .651$.

The Higgs boson was finally discovered at the LHC in 2013, with $m_H \approx 125\text{GeV}$. This therefore predicts a Higgs self-coupling

$$\lambda \approx .25, \tag{10.23}$$

although this has so far not been measured directly (doing this is one of the main long-term goals of the LHC, and perhaps a future collider as well).

The full gauge and Higgs Lagrangian expressed in terms of W_μ, Z_μ, A_μ, and h is not so pleasant. It is convenient to first define the Abelian field strengths

$$\begin{aligned} F_{\mu\nu} &= \partial_\mu A_\nu - \partial_\nu A_\mu \\ Z_{\mu\nu} &= \partial_\mu Z_\nu - \partial_\nu Z_\mu, \end{aligned} \tag{10.24}$$

as well as

$$W_{\mu\nu} = D_\mu W_\nu - D_\nu W_\mu \tag{10.25}$$

with

$$D_\mu W_\nu = (\partial_\mu - ieA_\mu) W_\nu. \tag{10.26}$$

We then have

$$\begin{aligned} \frac{1}{\sqrt{2g_2^2}} \left(F^1_{\mu\nu} - iF^2_{\mu\nu}\right) &= W_{\mu\nu} - ie\cot\theta_W \left(Z_\mu W_\nu - Z_\nu W_\mu\right) \\ F^3_{\mu\nu} &= e\left(F_{\mu\nu} + \cot\theta_W Z_{\mu\nu}\right) - \frac{ie^2}{\sin^2\theta_W}\left(W^\dagger_\mu W_\nu - W^\dagger_\nu W_\mu\right) \\ B_{\mu\nu} &= g_1 \cos\theta_W F_{\mu\nu} - g_1 \sin\theta_W Z_{\mu\nu}, \end{aligned} \tag{10.27}$$

from which it isn't too hard to crank out the Lagrangian:

$$\begin{aligned} \mathcal{L}_{gauge} = &-\frac{1}{2g_3^2}\text{Tr}\left(G_{\mu\nu}G^{\mu\nu}\right) - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - \frac{1}{4}Z_{\mu\nu}Z^{\mu\nu} \\ &- \frac{1}{2}\Big(W_{\mu\nu} - ie\cot\theta_W\left(Z_\mu W_\nu - Z_\nu W_\mu\right)\Big)^\dagger \Big(W^{\mu\nu} - ie\cot\theta_W\left(Z^\mu W^\nu - Z^\nu W^\mu\right)\Big) \\ &+ ie\left(F^{\mu\nu} + \cot\theta_W Z^{\mu\nu}\right) W^\dagger_\mu W_\nu - \frac{e^2}{2}\sin^2\theta_W \left(W^\dagger_\mu W^\mu W^\dagger_\nu W^\nu - W^\dagger_\mu W^{\mu\dagger} W_\nu W^\nu\right) \\ &- \left(m_W^2 W^\dagger_\mu W^\mu + \frac{1}{2}m_Z^2 Z_\mu Z^\mu\right)\left(1 + \frac{h}{v}\right)^2 \\ &- \frac{1}{2}\partial_\mu h \partial^\mu h - V(h). \end{aligned} \tag{10.28}$$

As you can see, there are a large number of interaction terms involving the Higgs, the W, and the Z - computing higher loop amplitudes in the standard model is a full-time occupation!

10.2 Quark and lepton fields

We now discussed the matter fields, i.e. the quarks and leptons. Here there is a qualitatively new feature: the couplings of quark and lepton fields to $SU(2)\times U(1)_Y$ depends on their helicity, which is usually described by saying that the standard model is a **chiral gauge theory**. There are five types of matter field, which have the cuddly names of $q_L, u_R, d_R, \ell_L, e_R$. The first three are quark fields, which transform in the fundamental representation of the QCD $SU(3)$ gauge group, while the last two are lepton fields that are color singlets. In our notation all of these fields are four-component spinors, with subscripts indicating which Weyl component is nonzero (so e.g. the lower two spinor components of q_L vanish). The right-handed spinors are all singlets under the $SU(2)$ gauge group, while the left-handed spinors transform in the fundamental. There are three copies of each of these fields, so we can write them as $(q^i_L, u^i_R, d^i_R, \ell^i_L, e^i_R)$, with $i = 1, 2, 3$. The fields for

each i are called a **generation**, so the standard model has three generations of quarks and leptons. We can decompose the three generations into the familiar quarks and leptons as

$$\begin{aligned} q_L^1 &= \begin{pmatrix} u_L \\ d_L \end{pmatrix} \qquad u_R^1 = u_R \qquad d_R^1 = d_R \qquad \ell_L^1 = \begin{pmatrix} \nu_e \\ e_L \end{pmatrix} \qquad e_R^1 = e_R \\ q_L^2 &= \begin{pmatrix} c_L \\ s_L \end{pmatrix} \qquad u_R^2 = c_R \qquad d_R^2 = s_R \qquad \ell_L^2 = \begin{pmatrix} \nu_\mu \\ \mu_L \end{pmatrix} \qquad e_R^2 = \mu_R \\ q_L^3 &= \begin{pmatrix} t_L \\ b_L \end{pmatrix} \qquad u_R^3 = t_R \qquad d_R^3 = b_R \qquad \ell_L^3 = \begin{pmatrix} \nu_\tau \\ \tau_L \end{pmatrix} \qquad e_R^3 = \tau_R \end{aligned} \tag{10.29}$$

Using (10.11) we can then infer the hypercharge assignments of these fields from the known electric charges of these particles, leading to the following table of gauge and baryon/lepton charges for the standard model matter fields:

	$SU(3)$	$SU(2)$	$U(1)_Y$	B	L
q_L^i	3	2	$\frac{1}{6}$	$\frac{1}{3}$	0
u_R^i	3	1	$\frac{2}{3}$	$\frac{1}{3}$	0
d_R^i	3	1	$-\frac{1}{3}$	$\frac{1}{3}$	0
ℓ_L^i	1	2	$-\frac{1}{2}$	0	1
e_R^i	1	1	-1	0	1

With these gauge charge assignments, we can write the fermion kinetic terms as

$$\mathcal{L}_{matter} = -i\overline{q}_{L,i}\not{D}q_L^i - i\overline{u}_{R,i}\not{D}u_R^i - i\overline{d}_{R,i}\not{D}d_R^i - i\overline{\ell}_{L,i}\not{D}\ell_L^i - i\overline{e}_{R,i}\not{D}e_R^i, \tag{10.30}$$

where we have implicitly summed over generation labels as well as spin and gauge indices.

To see the interactions of these quarks and leptons with the standard model gauge bosons we need to expand out the covariant derivatives. The interactions with the W come from the covariant derivative on an $SU(2)$ fundamental via

$$D_\mu \supset \partial_\mu - \frac{ig_2}{\sqrt{2}} \begin{pmatrix} 0 & W_\mu \\ W_\mu^\dagger & 0 \end{pmatrix}, \tag{10.31}$$

while the photon interactions are the usual QED ones determined by the electric charge. The interactions with the Z boson are trickier since they depend on θ_W, and in particular the Z boson can talk to right-handed quarks and leptons since it mixes with the B_μ. One way to describe it is to say that the Lagrangian includes an interaction

$$\mathcal{L} \supset Z_\mu J_Z^\mu, \tag{10.32}$$

where

$$\begin{aligned} J_Z^\mu \equiv -e\Big[&\left(\frac{1}{2}\cot\theta_W - \frac{1}{6}\tan\theta_W\right)\overline{u}_L\gamma^\mu u_L - \left(\frac{1}{2}\cot\theta_W + \frac{1}{6}\tan\theta_W\right)\overline{d}_L\gamma^\mu d_L - e\tan\theta_W\left(\frac{2}{3}\overline{u}_R\gamma^\mu u_R - \frac{1}{3}\overline{d}_R\gamma^\mu d_R\right) \\ &+ \frac{1}{2}\left(\cot\theta_W + \tan\theta_W\right)\overline{\nu}_L\gamma^\mu\nu_L + \frac{1}{2}\left(\tan\theta_W - \cot\theta_W\right)\overline{e}_L\gamma^\mu e_L + \tan\theta_W\overline{e}_R\gamma^\mu e_R\Big] \end{aligned} \tag{10.33}$$

is called the **electroweak neutral current** (in this expression we have implicitly summed over color and generation number).

We can also use these charge assignments to comment on the choice of discrete central quotient K in the standard model gauge group. What we will show is that there is a discrete central subgroup of $SU(3) \times SU(2) \times U(1)$ that acts trivially on all matter fields. Indeed any central element has the form

$$g = (e^{2\pi i n_3/3}, e^{\pi i n_2}, e^{i\omega/6}), \tag{10.34}$$

with $n_3 \in \{0,1,2\}$, $n_2 \in \{0,1\}$, and $\omega \in [0, 12\pi)$. In the homework you will show that in order for this group element to act trivially on all quark and lepton fields (and also the Higgs, which has the same gauge representation as ℓ_L^i), it must live in a $\mathbb{Z}_6$ subgroup of $SU(3) \times SU(2) \times U(1)$ that is generated by

$$g = (e^{2\pi i/3}, -1, e^{i\pi/3}). \tag{10.35}$$

Therefore the most minimal choice of the Standard Model gauge group is

$$G = \frac{SU(3) \times SU(2) \times U(1)_Y}{\mathbb{Z}_6}. \tag{10.36}$$

It must be acknowledged however that at present we are not able to test experimentally whether or not this quotient is present; to do so we would need to either create some kind of magnetic monopole, which could be quite heavy compared to the reach of current colliders, or else observe some kind of quantum gravity fluctuation that changes the topology of the universe. On the other hand we could much more easily rule out the quotient by discovering a new matter field that transforms nontrivially under $\mathbb{Z}_6$. Why then should we provisionally adopt the hypothesis that the quotient is present? For the same reason that we say the electromagnetic gauge group is $U(1)$ instead of $\mathbb{R}$: either could be true, but if it is the latter then the observed quantization of charge is an accident. Similarly if the standard model gauge group does not include the quotient, then it is a coincidence that all the matter fields are invariant under it and unnecessary coincidences are undesirable.

10.3 Yukawa sector

The matter fields we have constructed so far are all massless, which is certainly not what is observed. At first this may seem like a serious problem, since adding a direct mass term like $\overline{u}_R u_L$ is forbidden by $SU(2)$ gauge invariance. Fortunately however we can get masses that are gauge-invariant by using Yukawa terms that couple the quark and lepton fields to the Higgs field: the gauge transformation of the Higgs field can soak up the $SU(2)$ charge, and then since this field has a vev we can still get a mass term. The basic Yukawa terms we can write down have the form

$$\overline{q}_L \phi u_R \qquad \overline{q}_L \widetilde{\phi}^* d_R \qquad \overline{\ell}_L \widetilde{\phi}^* e_R. \tag{10.37}$$

In the first of these the spinor and color indices are contracted between $\overline{q}_L$ and u_R, while the $SU(2)$ indices are contracted between the antifundamental $\overline{q}_L$ and the fundamental ϕ. It is easy to check that the hypercharges add up to zero. To construct the second and third terms we need to remember that we can convert an $SU(2)$ fundamental into an $SU(2)$ antifundamental using the ϵ tensor. This follows from the $SU(2)$-invariance

$$D_{n'}{}^{n}(g) D_{m'}{}^{m}(g) \epsilon^{n'm'} = \epsilon^{nm} \tag{10.38}$$

of the ϵ tensor. Indeed note that if

$$\phi'_n = D_n{}^m(g) \phi_m, \tag{10.39}$$

then defining

$$\widetilde{\phi}^n \equiv \epsilon^{nm} \phi_m \tag{10.40}$$

we have

$$\begin{aligned} \widetilde{\phi}^{n\prime} &= \epsilon^{nm} D_m{}^{m'}(g) \phi_{m'} \\ &= D_\ell{}^n(g^{-1}) D_k{}^\ell(g) \epsilon^{km} D_m{}^{m'}(g) \phi_{m'} \\ &= D^{*n}{}_m(g) \widetilde{\phi}^m. \end{aligned} \tag{10.41}$$

Thus $\widetilde{\phi}$ is an antifundamental. Taking the complex conjugate then brings us back to a fundamental:

$$\widetilde{\phi}^{*\prime}_n = D_n{}^m(g) \widetilde{\phi}^*_m. \tag{10.42}$$

In the second and third Yukawa terms we have thus contracted the antifundamental $SU(2)$ index on $\overline{q}_L$ or $\overline{\ell}_L$ with the fundamental index on $\widetilde{\phi}^*$. What this trick buys us is that the hypercharges still add up to zero. Note that there is no Yukawa term involving ϕ for leptons: this is because in the standard model we do not have a right-handed neutrino field ν_R.

So far we did not write down the couplings for these Yukawa terms. In general there is no reason for these couplings to be diagonal in the generation index i, so we in general have

$$\mathcal{L}_{Yukawa} = -iQ^i{}_j \overline{q}_{L,i} \phi u_R^j - i\widetilde{Q}^i{}_j \overline{q}_{L,i} \widetilde{\phi}^* d_R^j - iL^i{}_j \overline{\ell}_{L,i} \widetilde{\phi}^* e_R^j + \text{h.c.} \tag{10.43}$$

On the other hand, there is nothing to stop is from trying to remove some of these couplings by redefining each matter field by a unitary transformation on the generation index. It is a general theorem (called the **singular value decomposition**) that any matrix has the form

$$M = VDU, \tag{10.44}$$

with D diagonal with non-negative entries and U and V unitary, so applying this Q by absorbing U into u_R and V into q_L we can diagonalize Q. Similarly we can diagonalize L by doing a redefinition of ℓ_L and e_R. The diagonal elements determine the masses of the the electron, muon, τ, up, charm, and strange quarks once we set ϕ to its expectation value. We do not however have enough remaining freedom to also diagonalize $\widetilde{Q}$: we can redefine d_R to remove U, but there is a remaining unitary transformation V that cannot be removed since we have already redefined q_L to diagonalize $\widetilde{Q}$. We thus arrive at the Yukawa Lagrangian

$$\mathcal{L}_{Yukawa} = -i\sum_i y_i^u \overline{q}_L^i \phi u_R^i - i\sum_i y_i^d \overline{\widetilde{q}}_{L,i} \widetilde{\phi}^* d_R^i - i\sum_i y_i^e \overline{\ell}_{L,i} \widetilde{\phi}^* e_R^i + \text{h.c.}, \tag{10.45}$$

where

$$\widetilde{q}_L \equiv V_{CKM}^\dagger q_L \tag{10.46}$$

with V_{CKM} a unitary matrix that is called the the **Cabbibo-Kobayashi-Maskawa matrix**.[85] The CKM matrix is responsible for some of the most interesting features of the standard model. Returning first to the quark and lepton masses however, substituting in the Higgs expectation value we have

$$\begin{aligned} m_i^u &= \frac{y_i^u v}{\sqrt{2}} \\ m_i^d &= \frac{y_i^d v}{\sqrt{2}} \\ m_i^e &= \frac{y_i^e v}{\sqrt{2}}. \end{aligned} \tag{10.47}$$

In particular given the top quark mass $m_3^u \approx 173\,\text{GeV}$ we have $y_3^u \approx 1$, which is already rather large (but still under perturbative control due to loop factors like $\frac{1}{16\pi^2}$).

One way to think about the CKM matrix is that it describes a tension between two identities of quark fields: they would like to fit into nice $SU(2)$ doublets as $\begin{pmatrix} u \\ d \end{pmatrix}$, but they would also like to be eigenstates of the mass matrix. We can pull off both of these for one of them but not the other, and we have arbitrarily chosen the down-type quarks to be the ones that have clashing identities. It is up to us which identity to manifestly express in any particular equation. For example we could choose to use $\widetilde{q}_L$ in the bottom half of the $SU(2)$ doublet and q_L in the top, in which case the quark doublets become

$$\begin{pmatrix} u_L \\ V_{11}d_L + V_{12}s_L + V_{13}b_L \end{pmatrix} \qquad \begin{pmatrix} c_L \\ V_{21}d_L + V_{22}s_L + V_{23}b_L \end{pmatrix} \qquad \begin{pmatrix} t_L \\ V_{31}d_L + V_{32}s_L + V_{33}b_L \end{pmatrix} \tag{10.48}$$

[85] In another Nobel scandal the prize was given to Kobayashi and Maskawa but not Cabbibo, even though he introduced this matrix and explained its importance in the context of two generations and Kobayashi and Maskawa merely extended it to three.

with a diagonal mass matrix.

Perhaps the most important feature of the CKM matrix is that it is the only confirmed source of violation of $\mathcal{T}$ symmetry in the standard model.[86] We showed earlier that Yukawa theory is invariant under $\mathcal{T}$ symmetry, but it was crucial in that argument that the Yukawa coupling is real. This is because time-reversal is an antiunitary symmetry, so it takes the complex conjugate of any number appearing in the Lagrangian. In the Yukawa theory there is only one fermion field, so the reality of the coupling is a consequence of the hermiticity of the action. In the standard model however the CKM matrix mixes quarks of different generations, so it is only required to be unitary. This introduces phases into the Yukawa Lagrangian that will violate $\mathcal{T}$ symmetry unless they can be removed by field redefinitions. We have already done quite a few field redefinitions to diagonalize the other two Yukawa terms with positive y_i, so the only remaining field redefinitions available are the rephasings

$$\begin{aligned} q_L^{i\prime} &= e^{i\theta_i} q_L^i \\ u_R^{i\prime} &= e^{i\theta_i} u_R^i \\ d_R^{i\prime} &= e^{i\phi_i} d_R^i. \end{aligned} \tag{10.49}$$

This looks like six rephasings, but actually if we choose all of the θ_i and ϕ_i to be equal that does not change V_{CKM} so we have only five phase redefinitions available. A unitary matrix with no phases is an orthogonal matrix, and the space of orthogonal matrices has dimension three. A general unitary matrix has nine parameters, and so there are six phases in the CKM matrix. By rephasing we can remove five of them, but not the sixth - the CKM matrix indeed breaks $\mathcal{T}$ symmetry![87] How exactly to parametrize this phase is highly convention-dependent, but the absolute values of the matrix are more sociologically stable and have all been measured to give[88]

$$\begin{pmatrix} |V_{11}| & |V_{12}| & |V_{13}| \\ |V_{21}| & |V_{22}| & |V_{23}| \\ |V_{31}| & |V_{32}| & |V_{33}| \end{pmatrix} \approx \begin{pmatrix} .975 & .225 & .003 \\ .225 & .973 & .042 \\ .009 & .041 & .999 \end{pmatrix} \tag{10.50}$$

The $\mathcal{T}$ violation is often described in convention-independent way using the **Jarlskog invariant** (invented by Cecilia Jarlskog), which is defined as

$$J \equiv \mathrm{Im}\left(V_{12} V_{23} V_{13}^* V_{22}^*\right) \tag{10.51}$$

and by construction is invariant under the rephasings (10.49). It has been measured to be

$$J \approx 3 \times 10^{-5}, \tag{10.52}$$

which is not zero and thus there is no rephasing that makes V real.

10.4 θ terms

There are three more terms we need to discuss to complete our presentation of the standard model Lagrangian. These are the θ-angle terms[89]

$$\mathcal{L}_\theta = \frac{\theta_3}{32\pi^2}\epsilon^{\mu\nu\alpha\beta}\mathrm{Tr}\left(G_{\mu\nu}G_{\alpha\beta}\right) + \frac{\theta_2}{32\pi^2}\epsilon^{\mu\nu\alpha\beta}\mathrm{Tr}\left(F_{\mu\nu}F_{\alpha\beta}\right) + \frac{\theta_1}{36\times 32\pi^2}\epsilon^{\mu\nu\alpha\beta}B_{\mu\nu}B_{\alpha\beta}. \tag{10.53}$$

[86] Particle physicists usually talk about this as $\mathcal{CP}$ violation instead of $\mathcal{T}$ violation, which is equivalent by the $\mathcal{CRT}$ theorem. $\mathcal{T}$ however is much more intuitive however: I dont know a way to connect $\mathcal{CP}$ violation to phases in the Lagrangian except for using the $\mathcal{CRT}$ theorem to turn it back into $\mathcal{T}$ violation. I believe the reason people like to discuss $\mathcal{CP}$ is mostly historical, although it is true that some of the experimental signatures are easier to describe in terms of $\mathcal{CP}$ as we will see later in the section.

[87] This conclusion relies crucially on the existence of three matter generations, on the homework you will show that with two all phases can be removed.

[88] Using only the absolute values you cannot directly check the unitarity of V, but you can check some of its consequences such as $\sum_j |V_{ij}|^2 = 1$.

[89] The bizarre factor of 36 in the $U(1)_Y$ case is there because we took the minimal charge of $U(1)_Y$ to be 1/6, gauge field is 6 times what it should have been.

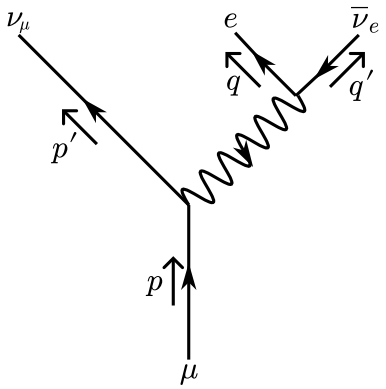

Figure 24: Electroweak decay of a muon by W boson exchange.

We will see when we discuss anomalies that we can actually set $\theta_2 = 0$ by a field redefinition, but θ_1 and θ_3 are genuine parameters of the standard model. Both of them violate $\mathcal{T}$ symmetry if they are not equal to zero or π (we will see later that the theory is periodic in θ with period 2π, so $\pi \to -\pi$ gives the same theory). Nonzero θ_3 would lead to an electric dipole moment of the neutron: since no such moment has been observed, this gives a bound

$$|\theta_3| \lesssim 10^{-10}. \tag{10.54}$$

Thus the current data is consistent with $\theta_3 = 0$. θ_1 is much more difficult to measure experimentally, for topological reasons that we will understand later. It leads to interesting effects if the topology of spacetime is different from $\mathbb{R}^4$, and it also would cause a magnetic monopole to acquire an electric charge via something called the **Witten effect**. So far we therefore do not have any observational information about θ_1, so as stated above the only known $\mathcal{T}$-violation in the standard model is from the CKM matrix. We will have more to say about these θ terms once we have understood anomalies and instantons.

10.5 Electroweak phenomenology

This is not a class in standard model phenomenology, but we will still consider two simple processes to illustrate some of the issues that arise.

10.5.1 Muon decay

Our first process is the electroweak decay of the muon. The diagram is shown in figure 24. The main new ingredient we will need for this calculation is the W-boson propagator in unitarity gauge, you will show on the homework that in momentum space this is given by

$$\Delta^W_{\mu\nu} = \frac{-i\left(\eta_{\mu\nu} + \frac{p^\mu p^\nu}{m_W^2}\right)}{p^2 + m_W^2 - i\epsilon}. \tag{10.55}$$

Evaluating the diagram (which also requires the interaction vertex from (10.31)) then gives a covariant matrix element

$$i\mathcal{M} = -\frac{ig_2^2}{2}\overline{u}(p')\gamma^\mu P_L u(p)\overline{u}(q)\gamma^\nu P_L v(q')\frac{\eta_{\mu\nu} + \frac{(p-p')^\mu(p-p')^\nu}{m_W^2}}{(p-p')^2 + m_W^2 - i\epsilon}. \tag{10.56}$$

Taking advantage of the fact that $m_\mu \ll m_W$ however we can simplify the W boson propagator to

$$\Delta^W_{\mu\nu} \approx \frac{-i\eta_{\mu\nu}}{m_W^2}, \tag{10.57}$$

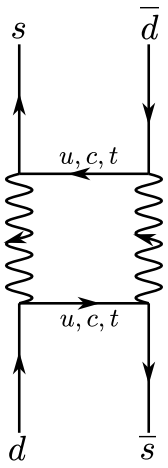

Figure 25: A loop diagram for $K - \overline{K}$ mixing.

which simplifies the amplitude to

$$i\mathcal{M} \approx -i\frac{g_2^2}{2m_W^2}\overline{u}(p')\gamma^\mu P_L u(p)\overline{u}(q)\gamma^\nu P_L v(q'). \tag{10.58}$$

This is the same amplitude we would have gotten starting with the effective Fermi interaction

$$\mathcal{L}_{Fermi} = -2\sqrt{2}G_F(\overline{e}_L\gamma^\alpha\nu_e)(\overline{\nu}_\mu\gamma_\alpha\mu_L) + \text{h.c.}, \tag{10.59}$$

with

$$G_F \equiv \frac{g_2^2}{4\sqrt{2}m_W^2} \approx 1.1 \times 10^{-5}\text{GeV}^{-2}. \tag{10.60}$$

Squaring this amplitude, averaging/summing over initial/final spins, and integrating over final momenta to get the total decay rate is not too interesting, since $m_e \ll m_\mu$ we know by dimensional analysis that the result must be proportional to $G_F^2 m_\mu^5$. It turns out to be

$$\Gamma = \frac{G_F^2 m_\mu^5}{192\pi^3}, \tag{10.61}$$

which gives a lifetime of about 2×10^{-6} s. This is in good agreement with experiment.

10.5.2 K-$\overline{K}$ mixing

The other process we will consider is the electroweak mixing between the neutral kaons K^0 with quark content $d\overline{s}$ and $\overline{K}^0$ with quark content $\overline{d}s$. In QCD these particles are strictly orthogonal quantum states since they have different down and strange flavor numbers, but flavor number symmetries are broken in the electroweak theory so there can be a mixing between the two of them. This process is an example of what is conventionally called a **flavor changing neutral current (FCNC) interaction**, since the QCD currents that creates the neutral kaons, $\overline{s}\gamma^\mu\gamma d$ and $\overline{d}\gamma^\mu\gamma s$, mix different flavors but are electrically neutral. The important point however is that in the standard model such interactions are suppressed for *three* independent reasons. The first reason is that this mixing cannot happen at tree level. This is because the Z vertex is diagonal in flavor while the W vertex carries away electric charge and thus cannot couple to a neutral current. The first thing that works is a one-loop diagram, which is shown in figure 25. We will not compute it in detail, but it has one particularly elegant feature that is worth emphasizing. This is that

once we take the CKM matrix into account, all three kinds of up-type quarks can run into the loop. This is most clear if we put the CKM matrix into the gauge interactions, by writing the $SU(2)$ quark doublet as in (10.48). The four interactions thus contribute something like

$$\sum_i V_{1i} V_{i2}^\dagger \sum_j V_{2j} V_{j1}^\dagger \tag{10.62}$$

to the diagram. If this were the only dependence on the internal quark flavor labels i, j, which would for example be the case if all the up-type quarks had equal masses, then this would actually vanish due to the unitarity of the CKM matrix! This is called the **GIM mechanism**, named after Glashow, Iliopoulos, and Maiani, and it further suppresses this diagram on top of the loop suppression. Of course the up-type quark masses are not all equal, so the suppression is not complete. Roughly speaking we can estimate the GIM suppression as follows: naively this loop diagram is quadratically divergent in the Fermi theory where we integrate out the W boson, so we should expect something of order $G_F^2 m_W^2$ since m_W is the cutoff scale for the Fermi theory. But in order to get a nonzero answer we need to bring in a power of an up-type quark mass (and in fact it needs to be squared due to the chiral nature of the interaction). Due to the smallness of the CKM matrix elements involving the top quark, the dominant contribution ends up having the charm quark at least one of the internal legs, with either the charm or the up on the other. This gives a factor of m_c^2/m_W^2, which substantially suppresses the diagram. Finally the third suppression arises because the CKM matrix elements that mix generations themselves are small, for the dominant uc or cc channels they give another suppression by $.225^2$. This triple suppression by a loop factor, the GIM mechanism, and the CKM matrix is very important in studying proposals for particle physics beyond the standard model, since pretty much any new physics breaks at least one of these and thus gives neutral kaon oscillation which is too large unless the new physics is at a rather high energy scale.

The presence of the CKM matrix in this diagram also suggests that we can use it to probe $\mathcal{T}$ violation in the standard model, and indeed we can: it leads to an oscillation between $\mathcal{CR}$-even and $\mathcal{CR}$-odd superpositions of K^0 and $\overline{K}^0$ that would be strictly forbidden were $\mathcal{T}$ an exact symmetry.

10.6 Neutrino masses

Although the standard model predicts massless neutrinos, flavor oscillations of neutrinos have actually been observed in a number of experiments. The easiest way to explain this is that neutrinos are actually massive. One way to accomplish this is to add a ν_R field to the standard model which is neutral under all gauge symmetries, and then use the Higgs to write down a mass term of the form $\overline{\ell}_L \phi \nu_R$. Since there are now two lepton Yukawa terms, the situation becomes exactly analogous to the quark Yukawa terms, and in particular there is now a lepton analogue of the CKM matrix which is called the **PMNS matrix**, for Pontecorvo, Maki, Nakagawa, and Sakata. This matrix then leads to oscillations between neutrinos of different type. This is the only way to include neutrino masses in the standard model without introducing non-renormalizable terms. If we are willing to do the latter, then there is also a term called the **Weinberg operator**, which is essentially $(\widetilde{\phi}\ell_L)^T \mathcal{C} (\widetilde{\phi}\ell_L)$, which gives a Majorana-type mass to neutrinos without needing to introduce any new fields such as ν_R. The Weinberg operator also does not need to be diagonal in generation number and thus still introduces a PMNS matrix. So far it is not known which of these options is a better description of the observed neutrino oscillations, hopefully we will find out soon!

10.7 Homework

1. In this problem we will define and study unitarity gauge for a general Higgsing of a gauge group G to a subgroup H with Lagrangian

$$\mathcal{L} = -\frac{1}{4}\sum_a \frac{1}{g_a^2} F^a_{\mu\nu} F_a^{\mu\nu} - (D_\mu \phi)^\dagger D^\mu \phi - V(\phi) \tag{10.63}$$

and covariant derivative

$$D_\mu \phi = \partial_\mu \phi - i\tau_a A^a_\mu \phi. \tag{10.64}$$

Writing the expectation value of ϕ as

$$\langle \phi \rangle = \frac{v}{\sqrt{2}}, \tag{10.65}$$

the subgroup H is defined by

$$H = \{h \in G | D(h)v = v\}. \tag{10.66}$$

The definition of unitarity gauge is that in each gauge orbit $\phi(x) \sim D(g(x))\phi(x)$ we choose a representative that maximizes the quantity $\mathrm{Re}\left(\phi^\dagger v\right)$. In other words at each point in spacetime we choose $g_\phi(x) \in G$ such that the function

$$f_\phi(g) \equiv \mathrm{Re}\left(\phi^\dagger(x) D(g) v\right) \tag{10.67}$$

is maximized at $g = g_\phi(x)$, and then we define the fields in unitarity gauge as

$$\begin{aligned} \widetilde{\phi}(x) &\equiv D(g_\phi(x)^{-1})\phi(x) \\ \widetilde{A}_\mu &\equiv g_\phi^{-1} A_\mu(x) g_\phi + i g_\phi^{-1} \partial_\mu g_\phi. \end{aligned} \tag{10.68}$$

Such a maximum always exists since we are maximizing a continuous function over a compact space.

(a) Show that the unitarity gauge Higgs field obeys

$$\mathrm{Im}\left(\widetilde{\phi}^\dagger(x) T_a v\right) = 0, \tag{10.69}$$

where T_a are the generators of $\mathfrak{g}$. Also show that if $\phi = v$ then we can take $\widetilde{\phi} = v$ as well, so ϕ and $\widetilde{\phi}$ have the same expectation value.

(b) Show that multiplying $g_\phi(x)$ on the right by an element of H always gives another maximum of $f_\phi(g)$, and argue that this means that in unitarity gauge H is a residual gauge symmetry that is not fixed. In particular argue that if we decompose the gauge fields as $\mathfrak{g} = \mathfrak{h} + \mathfrak{h}^\perp$, then the unbroken gauge fields in $\mathfrak{h}$ transform as gauge fields under local H transformations, while the broken gauge fields in $\mathfrak{h}^\perp$ transform in a linear representation of H.

(c) We could still hope that there is a unique maximum if we re-interpret $f_\phi(g)$ as a function on coset space G/H, but in general this is not true.[90] It *is* true however in a small neighborhood of $[g_\phi(x)] \in G/H$ and for $\widetilde{\phi}$ valued in some neighborhood of v, which is enough to justify using unitarity gauge in perturbation theory. Confirm this by expanding

$$f_{\widetilde{\phi}}(e^{i\theta^a T_a}) = \mathrm{Re}\left(\widetilde{\phi}^\dagger e^{i\theta^a \tau_a} v\right) \tag{10.70}$$

to second order in θ^a and then arguing that for $\widetilde{\phi}$ close to v, $f_{\widetilde{\phi}}$ is strictly less than its maximum at $\theta^a = 0$ unless we have

$$\theta^a \tau_a v = 0. \tag{10.71}$$

This shows that $\theta^a T_a \in \mathfrak{h}$, and thus $e^{i\theta^a \mathcal{T}_a} \in H$.

[90] This is another avatar of the Gribov problem in gauge fixing.

(d) Parametrizing the Higgs field in unitarity gauge as

$$\widetilde{\phi} = \frac{1}{\sqrt{2}}(v + h), \tag{10.72}$$

write the Lagrangian in terms of h and $\widetilde{A}_\mu$. Make sure to use the condition (10.69). Show that the gauge field mass matrix for the canonically-normalized gauge field $\hat{A}^a_\mu = \frac{\widetilde{A}^a_\mu}{g_a}$ is

$$\mu^2_{ab} = \frac{g_a g_b}{2} v^\dagger \{\tau_a, \tau_b\} v, \tag{10.73}$$

and argue that this is positive, real, and symmetric, and thus that it can be diagonalized by an orthogonal change of basis on $\mathfrak{g}$. Also show that c^a is a null eigenvector of μ^2_{ab} if and only if $\sum_a g_a c^a T_a \in \mathfrak{h}$, so the gauge field remains massless in the unbroken directions.

(e) Consider the case of $G = U(1)$ with Higgs field ϕ transforming with charge $p \in \mathbb{Z}$. What is g_ϕ and $\widetilde{\phi}$, and what is μ^2? Does this agree with what we called unitarity gauge in our previous discussion of the abelian Higgs model?

(f) For the standard model case $G = SU(2) \times U(1)_Y$ with a Higgs field in the $(2, -1/2)$ representation, assuming that the Higgs vev is

$$\langle \phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} v \\ 0 \end{pmatrix} \tag{10.74}$$

with $v > 0$ show that in unitarity gauge the Higgs field has the form (10.8). Also show that the gauge boson mass matrix (10.73) is

$$\mu^2_{ab} = \frac{v^2}{4} \begin{pmatrix} g_2^2 & 0 & 0 & 0 \\ 0 & g_2^2 & 0 & 0 \\ 0 & 0 & g_2^2 & -g_1 g_2 \\ 0 & 0 & -g_1 g_2 & g_1^2 \end{pmatrix}, \tag{10.75}$$

with the first three rows/columns being the $SU(2)$ generators and the last row/column being the $U(1)_Y$ generators.

2. Derive the expression (10.55) for the W-boson propagator in unitarity gauge.

3. Show that the subgroup of $SU(3) \times SU(2) \times U(1)$ generated by the (10.35) is the full set of gauge transformations that act trivially on all quark and lepton fields.

4. Argue that if there were only two generations then we could remove all of the $\mathcal{T}$-violating phases from the CKM matrix.

5. Argue that the standard model of particle physics has twenty dimensionless parameters.[91] How many are added if we include Dirac or Majorana neutrino masses? Hint: you do not want to count parameters that can be removed by field redefinitions.

[91] People often say there are nineteen, but that is because they do not understand that the $U(1)$ θ-angle is physical.

11 Anomalies

We now turn to the topic of "anomalous" symmetries in field theory. This can be a somewhat confusing topic, in substantial part because the traditional understanding of it is rather muddled due to various historical baggage. Indeed the term "anomaly" is used for a number of distinct phenomena. I will do my best to give a more coherent treatment.

11.1 What is an anomaly?

In order to give a section about anomalies we need to define them. The definition we will use is the following: a quantum field theory with a global symmetry group G has an **candidate anomaly** for that symmetry if the partition function of the theory on some spacetime manifold M as a function of a background gauge field A for the symmetry is not gauge invariant:[92]

$$Z[A_g] \neq Z[A], \tag{11.1}$$

where

$$A_{\mu,g} = g\left(A_\mu - ig^{-1}\partial_\mu g\right)g^{-1}. \tag{11.2}$$

There are many ways to not be equal, but in most cases of interest this non-equality has a rather specific form: it comes in the form of a phase that is a local integral of A and g:

$$Z[A_g] = e^{i\alpha(A,g)}Z[A], \tag{11.3}$$

with

$$\alpha[A,g] \equiv \int d^dx \mathcal{A}(A,g). \tag{11.4}$$

The form of the candidate anomaly $\alpha[A,g]$ is constrained by the group law of gauge transformations: we must have

$$e^{i\alpha[A,g'g]}Z[A] = Z[A_{g'g}] = Z[(A_g)_{g'}] = e^{i\alpha[A_g,g']}Z[A_g] = e^{i\alpha[A_g,g']+i\alpha[A,g]}Z[A], \tag{11.5}$$

and thus

$$\alpha[A,g'g] = \alpha[A_g,g'] + \alpha[A,g] + 2\pi\mathbb{Z}, \tag{11.6}$$

which is called the **Wess-Zumino consistency condition**.

This candidate anomaly is not so interesting however if it can be removed by redefining the partition function as

$$Z'[A] \equiv e^{i\beta[A]}Z[A] \tag{11.7}$$

with β being a local functional

$$\beta[A] = \int d^dx \mathcal{B}[A] \tag{11.8}$$

of A and its derivatives. This is because we are always free to include such local terms in the definition of the action, so if by doing so we can restore naive gauge invariance $Z[A_g] = Z[A]$ then we might as well do so. We can see how this redefinition changes the candidate anomaly as follows:

$$Z'[A_g] = e^{i\beta[A_g]}Z[A_g] = e^{i\beta[A_g]+i\alpha[A,g]}Z[A] = e^{i\beta[A_g]+i\alpha[A,g]-\beta[A]}Z'[A], \tag{11.9}$$

and thus

$$\alpha'[A,g] = \alpha[A,g] + \beta[A_g] - \beta[A] + 2\pi\mathbb{Z}. \tag{11.10}$$

[92] We have not yet defined gauge fields on general manifolds M, we will do so in the next section. In general the gauge transformation rule is a bit more subtle than what we write here, after all if G is a finite group what we write here has no content, but the idea is the same: there is an anomaly if the partition function assigns different values to gauge-equivalent background gauge field configurations.

We therefore say that the symmetry has a **true anomaly**, or just an **anomaly**, if it has a candidate anomaly that cannot be removed by such a redefinition.[93] Classifying local functionals obeying (11.6) up to an equivalence of the form (11.10) turns out to be a problem it what is called group cohomology, although we will not explore this further here.

11.1.1 A simple example

Having given a definition, we should now give an example. Consider a free massless scalar field θ in $1+1$ dimensions, obeying a periodicity condition

$$\theta \sim \theta + 2\pi. \tag{11.11}$$

The theory has a global shift symmetry

$$\theta' = \theta + \lambda, \tag{11.12}$$

and we will turn on a background gauge field A_μ for this symmetry like this:

$$\mathcal{L} = -\frac{1}{2}\left(\partial^\mu\theta - A^\mu\right)\left(\partial_\mu\theta - A_\mu\right) + \frac{\theta}{4\pi}\epsilon^{\mu\nu}F_{\mu\nu}. \tag{11.13}$$

You can think of the field θ as the phase of a more conventional scalar field, and the first term in the Lagrangian is precisely the kinetic term we had for such a phase in the Abelian Higgs model (although I remind you that here A_μ is a background field instead of a dynamical gauge field). The second term is perhaps more mysterious, in particular it looks like it does not respect the periodicity (11.11). Quantum mechanically however it actually does: we will see in the next section that[94]

$$\frac{1}{2}\int_M \epsilon^{\mu\nu}F_{\mu\nu} = 2\pi\mathbb{Z} \tag{11.14}$$

for any spacetime M, so a shift of ϕ by $2\pi m$ does not change the quantity e^{iS} in the path integral. The equation of motion for this theory in the presence of the background field is

$$\partial_\mu(\partial^\mu\theta - A^\mu) = -\frac{1}{4\pi}\epsilon^{\mu\nu}F_{\mu\nu}. \tag{11.15}$$

Now let's consider what happens to the partition function

$$Z[A] = \int \mathcal{D}\theta e^{iS[\theta,A]} \tag{11.16}$$

under a gauge transformation

$$A_{e^{i\Omega}} = A + \partial_\mu\Omega. \tag{11.17}$$

The natural way to try to compute $Z[A_{e^{i\Omega}}]$ is to change variables in the path integral as

$$\theta \to \theta + \Omega, \tag{11.18}$$

after which we can use the gauge invariance of the first term in $\mathcal{L}$ to absorb the gauge transformation of A_μ. The second term however now picks up a nontrivial gauge transformation, giving us our first calculation of a candidate anomaly:

$$Z[A_{e^{i\Omega}}] = e^{i\alpha[A,e^{i\Omega}]}Z[A], \tag{11.19}$$

[93] What I am calling an anomaly here is these days usually called an "'t Hooft anomaly", to distinguish it from the various other things that are also called anomalies. My view is that all of those things are downstream of this definition, so with all respect to 't Hooft it is easier to just call it an anomaly (and anyways the way 't Hooft thought about it was not quite the same as the modern approach).

[94] This statement is essentially the Dirac quantization of magnetic charge, which we studied in problem 10.4 last semester. If M is noncompact then we also need to impose boundary conditions at infinity, for example for $M = \mathbb{R}^2$ the requirement is that a large Wilson loop needs to become trivial as we take its radius to infinity.

with

$$\alpha[A, e^{i\Omega}] = \frac{1}{4\pi}\int d^2x\, \Omega\, \epsilon^{\mu\nu} F_{\mu\nu}. \tag{11.20}$$

To confirm that this is indeed an anomaly, we need to show that we cannot remove it by a local redefinition of the action as in (11.10). Let's check: in order to remove it, we need

$$\alpha[A, e^{i\Omega}] = \beta[A_{e^{i\Omega}}] - \beta[A] \tag{11.21}$$

for some local functional β. Integrating by parts in the anomaly we can rewrite it as

$$\alpha[A, e^{i\Omega}] = -\frac{1}{2\pi}\int d^2x \partial_\mu \Omega \epsilon^{\mu\nu} A_\nu, \tag{11.22}$$

which at first looks like indeed it wants to be of the form (11.21), but actually the term we would want to write is

$$\beta[A] = -\frac{1}{4\pi}\int d^2x A_\mu A_\nu \epsilon^{\mu\nu} = 0 \tag{11.23}$$

so we are out of luck.

You might complain that in the example the anomaly was self-inflicted - why did we add this weird extra term to the Lagrangian? On the other hand the equation of motion (11.15) is perfectly invariant under the gauge transformation

$$\begin{aligned} \theta' &= \theta + \Omega \\ A'_\mu &= A_\mu + \partial_\mu \Omega, \end{aligned} \tag{11.24}$$

so classically there is nothing wrong with the gauge symmetry of this theory. We will also see later in the section that terms like this can emerge in the infrared from theories that are quite natural in the UV, and in particular we will see that this happens in QCD. So we had better understand how to think about theories with anomalies.

There are three conceptual points I want to make about this anomaly:

- It has nothing to do with difficulty in finding a UV regulator that is consistent with gauge invariance.
- There are no fermions in this theory.
- There is nothing wrong with this theory, or with its global shift symmetry.

I mention these because most QFT textbooks strongly give the impression that anomalies are a phenomenon that arises from the difficulty of regulating chiral fermions in a way that preserves symmetries that are there in the continuum, and also that they ruin the symmetry that is anomalous. A problem regulating fermions is indeed one way to get an anomaly, as we will see shortly, but it is only one way out of several. Moreover we will see that anomalies can indeed sometimes lead to a symmetry being violated, but not always. The essential feature of an anomaly is not the fermions or the regulator, or the possible destruction of a symmetry; instead it is the failure of the partition function to be invariant under background gauge transformations.

11.1.2 What do anomalies do?

Having defined anomalies and given an example, we should now say why they are interesting. There are three main reasons:

(1) The presence of an anomaly is an obstruction to gauging the global symmetry G. By "gauging", I mean turning on a background gauge field for the symmetry and then making it dynamical by integrating over it in the path integral and dividing by the volume of the gauge group,[95]

$$\int \frac{\mathcal{D}A}{\mathrm{Vol}(\mathcal{G})} Z[A]. \tag{11.25}$$

[95] Often when we do this we included extra "kinetic" terms for the gauge field in the action; here we are thinking of these as local modifications of $Z[A]$ as in (11.7).

The reason the anomaly is an obstruction to such gauging is that we cannot view a gauge transformation as a redundancy of description if the integrand in the path integral is not gauge-invariant, and we saw that in a gauge theory if we do not view the gauge symmetry as a redundancy then the initial value problem is not well-defined (in quantum mechanical terms the Hamiltonian is not well-defined on the larger Hilbert space that includes gauge-variant states).[96]

(2) Sometimes there is a normal subgroup H of G that is not affected by the anomaly. We are then free to gauge it in the usual sense, but once we do then the naive remaining global symmetry group G/H is often broken by the anomaly. We will see that this is what happens to the $U(1)_A$ symmetry of QCD.

(3) The existence of an anomaly is a statement about the partition function of a field theory. It does *not* depend on how we evaluate the path integral. This has a very important consequence: if we integrate out some heavy degrees of freedom in the theory to get a low-energy effective action, then that action must have some kind of dynamics which is able to reproduce any anomaly that we found by doing calculations directly in the high-energy theory. This is called **anomaly matching**: if a UV theory has a calculable anomaly, then whatever happens in the IR must be able to reproduce that anomaly. This is powerful because in general the relationship between the UV and the IR involves strong coupling dynamics that are often intractable for concrete calculation, but any proposal for what happens in the IR can be immediately tested by asking if it is consistent with anomaly matching.

We will illustrate all three of these points below in concrete examples.

11.2 Path integral calculation of the Abelian chiral anomaly

We now turn to the first kind of anomaly that was discovered, what we will call the **Abelian chiral anomaly**. The theory we consider is N_f free Dirac fermions in d spacetime dimensions, with Lagrangian

$$\mathcal{L} = -i\overline{\psi}_i \not{\partial} \psi^i \tag{11.26}$$

with $i = 1, 2, \ldots, N_F$. We will take d to be even. This theory has a large flavor symmetry group. In our discussion of massless QCD we discussed a flavor symmetry $U(N_f)_L \times U(N_F)_R$, but since we here do not have dynamical gauge fields there is actually a larger $U(2N_F)$ flavor symmetry since we can mix the left-handed spinors ψ_L^i with the left-handed spinors $B^*\psi_R^*$, where

$$B = \begin{pmatrix} 0 & -i\sigma_2 \\ i\sigma_2 & 0 \end{pmatrix} \tag{11.27}$$

is the matrix appearing in the Majorana constraint condition $\psi^* = B\psi$. Such mixing wasn't a symmetry in QCD since ψ_L and $B^*\psi_R^*$ are in conjugate $SU(3)$ gauge representations. To study in general the possible anomalies of this larger symmetry we should therefore turn on an arbitrary background $U(2N_F)$ gauge field and then study the gauge transformation of the resulting partition function under $U(2N_F)$ gauge transformations. We will do this calculation later in the section, but unfortunately it is somewhat tedious. We therefore will illustrate the basic idea in a more elegant way by first considering the special case where we only turn on gauge fields for some arbitrary subgroup H of the vector $U(N_f)$ symmetry

$$\psi' = e^{i\theta^a \tau_a}\psi, \tag{11.28}$$

together with *one* axial symmetry

$$\psi' = e^{i\Omega\tau\gamma}\psi, \tag{11.29}$$

[96] It is worth mentioning however that if we nonetheless integrate over A the theory we get is not necessarily sick provided that we do *not* divide by the volume of the gauge group. What can happen instead is that the gauge transformation itself becomes a new degree of freedom. In such cases the resulting theory is not pathological, it just has extra degrees of freedom beyond what we would have expected from a gauge field (and in particular the resulting "gauge field" is often massive).

such that

$$[\tau, \tau_a] = 0 \tag{11.30}$$

for all $T_a \in \mathfrak{h}$. This last condition is why the anomaly is called "Abelian", we only turn on a background for one axial symmetry and it is required to commute with all the (possibly non-Abelian) vector symmetries. This is not merely a warmup exercise, we will see shortly that some of the most important anomaly physics is contained in this special case.[97]

This calculation is most easily discussed in Euclidean signature, where the partition function we are interested in is[98]

$$Z[A, B] = \int \mathcal{D}\psi \mathcal{D}(-i\overline{\psi}) e^{-S_E[\psi, \overline{\psi}, A, B]} = \det\left(-i\not{D}\right), \tag{11.31}$$

with

$$S_E[\psi, \overline{\psi}, A, B] = \int d^d x \overline{\psi} \not{D} \psi. \tag{11.32}$$

Here

$$D_\mu \psi = \partial_\mu \psi - i A^a_\mu \tau_a \psi - i B_\mu \tau \gamma \psi, \tag{11.33}$$

where B_μ is the background axial gauge field and A^a_μ is the background H gauge field. We will focus on gauge transformations of $B_\mu = 0$, so the anomaly that we will compute has the form

$$Z[A, d\Omega] = e^{i\alpha[A, e^{i\Omega}]} Z[A, 0]. \tag{11.34}$$

From the path integral we have

$$\begin{aligned} Z[A, d\Omega] &= \int \mathcal{D}\psi \mathcal{D}(-i\overline{\psi}) e^{-S_E[\psi, \overline{\psi}, A, d\Omega]} \\ &= \int \mathcal{D}\psi_{e^{i\Omega}} \mathcal{D}(-i\overline{\psi}_{e^{i\Omega}}) e^{-S_E[\psi_{e^{i\Omega}}, \overline{\psi}_{e^{i\Omega}}, A, d\Omega]} \\ &= \int \mathcal{D}\psi_{e^{i\Omega}} \mathcal{D}(-i\overline{\psi}_{e^{i\Omega}}) e^{-S_E[\psi, \overline{\psi}, A, 0]}, \end{aligned} \tag{11.35}$$

where in going from the first line to the second we changed the integration variable and in going from the second to the third we used the gauge invariance of S_E. Therefore the only possible Ω-dependence on the right-hand side, and thus the only possibly anomaly, can come from the gauge transformation of the path integral measure. We will spend the rest of this subsection regulating the measure carefully enough that we can extract it.[99]

We'll begin by first trying to compute the transformation of the measure without any regard for UV regularization. Introducing a notation $\psi_n(x)$ where the index n accounts for both spinor and flavor indices, we can view a symmetry transformation

$$\psi'_n(x) = U_{nm}(x) \psi_m(x) \tag{11.36}$$

as a big change of variables matrix

$$\mathcal{U}_{xn,ym} \equiv U_{nm} \delta^d(x - y) \tag{11.37}$$

[97] One reason why considering subgroups is natural is that in most applications we add other fields and interactions to the theory, and these typically break the full $U(2N_F)$ flavor symmetry to some smaller subgroup.

[98] The factor of $-i$ in the measure for $\overline{\psi}$ arises because in Euclidean signature we defined $\overline{\Psi} = \Psi^\dagger \gamma^0_E = i\Psi^\dagger \gamma^0$, while the measure was defined originally in terms of $\Psi^\dagger \gamma^0$. This overall phase rarely matters however, and in particular it has no effect on the anomaly.

[99] This approach to computing anomalies is called the **Fujikawa method**, and it can be compared with the direct perturbative approach we take for the general chiral anomaly later in the section. On major advantage of the Fujikawa approach is that it makes it clear that there are no higher-order corrections to the anomaly.

in the path integral. This symmetry acts on ψ viewed as an infinite-dimensional vector as

$$\begin{aligned}\psi' &= \mathcal{U}\psi \\ \overline{\psi}' &= \overline{\psi}\,\overline{\mathcal{U}},\end{aligned} \tag{11.38}$$

with

$$\overline{\mathcal{U}}_{xn,ym} = (\gamma_E^0 U^\dagger \gamma_E^0)_{nm}\delta^d(x-y). \tag{11.39}$$

The transformation of the path integral measure is thus

$$\mathcal{D}\psi'\mathcal{D}(i\overline{\psi}') = \frac{1}{\det(\mathcal{U})\det(\overline{\mathcal{U}})}\mathcal{D}\psi\mathcal{D}(i\overline{\psi}). \tag{11.40}$$

For vector gauge transformations in H we have

$$U = e^{i\theta^a\tau_a}, \tag{11.41}$$

which commutes with γ_E^0, so we have $\overline{\mathcal{U}} = \mathcal{U}^\dagger$ and thus $\det(\mathcal{U})\det(\overline{\mathcal{U}}) = 1$. Therefore the measure is invariant under the vector gauge transformations. For axial gauge transformations on the other hand we have

$$U = e^{i\Omega\tau\gamma}, \tag{11.42}$$

which obeys

$$\gamma_E^0 e^{-i\Omega\tau\gamma}\gamma_E^0 = e^{i\Omega\tau\gamma}. \tag{11.43}$$

Thus for the axial gauge transformation we have $\overline{\mathcal{U}} = \mathcal{U}$, and thus a potentially anomaly of the form

$$e^{i\alpha} = e^{-2\log\det\mathcal{U}} = e^{-2i\int d^dx\Omega(x)\mathrm{Tr}(\gamma\tau)\delta^d(x-x)}. \tag{11.44}$$

Thus we can extract the anomaly density

$$\mathcal{A}(x) = -2\Omega(x)\mathrm{Tr}\,(\gamma\tau)\,\delta^d(x-x). \tag{11.45}$$

Unfortunately as written this is a rather useless expression: the trace of γ is zero but $\delta^d(0)$ is infinity. To get something more useful we need to regulate this calculation.

To come up with a sensible version of (11.45), we can make our lives easier by 1) making sure to continue preserving vector gauge invariance and 2) constructing a regulator only using the differential operator $\not{D}$, since that is the operator whose determinant we are ultimately computing.[100] The choice we will make is

$$\mathcal{A}(x) = -2\Omega(x)\mathrm{Tr}\left(\gamma\tau e^{\not{D}_x^2/\Lambda^2}\right)\delta^d(x-y)|_{y\to x}. \tag{11.46}$$

This is a version of what is called a **heat-kernel regulator**, with UV cutoff Λ, and it can be derived directly from a heat-kernel regulated version of the determinant on the right-hand side of (11.31). Introducing a momentum-space representation of the δ-function, we have

$$\begin{aligned}\mathcal{A}(x) &= -2\Omega(x)\int\frac{d^dk}{(2\pi)^d}\mathrm{Tr}\left(\gamma\tau e^{\not{D}_x^2/\Lambda^2}\right)e^{ik\cdot(x-y)}|_{y=x} \\ &= -2\Omega(x)\int\frac{d^dk}{(2\pi)^d}\mathrm{Tr}\left(\gamma\tau e^{(i\not{k}+\not{D}_x)^2/\Lambda^2}\right) \\ &= -2\Omega(x)\Lambda^d\int\frac{d^dk}{(2\pi)^d}\mathrm{Tr}\left(\gamma\tau e^{-k^2+2ik\cdot D/\Lambda+\not{D}^2/\Lambda^2}\right)\end{aligned} \tag{11.47}$$

[100] Neither of these principles is strictly necessary: if we break the former we will get something that differs from the anomaly we compute by a local counterterm $\beta[A]$ that breaks the vector gauge invariance, while if we break the latter, for example by using $D_\mu D^\mu$ instead of $\not{D}^2$ below, then the regulation of the determinant produces an additional term that restores the anomaly. Whatever choices we make about the regulator, in the end they at most change the anomaly by some $\beta[A]$ so we might as well make a convenient choice.

In going from the first line to the second we have used that each derivative can either act on the gauge fields in powers of $\not{D}$ to its right or on $e^{ik\cdot(x-y)}$. In going from the second to the third we have rescaled k by Λ and expanded out the exponent. The idea is now to Taylor expand in the second two terms in the exponent. Any terms which give a power of $1/\Lambda$ that is greater than d can be ignored in the limit of large Λ, and we can also use that in order for the spinor trace to be nonzero we need to bring down at least d γ-matrices. The only term in the expansion that satisfies both requirements is to bring down $d/2$ powers of $\not{D}^2/\Lambda^2$, so the anomaly is given by

$$\begin{aligned}\mathcal{A}(x) &= -2\Omega(x)\int \frac{d^dk}{(2\pi)^d} e^{-k^2} \frac{1}{\left(\frac{d}{2}!\right)} \mathrm{Tr}\left(\gamma\tau \not{D}^d\right) \\ &= -\frac{\Omega(x)}{2^{d-1}\pi^{d/2}\left(\frac{d}{2}!\right)} \mathrm{Tr}\left(\tau D_{\mu_1}\dots D_{\mu_d}\right) \mathrm{Tr}\left(\gamma\gamma^{\mu_1}\dots\gamma^{\mu_d}\right), \end{aligned} \tag{11.48}$$

where in the second line we evaluated the Gaussian integral.[101] Everything in this expression is now finite, so we are on track for a nice answer. To finish up we need to figure out how to compute the spinor and flavor traces. I'll begin by noting that in Lorentzian signature we have the convenient identity

$$\mathrm{Tr}\left(\gamma\gamma^{\mu_1}\dots\gamma^{\mu_d}\right) = -2^{\frac{d}{2}} i^{\frac{d-2}{2}} \epsilon^{\mu_1\dots\mu_d}. \tag{11.49}$$

This follows from first noting that the left-hand side is completely antisymmetric due to the γ matrix algebra, and then noting that we can use

$$\gamma = \gamma^\dagger = i^{\frac{d-2}{2}} \gamma^{d-1\dagger}\dots\gamma^{0\dagger} \tag{11.50}$$

together with the unitarity of the γ-matrices to determinant the coefficient of proportionality. We have defined the Lorentzian ϵ tensor here to obey $\epsilon^{01\dots d-1} = -1$.[102] For the Euclidean γ matrices the only difference is that $\gamma^0_E = i\gamma^0$, so in Euclidean signature we instead have

$$Tr\left(\gamma\gamma^{\mu_1}\dots\gamma^{\mu_d}\right) = 2^{\frac{d}{2}} i^{\frac{d}{2}} \epsilon_E^{\mu_1\dots\mu_d}. \tag{11.51}$$

Here we have defined a Euclidean ϵ tensor obeying $\epsilon_E^{0\dots d-1} = 1$. Finally we note that in the flavor trace, since the covariant derivatives are all contracted with the ϵ tensor we can replace neighboring pairs by field strengths using

$$[D_\mu, D_\nu] = -iF^a_{\mu\nu}\tau_a. \tag{11.52}$$

Thus we at last have

$$\mathcal{A}(x) = -\frac{\Omega(x)}{2^{d-1}\pi^{d/2}\left(\frac{d}{2}!\right)} \mathrm{Tr}\left(\tau\tau_{a_1}\dots\tau_{a_{d/2}}\right) \epsilon_E^{\mu_1\dots\mu_d} F^{a_1}_{\mu_1\mu_2}(x)\dots F^{a_{d/2}}_{\mu_{d-1}\mu_d}(x), \tag{11.53}$$

and in particular for $d = 2$ we have

$$\mathcal{A} = -\frac{\Omega}{2\pi} \mathrm{Tr}\left(\tau\tau_a\right) \epsilon_E^{\mu\nu} F^a_{\mu\nu} \tag{11.54}$$

and for $d = 4$ we have

$$\mathcal{A} = -\frac{\Omega}{16\pi^2} \mathrm{Tr}\left(\tau\tau_a\tau_b\right) \epsilon_E^{\mu\nu\alpha\beta} F^a_{\mu\nu} F^b_{\alpha\beta}. \tag{11.55}$$

To continue these expressions back to Lorentzian signature we use[103]

$$\begin{aligned} dt_E &= idt \\ F^{a,E}_{0i} &= -iF^a_{0i} \\ \epsilon_E^{\mu_1\dots\mu_d} &= -\epsilon^{\mu_1\dots\mu_d}, \end{aligned} \tag{11.56}$$

[101] This step is where Euclidean signature is essential, in Lorentzian signature we would have had a divergent integral that required some subtle analytic continuation to justify.

[102] This is because the ϵ tensor naturally has indices down, as we will see in the next section, and raising the 0 index in Lorentzian signature flips the sign.

[103] The continuation of dt_E is relevant here because what are really comparing is $\int d^dx\mathcal{A}$ in Euclidean and Lorentzian signature, and the continuation for F_{0i} arises because $F_{0i}dx^0dx^i$ should be the same in Euclidean and Lorentzian signature.

so the Lorentzian anomaly is

$$\mathcal{A} = \frac{\Omega(x)}{2^{d-1}\pi^{d/2}(\frac{d}{2}!)} \mathrm{Tr}\left(\tau\tau_{a_1}\dots\tau_{a_{d/2}}\right)\epsilon^{\mu_1\dots\mu_d} F^{a_1}_{\mu_1\mu_2}(x)\dots F^{a_{d/2}}_{\mu_{d-1}\mu_d}(x) \tag{11.57}$$

in general,

$$\mathcal{A} = \frac{\Omega}{2\pi}\mathrm{Tr}\left(\tau\tau_a\right)\epsilon^{\mu\nu}F^a_{\mu\nu} \tag{11.58}$$

for $d = 2$, and

$$\mathcal{A} = \frac{\Omega}{16\pi^2}\mathrm{Tr}\left(\tau\tau_a\tau_b\right)\epsilon^{\mu\nu\alpha\beta}F^a_{\mu\nu}F^b_{\alpha\beta}. \tag{11.59}$$

for $d = 4$. Note that the form of the anomaly for $d = 2$ is the same as we got in our simple scalar example!

There is a useful alternative presentation of this anomaly as a violation of the current conservation law in the presence of a background gauge field. Since the axial current appears in the Lagrangian as

$$\mathcal{L} \supset B_\mu J^\mu, \tag{11.60}$$

and the gauge transformation is

$$\mathcal{L}' = \mathcal{L} + \partial_\mu\Omega J^\mu = -\Omega\partial_\mu J^\mu + \text{total derivative}, \tag{11.61}$$

we can compute the expectation value of the divergence of the current in the presence of the background gauge field A as

$$\langle\partial_\mu J^\mu(x)\rangle = iZ[A,0]^{-1}\frac{\delta}{\delta\Omega(x)}Z[A,d\Omega]|_{\Omega=0} = -\frac{\delta}{\delta\Omega(x)}\alpha[A,e^{i\Omega}]|_{\Omega=0}. \tag{11.62}$$

Thus for the abelian anomaly we have

$$\langle\partial_\mu J^\mu(x)\rangle = -\frac{1}{2^{d-1}\pi^{d/2}(\frac{d}{2}!)}\mathrm{Tr}\left(\tau\tau_{a_1}\dots\tau_{a_{d/2}}\right)\epsilon^{\mu_1\dots\mu_d}F^{a_1}_{\mu_1\mu_2}(x)\dots F^{a_{d/2}}_{\mu_{d-1}\mu_d}(x). \tag{11.63}$$

This suggests that, although the axial symmetry is perfectly fine in the theory with no background gauge fields turned on, if we do turn them on, or even worse make them dynamical, then there may be a problem with this symmetry.

11.2.1 The fate of $U(1)_A$ in massless QCD

Let's now apply the previous discussion to massless QCD in four dimensions. As a warmup we can first briefly consider what happens to the axial symmetry in $1+1$ dimensions with a single Dirac fermion in the presence of a background electromagnetic field under which it has charge one. This is a special case of the calculation that we just did, with the result

$$\langle\partial_\mu J^\mu\rangle = -\frac{1}{2\pi}\epsilon^{\mu\nu}F_{\mu\nu}. \tag{11.64}$$

We can compute the change in total axial charge from $t = -\infty$ to $t = +\infty$ by integrating this equation over all of space:

$$Q_\infty - Q_{-\infty} = -\frac{1}{2\pi}\int d^2x\epsilon^{\mu\nu}F_{\mu\nu}. \tag{11.65}$$

By equation (11.14) the quantity on the right-hand side is an even integer, but in general it is not zero so the charge is not conserved. Thus we see that axial symmetry is broken in the presence of a sufficiently interesting background gauge field. Moreover if we now make this gauge field dynamical, then we are summing over such configurations so the axial symmetry is destroyed (except for the $\mathbb{Z}_2$ subgroup, which remains a good symmetry since we have an even integer on the right-hand side of (11.67)).

Turning now to QCD with N_f massless quarks, a few sections ago we said that the $U(1)_A$ axial symmetry of the classical Lagrangian is destroyed by anomalies. We are now in the position to see this explicitly. Indeed first viewing the $SU(3)$ gauge field as a background field, the result of the previous section with $\tau = 1$ and $\tau_a = T_a$ tells us that

$$\langle \partial_\mu J^\mu_A \rangle = -\frac{N_f}{16\pi^2} \epsilon^{\mu\nu\alpha\beta} \mathrm{Tr}\left(G_{\mu\nu} G_{\alpha\beta}\right), \tag{11.66}$$

and thus

$$Q_\infty - Q_{-\infty} = -\frac{N_f}{16\pi^2} \int d^4x \epsilon^{\mu\nu\alpha\beta} \mathrm{Tr}\left(G_{\mu\nu} G_{\alpha\beta}\right). \tag{11.67}$$

It is far from obvious, but the quantity on the right-hand side of (11.67) is also an integer (and in fact it is an integer multiple of $2N_f$). The gauge field configurations for which it doesn't vanish are called instantons, and we will construct them in the next section. Thus once we make the gauge fields dynamical, the $U(1)_A$ symmetry is destroyed (except for its $\mathbb{Z}_{2N_f}$ subgroup, as we mentioned a few sections ago).

Having shown that $U(1)_A$ is broken to $\mathbb{Z}_{2N_F}$, we should check that this mechanism does not also destroy the conservation of the psuedovector currents that gave rise to the Goldstone bosons of chiral symmetry breaking. The point is that the color generators τ_a^{color} for $SU(3)$ act on different indices than the flavor generators τ_a^{flavor} for $U(N_f)_L \times U(N_f)_R$, so we have

$$\mathrm{Tr}\left(\tau_a^{flavor} \tau_b^{color} \tau_c^{color}\right) = \mathrm{Tr}\left(\tau_a^{flavor}\right) \mathrm{Tr}\left(\tau_b^{color} \tau_c^{color}\right) = 0, \tag{11.68}$$

since for the pseudovector generators we have $\tau_a^{flavor} = \sigma_a/2$ which is traceless. This didn't happen for $U(1)_A$ since that was the part with $\tau^{flavor} \propto I$.

11.2.2 Can we improve the current?

There is one particularly subtle aspect of this discussion that is worth highlighting. This is that the quantities which have appeared on the right-hand side of the current conservation equations have all been total derivatives,

$$\langle \partial_\mu J^\mu \rangle = \partial_\mu K^\mu. \tag{11.69}$$

For example for (11.64) we have

$$K^\mu = -\frac{1}{\pi} \epsilon^{\mu\nu} A_\nu. \tag{11.70}$$

Therefore it is quite tempting to define an "improved" current

$$\widetilde{J}^\mu \equiv J^\mu - K^\mu \tag{11.71}$$

so that

$$\partial_\mu \widetilde{J}^\mu = 0. \tag{11.72}$$

Why does this not restore the symmetry? An immediate warning sign is that K^μ is not gauge-invariant, but this is not necessarily fatal since to have a symmetry we really only need to have the charge

$$\widetilde{Q} = \int_\infty^\infty dx \widetilde{J}^0. \tag{11.73}$$

So the crucial question is whether or not this charge is gauge-invariant. Under a gauge transformation we have

$$\widetilde{J}^{0'} = \widetilde{J}^0 + \frac{1}{\pi} \partial_x \Omega, \tag{11.74}$$

so the charge will be gauge invariant if and only if

$$\int_{-\infty}^\infty dx \frac{1}{\pi} \partial_x \Omega = \frac{1}{\pi} \left(\Omega(\infty) - \Omega(-\infty)\right) \tag{11.75}$$

vanishes. Does it? Here it is crucial to decide whether we are treating the gauge group of electromagnetism as $U(1)$ or $\mathbb{R}$. If the latter, then we must have $\Omega \to 0$ at spatial infinity to get a gauge transformation we should quotient by. So indeed the charge is gauge-invariant, and we have a valid symmetry.[104] So when the gauge group is $\mathbb{R}$ this symmetry actually gives a counterexample to Noether's theorem: there is a continuous global symmetry but no local conserved current (this was pointed out by myself and Hirosi Ooguri a few years ago). On the other hand if the gauge group is $U(1)$, then we only need to have $e^{i\Omega} = 1$ at the spatial boundaries. The right-hand side of (11.75) thus does not need to vanish, it can be an arbitrary even integer. Therefore in the $U(1)$ case no improved symmetry can be defined and the axial symmetry is indeed destroyed (or more carefully broken to $\mathbb{Z}_2$).

This discussion has a direct parallel for $U(1)_A$ symmetry in massless QCD. You showed on the homework back in section 3 that the right-hand side of (11.66) is indeed a total derivative. We will discuss the gauge transformation properties of K^μ in the next section, but indeed it turns out that no gauge-invariant charge can be defined so the symmetry is lost (except for its $\mathbb{Z}_{2N_f}$ subgroup).

11.2.3 Anomaly matching and the neutral pion

Another important application of our anomaly result, and in fact the first anomaly to be discovered, arises when we turn on a background electromagnetic field for massless QCD. Focusing on $N_f = 2$, the electromagnetic gauge transformation is

$$\begin{aligned} u' &= e^{\frac{2i}{3}\Omega_e} u \\ d' &= e^{-\frac{i}{3}\Omega_e} d. \end{aligned} \tag{11.76}$$

The symmetry we are interested is the one generated by the neutral pion current

$$J^\mu_{P,3} = \overline{q} T_a \gamma \gamma^\mu q, \tag{11.77}$$

which acts on the quark fields as

$$\begin{aligned} u' &= e^{i\Omega\gamma/2} u \\ d' &= e^{-i\Omega\gamma/2} d. \end{aligned} \tag{11.78}$$

In our anomaly calculation we thus have

$$\tau = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix} \qquad \tau_e = \begin{pmatrix} \frac{2}{3} & 0 \\ 0 & -\frac{1}{3} \end{pmatrix}, \tag{11.79}$$

and therefore

$$\mathrm{Tr}\left(\tau \tau_e^2\right) = \frac{N_c}{6}, \tag{11.80}$$

where $N_c = 3$ is the number of colors (in the anomaly calculation color is just another flavor). As we already discussed there is no anomaly involving the QCD gauge field, so the full anomaly is

$$\mathcal{A} = \frac{N_c \Omega}{96\pi^2} \epsilon^{\mu\nu\alpha\beta} F_{\mu\nu} F_{\alpha\beta}, \tag{11.81}$$

with $F_{\mu\nu}$ being the usual Maxwell field strength. This is called the **ABJ anomaly**, for Adler, Bell, and Jackiw.[105]

Now let's consider what happens in this theory at low energies. Due to confinement all of the quarks are bound into hadrons, most of which are heavy and thus can be integrated out, but because of chiral symmetry

[104] In this case the integer that measured charge non-conservation in the previous section is actually zero since no magnetic flux is allowed for an $\mathbb{R}$ gauge field.

[105] When I arrived at MIT my office was in between those of Goldstone and Jackiw, which was somewhat intimidating...

breaking there are three massless pion fields ξ^a. The first two are charged under electromagnetism, but ξ^3 is neutral since the current (11.77) is invariant under (11.76). Since the anomaly (11.81) is a feature of the full partition function, the low-energy pion theory must have the same gauge non-invariance. And in fact we have already seen how to accommodate this kind of thing in the Goldstone theory in our first example of an anomaly above: we simply need the low-energy chiral Lagrangian to include a term

$$\mathcal{L}_{chiral} \supset \frac{N_c}{192\pi^2}\xi^3\epsilon^{\mu\nu\alpha\beta}F_{\mu\nu}F_{\alpha\beta}, \tag{11.82}$$

since then the pion gauge transformation[106]

$$\xi^{3\prime} = \xi^3 + 2\Omega \tag{11.84}$$

will reproduce the form of the anomaly. This is our first concrete example of anomaly matching. Typically this term is rewritten using the canonically normalized field $\pi^0 = \xi^3/f_\pi$, in which case it has the form[107]

$$\mathcal{L}_{chiral} \supset \frac{N_c}{192\pi^2 f_\pi}\pi^0\epsilon^{\mu\nu\alpha\beta}F_{\mu\nu}F_{\alpha\beta}. \tag{11.85}$$

This term is somewhat surprising from the point of view of the chiral effective theory, since it does not have any derivatives acting on π^0. It therefore cannot arise from the chiral Lagrangian we constructed before, since that always has at least two derivatives on $U = e^{i\xi^a T_a}$. In fact it arises from the Wess-Zumino-Witten term that we did not have time to discuss. Historically this was important because in real QCD with massive quarks this term gives rise to a decay rate for $\pi^0 \to \gamma\gamma$ which is of order

$$\Gamma \sim \frac{N_c^2\alpha^2 m_\pi^3}{f_\pi^2}, \tag{11.86}$$

where $\alpha \approx 1/137$ is the electromagnetic fine structure constant and the power of m_π^3 is supplied by dimensional analysis. Including the $O(1)$ factors it is

$$\Gamma \approx \frac{N_c^2\alpha^2 m_\pi^3}{2304\pi^3 f_\pi^2} \approx \left(\frac{N_c}{3}\right)^2 \times 1.2 \times 10^{16}\mathrm{s}^{-1}, \tag{11.87}$$

which is in good agreement with experiment. If we have already measured m_π and f_π then it tells us the value of N_c, while if we have already inferred $N_c = 3$ from the existence of baryons then we can use it to measure f_π (this is why f_π is called the pion decay constant). The way that this anomaly was originally discovered was that the naive chiral theory not including the WZW term predicted a π^0 decay rate that was much smaller than the observed value, so theorists thought hard to come up with an explanation of why.

11.3 General chiral anomaly

We now turn to calculating the full flavor anomaly, which in the theory (11.26) involved the symmetry $U(2N_f)$. In fact however there is no reason to restrict to an even number of left-handed spinors ψ_L^i, so to be fully general we will just take $i = 1, 2, \ldots N_L$ and discuss a $U(N_L)$ flavor symmetry (the previous case is $N_L = 2N_f$). We will turn on background gauge fields for all of $U(N_L)$. The Fujikawa approach is not so pleasant here, so we will instead use the traditional perturbative method by which such anomalies were first understood.[108] The idea is to directly study the conservation equation for the currents

$$J_a^\mu = -\overline{\psi}_L\tau_a\gamma^\mu\psi_L \tag{11.88}$$

[106]This transformation follows from the chiral effective theory flavor transformation

$$e^{i\xi^{3\prime}T_3} = e^{i\Omega T_3}e^{i\xi^3 T_3}e^{i\Omega T_3}. \tag{11.83}$$

[107]Beware that this formula is written in my convention where $f_\pi \approx 46$MeV, if you compare to something else you need to take this into account.

[108]The Fujikawa approach is still possible, but one has to directly regulate the Pfaffian arising from the Gaussian path integral and there are some subtle points to consider.

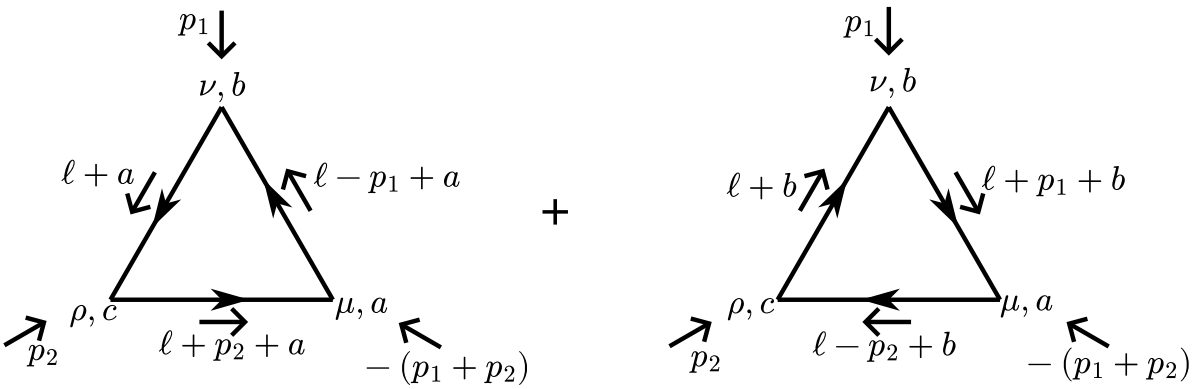

Figure 26: The two triangle diagrams computing the three-point function of a fermion current, with all external and internal momenta labeled.

in the presence of the background gauge fields. When the background fields are small we can expand the expectation value of one of the currents as

$$\langle J_a^\mu(x)\rangle = i\int d^dy A_\nu^b(y)\langle T J_a^\mu(x) J_b^\nu(y)\rangle - \frac{1}{2}\int d^dy d^dz A_\nu^b(y) A_\rho^c(z)\langle T J_a^\mu(x) J_b^\nu(y) J_c^\rho(z)\rangle + \dots, \tag{11.89}$$

so to extract any possible anomalies we "just" need to calculate these current correlators in free field theory and then take the divergence to see what we get. The reason this is nontrivial however is that the integrals on the right-hand side include points where the current locations coincide, so in particular contact terms, meaning δ-function type singularities in the correlators, can contribute. Contact terms are notoriously scheme-dependent, which means that we may (and in fact will) encounter some subtlety in computing them. Indeed the term which is linear in A leads to an anomaly in $d = 2$ that generalizes the abelian one we found above while the term which is quadratic in A leads to an anomaly in $d = 4$. We will here focus on $d = 4$, so the rest of this subsection will consist of an analysis of contact terms in the current three-point function.

In free field theory there are two ways of contracting the currents (excluding self-contractions that do not contribute to the renormalized currents), as shown in figure (26). The correlation function is

$$\begin{aligned}\langle T J_a^\mu(x) J_b^\nu(y) J_c^\rho(z)\rangle =& \operatorname{Tr}\left(S_F(x-y)\gamma^\nu\tau_b S_F(y-z)\gamma^\rho\tau_c S_F(z-x)\gamma^\mu\tau_a P_L\right)\\ &+ \operatorname{Tr}\left(S_F(x-z)\gamma^\rho\tau_c S_F(z-y)\gamma^\nu T_b S_F(y-x)\gamma^\mu\tau_a P_L\right),\end{aligned} \tag{11.90}$$

which we can write in momentum space as

$$\langle T J_a^\mu(x) J_b^\nu(y) J_c^\rho(z)\rangle = \int \frac{d^dp_1}{(2\pi)^d}\frac{d^dp_1}{(2\pi)^d} J_{abc}^{\mu\nu\rho}(p_1,p_2) e^{ip_1\cdot y + ip_2\cdot z - i(p_1+p_2)\cdot x} \tag{11.91}$$

with

$$\begin{aligned}J_{abc}^{\mu\nu\rho}(p_1,p_2) = -i\int\frac{d^d\ell}{(2\pi)^d}\Bigg[&\operatorname{Tr}\left(\tau_a\tau_b\tau_c\right)\frac{\operatorname{Tr}\left((\not\ell - \not p_1 + \not a)\gamma^\nu(\not\ell + \not a)\gamma^\rho(\not\ell + \not p_2 + \not a)\gamma^\mu P_L\right)}{((\ell - p_1 + a)^2 - i\epsilon)((\ell + a)^2 - i\epsilon)((\ell + p_2 + a)^2 - i\epsilon)}\\ &+ \operatorname{Tr}\left(\tau_b\tau_a\tau_c\right)\frac{\operatorname{Tr}\left((\not\ell - \not p_2 + \not b)\gamma^\rho(\not\ell + \not b)\gamma^\nu(\not\ell + \not p_1 + \not b)\gamma^\mu P_L\right)}{((\ell - p_2 + b)^2 - i\epsilon)((\ell + b)^2 - i\epsilon)((\ell + p_1 + b)^2 - i\epsilon)}\Bigg].\end{aligned} \tag{11.92}$$

We have somewhat mysteriously included here independent shifts a^μ and b^μ of the loop momenta in the two terms; naively these should have no effect since the current three-point function (11.90) is finite, but the two

terms separately are linearly divergent at large ℓ so a shift can have a nontrivial effect on the finite part of each. This linear divergence does cancel between the two terms, but the residual finite part still depends on these shifts as we will see. As a toy model of this, we can consider the definite integral

$$\int_0^\Lambda dx \frac{x}{\sqrt{x^2+1}} = \sqrt{\Lambda^2+1} - 1 = \Lambda - 1 + O(1/\Lambda). \tag{11.93}$$

If we shift the integration variable as $x = x' - a$ without changing the cutoff scheme, we instead have

$$\int_a^\Lambda dx' \frac{(x'-a)}{\sqrt{(x'-a)^2+1}} = \sqrt{(\Lambda-a)^2+1} - 1 = \Lambda - (1+a) + O(1/\Lambda). \tag{11.94}$$

Thus we see that the finite part can depend on a seemingly innocuous shift, and if we take the difference of two such integrals we get a finite integral whose result depends on the choices we made for each. We can therefore view a and b as allowing us to simultaneously consider a wide variety of cutoff schemes; we will interpret them later in terms of extra local counterterms $\beta[A]$. To compute the divergence of the first current we should contract $J^{\mu\nu\rho}_{abc}$ with $-i(p_1^\mu + p_2^\mu)$, and by noting that

$$p_1 + p_2 = (\ell + p_2 + a) - (\ell - p_1 + a) = (\ell + p_1 + b) - (\ell - p_2 + b) \tag{11.95}$$

we can use this to cancel one of the denominators at the cost of doubling the number of terms:

$$\begin{aligned} -i(p_1+p_2)_\mu J^{\mu\nu\rho}_{abc} = -i \int \frac{d^d\ell}{(2\pi)^d} \Bigg[&\mathrm{Tr}\,(\tau_a\tau_b\tau_c) \left(\frac{\mathrm{Tr}\left((\ell\!\!\!/ - p\!\!\!/_1 + a\!\!\!/)\gamma^\nu(\ell\!\!\!/ + a\!\!\!/)\gamma^\rho P_L\right)}{(\ell - p_1 + a)^2(\ell + a)^2} - \frac{\mathrm{Tr}\left((\ell\!\!\!/ + a\!\!\!/)\gamma^\rho(\ell\!\!\!/ + p\!\!\!/_2 + a\!\!\!/)\gamma^\nu P_L\right)}{(\ell + a)^2(\ell + p_2 + a)^2} \right) \\ &+ \mathrm{Tr}\,(\tau_b\tau_a\tau_c) \left(\frac{\mathrm{Tr}\left((\ell\!\!\!/ - p\!\!\!/_2 + b\!\!\!/)\gamma^\rho(\ell\!\!\!/ + b\!\!\!/)\gamma^\nu P_L\right)}{(\ell - p_2 + b)^2(\ell + b)^2} - \frac{\mathrm{Tr}\left((\ell\!\!\!/ + b\!\!\!/)\gamma^\nu(\ell\!\!\!/ + p\!\!\!/_1 + b\!\!\!/)\gamma^\rho P_L\right)}{(\ell + b)^2(\ell + p_1 + b)^2} \right) \Bigg]. \end{aligned} \tag{11.96}$$

Here we have also done a Wick rotation $\ell^0 = i\ell^0_E$ and then dropped the $i\epsilon$ factors. To proceed further it is useful to write

$$\mathrm{Tr}\,(\tau_a\tau_b\tau_c) = D_{abc} + \frac{i}{2} C(\alpha) C_{abc}, \tag{11.97}$$

where

$$D_{abc} \equiv \frac{1}{2}\mathrm{Tr}\,(\{\tau_a, \tau_b\}\tau_c) \tag{11.98}$$

are called the **anomaly coefficients**. The terms proportional to C_{abc} have a simple interpretation: they are there to provide the terms on the right-hand side of the Ward identity

$$\frac{\partial}{\partial x^\mu}\langle T J^\mu_a(x) J^\nu_b(y) J^\rho_c(z)\rangle = i\delta^d(x-y) C^d{}_{ab} \langle T J^\nu_d(y) J^\rho_c(z)\rangle + i\delta^d(x-z) C^d{}_{ac} \langle T J^\nu_b(y) J^\rho_d(z)\rangle, \tag{11.99}$$

and thus do not really constitute a violation of the symmetry. In fact if we move them to the left-hand side of the current conservation equation they have the beautiful effect of converting the non-covariant derivative $\partial_\mu J^\mu_a$ into a covariant derivative

$$D_\mu J^\mu_a = \partial_\mu J^\mu_a - C^c{}_{ba} A^b_\mu J^\mu_c, \tag{11.100}$$

which anyways is the thing that should vanish if the symmetry is preserved. Therefore if there is an anomaly

it comes from the D_{abc} part of this equation:

$$\begin{aligned}-i(p_1+p_2)_\mu J^{\mu\nu\rho}_{abc} \supset & - iD_{abc}\int\frac{d^d\ell}{(2\pi)^d}\left[\frac{\mathrm{Tr}\left((\ell\!\!\!/-p\!\!\!/_1+a\!\!\!/)\gamma^\nu(\ell\!\!\!/+a\!\!\!/)\gamma^\rho P_L\right)}{(\ell-p_1+a)^2(\ell+a)^2}-\frac{\mathrm{Tr}\left((\ell\!\!\!/+a\!\!\!/)\gamma^\rho(\ell\!\!\!/+p\!\!\!/_2+a\!\!\!/)\gamma^\nu P_L\right)}{(\ell+a)^2(\ell+p_2+a)^2}\right.\\ & \left.+\frac{\mathrm{Tr}\left((\ell\!\!\!/-p\!\!\!/_2+b\!\!\!/)\gamma^\rho(\ell\!\!\!/+b\!\!\!/)\gamma^\nu P_L\right)}{(\ell-p_2+b)^2(\ell+b)^2}-\frac{\mathrm{Tr}\left((\ell\!\!\!/+b\!\!\!/)\gamma^\nu(\ell\!\!\!/+p\!\!\!/_1+b\!\!\!/)\gamma^\rho P_L\right)}{(\ell+b)^2(\ell+p_1+b)^2}\right]\\ = & - iD_{abc}\Big[\mathrm{Tr}\left(\gamma^\alpha\gamma^\nu\gamma^\beta\gamma^\rho P_L\right)I_{\alpha\beta}(a-b-p_1,b,p_1+b)\\ & +\mathrm{Tr}\left(\gamma^\alpha\gamma^\rho\gamma^\beta\gamma^\nu P_L\right)I_{\alpha\beta}(b-a-p_2,a,p_2+a)\Big],\end{aligned}\tag{11.101}$$

where in the second equality we recognized (following Weinberg's book) that the first and fourth terms are the difference of the same integral with a shifted argument and so are the second and third. The integral is defined as

$$I_{\alpha\beta}(k,c,d)\equiv\int\frac{d^d\ell}{(2\pi)^d}\left(f_{\alpha\beta}(\ell+k,c,d)-f_{\alpha\beta}(\ell,c,d)\right)\tag{11.102}$$

with

$$f_{\alpha\beta}(k,c,d)\equiv\frac{(k+c)_\alpha(k+d)_\beta}{(k+c)^2(k+d)^2}.\tag{11.103}$$

This integral would vanish if the two terms were separately convergent, but for $d=4$ they are both quadratically divergent. One approach we could use to evaluate the integral is to combine the terms into one fraction and then evaluate the integral using dimensional regularization. This works, but it is unpleasant since the denominator has four factors so we need an integral over four Feynman parameters. Instead we will use a clever trick from Weinberg's book, which is to Taylor expand the integrand in k. The zeroth order term cancels, removing the quadratic divergence, and the other terms are all total derivatives so they can be reduced to surface integrals at some large sphere $\ell^2=\Lambda^2$ (remember we have already Wick rotated to Euclidean signature). For $d=4$ this sphere has a volume of order Λ^3, so we only need to keep terms that are at most of order $1/\Lambda^3$. These can only be generated by the first two terms in the Taylor series, so for $d=4$ we have

$$\begin{aligned}I_{\alpha\beta}(k,c,d) &= \int\frac{d^4\ell}{(2\pi)^4}k^\lambda\frac{\partial}{\partial\ell^\lambda}\left(f_{\alpha\beta}+\frac{1}{2}k^\sigma\frac{\partial f_{\alpha\beta}}{\partial\ell^\sigma}\right)\\ &= \lim_{\Lambda\to\infty}\frac{1}{16\pi^4\Lambda^3}\int d\Omega_3 k^\lambda n_\lambda\left(f_{\alpha\beta}+\frac{1}{2}k^\sigma\frac{\partial f_{\alpha\beta}}{\partial\ell^\sigma}\right)\Big|_{\ell^2=\Lambda^2}\end{aligned}\tag{11.104}$$

Setting $\ell^\mu=\Lambda n^\mu$ and expanding the right-hand side in $1/\Lambda$, we can use the integrals

$$\int d\Omega_3 n_\mu n_\nu=\frac{\pi^2}{2}\eta_{\mu\nu}\tag{11.105}$$

$$\int d\Omega_3 n_\mu n_\nu n_\alpha n_\beta=\frac{\pi^2}{12}\left(\eta^{\mu\nu}\eta^{\alpha\beta}+\eta^{\mu\alpha}\eta^{\nu\beta}+\eta^{\mu\beta}\eta^{\nu\alpha}\right),\tag{11.106}$$

which are determined by symmetry up to the constant multiple that can be determined by contracting with the Euclidean metric and using $\int d\Omega_3=2\pi^2$, to find

$$I_{\alpha\beta}(k,c,d)=\frac{1}{96\pi^2}\left[2k_\alpha d_\beta+2k_\beta c_\alpha-k_\alpha c_\beta-k_\beta d_\alpha+k_\alpha k_\beta-\eta_{\alpha\beta}k\cdot(k+c+d)\right].\tag{11.107}$$

Returning the the current conservation equation (11.101), for the terms not involving γ we can use the symmetry of $\mathrm{Tr}\left(\gamma^\alpha\gamma^\nu\gamma^\beta\gamma^\rho\right)$ under $\alpha\leftrightarrow\beta$ and $\nu\leftrightarrow\rho$ to see that the possible anomaly contribution is proportional to

$$I_{\alpha\beta}(a-b-p_1,b,p_1+b)+I_{\beta\alpha}(a-b-p_1,b,p_1+b)+I_{\alpha\beta}(b-a-p_2,a,p_2+a)+I_{\beta\alpha}(b-a-p_2,a,p_2+a).\tag{11.108}$$

Using (11.107) it is not hard to check that this vanishes if and only if we take

$$b = -a. \tag{11.109}$$

We will see in a moment that this choice ensures that the nonchiral part of the divergence with respect to the other two currents in the correlator also vanishes. Proceeding with the part involving γ, using (11.49) we have

$$\begin{aligned} -i(p_1+p_2)_\mu J^{\mu\nu\rho}_{abc} \supset & -2D_{abc}\epsilon^{\alpha\nu\beta\rho}\left(I_{\alpha\beta}(2a-p_1,-a,p_1-a) - I_{\alpha\beta}(-2a-p_2,a,p_2+a)\right) \\ = & -\frac{1}{8\pi^2}D_{abc}\epsilon^{\alpha\nu\beta\rho}a_\alpha(p_1+p_2)_\beta. \end{aligned} \tag{11.110}$$

From this result for the anomaly you may simply be tempted to choose

$$a \propto p_1 + p_2, \tag{11.111}$$

which would cause it to vanish. We need to be careful however, since the regulator we used was not symmetric between the three currents in the correlator. Looking at figure 26, to use this result for the other currents we should change a and b so that the momentum of the propagator opposite to the current we are differentiating is $\ell + a$ in the first diagram and $\ell + b$ in the second. If we want to take the divergence of $J^\nu_b(y)$ using the same regulator (i.e. without shifting the loop momentum) we should therefore set

$$\begin{aligned} a' &= a + p_2 \\ b' &= b - p_2, \end{aligned} \tag{11.112}$$

while if we want to take the divergence of $J^\rho_c(z)$ we should set

$$\begin{aligned} a'' &= a - p_1 \\ b'' &= b + p_1. \end{aligned} \tag{11.113}$$

We have written the b redefinitions to check that these redefinitions preserve the condition $b = -a$, which indeed they do, so we have succeeded in removing the nonchiral part of the anomaly from all three currents. Having done so, we can by symmetry write the current conservation for all three currents as

$$\begin{aligned} -i(p_1+p_2)_\mu J^{\mu\nu\rho}_{abc} &\supset -\frac{1}{8\pi^2}D_{abc}\epsilon^{\alpha\nu\beta\rho}a_\alpha(p_1+p_2)_\beta \\ ip_{1,\nu} J^{\mu\nu\rho}_{abc} &\supset -\frac{1}{8\pi^2}D_{abc}\epsilon^{\alpha\rho\beta\mu}(a+p_2)_\alpha(-p_1)_\beta \\ ip_{2,\rho} J^{\mu\nu\rho}_{abc} &\supset -\frac{1}{8\pi^2}D_{abc}\epsilon^{\alpha\mu\beta\nu}(a-p_1)_\alpha(-p_2)_\beta. \end{aligned} \tag{11.114}$$

To remove the chiral part from $J^\nu_b(y)$ we apparently should set

$$a + p_2 \propto p_1, \tag{11.115}$$

while to remove it from $J^\rho_c(z)$ we should set

$$a - p_1 \propto p_2. \tag{11.116}$$

Unfortunately there is no choice of a that achieves (11.111), (11.115), and (11.116) at the same time. For example if we choose

$$a = -(p_1 + p_2) \tag{11.117}$$

we can arrange for the first two currents to be conserved but then (11.116) does not hold, while if we choose

$$a = p_1 - p_2 \tag{11.118}$$

then the second two currents are conserved but the first is not. Similarly if we choose

$$a = p_1 + p_2 \tag{11.119}$$

then the first and third currents are conserved but the second is not. This is the main lesson: *For any three currents for which $D_{abc} \neq 0$, at least one of the three must be anomalous.* In particular if any of these currents generate symmetries that we wish to gauge, then we *must* choose a_μ to preserve the conservation of those currents.

Once we make a choice for a_μ we finally compute the current expectation value (11.89) in the presence of background gauge fields. For example say that J_b^ν and J_c^ρ are gauge currents, so that we much choose (11.118). We then have

$$\langle D_\mu J_a^\mu \rangle = -\frac{1}{8\pi^2} D_{abc} \epsilon^{\alpha\nu\beta\rho} \partial_\alpha A_\nu^b \partial_\beta A_\rho^c + \ldots, \tag{11.120}$$

where the neglected terms are higher order in A and thus would in principle require a calculation of more diagrams. But in fact since we are regulating to preserve gauge invariance there is a unique possible answer:

$$\langle D_\mu J_a^\mu \rangle = -\frac{1}{32\pi^2} D_{abc} \epsilon^{\alpha\nu\beta\rho} F_{\alpha\nu}^b F_{\beta\rho}^c. \tag{11.121}$$

Alternatively if none of the currents are gauged then it is perhaps more natural to choose an a that treats all currents symmetrically. This means that we want

$$a = \alpha p_1 + \beta p_2 \qquad a' = \alpha p_2 - \beta(p_1 + p_2) \qquad a'' = -\alpha(p_1 + p_2) + \beta p_1, \tag{11.122}$$

which holds for $\alpha = -\beta = 1/3$ and thus

$$a = \frac{1}{3}(p_1 - p_2). \tag{11.123}$$

With this choice we have

$$\langle D_\mu J_a^\mu \rangle = -\frac{1}{24\pi^2} D_{abc} \epsilon^{\alpha\nu\beta\rho} \partial_\alpha A_\nu^b \partial_\beta A_\rho^c + \ldots. \tag{11.124}$$

The higher order terms can no longer be determined using gauge invariance since this regulator does not preserve it, but they can still be determined from the quadratic part by solving the Wess-Zumino consistency conditions (11.6).

Let's quickly check this calculation by comparing to our result for the abelian anomaly obtained by the Fujikawa method. To bring back the Dirac structure we have N_f left-handed spinors ψ_L, on which the axial generator acts as τ and the gauge generators act as τ_a, and N_f left handed spinors $B^* \psi_R^*$ on which the the axial generator acts as τ^* and the gauge generators act as $-\tau_a^*$. We thus have

$$D_{abA} = \frac{1}{2}\mathrm{Tr}_{N_f}\left(\{\tau, \tau_a\}\tau_b\right) + \frac{1}{2}\mathrm{Tr}_{N_f}\left(\{\tau^*, \tau_a^*\}\tau_b^*\right) = 2\mathrm{Tr}_{N_f}\left(\tau\tau_a\tau_b\right), \tag{11.125}$$

where we have used $[\tau, \tau_a] = 0$ and also that τ and τ_a are hermitian. Here Tr_{N_f} indicates the trace over the N_f-dimensional representation of τ and τ_a on ψ_L, as opposed to the $2N_f$-dimensional trace appearing in the definition of D_{abA}. Thus (11.121) agrees with (11.63).

11.4 Cancellation of gauge anomalies in the standard model

Having worked so hard to compute the general chiral anomaly, we should do something with it. The first thing we will do is make sure that anomalies do not ruin the gauge symmetries of the standard model of particle physics. In other words we need to check that

$$D_{abc} = \frac{1}{2}\mathrm{Tr}\left(\{\tau_a, \tau_b\}\tau_c\right) = 0 \tag{11.126}$$

for all gauge generators. In showing this it is useful to first note that if there is an invertible matrix such that

$$(i\tau_a)^* = S(i\tau_a)S^{-1}, \tag{11.127}$$

for all the generators τ_a in some Lie algebra $\mathfrak{h} \subseteq U(N_L)$, meaning that the fermion representation of this subalgebra (and the subgroup it generates) is equivalent to its complex conjugate, then we have

$$D_{abc} = 0 \tag{11.128}$$

for all $\tau_a, \tau_b, \tau_c \in \mathfrak{h}$. This is because since τ_a is hermitian, (11.127) is equivalent to

$$\tau_a^T = -S\tau_a S^{-1}, \tag{11.129}$$

from which we have

$$\mathrm{Tr}\left((\tau_a\tau_b + \tau_b\tau_a)\tau_c\right) = \mathrm{Tr}\left((\tau_a^T\tau_b^T + \tau_b^T\tau_a^T)\tau_c^T\right) = -\mathrm{Tr}\left((\tau_a\tau_b + \tau_b\tau_a)\tau_c\right). \tag{11.130}$$

A representation for which we can find such an S is called either **real** or **pseudoreal**, depending on whether or not there is a basis where $i\tau_a$ is real. In particular this is true for the fundamental representation of $SU(2)$, where we can choose S to be the ϵ tensor.

Turning now to the standard model, the vanishing of D_{abc} is automatic for three $SU(3)$ generators since these are all vector generators. More concretely each pair of left-handed quarks which live in the same Dirac spinor transform in a real representation $3 \oplus \overline{3}$ of $SU(3)$ (the matrix S just exchanges the two left-handed spinors). The anomaly also vanishes for three $SU(2)$ generators since the fundamental of $SU(2)$ is pseudoreal. Moreover since D_{abc} is a trace of a product of generators, and thus a singlet under the adjoint action of the gauge group (which acts on each generator as $D(g)\tau_a D(g^{-1})$ and thus cancels in the trace), the anomaly can only be nonzero if the tensor product of the three generators contains a singlet. This means that D_{abc} must also vanish if we have one $SU(3)$ or one $SU(2)$ generator. Therefore the only anomaly coefficients we actually need to compute are those with zero or two $SU(3)$ or $SU(2)$ generators. Here are the computations:

- $SU(3) - SU(3) - U(1)$: we have

$$D_{abY} = 3\mathrm{Tr}\left(T_a T_b\right)\left(\frac{2}{6} - \frac{2}{3} + \frac{1}{3}\right) = 0, \tag{11.131}$$

  where the 3 comes from the number of generations, T_a are $SU(3)$ fundamental generators, the $2/6$ comes from the left-handed quark doublet, the $-2/3$ comes from the right-handed up-type quark, and the $1/3$ comes from the right-handed down type quark.

- $SU(2) - SU(2) - U(1)$: we have

$$D_{abY} = 3\mathrm{Tr}\left(T_a T_b\right)\left(\frac{3}{6} - \frac{1}{2}\right) = 0, \tag{11.132}$$

  where $3/6$ comes from the left-handed quarks and $1/2$ comes from the left-handed leptons.

- $U(1) - U(1) - U(1)$: we have

$$D_{YYY} = 3 \times \left(6 \times \left(\frac{1}{6}\right)^3 - 3\left(\frac{2}{3}\right)^3 + 3 \times \left(\frac{1}{3}\right)^3 - 2\left(\frac{1}{2}\right)^3 + 1\right) = 0. \tag{11.133}$$

Thus we see that the standard model is indeed free of gauge anomalies!

11.5 Anomalies involving baryon and lepton number symmetry

The standard model Lagrangian has two important $U(1)$ global symmetries, baryon number and lepton number. To what extent are these broken by anomalies? These are nonchiral symmetries, so we do not need to check their anomaly with $SU(3)$. For the electroweak gauge symmetries we have:

- $SU(2) - SU(2) - U(1)_B$: we have

$$D_{abB} = 3\mathrm{Tr}(T_a T_b)\left(\frac{3}{3}\right). \tag{11.134}$$

- $U(1)_Y - U(1)_Y - U(1)_B$: we have

$$D_{YYB} = 3\left(6 \times \left(\frac{1}{6}\right)^2 \times \frac{1}{3} - 3 \times \left(\frac{2}{3}\right)^2 \times \frac{1}{3} - 3\left(\frac{1}{3}\right)^2 \times \frac{1}{3}\right) = -\frac{3}{2}. \tag{11.135}$$

- $SU(2) - SU(2) - U(1)_L$: we have

$$D_{abL} = 3\mathrm{Tr}\left(T_a T_b\right)(1) \tag{11.136}$$

- $U(1)_Y - U(1)_Y - U(1)_L$: we have

$$D_{YYL} = 3\left(2 \times \left(\frac{1}{2}\right)^2 - 1\right) = -\frac{3}{2}. \tag{11.137}$$

Note in particular that the anomalies are the same for B and L: this means that $B - L$ is a continuous global symmetry of the standard model. On the other hand $B + L$ is anomalous, which is an overall phase rotation of all the quarks and leptons. Due to the form (11.121) of the anomaly in this case, we can use such a redefinition to remove θ_2 or θ_1 from the action (but not both).

In the case of $U(1)_A$ we saw that the anomaly did not completely destroy the symmetry, instead it broke it to $\mathbb{Z}_{2N_F}$. We can ask if a similar statement is true for baryon number. Rewriting the anomaly in terms of a baryon number background gauge transformation, from (11.121) we have[109]

$$\begin{aligned}\alpha[A, e^{i\Omega}] &= -\frac{\Omega N_g}{32\pi^2}\epsilon^{\alpha\nu\beta\rho}\left(\mathrm{Tr}\left(F_{\alpha\nu}F_{\beta\rho}\right) - \frac{1}{2}B_{\alpha\nu}B_{\beta\rho}\right) \\ &= -\Omega N_g\left(n_2 - 18n_1\right),\end{aligned} \tag{11.138}$$

where we will see next time that n_2 and n_1 can be arbitrary integers on a general spacetime manifold M and n_2 can be nonzero already on $\mathbb{R}^4$. Here N_g is the number of generations. Since Ω has periodicity 2π (that is why we assigned baryon number $1/3$ to quarks), this means that in order for this to be equal to 2π times an integer we need $\Omega = 2\pi m/N_g$ where m is an integer. Thus we see that in the standard model with N_g generations baryon number is violated except for a subgroup $\mathbb{Z}_{N_g}$. Interestingly this means that with $N_g > 1$ the proton is stable, while for $N_g = 1$ it is not.

11.6 Fermion masses and decoupling

There is a somewhat puzzling feature of anomalies that is worth mentioning. The chiral fermion anomalies we discussed so far all arose in theories where those fermions do not have a bare mass term. In the abelian case this was because we considered an axial symmetry involving $e^{i\gamma\theta}$ that would be explicitly broken by a mass term, while in the standard model bare mass terms are not allowed by gauge invariance. But from the path integral point of view we interpreted the anomalies as arising from a UV-regulator problem in the

[109] In this calculation I do not include the $\mathbb{Z}_6$ quotient in the gauge group. If we include it it is trickier to identify the integer, but I think the unbroken subgroup is still $\mathbb{Z}_{N_g}$.

measure, and fermion masses should not matter for a UV regulator problem at least as long as $m \ll \Lambda$. And moreover in the standard model the quarks and leptons are indeed massive due to the Higgs mechanism, and yet they contribute to anomalies e.g. for B and L symmetry. On the other hand from the point of view of anomaly matching there is clearly something strange about having a fermion that contributes to an anomaly but whose mass can be increased arbitrarily: if we make the fermion sufficiently heavy we usually expect that it should decouple and have no effect on the infrared physics of the theory. Otherwise we would be able to use the low-energy anomaly to discover an arbitrarily heavy particle! How can we resolve this tension?

The first thing we will note is that in the case of the general chiral anomaly, a fermion which is allowed by some symmetry to have a bare mass term cannot contribute to any anomaly for that symmetry. Conceptually this is because if a mass term is allowed we can regulate the fermion determinant using something like Pauli-Villars, which does not break the symmetry. We can confirm this directly from our result for the general chiral anomaly: in our language where all fermion fields are left-handed, the most general mass term allowed by Lorentz invariance has the form

$$\mathcal{L}_{mass} = -\frac{i}{2}\psi^{i,T}\mathcal{C}M_{ij}\psi^j_L + \text{h.c.}, \tag{11.139}$$

where $i, j = 1, \ldots, N_L$ label the left-handed spinors. We can take the mass matrix M_{ij} to be symmetric since $\psi^{i,T}\mathcal{C}\psi^j$ is symmetric in i and j due to the antisymmetry of $\mathcal{C}$ and the anticommutation of the fermions. Now let's say that $H \subset U(N_L)$ is the symmetry group of the theory that we want to test for anomalies. If this mass term is invariant under H then we must have

$$\tau_a^T M + M\tau_a = 0 \tag{11.140}$$

for all generators τ_a such that $T_a \in \mathfrak{h}$. Group theoretically this equation tells us that M is an intertwiner between the representation α of H that the ψ^i_L transform under and its inverse transpose (or equivalently its conjugate $\overline{\alpha}$). We would like to apply Schur's lemma to constrain the form of M, but in general the representation α is not irreducible. We can decompose it however into a direct sum of irreducible representations:

$$\alpha = \oplus_r \alpha_r, \tag{11.141}$$

after which we can re-interpret M as a list of matrices $M^{\{rr'\}}_{n_r m_{r'}}$, with $M^{\{rr'\}}$ giving a list of intertwiners between irreducible representations r and $\overline{r'}$. We may then invoke Schur's lemma to tell us that each $M^{\{rr'\}}$ is either zero or invertible. Thus if we define the contribution of the fermions in a particular irrep r to the anomaly coefficient as

$$D^r_{abc} \equiv \frac{1}{2}\text{Tr}\left(\{\tau_a^{\{r\}}, \tau_b^{\{r\}}\}\tau_c^{\{r\}}\right), \tag{11.142}$$

then for any r and r' for which $M^{\{rr'\}} \neq 0$ we have

$$\begin{aligned} D^r_{abc} =& \frac{1}{2}\text{Tr}\left(\{M^{\{r'r\}}\tau_a^{\{r\}}(M^{\{r'r\}})^{-1}, M^{\{r'r\}}\tau_b^{\{r\}}(M^{\{r'r\}})^{-1}\}M^{\{r'r\}}\tau_c^{\{r\}}(M^{\{r'r\}})^{-1}\right) \\ =& -\frac{1}{2}\text{Tr}\left(\{\tau_a^{\{r'\}}, \tau_b^{\{r'\}}\}\tau_c^{\{r'\}}\right) \\ =& -D^{r'}_{abc}. \end{aligned} \tag{11.143}$$

Therefore if $r = r'$ the representation does not contribute, while if $r \neq r'$ then the contributions of the two irreps cancel in the full anomaly coefficient $D_{abc} = \sum_r D^r_{abc}$. This result ensures that fermions that we can decouple by taking their bare mass to infinity do not contribute to any chiral anomalies.

The situation becomes more interesting however for fermions that get their mass from the expectation value of another field, as in the standard model. Then the above argument does not work, so in general massive fermions can indeed contribute to anomalies (as they do in the standard model). These contributions still must be consistent with the decoupling principle however, and the way this happens can be subtle. For example in the standard model itself we cannot make the quarks and leptons arbitrarily massive without

also increasing the Higgs expectation value v, and this increases the masses of the W and Z bosons as well. Thus the limit that decouples the fermions is the same limit that removes any evidence of $SU(2) \times U(1)$ symmetry from the theory, except for its unbroken and non-anomalous subgroup $U(1)_e$. In particular the topologically nontrivial $SU(2) \times U(1)$ field configurations that activate the baryon and lepton number anomalies become strongly suppressed by the W and Z masses. But this is by no means the only way to resolve the tension between heavy fermions contributing to anomalies and decoupling. For example if we instead view $SU(2) \times U(1)$ as background gauge fields, then even when v is large there are three massless Goldstone bosons $\widetilde{\pi}_{\pm,0}$ arising from the Higgs modes that would be eaten by the W and Z if we gauged the symmetry. In situations of this type it is possible for fermions that are arbitrarily heavy to still contribute to anomalies that in the infrared theory are reproduced by the Goldstone boson effective theory. The reason this is not a violation of decoupling is that a low-energy observer cannot detect these anomalous terms (such as (11.85)) without creating some kind of configuration that varies over large distances in the vacuum space G/H (this is reflected by the $1/f_\pi$ in (11.85)).

12 Differential forms and the topology of gauge fields

So far we have only considered gauge fields configurations on $\mathbb{R}^d$, where we defined a gauge field A as a one-form valued in the Lie algebra $\mathfrak{g}$ of the gauge group G. This is fine as far as it goes, but to understand the full range of interesting applications of field theory we need to define gauge theory on general spacetime manifolds M. There are at least three reasons for this:

- We can have an effective field theory description that is only valid in part of space, for example if we build a piece of material with the topology of a ring. We would like to be able to study what happens in this ring without making assumptions about the rest of spacetime where the effective theory is not valid.
- Even for field theory in $\mathbb{R}^d$ we have a choice of what boundary conditions to impose at infinity, and there are interesting dynamical effects which are most easily demonstrated by imposing boundary conditions which replace $\mathbb{R}^d$ by $\mathbb{S}^d$. We will see that gauge fields on $\mathbb{S}^d$ can have topologically non-trivial gauge configurations, and also that this leads to important consequences for particles physics on $\mathbb{R}^d$.
- Once we include gravity, spacetime itself becomes dynamical so general topologies must be included.

In this section we will explain how to define gauge fields on general manifolds M, and then discuss some of the remarkable effects which can arise from this. Along the way we will introduce the idea of differential forms, which give a powerful notation for discussing the topology of gauge fields.

12.1 What is a gauge field on a manifold?

The reason the construction of a gauge field on general M is nontrivial is that manifolds are built by gluing together patches U_α, and we need to say what happens to the gauge field in the gluing. Perhaps the most obvious thing we could do is require A to be a globally defined one-form on M, but this is too restrictive since it is not compatible with our understanding that a gauge transformation should be viewed as a redundancy of description. Since a gauge transformation is a redundancy, we should allow for the gauge field in a patch U_α be related to the gauge field in a patch U_β by a nontrivial gauge transformation:[110]

$$A_\alpha = g_{\alpha\beta}\left(A_\beta - ig_{\alpha\beta}^{-1}dg_{\alpha\beta}\right)g_{\alpha\beta}^{-1}. \tag{12.1}$$

Here $g_{\alpha\beta} : U_\alpha \cap U_\beta \to G$ is a new kind of object for us: it is a "gluing" gauge transformation that tells us how to relate the gauge fields in different patches. Similarly matter fields that are charged under the gauge

[110] Note that in this equation α and β label patches, they are *not* Lorentz indices. The quantity A_α is a Lie-algebra-valued one-form defined in the patch U_α, and $dg_{\alpha\beta}$ is a one-form that is the exterior derivative of $g_{\alpha\beta}$. We will review differential forms in the next subsection.

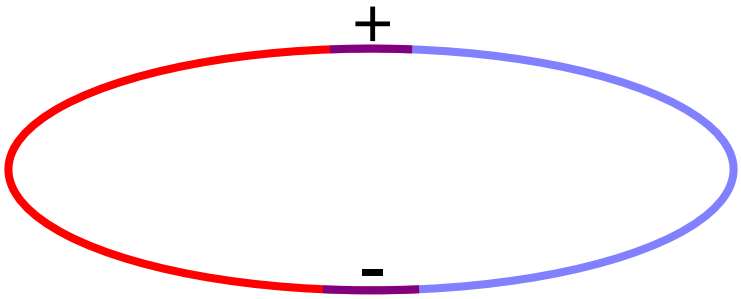

Figure 27: A simple nontrivial $\mathbb{Z}_2$ bundle over a circle. The red patch and the blue patch overlap in the two purple regions, one of which has transition function $+1$ and the other of which has transition function -1. Doing a gauge transformation by -1 in either region flips both signs, but it cannot remove the relative sign.

symmetry must obey the gluing relation

$$\phi_\alpha(x) = D(g_{\alpha\beta}(x))\phi_\beta(x). \tag{12.2}$$

There are two important conditions that gluing gauge transformations need to obey. First of all if we transition from α to β and then back to α we should get back the same local gauge field A_α and matter field ϕ_α, so we need to have

$$g_{\beta\alpha} = g_{\alpha\beta}^{-1}. \tag{12.3}$$

Secondly if we go in a triple overlap $U_\alpha \cap U_\beta \cap U_\gamma$ from α to β to γ and then back to α the same should be true:

$$g_{\alpha\gamma}g_{\gamma\beta}g_{\beta\alpha} = 1. \tag{12.4}$$

These two conditions then imply that higher overlaps also work, for example in a quadruple overlap we have

$$g_{\alpha\delta}g_{\delta\gamma}g_{\gamma\beta}g_{\beta\alpha} = (g_{\alpha\delta}g_{\delta\gamma}g_{\gamma\alpha})(g_{\alpha\gamma}g_{\gamma\beta}g_{\beta\alpha}) = 1. \tag{12.5}$$

An open cover $\{U_\alpha\}$ of M together with a set of transition functions $g_{\alpha\beta} : U_\alpha \cap U_\beta \to G$ obeying (12.3) and (12.4) is called a **principal G-bundle over M**. Two principal G-bundles $(\{U_\alpha\}, \{g_{\alpha\beta}\})$ and $(\{U'_{\alpha'}\}, \{g'_{\alpha'\beta'}\})$ over M are said to be **equivalent** if we can define transition functions $k_{\alpha'\beta} : U'_{\alpha'} \cap U_\beta \to G$ such that the union cover $\{U_\alpha\} \cup \{U'_{\alpha'}\}$ together with $g_{\alpha\beta}, k_{\alpha'\beta}, g'_{\alpha'\beta'}$ together also define a principal G-bundle over M. We won't prove it here, but all equivalence classes of G-bundles over M can be constructed using any fixed covering $\{U_\alpha\}$ with the property that the U_α are all contractible.[111] Two bundles with the same open cover $\{U_\alpha\}$ are equivalent if and only if their transition functions are related as

$$g'_{\alpha\beta} = h_\alpha g_{\alpha\beta} h_\beta^{-1} \tag{12.6}$$

for some maps $h_\alpha : U_\alpha \to G$. A principal G-bundle is said to be **trivial** if it is equivalent to a bundle that covers M with only one patch $U_\alpha = M$. If if we fix the cover, for example to some list of contractible U_α, then the bundle is trivial if and only if we can set all the transition functions to 1 by some appropriate choice of h_α. The simplest example of a nontrivial G-bundle is shown in figure 27: we take $M = \mathbb{S}^1$ and $G = \mathbb{Z}_2$, we take $\{U_\alpha\}$ to be a pair of intervals U_1, U_2 that have nontrivial intersection at both ends, and then we take g_{12} to be $+1$ in one of the intersections and -1 in the other. By choosing $h_1 = 1$ and $h_2 = -1$ (or vice versa) we can switch which intersection has the minus sign, but we can't get rid of it. This is the simplest example of a nontrivial $\mathbb{Z}_2$ gauge field, and indeed in general when G is a discrete group the definition of

[111] "Contractible" means that a space is homotopy-equivalent to a point. The thing to show is that any G bundle over a contractible space is trivial, which follows from a simple homotopy argument.

a gauge field configuration on M is precisely a principal G-bundle.[112] The other two nontrivial bundles we will study in this section are generalizations of this example to $\mathbb{S}^2$ and $\mathbb{S}^4$.

When G is continuous, i.e. has manifold dimension greater then zero, then there is more to a gauge field than just a principal G-bundle. Namely we also need to specify a one-form gauge field A_α valued in the Lie algebra $\mathfrak{g}$ in each patch such that the overlap rule (12.1) holds. A set $\{A_\alpha\}$ of such one-forms is called a **connection** on the principal G bundle defined by $(\{U_\alpha\}, \{g_{\alpha\beta}\})$, so the modern definition of a gauge field is that it is a connection on principal G-bundle. In the discrete case we have $\mathfrak{g} = 0$, so we are just left with the bundle. Similarly a matter field is defined as an assignment of a field ϕ_α to each patch U_α obeying overlap rule (12.2), where D is a representation of G. A set $\{\phi_\alpha\}$ of such matter fields is called a **section of an associated vector bundle**. Under a local gauge transformation (12.6) the gauge and matter fields transform as

$$\begin{aligned} A'_\alpha &= h_\alpha \left(A_\alpha - i h_\alpha^{-1} dh_\alpha\right) h_\alpha^{-1} \\ \phi'_\alpha &= D(h_\alpha)\phi_\alpha. \end{aligned} \tag{12.7}$$

There is a lot more that could be said about the geometry of these definitions, but for our purposes this will be enough to be getting on with.

12.2 Differential forms

When discussing the role of topology in gauge theory, differential form notation is so convenient that it is silly not to use it (and in fact we already used the exterior derivative "d" in the previous subsection). We therefore will now explain how it works. A **differential p-form**, or a p-form for short, is a completely antisymmetric tensor $\omega_{\mu_1\ldots\mu_p}$ with p lowered indices. In manipulating differential forms it is useful to introduce an antisymmetrization operation

$$T_{[\mu_1\ldots\mu_p]} \equiv \frac{1}{p!}\sum_{\pi\in S_p} (-1)^{|\pi|} T_{\mu_{\pi(1)}\ldots\mu_{\pi(p)}}, \tag{12.8}$$

where $T_{\mu_1\ldots\mu_p}$ is an arbitrary tensor with p lowered indices, not necessarily antisymmetric, and $|\pi|$ indicates the parity of the permutation π (0 for even and 1 for odd), in terms of which a p-form obeys

$$\omega_{[\mu_1\ldots\mu_p]} = \omega_{\mu_1\ldots\mu_p}. \tag{12.9}$$

The vector space of p-forms at a point $x \in M$ is $\binom{d}{p}$ -dimensional for $0 \le p \le d$, and there are no nonzero p-forms with $p > d$. A d-form is often called a **top form**, and the set of p-forms on a manifold M is denoted $\Omega^p(M)$.

12.2.1 Wedge product and exterior derivative

A natural operation on differential forms is the wedge product $\wedge : \Omega^p(M) \times \Omega^q(M) \to \Omega^{p+q}(M)$, which is defined by

$$(\omega \wedge \sigma)_{\mu_1\ldots\mu_p\nu_1\ldots\nu_q} \equiv \frac{(p+q)!}{p!q!}\omega_{[\mu_1\ldots\mu_p}\sigma_{\nu_1\ldots\nu_q]}. \tag{12.10}$$

We can string together wedge products as

$$\left(\omega^{(1)} \wedge \ldots \wedge \omega^{(n)}\right)_{\mu_1^{(1)}\ldots\mu_{p_1}^{(1)}\ldots\mu_1^{(n)}\ldots\mu_{p_n}^{(n)}} = \frac{(p_1+\ldots+p_n)!}{p_1!\ldots p_n!}\omega^{(1)}_{[\mu_1^{(1)}\ldots\mu_{p_1}^{(1)}}\ldots\omega^{(n)}_{\mu_1^{(n)}\ldots\mu_{p_n}^{(n)}]}. \tag{12.11}$$

[112] With a little algebraic topology you can convince yourself that something stronger is true: for any finite group G a principal G bundle on a connected manifold M is equivalent to a homomorphism from $\pi_1(M)$ to G, with two homomorphisms identified if they are related by conjugation by an element of G. This homomorphism assigns a G-holonomy to each loop in M, and the quotient by conjugation accounts for possibility of a gauge transformation at the base of the loop.

This shows that the wedge product is associative, but it is not in general commutative:

$$\omega \wedge \sigma = (-1)^{pq} \sigma \wedge \omega. \tag{12.12}$$

Another natural operation is the exterior derivative $d : \Omega^p(M) \to \Omega^{p+1}(M)$, which is defined by:

$$(d\omega)_{\mu_0 \ldots \mu_p} = (p+1) \partial_{[\mu_0} \omega_{\mu_1 \ldots \mu_p]}. \tag{12.13}$$

It is easy to show that this obeys the important relation

$$d^2 = 0. \tag{12.14}$$

The exterior derivative distributes over the wedge product as

$$d(\omega \wedge \sigma) = d\omega \wedge \sigma + (-1)^p \omega \wedge d\sigma. \tag{12.15}$$

We can nicely encapsulate these formulas by noting that they all follow from the simple definitions

$$\begin{aligned} \omega &= \frac{1}{p!} \omega_{\mu_1 \ldots \mu_p} dx^{\mu_1} \wedge \ldots \wedge dx^{\mu_p} \\ d\omega &= dx^\mu \partial_\mu \wedge \omega, \end{aligned} \tag{12.16}$$

together with the assumption that exchanging two dx^μ's in the wedge product costs a minus sign.

12.2.2 Integration and Stokes' Theorem

One of the most important applications of differential forms is that any top form ω on an orientable manifold M can be integrated over M to produce a real number. To understand this we first note that in each coordinate patch U on a manifold M of dimension d we can define the quantity

$$\hat{\epsilon} \equiv dx^1 \wedge \ldots \wedge dx^d, \tag{12.17}$$

whose components are

$$\hat{\epsilon}_{\mu_1 \ldots \mu_d} = d! \delta^1_{[\mu_1} \ldots \delta^d_{\mu_d]}. \tag{12.18}$$

This quantity is not in general a d-form on M, since dx^μ is only defined on U and the definitions in different patches do not need to coincide. Since the space of d-forms is one-dimensional, any p-form on M restricts to a p-form on U that is a multiple of $\hat{\epsilon}$:

$$\omega(x)|_U = a(x) \hat{\epsilon}(x). \tag{12.19}$$

We may then define the integral of ω over U to be

$$\int_U \omega \equiv \int_U dx^1 \ldots dx^d a(x), \tag{12.20}$$

where the integral on the right-hand side is the usual Riemann integral. The motivation for this definition is that under a coordinate transformation on U we have

$$\hat{\epsilon}' = \det \left(\frac{\partial x'}{\partial x} \right) \hat{\epsilon}, \tag{12.21}$$

which is the usual Jacobian transformation of the integration measure on $\mathbb{R}^d$ as long as the determinant is positive. We would like to use this to extend the integral from U to M, but for this to work we need to make sure that the transformations between the patches all have positive Jacobian determinant. In other words we need M to be **orientable**, which means that we can cover it with a collection of patches U_α whose

transition functions all have positive Jacobian determinant. This is not true for all manifolds, for example the Mobiüs strip is not orientable, but it is true for $\mathbb{S}^d$ and that is all we will need.[113]

To complete the definition of the integral we also need the idea of a **partition of unity subordinate to** $\{U_\alpha\}$, where $\{U_\alpha\}$ is an open cover of M. This means a set of smooth scalar functions $\rho_\alpha : M \to \mathbb{R}$ such that

(i) $\rho_\alpha(x) = 0$ for all $x \notin U_\alpha$

(ii) $\sum_\alpha \rho_\alpha(x) = 1$ for all $x \in M$.

It is a theorem that for any manifold M and open cover $\{U_\alpha\}$ there exists a (highly non-unique) partition of unity subordinate to $\{U_\alpha\}$.[114] Finally we need a way to actually choose an orientation on M. There are various ways to do this; the choice we will use is to say that an **orientation** on M is a nowhere-vanishing d-form Λ. Such a d-form exists if and only if M is orientable, and we can flip the orientation of M by flipping the sign of Λ.[115] We say that an open cover $\{U_\alpha\}$ is **compatible** with an orientation Λ if in each patch we have

$$\Lambda|_{U_\alpha}(x) = \Lambda_\alpha(x)\hat{\epsilon}_\alpha \tag{12.22}$$

with $\Lambda_\alpha(x) > 0$. The integral of a d-form ω over a manifold M with orientation Λ is then defined by

$$\int_M \omega \equiv \sum_\alpha \int_{U_\alpha} \rho_\alpha \omega, \tag{12.23}$$

where $\{U_\alpha\}$ is any cover of M that is compatible with Λ and the integral on the right-hand side is defined by (12.20)

Perhaps the most important general result about integration of d-forms on oriented d-manifolds is **Stokes' Theorem**. This says that for any $(d-1)$-form ω on an oriented d-manifold M with boundary ∂M, if we choose an orientation on ∂M that is compatible with the orientation Λ on M then

$$\int_M d\omega = \int_{\partial M} \omega. \tag{12.24}$$

What it means for the orientations to be compatible is that in a coordinate patch U_α in which M is the region $x^1 < 0$, with ∂M being at $x^1 = 0$, the orientations Λ on M and Λ_∂ on ∂M obey

$$\Lambda|_{U_\alpha} = \beta dx^1 \wedge \Lambda_\partial|_{U_\alpha} \tag{12.25}$$

with $\beta(x) > 0$. The usual integration theorems of vector calculus are all special cases of this theorem.

Metrics and the Hodge star

You may object that this definition of integration did not make any use of the metric notion of volume. Indeed this is true, but this is really an advantage since it allows integration to be defined even in situations where no metric exists. On the other hand we would still like to know in what sense this definition coincides with the usual one when we do have a metric. The point is that given a Euclidean or Lorentzian metric g on M we can define in each patch U_α a quantity

$$\epsilon_\alpha \equiv \sqrt{\pm g}\hat{\epsilon}_\alpha, \tag{12.26}$$

[113] It is possible to define integration on unorientable manifolds, but not of d-forms. Instead what you integrate are called densities. This is important in some string theory applications of field theory.

[114] This theorem is one of the places where it is necessary to use the requirements that manifolds are second countable and Hausdorff.

[115] The proof uses a partition of unity: given an open cover $\{U_\alpha\}$ whose transition functions all have positive Jacobian, we can take $\Lambda(x) = \sum_\alpha \rho_\alpha(x)\hat{\epsilon}_\alpha$. Conversely if Λ is nowhere vanishing, then on any open cover $\{U_\alpha\}$ we can re-order the coordinates in each patch so that $\Lambda(x) = \Lambda_\alpha(x)\hat{\epsilon}_\alpha$ with $\Lambda_\alpha(x) > 0$.

where g is the determinant of the metric and $\pm$ indicates Euclidean/Lorentzian signature. This definition is useful because if M is oriented then the coordinate transformation

$$g' = \det\left(\frac{\partial x'}{\partial x}\right)^{-2} g \tag{12.27}$$

of the metric determinant cancels the coordinate transformation (12.21) of $\hat{\epsilon}$ (an orientation on M is needed to ensure that $\det\left(\frac{\partial x'}{\partial x}\right)$ is positive). ϵ is thus well-defined as a global d-form on M, and since it is nonzero we can (and usually do) use it to define the orientation of M. ϵ is usually called the **volume form** induced by the metric g. Given an oriented Euclidean/Lorentzian manifold with a metric g and volume form ϵ, we can define an integral of any scalar function ϕ on M as

$$\int_M d^d x \sqrt{\pm g}\phi \equiv \int_M \phi\epsilon. \tag{12.28}$$

One application of this formula is the divergence theorem for a vector field V^μ on an oriented manifold M with metric g. Given such a V^μ we can define a $d-1$ form

$$\omega \equiv V \cdot \epsilon, \tag{12.29}$$

with the right-hand side defined by

$$(V \cdot \epsilon)_{\mu_1 \ldots \mu_{d-1}} \equiv V^\nu \epsilon_{\nu\mu_1\ldots\mu_{d-1}}. \tag{12.30}$$

A short calculation shows that

$$d\omega = (\nabla_\mu V^\mu)\,\epsilon, \tag{12.31}$$

where ∇_μ is the covariant derivative with respect to the metric g. If M has a boundary then we can define a normal form n_μ at that boundary by

$$\epsilon = n \wedge \epsilon_\partial, \tag{12.32}$$

where ϵ is the volume form on M and ϵ_∂ is the boundary volume form constructed using the boundary induced metric $\gamma_{\mu\nu}$ and whose orientation is compatible to that of ϵ in the sense of Stokes theorem. In particular this means that in coordinates where the boundary is at $x^1 = 0$, with M lying in $x^1 < 0$, then n_μ is outward-pointing[116] and

$$\epsilon_\partial = \sqrt{|\gamma|} dx^2 \wedge \ldots \wedge dx^d. \tag{12.33}$$

Applying Stokes' Theorem to ω we get the divergence theorem

$$\int_M d^d x \sqrt{\pm g} \nabla_\mu V^\mu = \int_{\partial M} d^{d-1} x \sqrt{|\gamma|} n_\mu V^\mu. \tag{12.34}$$

The ϵ tensor on an oriented manifold M can also be used to define a convenient map $\star : \Omega^p(M) \to \Omega^{d-p}(M)$ called the **Hodge star** map:

$$(\star\omega)_{\mu_1\ldots\mu_{d-p}} \equiv \frac{1}{p!}\epsilon^{\nu_1\ldots\nu_p}{}_{\mu_1\ldots\mu_{d-p}}\omega_{\nu_1\ldots\nu_p}. \tag{12.35}$$

It is not too hard to show that

$$\star\star\,\omega = \pm(-1)^{p(d-p)}\omega, \tag{12.36}$$

and there is also the very useful fact that if ω and σ are p-forms then

$$\omega \wedge \star\sigma = \sigma \wedge \star\omega = (-1)^{p(d-p)} \star \omega \wedge \sigma = \left(\frac{1}{p!}\omega_{\mu_1\ldots\mu_p}\sigma^{\mu_1\ldots\mu_p}\right)\epsilon. \tag{12.37}$$

[116] Beware that in Lorentzian signature this means that if n^μ is timelike then it is inward-pointing.

These are most easily derived using the ϵ-tensor contraction identity

$$\epsilon^{\alpha_1\ldots\alpha_p\mu_{p+1}\ldots\mu_d}\epsilon_{\beta_1\ldots\beta_p\mu_{p+1}\ldots\mu_d} = \pm p!(d-p)!\delta^{\alpha_1}_{[\beta_1}\ldots\delta^{\alpha_p}_{\beta_p]}, \tag{12.38}$$

which you can check by writing out the definitions. Using this contraction identity we can also give a formula relating the two notions of integration on an oriented manifold M:

$$\int_M \omega = \pm\frac{1}{d!}\int_M d^dx\sqrt{\pm g}\epsilon^{\mu_1\cdots\mu_d}\omega_{\mu_1\ldots\mu_d}. \tag{12.39}$$

12.3 Gauge theory using differential forms

The reason we have made this aside to set up differential form notation is that it really is a superior way to discuss gauge theory. For example the electromagnetic gauge potential is a one-form A and the electromagnetic field strength is the two-form

$$F \equiv dA. \tag{12.40}$$

The Maxwell action coupled to a background current one-form J on an oriented Lorentzian d-manifold M is

$$S = \int_M \left[-\frac{1}{2e^2}F\wedge\star F + A\wedge\star J\right], \tag{12.41}$$

which can be compared to our previous expression using (12.37). Varying with respect to A we have

$$\begin{aligned}
\delta S &= \int_M \left[-\frac{1}{2e^2}\left(d\delta A\wedge\star F + F\wedge\star d\delta A\right) + \delta A\wedge\star J\right] \\
&= \int_M \left[-\frac{1}{e^2}d\delta A\wedge\star F + \delta A\wedge\star J\right] \\
&= \int_M \left[-\frac{1}{e^2}d\left(\delta A\wedge\star F\right) + \delta A\wedge\left(-\frac{1}{e^2}d\star F + \star J\right)\right] \\
&= \int_M \delta A\wedge\left(-\frac{1}{e^2}d\star F + \star J\right) - \frac{1}{e^2}\int_{\partial M}\delta A\wedge\star F.
\end{aligned} \tag{12.42}$$

In going from the first line to the second we used the first equality in (12.37), in going from the second to the third we used (12.15), and in going from the third to the fourth we used Stokes' theorem (12.24). Thus Maxwell's equations can be written as

$$\frac{1}{e^2}d\star F = \star J \tag{12.43}$$

together with the Bianchi identity

$$dF = d^2A = 0, \tag{12.44}$$

and to make the boundary term vanish at the spatial boundary $\Gamma\subset\partial M$ we should either adopt Dirichlet boundary conditions $\delta A|_\Gamma = 0$ or Neumann boundary conditions $\star F|_\Gamma = 0$. For $d = 4$ we can also have a θ term, which in differential form notation is[117]

$$S_\theta = -\frac{\theta}{8\pi^2}\int_M F\wedge F. \tag{12.45}$$

The easiest way to get this from our previous expression for the θ term is using (12.39), with the minus sign arising because we are in Lorentzian signature.

[117] We will see shortly that this is the correct normalization assuming that M is required to be a "spin manifold", meaning a manifold on which spinors can be defined. If general orientable M are allowed, i.e. if all fields are bosonic, then we should instead write $\frac{1}{4\pi^2}$. We used the normalization appropriate for spin manifolds in our discussion of the standard model, which of course is a theory with fermions.

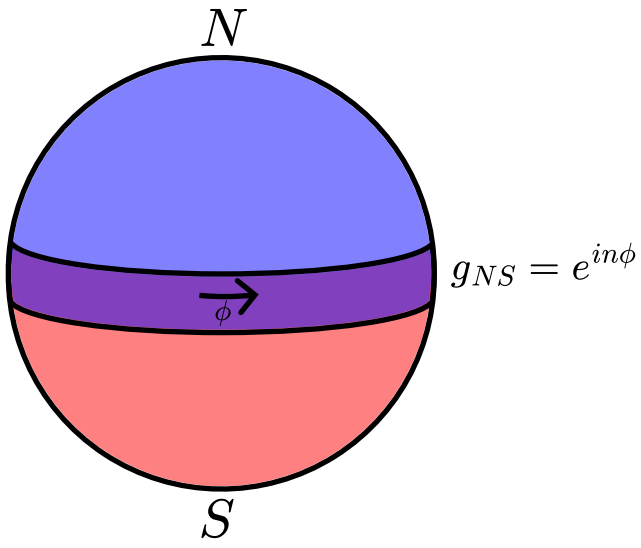

Figure 28: The Dirac monopole on $\mathbb{S}^2$. The transition function between the northern and southern hemispheres wraps n times around $U(1)$.

We can also describe non-abelian gauge theory using differential forms: A is a Lie-algebra valued one-form (sticking in one patch for now), with gauge transformation

$$A' = g(A - ig^{-1}dg)g^{-1}, \tag{12.46}$$

and the field strength is

$$F = dA - iA \wedge A \tag{12.47}$$

with gauge transformation

$$F' = gFg^{-1}. \tag{12.48}$$

In (12.47) the wedge product of Lie-algebra valued quantities is defined by

$$A \wedge A = T_a T_b A^a \wedge A^b, \tag{12.49}$$

so we antisymmetrize in Lorentz indices but multiply the Lie algebra matrices in order. Using the same convention we can write the Lorentzian Yang-Mills action (including a θ term for $d = 4$) as

$$S = -\int_M \left[\frac{1}{g^2}\mathrm{Tr}\,(F \wedge \star F) + \frac{\theta}{8\pi^2}\mathrm{Tr}\,(F \wedge F)\right]. \tag{12.50}$$

This action is well-defined on all of M because the traces make the Lagrangian gauge-invariant under (12.48), so it is a globally-defined d-form without any extra transition function as we move from patch to patch.

12.4 The Dirac monopole

We now consider the simplest topologically nontrivial gauge field, the Dirac monopole on $\mathbb{S}^2$. This describes the gauge field configuration on a spatial two-sphere surrounding a magnetic monopole. The idea is to find a $U(1)$ gauge field configuration such that

$$\int_{\mathbb{S}^2} F \neq 0. \tag{12.51}$$

Such a configuration must be a connection on a nontrivial $U(1)$ bundle, since if globally we have $F = dA$ then we have

$$\int_{\mathbb{S}^2} F = \int_{\mathbb{S}^2} dA = \int_{\partial \mathbb{S}^2} A = 0 \tag{12.52}$$

by Stokes' theorem. We can build a connection that does work using two patches, which in spherical coordinates we can take to be a "northern hemisphere" U_N with $0 \leq \theta < \pi/2 + \epsilon$ and a "southern hemisphere" U_S with $\pi/2 - \epsilon < \theta \leq \pi$ (see figure 28 for an illustration). Here we are using spherical coordinates (θ, ϕ), with θ being the polar angle and ϕ being the azimuthal angle. We can look for gauge fields in these two patches such that the field strength is proportional to the volume form on $\mathbb{S}^2$ with round metric,

$$F = B \sin\theta d\theta \wedge d\phi. \tag{12.53}$$

A gauge potential that locally obeys $F = dA$ is

$$A = -(B\cos\theta + C)d\phi, \tag{12.54}$$

where C is arbitrary. For this gauge field to be smooth at the north pole ($\theta = 0$) we need $C = -B$, while for this gauge field to be smooth at the south pole ($\theta = \pi$) we need $C = B$. Thus in our two patches we have

$$\begin{aligned} A_N &= -B\,(\cos\theta - 1)\,d\phi \\ A_S &= -B\,(\cos\theta + 1)\,d\phi. \end{aligned} \tag{12.55}$$

I emphasize that A_N does not need to be smooth at the south pole and A_S does not need to be smooth at the north pole. On the other hand we do need to check that the relationship between the two gauge potentials comes from a valid gluing transformation $g_{NS} : U_N \cap U_S \to U(1)$, meaning that we need

$$A_N - A_S = 2Bd\phi = -ig_{NS}^{-1} dg_{NS}. \tag{12.56}$$

The solution of this equation is

$$g_{NS} = g_0 e^{2iB\phi}, \tag{12.57}$$

where g_0 is an arbitrary phase, but this function is only smooth on $U_N \cap U_S$ if

$$B = \frac{n}{2} \tag{12.58}$$

with $n \in \mathbb{Z}$. In particular we have

$$\int_{\mathbb{S}^2} F = 2\pi n, \tag{12.59}$$

which is called the the **Dirac quantization** of magnetic flux. There is a nice topological interpretation of the integer n. Namely the gluing transformation is

$$g_{NS} = e^{in\phi}, \tag{12.60}$$

so viewing this as a map from $\mathbb{S}^1$ (the equator) to $\mathbb{S}^1$ (the gauge group) the integer n counts the number of times that the domain circle wraps around the target circle. In mathematical language, monopole number classifies elements of the fundamental group $\pi_1(U(1))$.

You may be wondering to what extent the Dirac quantization we found is an accident of the particular configuration we studied. In fact it is not: for any $U(1)$ gauge field on any oriented two-dimensional manifold M, we have

$$\int_M F = 2\pi n \tag{12.61}$$

with $n \in \mathbb{Z}$. There are various ways to explain this, my favorite is as follows. Begin by covering M with patches V_α that we have "shrunk" so that they fit together to tile M as in figure 29 (mathematically this

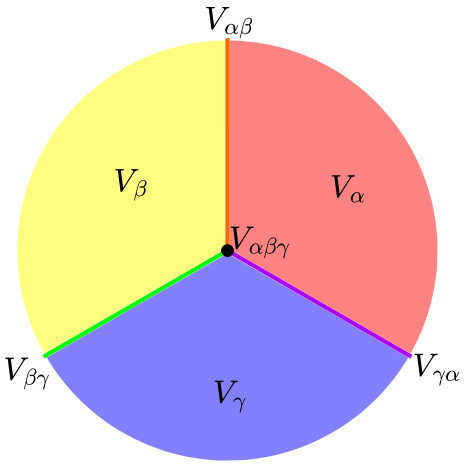

Figure 29: The geometry near a triple overlap $V_{\alpha\beta\gamma}$ of patches in a two-dimensional manifold. The orientations on the double and triple overlaps are assigned so that (12.66) and (12.67) hold.

tiling gives M the structure of a CW complex). The partition of unity construction can then be simplified to

$$\int_M F = \sum_\alpha \int_{V_\alpha} F. \tag{12.62}$$

Within each patch we can write $F = dA_\alpha$, with the A_α in the different patches being related by

$$A_\alpha = A_\beta + d\Omega_{\alpha\beta}. \tag{12.63}$$

The triple overlap rule (12.4) implies that for any three patches we have

$$\Omega_{\alpha\beta} + \Omega_{\beta\gamma} + \Omega_{\gamma\alpha} = 2\pi n_{\alpha\beta\gamma} \tag{12.64}$$

where $n_{\alpha\beta\gamma}$ are integers. $\Omega_{\alpha\beta}$ and $g_{\alpha\beta}$ are related by

$$g_{\alpha\beta} = e^{i\Omega_{\alpha\beta}}. \tag{12.65}$$

We can write the boundary of each V_α as a sum over intersections $V_{\alpha\beta}$ as

$$\partial V_\alpha = \sum_\beta V_{\alpha\beta} \tag{12.66}$$

and we can write the boundary of each $V_{\alpha\beta}$ as a sum over triple intersections:

$$\partial V_{\alpha\beta} = \sum_\gamma V_{\alpha\beta\gamma}. \tag{12.67}$$

We can then use these expressions together with Stokes' theorem to evaluate the integral of F over M:[118]

$$\begin{aligned}\int_M F &= \sum_\alpha \int_{V_\alpha} dA_\alpha = \sum_\alpha \int_{\partial V_\alpha} A_\alpha = \sum_{\alpha,\beta} \int_{V_{\alpha\beta}} A_\alpha = \sum_{\alpha<\beta} \int_{V_{\alpha\beta}} (A_\alpha - A_\beta) = \sum_{\alpha<\beta} \int_{V_{\alpha\beta}} d\Omega_{\alpha\beta} \\ &= \sum_{\alpha<\beta} \int_{\partial V_{\alpha\beta}} \Omega_{\alpha\beta} = \sum_{\alpha<\beta} \sum_\gamma \int_{V_{\alpha\beta\gamma}} \Omega_{\alpha\beta} = \sum_{\alpha<\beta<\gamma} \int_{V_{\alpha\beta\gamma}} (\Omega_{\alpha\beta} + \Omega_{\beta\gamma} + \Omega_{\gamma\alpha}) = 2\pi \sum_{\alpha<\beta<\gamma} \int_{V_{\alpha\beta\gamma}} n_{\alpha\beta\gamma}.\end{aligned} \tag{12.68}$$

[118] In the fourth and eighth equalities we relabeled the sum to put the patch labels in order, it may be helpful to refer to figure 29 to confirm the orientations.

In the second line the integrals over the points $V_{\alpha\beta\gamma}$ are just signs to keep track of orientation, so the final expression is indeed $2\pi n$ for some integer n! This basic chain of operations is what lies behind most of the applications of topology in field theory.

12.5 Topology of non-abelian gauge fields in $3+1$ dimensions

We now turn to non-abelian gauge theory in $3+1$ dimensions. In our discussion of anomalies we claimed that the quantity

$$c_2[A] \equiv \frac{1}{8\pi^2}\int_M \mathrm{Tr}\,(F\wedge F) = \pm\frac{1}{32\pi^2}\int_M d^4x\sqrt{\pm g}\epsilon^{\mu\nu\alpha\beta}\mathrm{Tr}\,(F_{\mu\nu}F_{\alpha\beta})\,, \tag{12.69}$$

which is called the **second Chern class**, is an integer for any four-manifold M and simple gauge group G. The full justification for this integrality is rather subtle; if you ask a mathematician they will tell you that you need to learn about classifying spaces and the splitting principle, perhaps with some category theory thrown in for good measure. In fact the Stokes' theorem method of equation (12.68) is up to the task, but it takes some thought to figure out how to make it work so instead of explaining the details here I will post a paper about it later this summer. Here we will instead content ourselves with telling the first part of the story and then specializing to the case of $M=\mathbb{S}^4$, where we can finish the computation in a more straightforward way. In the next subsection we will see that there are solutions of the Yang-Mills equations of motion on $\mathbb{R}^4$ called **instantons** that have $c_2[A]=1$.

Based on our experience with Dirac quantization, we can guess that the explanation of the integrality of $c_2[A]$ begins by showing that the integrand is the exterior derivative of something. In fact you already showed this on a previous homework, in differential form notation we have

$$\mathrm{Tr}\,(F\wedge F) = d\omega_{CS}, \tag{12.70}$$

with

$$\omega_{CS} \equiv \mathrm{Tr}\left(A\wedge F + \frac{i}{3}A\wedge A\wedge A\right). \tag{12.71}$$

ω_{CS} is called the **Chern-Simons three-form**. Its gauge transformation is given by

$$\omega'_{CS} = \omega_{CS} + d\left(i\mathrm{Tr}\,(\omega\wedge A)\right) - \frac{1}{3}\mathrm{Tr}\,(\omega\wedge\omega\wedge\omega)\,, \tag{12.72}$$

where

$$\omega \equiv g^{-1}dg \tag{12.73}$$

is the Maurer-Cartan one-form. In checking this equation (and also (12.70)) it is useful to first show that

$$d\omega = -\omega\wedge\omega. \tag{12.74}$$

Beginning as in (12.68), we therefore have

$$c_2[A] = \frac{1}{8\pi^2}\sum_{\alpha<\beta}\int_{V_{\alpha\beta}}\left(d\left(i\mathrm{Tr}\,(\omega_{\alpha\beta}\wedge A_\beta)\right) - \frac{1}{3}\mathrm{Tr}\,(\omega_{\alpha\beta}\wedge\omega_{\alpha\beta}\wedge\omega_{\alpha\beta})\right) \tag{12.75}$$

with

$$\omega_{\alpha\beta} = g_{\alpha\beta}^{-1}dg_{\alpha\beta}. \tag{12.76}$$

Proceeding from here is where things get tricky on general M, so we will now specialize to the case of $M=\mathbb{S}^4$ covered by northern and southern hemisphere patches as in figure 28. This case is special because there is only one double overlap V_{NS} and it obeys $\partial V_{NS}=0$, so by Stokes theorem we simply have

$$c_2[A] = -\frac{1}{24\pi^2}\int_{\mathbb{S}^3}\mathrm{Tr}\,(\omega_{NS}\wedge\omega_{NS}\wedge\omega_{NS}) \tag{12.77}$$

where $\mathbb{S}^3$ is the equator of the $\mathbb{S}^4$ with its orientation defined by viewing it as the boundary of V_N. The quantity on the right-hand side here is a famous topological invariant: given any smooth map $g : \Sigma_3 \to G$, with Σ_3 is a closed oriented three-manifold and G a compact matrix group, defining $\omega = g^{-1}dg$ as before we have

$$\nu_3[g] \equiv \frac{1}{24\pi^2}\int_{\Sigma_3} \mathrm{Tr}\,(\omega\wedge\omega\wedge\omega) \in \mathbb{Z}. \tag{12.78}$$

Rather surprisingly $\nu_3[g]$ doesn't seem to have a standard name, I will refer to it as the **third Maurer-Cartan topological charge**. The method of (12.68) can be used to show that $\nu_3[g]$ is indeed an integer, somewhat more easily than for the case of $c_2[A]$ on general M, but I will still leave the details to my upcoming paper. What is easier is just to show that it is a topological invariant in the sense that it is unchanged by small changes of the function g. Indeed under a small change δg of g we have

$$\delta\omega = \omega g^{-1}\delta g - g^{-1}\delta g\omega + d(g^{-1}\delta g), \tag{12.79}$$

and thus

$$\delta\mathrm{Tr}\,(\omega\wedge\omega\wedge\omega) = 3\mathrm{Tr}\left(\left(\omega g^{-1}\delta g - g^{-1}\delta g\omega + d(g^{-1}\delta g)\right)\wedge\omega\wedge\omega\right) = d\left(3\mathrm{Tr}\left(g^{-1}\delta g\omega\wedge\omega\right)\right), \tag{12.80}$$

where we made liberal use of the cyclicity of the trace and also used (12.74). Thus we have

$$\delta\nu_3[g] = \frac{1}{24\pi^2}\int_{\Sigma_3} d\left(3\mathrm{Tr}\left(g^{-1}\delta g\omega\right)\right) = 0 \tag{12.81}$$

by Stokes' theorem. A similar argument shows that ν_3 is additive in the sense that

$$\nu_3[g_1 g_2] = \nu_3[g_1] + \nu_3[g_2]. \tag{12.82}$$

Returning to the particular case of $\mathbb{S}^3$, since g_{NS} is a map from $\mathbb{S}^3$ to the gauge group G, and ν_3 is unchanged by continuous deformations of g_{NS}, we can view it as giving us information about the third homotopy group $\pi_3(G)$. In fact for any simple Lie group we have $\pi_3(G) \cong \mathbb{Z}$, so ν_3 fully captures the topological information in g_{NS}. In this case ν_3 is often called the **winding number** of the map g_{NS}, we will see more justification for this term in a moment (although really "wrapping number" would be better).

To give a concrete example of g_{NS} we need to assume something about the gauge group G. To begin with we will just take $G = SU(2)$, which has the topology of $\mathbb{S}^3$ since we can uniquely represent any element of $SU(2)$ as

$$g = b^0\sigma_0 + ib^i\sigma_i, \tag{12.83}$$

with σ_i the Pauli matrices, $\sigma_0 = I$, and b^μ a unit vector in Euclidean $\mathbb{R}^4$. To get a nontrivial winding for g_{NS}, we can take g_{NS} to be the identity map

$$g_{NS}(x) = x^0\sigma_0 + ix^i\sigma_i, \tag{12.84}$$

with x being a unit vector in $\mathbb{R}^4$ that parametrizes the $\mathbb{S}^3$ equator. We can think of this g_{NS} as wrapping $SU(2)$ exactly once. We can parametrize the hemisphere with $x^0 > 0$ using x^i with $0 \le x^i x^i \le 1$, in terms of which we have

$$g_{NS}(\vec{x}) = \sqrt{1-x^2} + ix^i\sigma_i. \tag{12.85}$$

Computing ν_3 directly is a bit tedious. We can streamline the calculation by realizing that g_{NS} is rotationally invariant, so we must have

$$\mathrm{Tr}\,(\omega_{NS}\wedge\omega_{NS}\wedge\omega_{NS}) = A\epsilon \tag{12.86}$$

where[119]

$$\epsilon = \frac{1}{|x^0|}dx^1\wedge dx^2\wedge dx^3 \tag{12.87}$$

[119]To derive this we can note that the induced metric on the sphere is $\left(\delta^{ij} + \frac{x^i x^j}{1-x^2}\right)dx^i dx^j$, which has determinant $\frac{1}{1-x^2}$.

is the volume form on a round $\mathbb{S}^3$. To determine A it therefore is sufficient to compute ω_{NS} at the north pole. In general we have

$$\omega_{NS} = \left(\sqrt{1-x^2} - i\vec{x}\cdot\vec{\sigma}\right)\left(-\frac{\vec{x}\cdot d\vec{x}}{\sqrt{1-x^2}} + i\vec{\sigma}\cdot d\vec{x}\right), \tag{12.88}$$

so at the north pole

$$\omega_{NS} = i\vec{\sigma}\cdot d\vec{x} \tag{12.89}$$

and thus

$$\begin{aligned}\mathrm{Tr}\left(\omega_{NS}\wedge\omega_{NS}\wedge\omega_{NS}\right) &= -i\mathrm{Tr}\left(\sigma_n\sigma_m\sigma_\ell\right)dx^n\wedge dx^m\wedge dx^\ell \\ &= 2\epsilon_{nm\ell}dx^n\wedge dx^m\wedge dx^\ell \\ &= 12dx^1\wedge dx^2\wedge dx^3.\end{aligned} \tag{12.90}$$

Therefore $A = 12$, so we have

$$\mathrm{Tr}\left(\omega_{NS}\wedge\omega_{NS}\wedge\omega_{NS}\right) = \frac{12}{x^0}dx^1\wedge dx^2\wedge dx^3 \tag{12.91}$$

and thus

$$\nu_3 = \frac{1}{24\pi^2}\times 2\times 4\pi\times 12\int_0^1\frac{r^2dr}{\sqrt{1-r^2}} = 1. \tag{12.92}$$

The factor of two is here because we need to do the integral for both the northern ($x^0 > 0$) and southern ($x^0 < 0$) hemispheres. We can also get this directly from the volume of a unit $\mathbb{S}^3$:

$$\nu_3 = \frac{1}{24\pi^2}\times 12\times 2\pi^2 = 1. \tag{12.93}$$

Using the additivity (12.82) it is now easy to construct a configuration that realizes any winding number n: we simply take the transition function to be

$$g_{NS} = \left(x^0\sigma_0 + ix^i\sigma_i\right)^n, \tag{12.94}$$

which is a map that wraps $SU(2)$ n times. Finally for more general simple Lie groups we can simply choose an arbitrary $SU(2)$ subgroup and then perform the same construction, although in general the integer we get from the identity map of $\mathbb{S}^3$ into this subgroup may be larger than one (for $SU(N)$ however we can always get one since we can simply use an $SU(2)$ subblock).[120]

12.6 Instantons

The discussion of the previous subsection tells us that there exist on $\mathbb{S}^4$ gauge field configurations with nonzero second Chern-class. Namely we can pick an arbitrary gauge field configuration A_S in the southern hemisphere, use (12.84) to transition to the northern hemisphere, and then extend this A_N away from the equator however we like. It is not obvious, but gauge fields on $\mathbb{S}^4$ are actually in one-to-one correspondence with gauge field configurations of finite action on $\mathbb{R}^4$. To see this, we note that in order to have finite action a gauge field configuration A_μ on Euclidean $\mathbb{R}^4$ must approach pure gauge at large radius:

$$A = ig^{-1}dg + O(1/r^{1+\epsilon}), \tag{12.95}$$

with g independent of r. This ensures that the field strength tensor is of order $1/r^{2+2\epsilon}$ and the Lagrangian density is of order $1/r^{4+4\epsilon}$, and thus is integrable at large r. This gauge field configuration does not actually

[120] Something nicer is actually true: *any* map from $\mathbb{S}^3$ to G can be smoothly deformed into a map that is takes values only in an $SU(2)$ subgroup. So there is nothing else to find beyond what we have already constructed.

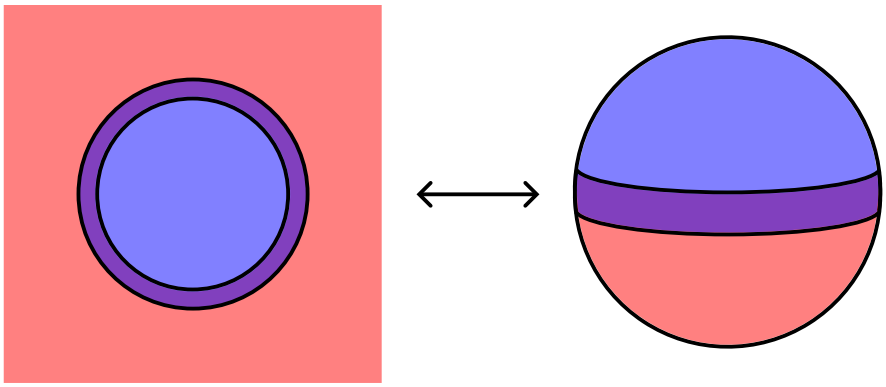

Figure 30: The relationship between finite-action configurations on $\mathbb{R}^4$ and gauge field configurations on $\mathbb{S}^4$: the "inner" and "outer" patches on $\mathbb{R}^4$ become hemispheres on $\mathbb{S}^4$, with the asymptotic gauge transformation g in the inner patch becoming the gluing gauge transformation g_{SN} at the equator.

obey our boundary conditions that $A|_{\partial M} = 0$, but we can turn it into one that does by introducing an "inner" patch $r < R + \epsilon$ and an "outer" patch $r > R - \epsilon$ and then using the gluing between the inner and outer patches to remove the gauge transformation g. Since the gauge field in the outer patch vanishes at infinity, we can extend this gauge field to a gauge field on $\mathbb{S}^4$ where we re-interpret these inner and outer patches as the northern and southern hemispheres and the south pole becomes the "point at infinity" in $\mathbb{R}^4$. See figure 30 for an illustration.

The existence of gauge field configurations on $\mathbb{R}^4$ with finite action and nonzero $c_2[A]$ is all we need to justify the main consequences of instantons that we discussed in the previous sections: the violation of $U(1)_A$ symmetry in massless QCD and $U(1)_B$ and $U(1)_L$ in the standard model due to anomalies and the electric dipole moment of the neutron arising from the θ-term. For some purposes however it is nice to have examples of such configurations that obey the Yang-Mills equation of motion. Such an example was discovered by Belavin, Polyakov, Schwarz, and Tyuptin, and it is called the **BPST instanton**. In constructing this solution there is a useful trick, which is to note that in Euclidean signature by (12.37) we have

$$\int_M F \wedge \star F \mp \int_M F \wedge F = \frac{1}{2} \int_M (F \mp \star F) \wedge \star (F \mp \star F) \geq 0, \tag{12.96}$$

and thus

$$\int F \wedge \star F \geq \left| \int_M F \wedge F \right|. \tag{12.97}$$

Thus if we find a gauge field conguration where this inequality is saturated it must be a minimum of the Euclidean action among all configurations with fixed $c_2[A]$. This implies that it must be a local minimum of the action under arbitrary variations, since no small variation can change $c_2[A]$, and therefore that it must be a solution of the Yang-Mills equation of motion. Moreover the inequality is saturated for any gauge field configuration obeying

$$\star F = \pm F, \tag{12.98}$$

so any configuration obeying (12.98) is a solution of Yang-Mills equation. Such gauge configurations are called **self-dual/anti self-dual**. As a guess for such a gauge configuration we can take (in the inner patch)

$$A = ia(r)\omega, \tag{12.99}$$

where $a \to 1$ as $r \to \infty$ and $\omega = g^{-1}dg$ with

$$g = \frac{x^0 + ix^i\sigma_i}{r} \tag{12.100}$$

being the same gauge transformation we used for g_{NS} in the previous section. Here it is actually playing the role of g_{SN}, since this way we get

$$c_2[A] = 1 \tag{12.101}$$

due to the minus sign in (12.77). The field strength for this gauge field is

$$F = i\big(a' dr \wedge \omega - a(1-a)\omega \wedge \omega\big). \tag{12.102}$$

To impose the condition (12.98) we need to know that

$$\star\,(dr \wedge \omega) = -\frac{r}{2}\omega \wedge \omega, \tag{12.103}$$

which is most easily derived by using rotational invariance to set $x^i = 0$ and $x^0 > 0$, in which case $\omega = \frac{i}{r}\sigma_i dx^i$. We thus have

$$\star F = i\left(\frac{2}{r}a(1-a)dr \wedge \omega - \frac{r}{2}a'\omega \wedge \omega\right), \tag{12.104}$$

so the solution will be self-dual/anti self-dual if

$$a' = \pm\frac{2}{r}a(1-a). \tag{12.105}$$

The self-dual solution (meaning the + sign) is

$$a[r] = \frac{r^2}{r^2 + R^2} \tag{12.106}$$

with R an arbitrary real number, which is indeed smooth at $r = 0$ and goes to one as $r \to \infty$. The anti self-dual solution is

$$a[r] = \frac{1}{1 + r^2/R^2}, \tag{12.107}$$

which vanishes as $r \to \infty$ and thus does not give a solution with nonzero second Chern-class. We therefore will use the self-dual solution, which at last gives us the BPST instanton solution:

$$A = i\frac{r^2}{r^2 + R^2}g^{-1}dg. \tag{12.108}$$

The parameter R is not fixed by the equations of motion, it describes the "size" of the instanton. We are also free to change our origin of coordinates, so there is a five-dimensional continuous family of these instantons. To the extent that we want to compute a semiclassical approximation to the path integral using them as a stationary point, we need to integrate over this family.

The historical development of this subject was almost the opposite of what we have done here. The BPST solution was found first, and initially the primary use of instantons was expected to be a semiclassical approximation to the path integral around these solutions similar to what is done for tunneling in quantum mechanics. In QCD however this hope did not pan out quantitatively: instantons are most important when their size R is of order the QCD scale, and then the semiclassical approximation is not good. The most important lesson from the existence of instantons has turned out to be qualitative: they are what "activates" the second Chern class for applications such as the violation of $U(1)_A$ symmetry or the contribution of the QCD θ angle to the electric dipole moment of the neutron. For these applications it is not so important to actually construct instanton solutions of the Yang-Mills equation of motion, since anyways the dominant contribution comes from the regime that is not semiclassical. What really matters is just that there are configurations in the path integral with $c_2[A] \neq 0$. On the other hand there are some situations where semiclassical instantons do give reliable quantitative answers, in particular in supersymmetric theories, so it is still good to know about them.

12.7 Is it necessary to include instantons?

Based on the end of the previous section, you may wonder if it is really necessary to include instanton configurations in the path integral. In other words why not just stick to gauge field configurations which are globally-defined one-forms on M? On the other hand such a restriction does not seem consistent with locality: locally there is no way for us to tell how many patches are used to define a gauge field configuration, so demanding that there is only one imposes a rather non-local constraint on the set of gauge field configurations. We can make this argument more precise using the principle of cluster decomposition, which we will use in the guise that that if we work in a large but finite spatial volume, vacuum expectation values of local operators near us should be independent of the volume. Indeed we can split spacetime into a big region R_1 where our operators are located and a complementary big region R_2. We would like to get the same expectation values of operators in the interior of R_1 whether or not we include R_2. From the path integral point of view this happens in field theory because

$$\langle\Omega|\mathcal{O}[\Phi_1]|\Omega\rangle = \frac{\int \mathcal{D}\phi_1\mathcal{D}\phi_2\mathcal{O}[\phi_1]e^{-S[\phi_1]-S[\phi_2]}}{\int \mathcal{D}\phi_1\mathcal{D}\phi_2 e^{-S[\phi_1]-S[\phi_2]}} = \frac{\int \mathcal{D}\phi_1\mathcal{O}[\phi_1]e^{-S[\phi_1]}}{\int \mathcal{D}\phi_1 e^{-S[\phi_1]}}, \tag{12.109}$$

where ϕ_1 are the fields in R_1 and ϕ_2 are the fields in R_2. In the second equality we factored out the integral over the fields in R_2.[121] Now let's suppose that we would like to introduce a function $f(n)$ in the path integral that weights configurations by instanton number. In particular we could exclude instantons altogether by choosing $f(n) = \delta_{n,0}$, but for now we will be open-minded. Attempting the same manipulation, we have

$$\langle\Omega|\mathcal{O}[\Phi_1]|\Omega\rangle = \frac{\sum_{n_1,n_2} f(n_1+n_2)\int \mathcal{D}\phi_1^{n_1}\mathcal{D}\phi_2^{n_2}\mathcal{O}[\phi_1^{n_1}]e^{-S[\phi_1^{n_1}]-S[\phi_2^{n_2}]}}{\sum_{n_1,n_2} f(n_1+n_2)\int \mathcal{D}\phi_1^{n_1}\mathcal{D}\phi_2^{n_2} e^{-S[\phi_1^{n_1}]-S[\phi_2^{n_2}]}}, \tag{12.110}$$

where $\phi_1^{n_1}$ are field configurations in R_1 with n_1 instantons and $\phi_2^{n_2}$ are field configurations in R_2 with n_2 instantons. For general $f(n)$ however we cannot factor out the integral over fields in R_2: we can do this if and only if we have

$$f(n_1+n_2) = f(n_1)f(n_2). \tag{12.111}$$

Choosing $n_2 = 0$ we see that

$$f(n) = f(n)f(0), \tag{12.112}$$

so either $f(n) = 0$ for all n (clearly not acceptable) or else $f(0) = 1$. For $n > 0$ we have

$$f(n) = f(n-1)f(1) = f(n-2)f(1)^2 = \ldots = f(1)^n, \tag{12.113}$$

while for $n < 0$ we have

$$f(-n) = \frac{f(0)}{f(n)} = f(1)^{-n}. \tag{12.114}$$

Thus for all n we have

$$f(n) = f(1)^n \equiv e^{-i\theta n}, \tag{12.115}$$

where we should take θ to be real so that in Lorentzian signature the action is real. This however is nothing but the usual θ term for Yang-Mills theory, so the only freedom we have to modify instanton contributions is the one we already know about! In particular there is no value of θ that suppresses $n \neq 0$.

[121] This factorization isn't quite correct due to terms in the action right at the interface between R_1 and R_2, but this effect is small when R_1 is large and $\mathcal{O}$ is far from the interface. Similarly in the following argument we could worry about what happens with instantons right at the interface between the regions, but at large volume this again should have a small effect on the expectation value.

12.8 Chern-Simons theory

So far our primary application of the Chern-Simons three-form

$$\omega_{CS}[A] = \mathrm{Tr}\left(A \wedge F + \frac{i}{3} A \wedge A \wedge A\right) \tag{12.116}$$

was in computing the second Chern class. It is also interesting however to consider this form as the Lagrangian for a three-dimensional gauge theory, which is called **Chern-Simons theory**. There is an obvious problem with this proposal however, which is that ω_{CS} is not gauge-invariant. Its gauge transformation is given by (12.72). On the other hand this gauge transformation involves a total derivative, as well as the third Maurer-Cartan form $\mathrm{Tr}\,(\omega \wedge \omega \wedge \omega)$ that is locally a total derivative, so there is still some hope that we could fix this. The most common approach works for three-manifolds M and gauge connections A such that we can find a four-manifold $\widetilde{M}$ with boundary such that

$$\begin{aligned} \partial \widetilde{M} &= M \\ \widetilde{A}|_{\partial \widetilde{M}} &= A. \end{aligned} \tag{12.117}$$

The idea is then that we simply define the Chern-Simons action on M for simple gauge group G to be

$$S_{CS}[A] \equiv \frac{k}{4\pi} \int_{\widetilde{M}} \mathrm{Tr}\left(\widetilde{F} \wedge \widetilde{F}\right), \tag{12.118}$$

where the quantity k is called the **Chern-Simons level**. The motivation for this definition is that the four dimensional integrand is manifestly gauge-invariant, and a naive application of Stokes' theorem turns it into the integral of $\frac{k}{4\pi}\omega_{CS}$ on M. This application is not really justified however, since $\mathrm{Tr}\left(\widetilde{F} \wedge \widetilde{F}\right) = d\omega_{CS}[\widetilde{A}]$ applies only patch by patch in $\widetilde{M}$; it is not a global statement. In particular this means that there is some dependence of this S_{CS} on our particular choice of $\widetilde{M}$ and $\widetilde{A}$, which is not consistent with it really being a local action on M. The idea however is that if we change to a different four manifold $\widetilde{M}'$ and extension $\widetilde{A}'$, then we have

$$S_{CS}[A]' - S_{CS}[A] = \frac{k}{4\pi} \int_{\widetilde{M}'} \mathrm{Tr}\left(\widetilde{F}' \wedge \widetilde{F}'\right) - \frac{k}{4\pi} \int_{\widetilde{M}} \mathrm{Tr}\left(\widetilde{F} \wedge \widetilde{F}\right) = \frac{k}{4\pi} \int_{\widetilde{M}^*} \mathrm{Tr}\left(\widetilde{F}^* \wedge \widetilde{F}^*\right) = 2\pi k c_2[\widetilde{A}^*], \tag{12.119}$$

where $\widetilde{M}^*$ is the closed four manifold constructed by gluing $-\widetilde{M}$ to $\widetilde{M}'$ on the common boundary M and $\widetilde{A}^*$ is the gauge field on $\widetilde{M}^*$ that is equal to A on $\widetilde{M}$ and A' on $\widetilde{M}'$. What this tells us is that the quantity

$$e^{iS_{CS}[A]} \tag{12.120}$$

appearing in the path integral will be independent of our choice of $\widetilde{M}$ and $\widetilde{A}$ provided that we take the level k to be an integer, since then we have

$$e^{iS_{CS}[A]'} = e^{iS_{CS}[A]} e^{2\pi i k c_2[\widetilde{A}*]} = e^{iS_{CS}[A]}. \tag{12.121}$$

There are two problems with this approach to defining the Chern-Simons action:

- It requires the existence of a four manifold $\widetilde{M}$ and gauge field $\widetilde{A}$ that extend M and A. In general these do not exist. There is a fix for this based on bordism theory, which is to show there is a natural number n such that n copies of M and A can be extended and then define the action using this extension divided by n. I find this rather ugly (why is it local?), and in any case it does not work for more general actions of this type since in general there is no n that works.
- It is not manifestly local on M, and although we checked that $e^{iS_{CS}}$ is independent of our choice of extension it is not clear that that is really all we need to check to confirm locality.

For these reasons it is useful to have an alternative definition of S_{CS} that is intrinsic to M. I will motivate the idea behind it by using a simpler example of the same phenomenon, which is the definition of a Wilson loop for a nontrivial gauge bundle. In differential form notation we can write our previous definition of a Wilson loop in representation α as

$$W_\alpha[C] = Pe^{i\int_C A^a \tau_a}. \tag{12.122}$$

In particular in the $U(1)$ case we can write the Wilson line of charge one as

$$W_1[C] = e^{i\int_C A}. \tag{12.123}$$

Here A is analogous to the Chern-Simons form: it is not gauge-invariant, but it is gauge-invariant up to a total derivative. This is good enough for trivial gauge bundles where we can cover M with a single patch, but for general gauge fields involving multiple patches it needs some refinement: when we do a gauge transformation $A'_\alpha = A_\alpha + d\Omega_\alpha$ the quantity Ω_α is only defined in U_α, so we cannot use Stokes' theorem to argue that (12.123) is gauge-invariant. Another way to think about this problem is to ask what we do in an overlap $U_\alpha \cap U_\beta$: do we use A_α or A_β? One fix for this problem is the one we just discussed for Chern-Simons theory: we write C as the boundary of an auxiliary disk $\widetilde{C}$, and then define

$$W_1[C] = e^{i\int_{\widetilde{C}} \widetilde{F}}, \tag{12.124}$$

where $\widetilde{A}$ is an extension of the gauge field on C to $\widetilde{C}$. What we will do instead is a more intrinsic fix: we split up the curve into segments C_α each contained in V_α where V_α is our tiling of spacetime as in figure 29, and then at the end of each segment we include an extra gauge transformation as the curve passes from V_α to V_β. Defining these endpoints by

$$\partial C_\alpha = \sum_\beta C_{\alpha\beta}, \tag{12.125}$$

we can take the Wilson line of charge n to be

$$W_n[C] = \exp\left[in\left(\sum_\alpha \int_{C_\alpha} A_\alpha - \sum_{\alpha<\beta}\int_{C_{\alpha\beta}} \Omega_{\alpha\beta}\right)\right]. \tag{12.126}$$

This is gauge-invariant under

$$\begin{aligned} A'_\alpha &= A_\alpha + d\Omega_\alpha \\ \Omega'_{\alpha\beta} &= \Omega_{\alpha\beta} + \Omega_\alpha - \Omega_\beta, \end{aligned} \tag{12.127}$$

since we have

$$\sum_\alpha \int_{C_\alpha} d\Omega_\alpha = \sum_\alpha \int_{\partial C_\alpha} \Omega_\alpha = \sum_{\alpha,\beta}\int_{C_{\alpha\beta}} \Omega_\alpha = \sum_{\alpha<\beta}\int_{C_{\alpha\beta}} (\Omega_\alpha - \Omega_\beta), \tag{12.128}$$

which cancels the gauge transformation of the second term in the exponent of (12.126). We can also use this definition to see why n must be an integer: if we consider an infinitesimally short loop that circles a triple overlap like that in figure 29, then we would like to get one for the Wilson loop. What we get from (12.126) is

$$W_n[C] = e^{-in(\Omega_{\alpha\beta}+\Omega_{\beta\gamma}+\Omega_{\gamma\alpha})}, \tag{12.129}$$

so using the triple overlap condition (12.64) we see this gives one only if we take n to be an integer.

The intrinsic definition of the Chern-Simons action proceeds in an analogous way to (12.126), but since the integral is higher dimensional we need more overlap terms. I will leave the details to my upcoming paper, but for the case of gauge group $U(1)$ the fully intrinsic definition of the Chern-Simons action is

$$\begin{aligned} S_{CS}[A] = \frac{k}{2\pi}\Bigg(&\sum_{\alpha_1}\int_{V_{\alpha_1}} A_{\alpha_1}\wedge F - \sum_{\alpha_1<\alpha_2}\int_{V_{\alpha_1\alpha_2}} \Omega_{\alpha_1\alpha_2}F \\ &+ 2\pi\sum_{\alpha_1<\alpha_2<\alpha_3}\int_{V_{\alpha_1\alpha_2\alpha_3}} n_{\alpha_1\alpha_2\alpha_3}A_{\alpha_3} - 2\pi\sum_{\alpha_1<\alpha_2<\alpha_3<\alpha_4}\int_{V_{\alpha_1\alpha_2\alpha_3\alpha_4}} n_{\alpha_1\alpha_2\alpha_3}\Omega_{\alpha_3\alpha_4}\Bigg). \end{aligned} \tag{12.130}$$

Here the orientations on the quadruple overlaps $V_{\alpha\beta\gamma\delta}$ are defined by

$$\partial V_{\alpha\beta\gamma} = \sum_{\gamma} V_{\alpha\beta\gamma\delta}, \tag{12.131}$$

with everything else defined as in our discussion of the Dirac monopole.

Chern-Simons theory is a simple example of a **topological field theory**, meaing a quantum field theory that does not require a spacetime metric $g_{\mu\nu}$. It has many applications, but unfortunately we do not have time to discuss any of them in detail so I will just mention a few:

- Chern-Simons theory is the low-energy theory of the integer and fractional quantum hall effects, which are quite remarkable phenomena in condensed matter physics where a certain electrical conductivity can take only certain discrete values. The integer quantum hall effect is described by $U(1)$ Chern Simons theory, with the level k giving the number of filled Landau levels. The fractional quantum hall effect is described by a mixed Chern-Simons theory with two gauge fields. There is also some evidence for non-abelian Chern-Simons theory in the $\nu = 5/2$ fractional quantum hall effect, with an $SU(2)$ gauge field at level $k = 2$ coupled to a $U(1)$ gauge field at level $k = 8$.
- String theory involves a number of gauge fields, including some where A is a more general p-form instead of a one-form, and its low-energy effective action includes a number of Chern-Simons type terms.
- Chern-Simons theory has many interesting applications in mathematics, for example Witten won the Fields medal for showing that the expectation value of a Wilson loop in Chern-Simons theory whose path ties a knot is related to a famous topological invariant of that knot called the Jones polynomial.

12.9 Wess-Zumino-Witten term

As a final application of topological methods we will briefly return to the Wess-Zumino-Witten term in chiral effective field theory, which we now have the tools to understand. It is based on the five-form version of the Maurer-Cartan topological charge, which is given by

$$\nu_5 \equiv \frac{i}{240\pi^3} \int_{\Sigma_5} \mathrm{Tr}\,(\omega \wedge \omega \wedge \omega \wedge \omega \wedge \omega) \tag{12.132}$$

where Σ_5 is an arbitrary oriented five-manifold. The method of equation (12.68) can again be used to show that this is an integer, see my upcoming paper. The Wess-Zumino-Witten term is a four-dimensional term in chiral effective theory whose relation to ν_5 is the same as the relationship of the Chern-Simons action in three-dimensions to the second Chern class. More concretely let $U : M_4 \to SU(3)$ be the Goldstone field for pions, kaons, and the η on a spacetime manifold M that we can realize as the boundary of some five-manifold $\widetilde{M}$ with boundary on which U extends to a map $\widetilde{U}$. Then the Wess-Zumino-Witten contribution to the Goldstone theory can be written as

$$S \supset \frac{in}{240\pi^3} \int_{\widetilde{M}} \mathrm{Tr}\,(\widetilde{\omega} \wedge \widetilde{\omega} \wedge \widetilde{\omega} \wedge \widetilde{\omega} \wedge \widetilde{\omega}) \tag{12.133}$$

with

$$\widetilde{\omega} = \widetilde{U}^{-1} d\widetilde{U}, \tag{12.134}$$

where n is an integer for the same reason it needed to be in Chern-Simons theory. It also has an intrinsic presentation analogous to (12.130), see again my upcoming paper. The WZW term turns out to be needed with $n \neq 0$ for anomaly matching (for example once we turn on a background gauge field it includes the term that leads to $\pi_0 \to \gamma\gamma$), and it also accounts for the observed process $KK \to \pi\pi\pi$. Mathematically we can think of it as probing the homotopy group $\pi_5(SU(3)) \cong \mathbb{Z}$.

Homework

1. Confirm the consistency of (12.16) with (12.10) and (12.13).
2. Confirm (12.37) and (12.38).
3. Confirm the parametrization (12.83) of $SU(2)$.
4. Confirm (12.79), (12.80), and (12.82)

13 What next?

We have reached the end of QFT III, but we have certainly not reached the end of quantum field theory. Indeed there are many important topics that I was not able to discuss, and if you stay in theoretical physics you will see that learning quantum field theory is a lifelong process. Here are a few of the things I wish I had had time to talk about:

- **Conformal field theories:** In our discussion of the renormalization group back in QFT I we saw that field theories that are invariant under rescaling of the spacetime coordinates are the natural starting and ending points of the renormalization group flow. These theories have many special properties, and in particular in $1+1$ spacetime dimensions they are rather well-understood. A good place to look to learn the basics of CFT is David Simmons-Duffin's TASI lectures, and there is also chapter 15 of Polchinski's string theory book and the "big yellow book" of Senechal, Di Francesco, and Mathieu.

- **Phases of gauge theory:** We have learned that gauge fields can be in deconfined, confined, and Higgs phases. But what are these phases really? This question turns out to be rather slippery to answer in general, since at first glance the confining and Higgs phases do not have any robust distinguishing features (to be contrasted with a broken phase of a global symmetry, which reliably has domain walls or Goldstone bosons). On the other hand these phases can sometimes have interesting topological structure in the infrared, and recently this has been understood rather systematically using the idea of "generalized global symmetries". Some references I like for this are the classic paper of Fradkin and Shenker on the phase diagram of $\mathbb{Z}_2$ lattice gauge theory, the paper "generalized global symmetries" by Gaiotto, Kapustin, Seiberg, and Willett, my own paper with Hirosi Ooguri on symmetries in field theory and gravity, and Shu-heng Shao's TASI lecture. I also very much like Polyakov's textbook "Gauge fields and strings", although parts of it are dated.

- **Large N expansion:** Quantum field theories sometimes simplify in the limit of a large number of fields. In particular this is true for the $O(N)$ vector model at large N, and also to some extent $SU(N)$ Yang-Mills theory. A nice reference for this is the chapter in Coleman's book "Aspects of Symmetry", which also has many other great things in it. Polyakov talks about this in his book as well.

- **Supersymmetry:** We did have one homework problem on supersymmetry in QFT II, but there is much more that could be said about it. In particular supersymmetry sometimes enables exact calculations in interacting field theories, including those with strong interactions and confinement. It is thus a powerful laboratory for testing ideas about quantum field theory. Useful references for this are Weinberg volume III for the physics of SUSY, the book of Wess and Bagger for formalism, and the books of Argyres and Terning.

- **Physics beyond the standard model:** The standard model has passed many experimental tests, but it also has several major deficiencies. Chief among these are its lack of gravity, dark matter, and neutrino masses, and it also has cosmological problems since it cannot explain the observed baryon number density of the universe or give rise to the density perturbations that led to the formation of stars and galaxies (this seems to require something like inflation). There are also the apparent fine tunings that make the Higgs mass and cosmological constant small in Planck units, as well as the QCD θ angle small. Many ideas have been proposed to address these problems, including supersymmetry, grand unification, axions, and more. Some of them are discussed in volumes II and III of Weinberg's book.