Lecture Notes on

The Standard Model

Kevin Zhou

kzhou7@gmail.com

These notes cover the Standard Model and some of its extensions. The primary sources were:

- Griffiths, *Introduction to Elementary Particles*. An excellent and readable introduction that forms the first section of these notes. Gets the reader up to speed with the history of particle physics and Feynman diagrams for the SM with a minimum of field theory.
- Christopher Thomas's Part III Standard Model course. A clear introduction.
- John March-Russell's Standard Model and Beyond course. A course from the Oxford MMath-Phys program giving a lot of wisdom on the motivation for and methodology of model building.
- Burgess and Moore, *The Standard Model: A Primer*. Covers the dynamics of the SM at tree level in detail, with clear discussions, and considers theories beyond the SM through the lens of effective field theory. It begins with a clean introduction to quantum field theory, but realistically one would need prior exposure to make sense of it. The book is especially useful because it describes intuitive ideas commonly used in collider physics which aren't well-covered in dedicated quantum field theory books.
- Georgi, *Weak Interactions and Modern Particle Theory*. Covers many important SM topics that are omitted in introductory textbooks, such as chiral perturbation theory. Written in the inimitable Georgi style: irreverent and direct, always going straight to the physics.
- Donoghue, Golowich, and Holstein, *Dynamics of the Standard Model*. An authoritative monograph and useful reference.

The sources cited in the notes on Quantum Field Theory were also used. The gauge theory conventions here differ from those used there, but match those of Peskin and Schroeder. The most recent version is here; please report any errors found to kzhou7@gmail.com.

Contents

1 Introduction

1.1 Particle and Interactions

First, we summarize some ancient history.

- In the early 1960s, the Eightfold Way was introduced, followed by the quark model. Quark confinement was postulated to explain why free quarks had not been seen, while quark color was added for consistency with the Pauli exclusion principle.
- By the late 1960s and early 1970s, deep inelastic scattering experiments at SLAC and CERN found evidence for substructure in the proton, pointlike particles called partons, analogous to how Rutherford had found substructure in the atom. However, there was still widespread skepticism over the quark model, so partons were not identified with quarks.
- In 1974, groups led by Ting at Brookhaven and Richter at SLAC simultaneously found a new particle, called the J/ψ, a meson which was both extremely heavy and relatively long-lived. This set off a flurry of theoretical activity called the November revolution.
- The quark model explained the new particle as a bound state of a charm and anti-charm quark; it has excited states in analogy with positronium, which are collectively called "charmonium". It also predicted many new mesons and baryons containing charm quarks, organized into multiplets by group theory, which were quickly found.
- At this point, the elementary particles could be organized into families containing two quarks and one lepton each, but measurements of CP violation motivated a third generation. In 1975, the tau lepton was found. Then, a few years later, the upsilon meson was found and postulated to be a bottom and anti-bottom quark. Throughout the 1980s, more B mesons were discovered, and today LHCb and Belle II are devoted to studying them.
- The top quark would complete the third generation, and measurements of $B^0/\overline{B}^0$ oscillations indicated it had a huge mass. It was thus too heavy to be produced until 1995, at Fermilab's Tevatron, leading to the table of masses below.

particle	mass	mass determined by
up	2.2 MeV	lattice computations of light meson/baryon masses
down	4.7 MeV	lattice computations of light meson/baryon masses
strange	95 MeV	lattice computations of light meson/baryon masses
charm	1.3 GeV	charmonium and D meson masses
bottom	4.2 GeV	bottomonium and B meson masses
top	170 GeV	production at LHC
e^-	0.5 MeV	hydrogen spectroscopy
μ^-	106 MeV	muonium hyperfine splitting
τ^-	1.78 GeV	production at BES III
$W^\pm$	80 GeV	production at LEP
Z	91 GeV	production at LEP
H^0	125 GeV	production at LHC

Note that quarks don't exist as free particles, so the definition of their mass is somewhat ambiguous. For instance, the up mass may be as high as 5 MeV under some definitions.

- Finally, the weak interaction (as understood by Fermi's effective four-fermion interaction) was suspected to be mediated by an intermediate vector boson. By 1960, Glashow had formulated a unified electroweak theory, though there was no mechanism to break the symmetry. In 1964, the Higgs mechanism was discovered, and in 1967, Weinberg and Salam showed how it could break electroweak symmetry, predicting the $W^{\pm}$ and Z bosons.
- In 1983, the intermediate vector bosons were discovered at CERN's super proton synchrotron. In the 1990s, LEP was constructed to perform precision tests of the electroweak theory.

Next, we summarize the Standard Model (SM) interactions.

- The basic QED and QCD vertices are simple: charged particles emit photons and colored particles emit gluons. There are also ggg and $gggg$ vertices.
- The fundamental neural weak vertex is $Zf\overline{f}$ for any fermion f. For example, a Z could mediate neutrino-electron scattering. The Z behaves a lot like the photon, except that it also couples to neutrinos.
- The fundamental leptonic charged weak vertex is $W\ell\nu$, i.e. a W boson can decay into a lepton and its corresponding antineutrino. This vertex mediates the decay of the muon.
- Finally, the quark charged weak vertex converts an up-type quark to a down-type quark, e.g. Wud'. However, the quarks are defined in the mass basis, while this interaction is diagonalized in a different basis. Then W emission can convert an up quark to a down quark, but also to a strange or bottom quark.
- The two bases are related by the CKM matrix; the magnitudes of the matrix elements are

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} 0.974 & 0.227 & 0.004 \\ 0.227 & 0.973 & 0.042 \\ 0.008 & 0.042 & 0.999 \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}.$$

  Then most weak decays stay in the same generation. Crossing between adjacent generations is rare but possible, while crossing between the first and third generation is rarer still.
- For example, a charged weak interaction mediates beta decay. A baryon can decay into another baryon while emitting a photon by emitting and reabsorbing a W, which emits a photon. (This photon is necessary by energy-momentum conservation.)
- It is now known that neutrinos have mass. Neutrinos are defined in the flavor basis, so the weak vertex is unchanged. Instead, the flavor eigenstates oscillate into each other due to their mismatch with the mass eigenstates. The analogue of the CKM matrix is the PMNS matrix.
- The force carriers have interactions WWZ, $WWZZ$, $WWWW$, $WW\gamma$, $WW\gamma Z$, and $WW\gamma\gamma$.
- In the SM, the Higgs interacts with all fermions by Yukawa couplings. It also interacts with the weak mediators with vertices WWH, ZZH, $WWHH$, and $ZZHH$, and with itself as HHH and $HHHH$. One-loop diagrams also provide effective vertices ggH, $\gamma\gamma H$, and γZH.

- At LEP, the Higgs was searched for by Z Bremsstrahlung,

$$e^+ + e^- \to Z^* \to Z + H.$$

  This ruled out Higgs masses below 114 GeV, while perturbative unitarity arguments disfavored masses above 200 GeV.

- At hadron colliders, the main production mechanism is actually a one-loop "gluon fusion" process $gg \to H$, which dominantly goes through a t loop, because the heavy t quark has the largest coupling with the Higgs. (This is an example of "non-decoupling". We usually expect heavy particles to be irrelevant, but they remain relevant because they must couple more strongly to the Higgs to get mass. Thus, Higgs measurements already rule out a heavy 4th generation that gets mass the same way as the other 3.)

- There are other important production mechanisms which occur at tree level, including:
  - $gg \to g^* \to t\bar{t}H$ (radiation off a top quark, measuring the top Yukawa)
  - $q\overline{q} \to q\overline{q}H$ (each quark radiates a W or Z, and the two fuse)
  - $q\overline{q} \to V^* \to VH$ (W or Z Bremsstrahlung)

  These processes have all been observed and match Standard Model expectations to around the 20% level. Note that direct quark fusion, $q\overline{q} \to H$, is unlikely because all Yukawa couplings besides the top are small.

- The reason gluon fusion is dominant, despite being loop-induced, is because (1) it doesn't involve any weak couplings, and (2) at proton-proton colliders, the constituents ("partons") are dominantly gluons and quarks, with a smaller contribution from antiquarks. (There are some antiquarks, because of QCD effects, and the specific amounts of each particle are characterized by parton distribution functions.)

- The most likely Higgs decays are to pairs of heavy particles, such as τ, b, W, Z, or t. All of these are possible (though suppressed) since the W, Z, and t may be produced as virtual particles which then decay, though the Higgs is light enough for $H \to t\bar{t}$ to be very rare.

- The discovery of the Higgs was made by the clean decay channels of $H \to \gamma\gamma$, which occurs through a loop diagram, and $H \to ZZ \to 4$ leptons, though couplings to the other particles were later measured as well.

- One way that experimentalists characterize their knowledge of the Higgs couplings is the "κ framework", which is an ad hoc scaling of each Higgs coupling by a factor κ_i. This violates gauge invariance and doesn't yield a consistent quantum field theory, so one can't compute higher order corrections, or get accurate kinematics. But measurements indicate $\kappa_i \approx 1$ to about 10% accuracy for the weak bosons and the heaviest fermions.

- At the same time, over the past few decades, more accurate calculations have reduced the theory uncertainty, with the standard now "next to next to next to leading order" (N^3LO), which is good to within a few percent. However, more accurate calculations will be needed for the "high luminosity" (HL) phase of the LHC, which will further shrink experimental uncertainties.

- We can also try to measure the Higgs trilinear coupling, which would help confirm the nature of its potential. Varying the trilinear coupling affects the rate of two Higgs production, though the uncertainties at the HL-LHC will remain of order 100%.
- Proposed future "Higgs factory" e^+e^- colliders, such as the ILC, could substantially improve the precision. For instance, they can operate at around 230 GeV to produce Higgs from Z Bremsstrahlung (slightly above $m_H + m_Z$, to remove phase space suppression), around 250 GeV to produce Higgs pairs, and at 350 GeV to produce top pairs and measure the top mass. It will also be possible to directly measure the Higgs width, since the kinematics is much cleaner.
- On all these fronts, we could do much better with a $\mu^+\mu^-$ collider, since muons couple much more strongly to the Higgs, but the technology to build one doesn't yet exist. It's not hard to get the muons (e.g. electron-positron colliders make the positrons by just directing a beam into a wall), but it seems challenging to form the muons into a beam before they decay.

Note. Before going on, it's nice to step back and appreciate the massive engineering effort that goes into particle colliders. The Large Hadron Collider (LHC) sits inside a 27 km circular tunnel, buried 100 m underground due to a combination of radiation shielding and political reasons. Inside the tunnel, beams of protons circulate in two separate tubes in opposite directions. There are thousands of superconducting NbTi magnets distributed throughout, to keep the beam going in a circle. This acceleration causes energy loss due to Bremsstrahlung, so the protons are pushed along by thousands of superconducting radiofrequency (SRF) cavities, whose oscillating fields are timed to accelerate the protons as they pass by. The whole system must be cooled with liquid helium to cryogenic temperatures, to maintain superconductivity. An unplanned rise in temperature can cause an explosive "magnet quench", which knocked the LHC out of commission in 2008.

The proton beams are organized into thousands of "bunches" of about 10^{11} protons each, spaced only 25 ns apart. They are focused to a transverse size of 16 microns (smaller for the upcoming "High Luminosity" LHC) to increase the chance of an interesting event when the bunches collide at the centers of the detectors. Maintaining this high beam quality is an entire field of study, involving hundreds of physicists and a number of dedicated journals. The beam is carefully focused using about a thousand specialized magnets, such as quadrupoles, sextupoles, octupoles, and decapoles. And of course, the entire beamline needs to be a vacuum as empty as outer space, to avoid scattering the beam protons.

The proton beam for the LHC needs to already be at a relatively high energy before entering, so the energy is built up through a series of smaller accelerators. Protons are injected into the LHC from the Super Proton Synchrotron (SPS), which discovered the W and Z bosons. The SPS in turn receives them from the Proton Synchrotron, which gets them from the Proton Synchrotron Booster, which gets them from Linac4, which gets them from ionizing hydrogen, which comes from a single little bottle of hydrogen gas. The beam itself degrades as more collisions happen, so it needs to be safely "dumped" and reformed every few hours, repeating this entire process.

The LHC supports many experiments. People often think of ATLAS and CMS, the general purpose experiments which analyze high-energy proton-proton collisions. But there's also ALICE (measuring collisions of lead nuclei, to study quark gluon plasma), LHCb (with a specialized detector to see hadrons containing b quarks, to study CP violation and flavor anomalies), LHCF and TOTEM (downstream of the ATLAS and CMS collision points, to study forward produced particles for cosmic ray physics), MoEDAL-MAPP (near LHCb, to search for produced magnetic monopoles), and soon FASER (far downstream of the ATLAS collision point, to study new light particles). There are also

some proposed small experiments, such as MATHUSLA (an above-ground detector, to see long-lived particles), MilliQan (millicharged particles), and CODEX-b (like MATHUSLA, but near LHCb).

The enormous detectors in these experiments use a variety of techniques to infer what happened in the collision. Powerful magnets bend the trajectories of charged particles, measuring their momentum. Calorimeters are solid components which absorb energy from the particles through the ionization or excitation of particles, which measures their energy. There are also "trackers", often filled with sparse gas, which allow the particles to pass through with lower energy loss. Particles going through the trackers leave a trail of ionized gas, which can be pushed towards a detector with an electric field. Modern trackers have exquisite sensitivity, allowing the paths of particles to be known to finer than millimeter precision. The rate at which the particles lose energy is also measured and can be described by the Bethe–Bloch formula. All of these pieces of information are used together to identify the particle and figure out what happened in the collision.

1.2 Symmetries and Conservation Laws

Next, we summarize considerations for computing decay rates and cross sections.

- The SM has three conserved quantities: charge, baryon number (or equivalently quark number) and lepton number. If we ignore neutrino oscillations, the individual lepton flavors are conserved. If we ignore charged weak interactions, the individual quark numbers are conserved. If we ignore all weak interactions, parity is conserved.
- Energy conservation forbids decays of particles into heavier particles. It places no restriction on scattering, since the incoming energy can be arbitrary.
- A decay is more likely if the products are much lighter than the decaying particle, because there is more available phase space volume; this is the reason neutron decay is so slow.
- Generally, a strong decay takes about 10^{-23} seconds, an electromagnetic decay takes 10^{-16}, and a weak decay takes at least 10^{-13}, higher if generation mixing occurs.
- Finally, the OZI/Zweig rule states that any diagram which can be cut in half by only cutting gluon lines is suppressed. This is because the reaction requires hard gluons, and QCD is weak at high energies.

Next, we give a qualitative overview of the discrete symmetries of the SM.

- In 1956, Sachs and Wu found that nature did not respect parity symmetry. In the beta decay of cobalt 60, it was found that the emitted electron came out opposite to the nuclear spin, which may be aligned with a magnetic field. This violates parity since spin and magnetic fields are axial vectors while the emission velocity is a true vector.
- Parity violation can also be seen from the helicity of the neutrino, which is Lorentz invariant assuming the neutrino is massless. Defining right-handed helicity to mean spin pointing along the direction of motion, we would expect neutrinos to be left-handed and right-handed with equal frequency by parity invariance. Instead, experiments find that all neutrinos are left-handed and all antineutrinos are right-handed, where the helicity of the neutrino is not directly measured but inferred from the helicities of the other products of a decay.
- In the absence of the weak force, parity places constraints on allowed particle decays.

  - Vector particles such as the photon have parity -1 and axial vectors have parity $+1$. Conversely, scalars have parity $+1$ and pseudoscalars have parity -1.
  - The individual quark numbers and lepton numbers are conserved, so we are free to assign any parity to them. By convention, we assign parity $+1$ to leptons and quarks (and thereby also to the proton and neutron). We will show below that this implies parity -1 for antiquarks and antileptons. In particular, this means that a bound state consisting of a fermion and its antiparticle has a factor of -1 in its intrinsic parity.
  - Note that parity cannot be defined for neutrinos in the Standard Model, as it would map a left-helicity neutrino to a right-helicity neutrino, which doesn't exist.
  - In a two-body decay with angular momentum ℓ, there is an extra factor of $(-1)^\ell$.

- For example, with $\ell = 0$, we expect to have pseudoscalar and vector mesons, corresponding to spin 0 and 1 respectively. These are indeed the lowest-energy meson octets; there are also higher-energy positive parity meson octets corresponding to excited states with $\ell = 1$.

- One early hint of parity violation was the 'theta–tau' puzzle in the 1950s. Two mesons, called the θ and the τ, decayed as

$$\theta^+ \to \pi^+ + \pi^0, \quad \tau^+ \to \pi^+ + \pi^0 + \pi^0.$$

  Every particle involved has spin 0, so there can be no orbital angular momentum. Then the θ and τ must have parity 1 and -1, but they had nearly the same mass. The resolution is that they are indeed the same particle, the K^+, and the first decay violates parity.

- Next, we turn to charge conjugation, which replaces particles with antiparticles by flipping the sign of all internal quantum numbers. Only particles that are their own antiparticles can be eigenstates of C, severely restricting its use.
  - The photon has $C = -1$, since it is sourced classically by a current which flips under C.
  - Consider a spin 1/2 particle and its antiparticle with total orbital angular momentum ℓ and total spin s. We get a factor of $(-1)^\ell$ from the orbital part, a factor of -1 from identical particle exchange, and a factor of $(-1)^{s+1}$ from the antisymmetry/symmetry of the singlet/triplet. Then $C = (-1)^{\ell+s}$.
  - For example, the neutron pion π^0 has $C = +1$, so it can't decay into an odd number of photons.

- For the strong interactions, where isospin is conserved, we may define the G-parity

$$G = Ce^{i\pi I_2}.$$

  This is more useful because charged mesons can have definite G-parity. For example, the charged pion π^+ is mapped to π^- and then back to π^+ by C, so it has definite G-parity.

Next, we turn to the subtler case of CP symmetry.

- As we've seen, the leptonic weak decays violate parity. For example, in the decay

$$\pi^+ \to \mu^+ + \nu_\mu$$

the antimuon is always left-handed, while it would be right-handed in the parity-flipped version. Similarly, in the decay

$$\pi^- \to \mu^- + \overline{\nu}_\mu$$

the muon is always right-handed. Thus C symmetry is also violated, since the charge conjugate would have a right-handed antimuon, but CP symmetry is not.

- One useful system for testing CP symmetry is the decay of the neutral kaon K^0 and $\overline{K}^0$. The K^0 and $\overline{K}^0$ mix by a $W^\pm$ loop, so neutral kaons found in the lab are mixtures of the two. Both K^0 and $\overline{K}^0$ have $P = -1$ and $C = +1$, and the states

$$|K_1\rangle = |K^0\rangle - |\overline{K}^0\rangle, \quad |K_2\rangle = |K^0\rangle + |\overline{K}^0\rangle$$

have $CP = +1$ and $CP = -1$ respectively.

- Assuming CP is conserved in the weak interactions, $|K_1\rangle$ and $|K_2\rangle$ decay in different ways. Since a pion has $CP = -1$, the most common decays are

$$K_1 \to 2\pi, \quad K_2 \to 3\pi.$$

The first decay is much faster because there is more phase space available. Therefore, a neutral kaon should quickly turn into an eigenstate $|K_2\rangle$ of CP. Concretely, this means that a beam of kaons should initially have many 2π decays, and later have only 3π decays.

- The neutral kaons provide another example where two bases mismatch, and the one we use is dictated by convenience. If we are studying strong interactions, we want the K^0 and $\overline{K}^0$, but if we are studying weak decays, we want the K_1 and K_2.

- In 1964, the Cronin–Fitch experiment established that the weak interaction does not conserve CP. This was done by taking a beam of neutral kaons, waiting for a time much greater than the lifetime of the K_1, and detecting residual 2π decays (about 0.2% of the total); this is only possible if the K_2 decay violates CP.

- In general, two particles can mix if they have approximately the same mass and the same relevant conserved quantities. For example, the particles always must have the same baryon number, but they don't have to have the same isospin if the mixing is by a weak process. In the case of neutrinos, lepton number is violated, but this is acceptable as the neutrino mass terms explicitly break lepton number conservation.

- Similarly, the B^0 and $\overline{B}^0$ mesons can mix. Oscillations between the B^0 and $\overline{B}^0$ were observed at Fermilab in 2006, and the decays of the B mesons have been observed to violate CP. Most measurements of CP violation are with neutral B-mesons or neutral kaons; many of the few other candidates are long-lived enough.

Next, we turn to T and CPT symmetry.

- T symmetry does not forbid decays, since no particle is in an eigenstate of T. It imposes detailed balance for reactions, but it is often difficult to set up a backwards reaction. For example, the reverse of the weak decay $\Lambda \to p^+ + \pi^-$ is difficult to observe because the proton and pion interact by the strong force. To remove contamination from other forces, we might turn to neutrinos, but it is hard to control them.

- As such, the most sensitive searches for T violation come from measurements of the electric dipole moment of elementary particles; such a dipole moment would violate both P and T. So far, all experiments have found the dipole moment to be zero within error.
- However, we do expect T violation to occur, since CPT must be a symmetry and CP is violated. T violation has been directly observed at BaBar at SLAC in 2012.
- One prediction of CPT symmetry is that every particle must have the same mass and lifetime as its antiparticle. (This is also true of C symmetry, but we know that to be broken.) This has been verified to great accuracy for many particles. Another result is that helicities must be symmetric about zero: if there exists a helicity λ state, there must also be a corresponding helicity $-\lambda$ state.
- One should always keep in mind that just about all of these exact symmetries, such as Lorentz symmetry, CPT symmetry, etc. are all strongly spontaneously broken in our universe. For example, the presence of the CMB breaks Lorentz symmetry. When we talk about verifying these symmetries, we always imagine experiments that are insensitive to the symmetry-breaking background.

Finally, we turn to the sources and consequences of CP and T violation.

- CP violation can be caused by complex phases in the CKM or PMNS matrices; the former is what accounts for CP violation in kaon and B-meson decay. For $n < 3$ generations such phases can always be removed by redefining the quark fields, so the observation of CP violation led to the prediction of a third generation, where there is one residual phase.
- CP violation can also come from $F\widetilde{F} = F \wedge F$ terms for the electromagnetic, weak, and strong forces; specifically, such a term breaks both P and CP. However, this term is more subtle since it is a total derivative, and hence a boundary term.
  - In electromagnetism, the term always integrates to zero for finite-energy configurations; there are no $U(1)$ instantons.
  - The chiral anomaly allows the weak theta term to be removed by a rotation of all quark fields, which is just the symmetry $U(1)_B$, as only left-chiral quarks couple to the weak force.
  - For the strong force, the term has a nontrivial effect, and induces a neutron electric dipole moment. Experiments indicate that this term is very small; the strong CP problem asks for a natural explanation.
- CP violation is said to "distinguish matter from antimatter". This comes from its presence in the Sakharov conditions for baryogenesis, the origin of a net matter-antimatter asymmetry in the universe. They are:
  - Violation of baryon number conservation, i.e. the existence of reactions $i \to f$ where i and f have different baryon number. This can be supplied by grand unified theories; it also occurs extremely rarely in the SM through sphalerons, as discussed in the notes on Quantum Field Theory.
  - C violation. This is necessary since otherwise the net baryon number produced by $i \to f$ will be balanced by $i^* \to f^*$. The SM has very strong C violation.

  - CP violation. This is necessary since otherwise $i \to f$ will be balanced by $i_P^* \to f_P^*$. The CP violation in the SM is quite small, and probably not enough to account for baryogenesis.
  - Departure from thermal equilibrium. Otherwise, $i \to f$ will be balanced by $f \to i$, by detailed balance. Equivalently, we can't go from $\mu_B = 0$ to $\mu_B \neq 0$. We could exit equilibrium, e.g. after a first-order phase transition. However, the electroweak and QCD phase transitions appears be smooth crossovers in the SM, which cannot do the job.

  As discussed in detail below, it can be ambiguous to define C or CP, or in some extreme cases impossible to define them at all. "C and CP violation" really stands for the absence of any symmetry which would relate processes, causing the net baryon number produced to cancel.

- One alternative possibility is "leptogenesis". In the SM, B and L are violated by nonperturbative effects while $B - L$ remains conserved; then leptons can be created, and turn into baryons. Note that the lepton number of the universe might or might not be zero, since we can't measure the neutrinos well.

Note. How complex phases in the CKM matrix cause observable CP violation. Consider a process and its CP-reverse. Under the standard electroweak theory, the matrix elements are

$$\mathcal{M} \sim |\mathcal{M}| e^{i\phi} e^{i\theta}, \quad \widetilde{\mathcal{M}} \sim |\mathcal{M}| e^{i\phi} e^{-i\theta}$$

where the CKM phase $e^{i\phi}$ is not conjugated. However, the magnitudes of the amplitudes are the same, so this has no observable effect. However, if the process can occur in multiple ways,

$$\mathcal{M} \sim |\mathcal{M}_1| e^{i\phi_1} e^{i\theta_1} + |\mathcal{M}_2| e^{i\phi_2} e^{i\theta_2}$$

then $|\mathcal{M}|$ and $|\widetilde{\mathcal{M}}|$ can differ because the terms can interfere differently. In the case of B-meson decay, there is a tree level process, and the next most significant contribution is from a "penguin diagram" involving a W loop. In fact, this is generic: one can show that the leading CP violation *must* involve loop-level processes; interference between just tree-level processes isn't enough. As a result, CP violation is in some sense always small, no matter how large the phases are, since it's always loop suppressed.

1.3 Bound States

Next, we briefly discuss nonrelativistic bound states.

- First, note that the potential energy and kinetic energy should be of the same order by the virial theorem. Thus a system is nonrelativistic if its binding energy is small compared with its mass energy.
- For example, light quark bound states are always relativistic, but charmonium $c\bar{c}$ and bottomonium $b\bar{b}$ are not. We don't count toponium $t\bar{t}$ since its lifetime is too short to be observed.
- The archetypical example of such a system is the hydrogen atom, where the energy levels are $\alpha^2 mc^2$ where m is the mass of the electron.
  - Fine structure comes from the lowest-order relativistic correction and the spin-orbit coupling of the electron. It can be calculated with the Dirac equation and contributes $\alpha^4 mc^2$.
  - The Lamb shift comes from QED effects and contributes $\alpha^5 mc^2$.

  - The hyperfine structure comes from the spin of the proton. The proton's magnetic moment is much smaller than that of the electron since it is much heavier. It has a spin-spin interaction as well as a spin-orbit coupling with the electron and contributes $(m/m_p)\alpha^4 mc^2$.

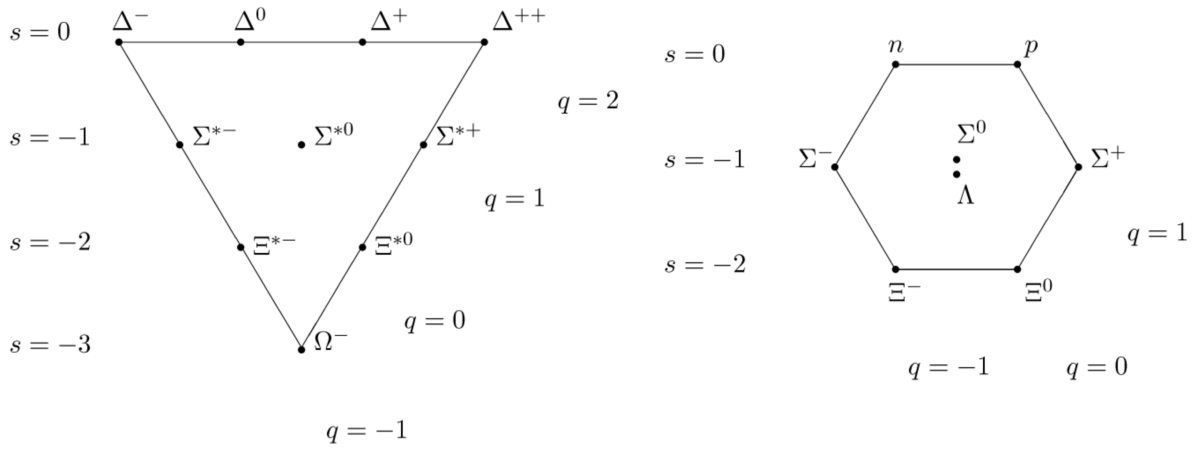
- Positronium behaves very similarly to hydrogen; using the reduced mass, its energy levels are like those of hydrogen with an electron half as massive.
  - One major difference is that the hyperfine structure is now of the same order as the fine structure.
  - Since both the particles move quickly, there is another $\alpha^4 mc^2$ correction due to the propagation time for the electromagnetic field.
  - The electron and positron can also temporarily annihilate into a virtual photon. Since the probability for this process is proportional to $|\psi(0)|^2$, at lowest order it only occurs for $\ell = 0$. Since the photon has spin one, it only occurs in the triplet configuration $s = 1$.
  - Finally, the electron and positron can annihilate. Positronium has C eigenvalue $(-1)^{\ell+s}$ while a state with n photons has $(-1)^n$, restricting the number of photons produced. By the same logic as above, annihilation only occurs at lowest order for $\ell = 0$, usually producing 2 photons for the spin singlet and 3 for the spin triplet. The decay of the triplet state (called ortho-positronium) is hence slower, by roughly a factor of α.

- Finally, we turn to 'quarkonium', a system of a quark and its antiquark. In this case, the energy levels are far enough apart that we regard excited states as entirely different particles.
  - The energy levels can be found numerically using the ansatz

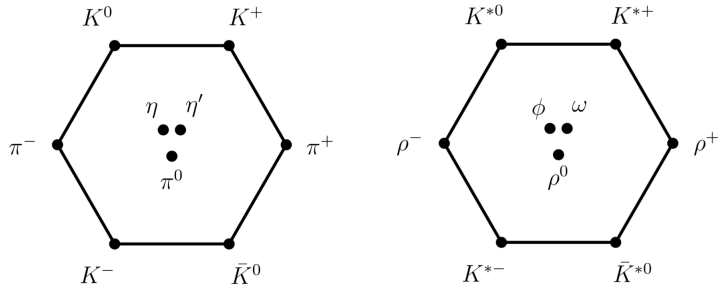
$$V(r) = -\frac{4}{3}\frac{\alpha_s}{r} + F_0 r$$

    where the factor of 4/3 is a color factor. The second factor is purely an empirical guess; the numerics also work well if we use $\log r$ or r^2.
  - When the J/ψ was discovered, in the "November revolution" of 1974, it was quickly identified as the 1^3S_1 state of charmonium, where this notation should be read as $n = 1$, $2s + 1 = 3$, $\ell = 0$, and $j = 1$, where j is the total angular momentum; it was known that $s = 1$ because it was produced through a virtual photon.
  - Shortly afterward, the 1^1S_0 and 2^3S_1 states were found, and eventually all of the $n = 1$ and $n = 2$ states. For $n \leq 2$, charmonium decays slowly by an OZI-suppressed process where the c and $\bar{c}$ annihilate to gluons, making it an exceptionally narrow resonance, but for $n > 2$ charmonium decays quickly via $\psi \to D^+D^-$.
  - In 1977, the upsilon meson Υ was quickly identified as the 1^3S_1 state of bottomonium, and states of bottomonium have been found up to $n = 6$.

Note. Before tackling the light mesons and baryons, we show the lowest energy meson nonets and the lowest energy baryon octet and decouplet. The baryons are shown below.



The pseudoscalar and vector mesons are shown below, with the pseudoscalars at left.



Next, we consider the light quark mesons. In this case, we can't say anything quantitative about the bound state masses, so we focus on the wavefunctions.

- For simplicity, we consider the ground state $n = 1$ and $\ell = 0$, so we only have to think about the quark flavor and spin. Since isospin and spin commute, the mesons can be organized into groups of definite flavor $\mathfrak{su}(3)$ representation and spin, e.g. the meson octets and singlets we've already seen. Excited states will give further meson multiplets.
- In our case, we expect $3 \times 3 \times 2 \times 2 = 36$ states in total. They are organized into:
  - A pseudoscalar octet with spin $J = 0$, containing 8 states.
  - A pseudoscalar singlet η' with spin $J = 0$, containing 1 state.
  - A vector octet with spin $J = 1$, containing 24 states.
  - A vector singlet with spin $J = 1$, containing 3 states.

  The octet and singlet can be superposed to form a nonet.
- Now, we can write down the wavefunctions of these states. Since we've restricted to $n = 1$ and $\ell = 0$, the position space wavefunction is rather trivial; the color wavefunction must simply be the color singlet, and the spin wavefunction is totally independent of the flavor wavefunction. So the only nontrivial part we need is the flavor wavefunction.
- We focus on the states with $I_3 = 0$, which occupy the center of the pseudoscalar nonet. The pions in this row form an isospin triplet, so the π^0 is the $I_3 = 0$ state of the triplet, $u\overline{u} - d\overline{d}$.

(Naively we would think there should be a plus sign here, because a minus sign indicates a spin singlet state, but the minus sign is correct, as explained below.)

- The remaining states with $I_3 = 0$ are

$$\eta = u\overline{u} + d\overline{d} - 2s\overline{s}, \quad \eta' = u\overline{u} + d\overline{d} + s\overline{s}$$

where the η' is just the $\mathfrak{su}(3)$ singlet.

- The situation is slightly different for the vector mesons; here the equivalents of the η and η' mix, to form the physical states

$$\omega = u\overline{u} + d\overline{d}, \quad \phi = s\overline{s}.$$

This occurs because $\mathfrak{su}(3)$ is broken by quark masses, and the strange quark is quite heavy. The reason this mixing doesn't happen for η and η' is that the singlet η' has a large contribution to its mass due to instanton effects.

- The ϕ is the strange quark analogue of the J/ψ. It decays slowly because it's too light to decay into two mesons with one s or $\overline{s}$ each, and an $s\overline{s} \to g^* \to \ldots$ decay is OZI suppressed; in fact, this was how the OZI rule was discovered.

- The mesons in the pseudoscalar nonet and vector nonet differ in mass, so part of the strong force must be spin-dependent. A good empirical model for the meson masses is

$$M = m_1 + m_2 + A\frac{\mathbf{S}_1 \cdot \mathbf{S}_2}{m_1 m_2}$$

where we divide the spins by the masses to get magnetic moments, and the constant A and the effective quark masses m_i are fit numerically.

- Heuristically, the effective quark masses account for the bare quark masses and the QCD field energy each quark carries around. The other term accounts for spin-spin coupling through color magnetic moments, which are inversely proportional to the masses. We've already accounted for the "color electric" force; it just binds the quarks together independent of their flavor and is counted in the effective masses.

Note. Why are the vector mesons heavier? Consider an analogy with positronium. The energy is lowest when the magnetic moments are aligned, but since the charges are opposite, this corresponds to the spins anti-aligned, giving a total of spin 0. The same reasoning holds for mesons, though it's color charge rather than electric charge that's opposite. For baryons, the situation is more complicated because the color charges differ by more than just a sign, but the same idea holds.

We can get plenty of insight from our simple model above. For example, the splitting between K and K^* is smaller than the splitting between π and ρ because it involves the strange quark, which has a larger mass. Another example is the Σ-Λ splitting. The Σ^0 and Λ have exactly the same quark content uds, but the latter has isospin zero, so the u and d quarks are antisymmetric in flavor and hence antisymmetric in spin. Since the u-d spin-spin interaction is the most important, the Λ is slightly lighter.

Note. Extra sign flips arise because there are two competing and incompatible sign conventions. We would like to define the antiparticles by charge conjugation, e.g. $|\overline{u}\rangle = C|u\rangle$, i.e. so the matrix elements of C are all positive. On the the other hand, we want to work with eigenstates of I_3 and I^2 under the Cordan–Shortley phase convention, under which the $I_\pm$ have real positive entries.

By definition, charge conjugation flips the isospin,

$$CI_3C^{-1} = -I_3.$$

As a result, it must exchange raising and lowering operators, so

$$CI_\pm C^{-1} = \alpha I_\mp$$

where $\alpha = \pm 1$, because the transformed $I_\pm$ operators must be Hermitian conjugates. We choose $\alpha = -1$, which implies

$$CI_1C^{-1} = -I_1, \quad CI_2C^{-1} = I_2.$$

Now consider the isospin doublet $\{|u\rangle, |d\rangle\}$ and their images under C, $|\overline{u}\rangle, |\overline{d}\rangle$. If we didn't care about sign conventions, we would have a new isospin doublet $\{|\overline{d}\rangle, |\overline{u}\rangle\}$, but

$$I_+|\overline{u}\rangle = I_+C|u\rangle = -CI_-|u\rangle = -C|d\rangle = -|\overline{d}\rangle.$$

Then for the Cordan–Shortley phase convention to apply we must introduce a relative sign, though a global sign is still arbitrary; we thus choose the isospin doublet $\{-|\overline{d}\rangle, |\overline{u}\rangle\}$. We can then use standard tables of Clebsch–Gordan coefficients to add isospin.

Note. The role of antisymmetrization of the wavefunction. At the level of quantum field theory, the wavefunction for any system of fermions must always be antisymmetrized, whether the fermions are the same 'type' or not, because all fermion creation operators anticommute,

$$\langle 0|a_x b_y b_y^\dagger a_x^\dagger|0\rangle = -\langle 0|b_y a_x b_y^\dagger a_x^\dagger|0\rangle.$$

In wavefunction notation, the state of n fermions lives in the totally antisymmetric subspace of $\mathcal{H}^{\otimes n}$, where the single-particle space $\mathcal{H}$ includes fermions of all positions, spins, flavors, and colors. The exchange operation swaps all of these properties, not just the positions.

However, if some particle has a property that none of the other particles share, it can be excluded from the antisymmetrization without any effect. For example, if an electron is far away from all the others, we can turn off the antisymmetrization because the only effect is to remove the exchange force, which that electron doesn't feel; the electron is 'distinguishable by its position'. Similarly, if only one electron in an atom has spin up, we can treat it naively because it won't violate the exclusion principle.

In the case of mesons, the constituents can always be treated as distinguishable because only one of them will be an antiquark. But most baryons contain quarks with the same flavors, in which case the antisymmetrization matters.

Finally, we turn to the light baryons, carefully accounting for the antisymmetrization. The wavefunction has four parts: position, color, flavor, and spin.

- The position wavefunction is more complicated; the orbital angular momentum must be described by two parameters (e.g. the angular momentum $\mathbf{L}$ of the first two particle about their center of mass, and the angular momentum $\mathbf{L}'$ of their center of mass and the third particle about the combined center of mass). We ignore these problems by restricting to $n = 1$ and $\ell = \ell' = 0$, so the position wavefunction is symmetric.

- The color wavefunction must always be the color singlet, i.e. the 1 in

$$3 \times 3 \times 3 = 1 + 8 + 8 + 10.$$

  This is antisymmetric and always the same, so we don't write it.

- Now the combined spin and flavor wavefunctions must be symmetric. In the case of spin,

$$2 \times 2 \times 2 = 2_{ma} + 2_{ma} + 4_s$$

  where the 4 is spin 3/2 and contains totally symmetric wavefunctions, and the 'ma' stands for 'mixed antisymmetry'. The decomposition of the remainder into $2 + 2$ is not unique; for example, we have

$$\left|\frac{1}{2}\frac{1}{2}\right\rangle_{12} = |10-01\rangle|1\rangle, \quad \left|\frac{1}{2}-\frac{1}{2}\right\rangle_{12} = |10-01\rangle|0\rangle$$

  which is antisymmetric in slots 1 and 2, but also a 2 antisymmetric in 2 and 3, and a 2 antisymmetric in 1 and 3.

- Finally, we turn to the flavor wavefunction, where

$$3 \times 3 \times 3 = 1_a + 8_{ma} + 8_{ma} + 10_s.$$

  We can build an allowed baryon multiplet with $10_s \times 4_s$, giving the spin 3/2 baryon decuplet.

- Many of the remaining states are forbidden, since we can't build symmetric combinations from them. However, we can build a baryon octet out of the mixed antisymmetric representations,

$$\psi \sim \psi_{12}(\text{spin})\psi_{12}(\text{flavor}) + \psi_{23}(\text{spin})\psi_{23}(\text{flavor}) + \psi_{13}(\text{spin})\psi_{13}(\text{flavor}).$$

  This accounts for all of the low-energy baryons.

- Similarly, we can construct a totally antisymmetric combination

$$\psi \sim \psi_{12}(\text{spin})(\psi_{31}(\text{flavor}) + \psi_{32}(\text{flavor})) + \text{cyclic}.$$

  This would be the baryon octet if quark color didn't exist; it appears for excited states with nonzero angular momentum.

- As an application, we can compute the magnetic moments of the baryons by

$$\boldsymbol{\mu} = \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 + \boldsymbol{\mu}_3, \quad \boldsymbol{\mu}_i = \frac{q}{m_i}\mathbf{S}_i$$

  where m_i is the effective quark mass. Specifically, we usually calculate the z–component of $\boldsymbol{\mu}$ in the spin up configuration, so we need the expectation value of $\sum_i \mu_i^z$.

- As an explicit example, consider the proton. The wavefunction is

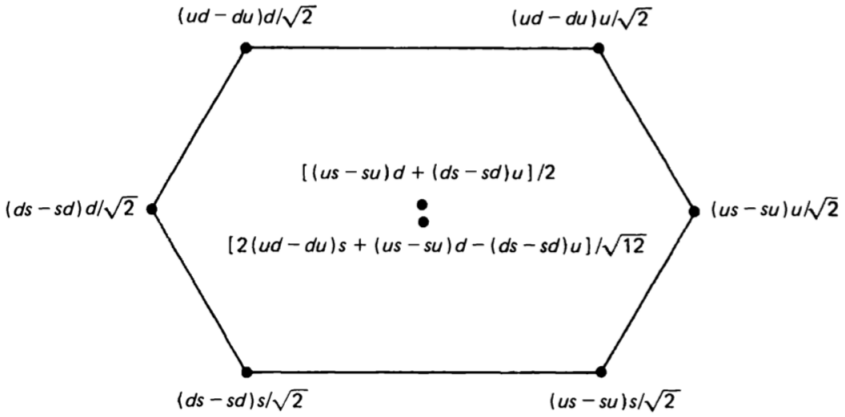
$$\left|p, +\frac{1}{2}\right\rangle = (|101\rangle - |011\rangle)(|udu\rangle - |duu\rangle) + (|110\rangle - |101\rangle)(|uud\rangle - |udu\rangle) + (|110\rangle - |011\rangle)(|uud\rangle - |duu\rangle)$$

  and we can expand out the 12 terms and evaluate the magnetic moment to be $(4\mu_u - \mu_d)/3$, within 1% of the experimental result.

- As with the mesons, we can compute the masses using the empirical formula
$$M = m_1 + m_2 + m_3 + A\left[\frac{\mathbf{S}_1 \cdot \mathbf{S}_2}{m_1 m_2} + \frac{\mathbf{S}_1 \cdot \mathbf{S}_3}{m_1 m_3} + \frac{\mathbf{S}_2 \cdot \mathbf{S}_3}{m_2 m_3}\right].$$
For example, for the baryon decouplet, all pairs of spins are 'parallel', so
$$(\mathbf{S}_1 + \mathbf{S}_2)^2 = S_1^2 + S_2^2 + 2\mathbf{S}_1 \cdot \mathbf{S}_2, \quad \mathbf{S}_1 \cdot \mathbf{S}_2 = \frac{\hbar^2}{4}$$
which implies that
$$M_{\Sigma^*} = 2m_u + m_s + \frac{\hbar^2 A'}{4}\left(\frac{1}{m_u^2} + \frac{2}{m_u m_s}\right).$$
These predictions are also within 1% of the experimental results, though we need to fit the quark masses differently.

Note. We can also arrive at the above result with more powerful machinery. We combine flavor and spin into an $\mathfrak{su}(6)$ symmetry and use the fact
$$6 \times 6 \times 6 = 56_s + 70_{ms} + 70_{ma} + 20_a.$$
Then the 56_s is exactly the set with the right symmetry. Restricting to $\mathfrak{su}(3) \oplus \mathfrak{su}(2)$ gives
$$56_s \to (10, 4) + (8, 2)$$
which are exactly the baryon decuplet and octet. The other possibly useful representation is the antisymmetric one, which breaks up as
$$20_a \to (8, 2) + (1, 4).$$
For the mesons, we have
$$6 \times \overline{6} = 35 + 1 \to (8, 1) + (8, 3) + (1, 3) + (1, 1)$$
which reproduces the two octets and singlets seen before. One might worry that combining a spacetime and internal symmetry in this way is forbidden by the Coleman–Mandula theorem, but there's no problem because we're working nonrelativistically. We can also handle magnetic moments; since the magnetic moment operator is in the adjoint 35, and 35×56 only contains 56 once, all of the moments can be expressed in terms of a single one, up to $\mathfrak{su}(6)$ breaking.

Example. We construct the spin wavefunctions using the usual $\mathfrak{su}(2)$ procedure. We can handle flavor with an ad hoc method. The 10_s is easy because the wavefunctions are totally symmetric and the quark content is fixed by the strangeness and isospin; for example, the Δ^0 is $|ddu\rangle + |dud\rangle + |udd\rangle$. The 1_a is simply the totally antisymmetric combination.



Now consider the 8_{ma} antisymmetric in the first two particles. The outer six states are found by taking the known quark content and simply antisymmetrizing the first two particles. One of the center states is part of an isospin triplet and can be found by isospin raising $(|ds\rangle - |sd\rangle)|d\rangle$. The other center state is found by orthogonality with this state and the 1_a.

1.4 Quantum Chromodynamics

Next, we perform some elementary computations in quantum chromodynamics. We begin with the cross-section for $e^+e^- \to$ hadrons.

- Quarks can be pair produced by $e^+e^- \to \gamma^* \to q\overline{q}$. As the high-energy quarks separate, they emit gluons which emit quark-antiquark pairs. Eventually, each group of particles turns into a "jet" of hadrons, whose direction is correlated with that of the original hard quark.
- Note that to make the jets colorless, a quark or antiquark needs to be transferred between them. This doesn't make much of a difference, since it will be much lower-energy than the original hard quarks.
- The quarks can also emit a hard gluon, $\gamma^* \to q\overline{q}g$. In this case, we get a three-jet event; such events were key in establishing that gluons existed.
- Neglecting the masses of all particles, the cross-section for this process is

$$\sigma = \frac{\pi}{3}\frac{Q^2\alpha^2}{E^2}$$

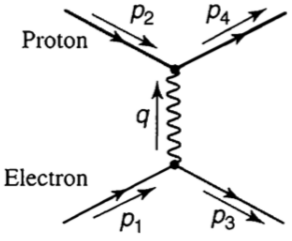
  where Q is the charge of the quark. Therefore,

$$R = \frac{\sigma(e^+e^- \to \text{hadrons})}{\sigma(e^+e^- \to \mu^+\mu^-)} = 3\sum_i Q_i^2$$

  where the 3 is for the three colors of quarks, and the sum is over quarks with masses much less than E. We expect Q to look like a step function, jumping up for every flavor of quark.
- There are a few complications. Each step should be smoothed out by the masses. We have neglected the interaction of the two final-state quarks, but this is very important near a resonance, where the cross-section has a peak. Above about 50 GeV, R quickly increases because of the Z^0 peak. But overall, the data fits reasonably well, and unambiguously establishes three quark colors.

Next, we turn to elastic electron-proton scattering, mediated by a photon.

- At the most naive level, suppose the proton is a Dirac point charge.



If the momentum transfer is q, the spin-averaged amplitude is

$$|\mathcal{M}|^2 = \frac{e^4}{q^4} L_e^{\mu\nu} L^p_{\mu\nu}, \quad L_e^{\mu\nu} = 2(p_1^\mu p_3^\nu + p_1^\nu p_3^\mu + \eta^{\mu\nu}(m^2 - p_1 \cdot p_3))$$

where the L factors come from the traces.

- In reality, the proton is much more complicated, and we can parametrize our ignorance with form factors. Letting p be the initial proton momentum, we may write the proton factor as

$$K^{\mu\nu} = -K_1 \eta^{\mu\nu} + \frac{K_2}{M^2} p^\mu p^\nu + \frac{K_4}{M^2} q^\mu q^\nu + \frac{K_5}{M^2}(p^\mu q^\nu + p^\nu q^\mu).$$

  We haven't written the antisymmetric combination, which would have coefficient K_3, since $L^{\mu\nu}$ is symmetric.

- Next, we can check that $q_\mu L^{\mu\nu} = 0$, which means that we can choose $K^{\mu\nu}$ so that $q_\mu K^{\mu\nu} = 0$ without affecting the result. This allows us to eliminate K_4 and K_5, giving

$$K^{\mu\nu} = K_1(q^2)\left(-\eta^{\mu\nu} + \frac{q^\mu q^\nu}{q^2}\right) + \frac{K_2(q^2)}{M^2}(p^\mu + q^\mu/2)(p^\nu + q^\nu/2)$$

  where $K_1(q^2)$ and $K_2(q^2)$ have absorbed the effects of K_4 and K_5, and depend on q. For example, for the original point charge model, $K_1 = -q^2$ and $K_2 = 4M^2$.

- The cross section is given by the Rosenbluth formula

$$\frac{d\sigma}{d\Omega} = \left(\frac{\alpha}{4ME\sin^2(\theta/2)}\right)^2 \frac{E'}{E}(2K_1 \sin^2(\theta/2) + K_2 \cos^2(\theta/2))$$

  where E and E' are the initial and final electron energies, and we have assumed $E \gg m$. As a check, when $E \ll M$, the point charge form factors work. Then our result reduces to the Mott formula, which describes electron scattering off a heavy pointlike target.

- The form factors $K_1(q^2)$ and $K_2(q^2)$ are measured by experiment and indicate the proton is not pointlike, as expected from QCD.

Next, we turn to the Feynman rules for QCD itself. The coupling is g_s, and we define $\alpha_s = g_s^2/4\pi$.

- Quarks are specified by both a spinor polarization and a color. We label the colors with mid-Latin letters and call them red, blue, and green.

- There is a gluon and two quark vertex, so the gluon colors must live in $3 \times \overline{3} = 8 + 1$. The elements of $3 \times \overline{3}$ have colors like $r\overline{r}$ ('red anti-red') and $b\overline{g}$ ('blue anti-green'). The color singlet $r\overline{r} + b\overline{b} + g\overline{g}$ is analogous to the meson singlet.

- One might wonder whether there is a ninth gluon. Theoretically, this is equivalent to the choice of gauge group $\mathfrak{su}(3)$ or $\mathfrak{u}(3)$. Since the ninth gluon would be a color singlet, it would not be confined, and would instead mediate a long-range force between color singlets; it would have an independent coupling since $\mathfrak{u}(3)$ is not semisimple. Such a force would appear as an anomalous contribution to gravity, and there was a brief excitement over this in 1986.

- The eight gluons can be put in correspondence with the eight Gell-Mann matrices λ^α, where

$$\lambda^1 = \begin{pmatrix} & 1 & \\ 1 & & \\ & & \end{pmatrix}, \quad \lambda^2 = \begin{pmatrix} & -i & \\ i & & \\ & & \end{pmatrix}, \quad \lambda^3 = \begin{pmatrix} 1 & & \\ & -1 & \\ & & \end{pmatrix}, \quad \lambda^4 = \begin{pmatrix} & & 1 \\ & & \\ 1 & & \end{pmatrix}$$

$$\lambda^5 = \begin{pmatrix} & -i & \\ & & \\ i & & \end{pmatrix}, \quad \lambda^6 = \begin{pmatrix} & & 1 \\ & & \\ 1 & & \end{pmatrix}, \quad \lambda^7 = \begin{pmatrix} & & \\ & -i & \\ & i & \end{pmatrix}, \quad \lambda^8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & & \\ & 1 & \\ & & -2 \end{pmatrix}$$

  which are normalized to match the Pauli matrices, with $\mathrm{tr}(\lambda^\alpha \lambda^\beta) = 2\delta^{\alpha\beta}$. The colors can be read off the columns and the anticolors off the rows, so that λ^1 essentially means 'red anti-blue plus blue anti-red'.

- We define $T^\alpha = \lambda^\alpha/2$, so the structure constants are

$$[T^\alpha, T^\beta] = i f^{\alpha\beta\gamma} T^\gamma.$$

  By direct calculation, we have

$$f^{123} = 1, \quad f^{147} = f^{246} = f^{257} = f^{345} = f^{516} = f^{637} = \frac{1}{2}, \quad f^{458} = f^{678} = \frac{\sqrt{3}}{2}$$

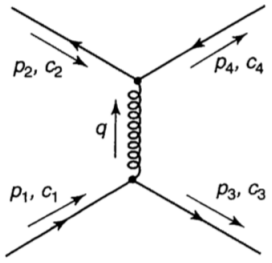
  with all other nonzero structure constants related by total antisymmetry.

- The Feynman rules for QCD are as follows.
  - Incoming quarks have a color and spin polarization $u^s(p)c$. Similarly, outgoing quarks have $c^\dagger$, incoming antiquarks have $c^\dagger$, and outgoing antiquarks have c.
  - Incoming gluons have a color and polarization $\epsilon_\mu(p)a^\alpha$, and outgoing gluons have $\epsilon^*_\mu(p)a^{\alpha *}$.
  - The propagators are the same as usual, with delta functions in color space.
  - The qqg vertex gives a factor of $-ig_s\lambda^a\gamma^\mu/2$.
  - The ggg vertex with colors α, β, and γ has a factor of $f^{\alpha\beta\gamma}$ along with other terms. The $gggg$ vertex is similar, with two structure constants.

- Many simple processes will have amplitudes that look just like the QED amplitudes, but with an additional 'color factor'. A useful rule for finding these factors is

$$\lambda^\alpha_{ij}\lambda^\alpha_{k\ell} = 2\delta_{i\ell}\delta_{jk} - \frac{2}{3}\delta_{ij}\delta_{k\ell}.$$

Example. Quark and antiquark scattering, $u + \overline{d} \to u + \overline{d}$. At lowest order, there is one diagram.



The amplitude is the same as in QED except for a color factor, so the potential is

$$V(r) = -f\frac{\alpha_s}{r}, \quad f = \frac{1}{4}(c_3^\dagger \lambda^\alpha c_1)(c_2^\dagger \lambda^\alpha c_4).$$

First, suppose the quark and antiquark are part of a color octet. For concreteness, let the incoming quark and antiquark be red and anti-blue, respectively. By color conservation, the outgoing quark and antiquark must also be red and anti-blue, respectively. Then

$$f = \frac{1}{4}\lambda_{11}^\alpha \lambda_{22}^\alpha = -\frac{1}{6}.$$

The color singlet state is $(r\overline{r} + b\overline{b} + g\overline{g})/\sqrt{3}$, so there are nine terms in all; for example, the part where the quarks come in $r\overline{r}$ and leave $b\overline{b}$ is $(1/4)(1/3)\lambda_{21}^\alpha \lambda_{12}^\alpha$. They can be compactly written as

$$f = \frac{1}{4}\frac{1}{3}\lambda_{ij}^\alpha \lambda_{ji}^\alpha = \frac{1}{12}\,\mathrm{tr}(\lambda^\alpha \lambda^\alpha) = \frac{4}{3}.$$

Then the force between a quark and antiquark is only attractive if they form a color singlet! This is nice, but only suggestive; after all, we worked to lowest order, which required asymptotic freedom, but confinement does not occur in this regime.

Note. In the case $u + \overline{u} \to u + \overline{u}$, we would also have the s-channel diagram. In the case where the incoming quarks form a color singlet, this is automatically zero since a singlet cannot couple to an octet.

Example. Quark and quark scattering, $u + d \to u + d$. The color factor is very similar,

$$f = \frac{1}{4}(c_3^\dagger \lambda^\alpha c_1)(c_4^\dagger \lambda^\alpha c_2)$$

where the labels on the c_i are as above. Now, $3 \times 3 = 6 + \overline{3}$, so we must consider the sextet and triplet configurations. They contain the symmetric and antisymmetric parts, respectively:

$$\left\{rr, bb, gg, \frac{rb+br}{\sqrt{2}}, \frac{bg+gb}{\sqrt{2}}, \frac{gr+rg}{\sqrt{2}}\right\}, \quad \left\{\frac{rb-br}{\sqrt{2}}, \frac{bg-gb}{\sqrt{2}}, \frac{gr-rg}{\sqrt{2}}\right\}.$$

For the sextet, we take rr, which gives

$$f = \frac{1}{4}\lambda_{11}^\alpha \lambda_{11}^\alpha = \frac{1}{3}.$$

For the triplet, we take $(rb - br)/\sqrt{2}$, which gives four terms,

$$f = \frac{1}{4}\frac{1}{2}(\lambda_{11}^\alpha \lambda_{22}^\alpha - \lambda_{21}^\alpha \lambda_{12}^\alpha - \lambda_{12}^\alpha \lambda_{21}^\alpha + \lambda_{22}^\alpha \lambda_{11}^\alpha) = -\frac{2}{3}.$$

Then the triplet is attractive and the sextet is not. There aren't triplets observed in nature, but note that the color singlet for three quarks is totally antisymmetric, so any two of the quarks form a color triplet. Then every quark in a color singlet baryon attracts every other quark, as expected.

Example. Pair annihilation. Consider the decay of charmonium. There are two tree-level QED diagrams, $c + \overline{c} \to \gamma + \gamma$, and three tree-level QCD diagrams, $c + \overline{c} \to g + g$. By angular momentum

addition, the amplitude is only nonzero if the charmonium is in the spin singlet state. One can show that the two amplitudes differ only by the color factor

$$f = \frac{1}{8}a_3^\alpha a_4^\beta (c_2^\dagger \{\lambda^\alpha, \lambda^\beta\} c_1) = \frac{1}{8\sqrt{3}} a_3^\alpha a_4^\beta \operatorname{tr}\{\lambda^\alpha, \lambda^\beta\} = \frac{1}{2\sqrt{3}} a_3^\alpha a_4^\alpha$$

where we used the fact that charmonium is in the color singlet state. Now we need to construct the singlet state for two gluons, i.e. the 1 in

$$8 \times 8 = 27 + 10 + \overline{10} + 8 + 8 + 1.$$

One can show that this state has the form $\sum_{i=1}^{8} |i\rangle|i\rangle/\sqrt{8}$ where $|i\rangle$ is the gluon state corresponding to the Gell-Mann matrix λ_i, so

$$f = \frac{1}{2\sqrt{3}} \frac{8}{\sqrt{8}} = \sqrt{2/3}.$$

The rate of the decay can be computed if $|\psi(0)|^2$ is known, since this gives the incident flux. Though we can't calculate this, we can calculate the ratio of the decay rate to gg to the decay rate to $\gamma\gamma$, calculated in QED.

Note. Consider two objects in color representations A and B. Their interaction is proportional to

$$T_a^A T_b^B = \frac{1}{2}(T_a^2 - T_a^{A\,2} - (T_a^B)^2)$$

where $T_a = T_a^A + T_a^B$ is a generator for total color. Then the attraction is strongest when the total state has the least color. The same reasoning goes for ordinary electromagnetic interactions or spin-spin interactions; the net effect will usually be to minimize or maximize the 'charge' of the composite state. This attraction is what leads to color confinement.

2 Symmetries

2.1 Chiral and Gauge Symmetries

We begin by reviewing some conventions for Dirac spinors.

- Let $\psi(x)$ be a Dirac spinor field. The Dirac equation is

  $$(i\not{\partial} - m)\psi = 0$$

  and the adjoint field $\overline{\psi} = \psi^\dagger \gamma^0$ satisfies

  $$\overline{\psi}(-i\overleftarrow{\not{\partial}} - m) = 0$$

  where the left arrow indicates the derivative acts to the left.

- The gamma matrices satisfy the anticommutation relations

  $$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}, \quad \eta = \mathrm{diag}(1, -1, -1, -1)$$

  where there is an implicit identity matrix on the right-hand side. In the chiral representation,

  $$\gamma^0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \gamma^i = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix}.$$

- Dirac masses are not fundamental; in this course we will be more concerned with massless fermions. Then the chirality projection operators become more important. We define

  $$\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad (\gamma^5)^2 = 1, \quad \{\gamma^5, \gamma^\mu\} = 0$$

  where the sign of γ^5 differs between references. Then if ψ solves the massless Dirac equation $\not{\partial}\psi = 0$, then $\gamma^5\psi$ does as well, $\not{\partial}(\gamma^5\psi) = 0$.

- We define the projection operators

  $$P_L = \frac{1 - \gamma^5}{2}, \quad P_R = \frac{1 + \gamma^5}{2}, \quad \psi_L = P_L\psi, \quad \psi_R = P_R\psi$$

  where it is straightforward to show the P_L and P_R project onto orthogonal subspaces,

  $$(P_{L,R})^2 = P_{L,R}, \quad P_L P_R = P_R P_L = 0, \quad P_L + P_R = 1.$$

  In the chiral representation, ψ_L/ψ_R has only the upper/lower two components nonzero.

- It's important to note that a lot of the facts above are conventional. For example, $(\psi_L)^*$ is clearly right-chiral in terms of its Lorentz transformation properties, because the left-chiral and right-chiral representations are conjugate, but it is annihilated by P_R because its bottom two components remain zero. When we consider the charge conjugation of fields, we will include a "charge conjugation matrix" whose purpose is to rearrange the components of the naive complex conjugate so that the familiar properties still hold.

- Note that the adjoints of the left-chiral and right-chiral fields satisfy

$$\overline{\psi}_L(x) = \overline{\psi}(x)P_R, \quad \overline{\psi}_R(x) = \overline{\psi}(x)P_L.$$

  Thus if we stick to only ψ and $\overline{\psi}$, then P_R projects right-chirality from both directions.

- A massless Dirac fermion has a $U(1)_L \times U(1)_R$ chiral symmetry. The Dirac Lagrangian is

$$\mathcal{L} = \overline{\psi}_L i\partial\!\!\!/ \psi_L + \overline{\psi}_R i\partial\!\!\!/ \psi_R - m(\overline{\psi}_R\psi_L + \overline{\psi}_L\psi_R)$$

  where we get a chirality flip from anticommuting past $\partial\!\!\!/$. Then when $m = 0$, we can rotate the phases of ψ_L and ψ_R independently. Adding the mass term requires the phases to be rotated the same way, breaking the symmetry to a $U(1)_V$ "vector" symmetry. Rotating the phases oppositely gives a $U(1)_A$ "axial" transformation.

Next, we review the process of gauging a symmetry.

- We would like to gauge the $U(1)_V$ symmetry, $\psi \to e^{i\alpha(x)}\psi$, but then

$$\overline{\psi}i\partial\!\!\!/\psi \to \overline{\psi}i\partial\!\!\!/\psi - (\overline{\psi}\gamma^\mu\psi)\partial_\mu\alpha.$$

  To do this, we introduce the covariant derivative,

$$D_\mu\psi = (\partial_\mu + igA_\mu)\psi, \quad A_\mu(x) \to A_\mu(x) - \frac{1}{g}\partial_\mu\alpha(x), \quad D_\mu\psi(x) \to e^{i\alpha(x)}D_\mu\psi(x).$$

  We cannot gauge the axial symmetry, even in the massless case, because of the chiral anomaly. Note that sources may differ on the sign of g, or pull a factor of g out from $\alpha(x)$.

- The kinetic term of the gauge field is

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \quad [D_\mu, D_\nu] = igF_{\mu\nu}, \quad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$$

  where $F_{\mu\nu}$ is gauge invariant.

- The procedure for non-abelian gauge symmetry is similar. The matter field now transforms in a unitary representation r of the gauge group G, with transformation

$$\psi_i(x) \to \exp(it^a\alpha^a(x))_{ij}\psi_j(x) = U_{ij}\psi_j(x), \quad \overline{\psi}_i(x) \to \overline{\psi}_j(x)\exp(-it^a\alpha^a(x))_{ji} = \overline{\psi}_j(x)(U^\dagger)_{ji}$$

  where the t^a are the Hermitian generators in this representation, and satisfy

$$[t^a, t^b] = if^{abc}t^c, \quad \operatorname{tr} t^a t^b = T(r)\delta^{ab}$$

  where $T(r)$ is the Dynkin index, which is $1/2$ for the fundamental representation.

- The covariant derivative is

$$(D_\mu)_{ij} = \partial_\mu\delta_{ij} + ig(t^aA^a_\mu)_{ij}, \quad (D_\mu\psi(x))_i \to (U(x)D_\mu\psi(x))_i$$

  where we have introduced a Lie algebra valued field A_μ that transforms as

$$A_\mu \to UA_\mu U^{-1} + \frac{i}{g}(\partial_\mu U)U^{-1}.$$

  where we now drop the matrix indices i and j.

- More generally, we define the covariant derivative of any object X similarly, but with A_μ in the appropriate representation; then the covariant derivative DX transforms just like X. For example, the infinitesimal transformation of A_μ itself is

$$A_\mu \to A_\mu - \frac{1}{g}(\partial_\mu \alpha + ig[A_\mu, \alpha]) = A_\mu - \frac{1}{g} D_\mu \alpha$$

  where the D_μ acts as if α is in the adjoint representation. Note that A_μ doesn't transform in any definite representation, much like how the connection in GR is not a tensor.

- Dropping the i and j indices, the field strength is

$$[D_\mu, D_\nu] = igt^a F^a_{\mu\nu}, \quad F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g f^{abc} A^b_\mu A^c_\nu.$$

  That is, as in general relativity, the commutator of two covariant derivatives is a tensor, not a differential operator. Then the field transforms in the adjoint representation, as

$$[D_\mu, D_\nu]\psi \to U[D_\mu, D_\nu]\psi = U[D_\mu, D_\nu]U^\dagger U\psi, \quad F_{\mu\nu} \to U F_{\mu\nu} U^\dagger.$$

  Then a gauge invariant kinetic term is

$$\mathcal{L} = -\frac{1}{4} F^a_{\mu\nu} F^{a\mu\nu} = -\frac{1}{2} \operatorname{tr} F_{\mu\nu} F^{\mu\nu}.$$

  In combination with the fermion kinetic term $\overline{\psi}(i\not{D} - m)\psi$, this is the most general renormalizable gauge invariant Lagrangian in dimension 4 with P and T symmetry.

- More generally, we can think of the field as an infinitesimal Wilson loop. In a sense, the most general gauge invariant observable is the trace of a Wilson loop.

- In some sources, for e.g. gauge group $SU(n)$, the gauge field A_μ is thought of as an $n \times n$ matrix rather than an abstract element of $\mathfrak{su}(n)$, leading to equations like

$$D_\mu F_{\nu\rho} = \partial_\mu F_{\nu\rho} + ig[A_\mu, F_{\nu\rho}].$$

  Another example is a matter field which transforms in the adjoint representation, with

$$\mathcal{L} = \operatorname{tr} \overline{\psi}(i\not{D} - m)\psi, \quad \psi \to U\psi U^{-1}$$

  where U is the same gauge transformation we would have in the fundamental representation, ψ is now a matrix, and A_μ again acts by commutator in the covariant derivative. Expressions like these are less mathematically general but can be easy to compute with.

Note. Symmetries will manifest in several ways below.

- The symmetry can be intact, e.g. the gauge symmetries $U(1)_{EM}$ and $SU(3)_C$.

- The symmetry can be anomalous, holding in the classical theory but not the quantum theory, e.g. the global axial symmetry $U(1)_A$.

- The symmetry can be explicitly broken in the Lagrangian, e.g. isospin $SU(2)$ or generally flavor $SU(6)$. This is useful as long as the symmetry is approximate.

- The symmetry can be spontaneously broken, i.e. the vacuum does not respect the symmetry though the Lagrangian does, e.g. $SU(2)_L \times U(1)_Y$ is spontaneously broken to $U(1)_{EM}$.

2.2 Discrete Symmetries

Before beginning with parity, we review discrete spacetime symmetries naively.

- Let W be an operator on a Hilbert space with inner product $(\cdot,\cdot)$. If W is unitary and linear,

$$(W\Phi, W\Psi) = (\Phi, \Psi), \quad W(\alpha\Phi + \beta\Psi) = \alpha W\Phi + \beta W\Psi.$$

  If W is anti-unitary and hence anti-linear,

$$(W\Phi, W\Psi) = (\Phi, \Psi)^*, \quad W(\alpha\Phi + \beta\Psi) = \alpha^* W\Phi + \beta^* W\Psi.$$

  Wigner's theorem states that groups of operators that preserve norms (and hence observable probabilities) must be unitary or anti-unitary.

- Let $W(\Lambda, a)$ be the operator on the state space that corresponds to a Poincare transformation consisting of a Lorentz transformation Λ followed by a translation a. Then

$$W(\Lambda_2, a_2)W(\Lambda_1, a_1) = W(\Lambda_2, \Lambda_1, \Lambda_2 a_1 + a_2).$$

  We also consider the improper Poincare transformations of parity and time reversal, defining

$$\widehat{P} = W(\mathcal{P}, 0), \quad \widehat{T} = W(\mathcal{T}, 0).$$

- We now consider an infinitesimal proper Poincare transformation,

$$\Lambda^\mu{}_\nu = \delta^\mu_\nu + \omega^\mu{}_\nu, \quad a^\mu = \epsilon^\mu.$$

  The corresponding quantum operator is expanded as

$$W(1+\omega, \epsilon) = 1 + \frac{i}{2}\omega_{\mu\nu}J^{\mu\nu} - i\epsilon_\mu P^\mu$$

  where the operators $J^{\mu\nu}$ and P^μ are Hermitian and are physically

$$\text{energy } H = P^0, \quad \text{momentum } \mathbf{P} = (P^1, P^2, P^3)$$

  and

$$\text{angular momentum } \mathbf{J} = (J^{23}, J^{31}, J^{12}), \quad \text{Lorentz boosts } \mathbf{K} = (J^{01}, J^{02}, J^{03}).$$

- Considering $\widehat{P}W(\Lambda, a)\widehat{P}^{-1}$ and $\widehat{T}W(\Lambda a)\widehat{T}^{-1}$ for an infinitesimal translation,

$$\widehat{P}iP^\mu\widehat{P}^{-1} = i\mathcal{P}_\nu{}^\mu P^\nu, \quad \widehat{T}iP^\mu\widehat{T}^{-1} = i\mathcal{T}_\nu{}^\mu P^\nu.$$

  In particular, focusing on the time component, we have

$$\widehat{P}iH\widehat{P}^{-1} = iH, \quad \widehat{T}iH\widehat{T}^{-1} = -iH.$$

  If $\widehat{P}$ were antilinear, then it would flip H in conjugation, implying a negative energy state for every positive energy state. This is unacceptable, as it would forbid the existence of a ground state, so $\widehat{P}$ is linear and hence unitary. Similarly, $\widehat{T}$ is anti-unitary.

- With the linearity and antilinearity established, we can now conjugate our other operators by $\widehat{P}$ and $\widehat{T}$ to see how they transform. We find

$$\widehat{P}\mathbf{P}\widehat{P}^{-1} = -\mathbf{P}, \quad \widehat{P}\mathbf{J}\widehat{P}^{-1} = \mathbf{J}, \quad \widehat{P}\mathbf{K}\widehat{P}^{-1} = -\mathbf{K}.$$

  and

$$\widehat{T}\mathbf{P}\widehat{T}^{-1} = -\mathbf{P}, \quad \widehat{T}\mathbf{J}\widehat{T}^{-1} = -\mathbf{J}, \quad \widehat{T}\mathbf{K}\widehat{T}^{-1} = \mathbf{K}.$$

  Moreover, upon applying the relations above, we find that parity acts on one-particle states by changing their momenta and angular momenta as implied above, along with a phase factor η_P which depends only on the particle species, called the intrinsic parity.

- Under the naive assumptions we have made above, $\widehat{P}$ and $\widehat{T}$ *automatically* commute with H. That is, our initial assumptions are equivalent to assuming that $\widehat{P}$ and $\widehat{T}$ violation don't occur! To allow it, we need to think more carefully.

First, we review the basics of representation theory, as covered in the notes on Group Theory.

- The representations of $SO(3)$ are indexed by a nonnegative integer s called the spin. The double/universal cover of $SU(2)$ are indexed by a half-integer, and representations of $SU(2)$ correspond to projective representations of $SO(3)$.

- If we include Lorentz boosts, we arrive at the connected Lorentz group $SO(3,1)_0$, whose double/universal cover is $SL(2,\mathbb{C})$. In general, such a double cover is called a spin group. The finite-dimensional representations are indexed by two half-integers (s_1, s_2), where $s_1 + s_2$ is called the spin. When s is half-integer, the representation is projective.

- Note that restricting to rotations does not produce the spin $s_1 + s_2$ representation of $SU(2)$. Instead, every spin from $|s_1 - s_2|$ to $s_1 + s_2$ in integer steps is represented.

- If $s_2 = 0$, the representation is said to be left-chiral or left-handed, and if $s_1 = 0$, the representation is said to be right-chiral or right-handed. Otherwise, chirality is not defined.

- Fields transform in finite-dimensional non-unitary representations of the Lorentz group, while particles transform in infinite-dimensional unitary representations of the Poincare group, which we take to be $SO(3,1)_0 \rtimes \mathbb{R}^4$. These representations and the vectors in them are labeled by several quantum numbers.
  - The mass M labels the physical mass of the particle.
  - When $M > 0$, the irreps are labeled by an integer spin s, so that for each momentum, there are $2s + 1$ spin states. These states can be indexed by the helicity h, i.e. the projection of spin along the direction of momentum; Lorentz transformations change h.
  - When $M = 0$, the irreps have one state for each momentum, indexed by the integer helicity h. Then the helicity is Lorentz invariant. A helicity h particle is also loosely called a massless spin $s = h$ particle.
  - To switch to quantum fields, a massive particle of spin s is embedded in a field of spin s, and a relativistic wave equation is used to eliminate the extra degrees of freedom.
  - Using the double cover $SL(2,\mathbb{C}) \rtimes \mathbb{R}^4$, the spin s and helicity h may be half-integer.

- Sometimes one hears that "for massless particles, chirality is the same thing as helicity". This is an oversimplification that can lead to confusion. Helicity is defined for particles, chirality is defined for fields, and the two can behave rather differently.

Next, we confront the issue of discrete symmetries, and their possible violation.

- We introduce parity and time reversal by going to the group $O(3,1)$. Ignoring the issue of projective representations, the assertion that the Hilbert space carries a representation of $O(3,1) \rtimes \mathbb{R}^{3,1}$ carries dynamical content, because it automatically implies $\widehat{P}$ and $\widehat{T}$ are conserved. That is, postulating a representation of a set of physical operations exists is a nontrivial statement about the dynamics, when one of the operations is time translation.

- For spinless particles, if we have a representation of $O(3,1) \rtimes \mathbb{R}^{3,1}$, then $\widehat{P}^2 = 1$ and particles have parity ± 1. Now consider a spinless theory that violates parity. In this case, we can still talk about parity for asymptotic states, because they are free; we define parity just as in the free theory. This is why we can speak about the change of parity in a scattering process.

- More generally, we must allow projective representations. For the Poincare group, it suffices to promote $SO(3,1)_0$ to $SL(2,\mathbb{C})$. There is a two-to-one map $\pi\colon SL(2,\mathbb{C}) \to SO(3,1)_0$, which can be extended to include the parity operation.

- However, there are two ways to incorporate parity; if $\mathcal{P} \in O(3,1)$ is parity, then $\pi^{-1}(\mathcal{P})$ contains two elements. Letting $\pi(P) = \mathcal{P}$, we have

$$\pi(P^2) = \mathcal{P}^2 = 1$$

  which implies that $P^2 = \pm 1$. This is a genuine physical ambiguity, and it isn't presently known which is the right option in reality.

- In the case $M > 0$, if $P^2 = 1$ then $\widehat{P}^2 = 1$, and for each s we have two representations, of intrinsic parities ± 1. If $P^2 = -1$, we instead have $\widehat{P}^2 = (-1)^F$ where F is the fermion number. Specifically, for integer s the intrinsic parities are ± 1 and for half-integer s the intrinsic parities are $\pm i$. This doesn't contradict the fact that $P^2 = -1$ because for integer s, -1 is represented as $+1$.

- However, Dirac fermions carry other conserved quantum numbers, and we may replace $\widehat{P}$ with $\widehat{P}e^{i\alpha Q}$ for any conserved charge Q to find the same experimental consequences; in the SM the conservation of electric charge, lepton number, and baryon number are sufficient to redefine parity so that $\widehat{P}^2 = 1$ in all cases. Stated another way, other conservation laws always rule out possible experimental tests between the situations above.

- On the other hand, if a Majorana fermion were discovered, it would carry no conserved charges, so it could distinguish between the possibilities. Specifically, if $\widehat{P}^2 = (-1)^F$, then no process which conserves parity can turn this particle into three copies of itself, since $(\pm i) \neq (\pm i)^3$.

- In the case $M = 0$, parity implies that irreps must contain helicities of $\pm\lambda$ in pairs; this is also a consequence of CP or CPT. However, if we also demand that parity does not change the values of internal quantum numbers, then there's no reasonable way to define parity for a theory with a single Weyl spinor. The helicities still come in pairs, but the pairing requires flipping internal quantum numbers; we instead call this symmetry CPT.

- In the real world, parity is not conserved, but with the exception of chiral theories (e.g. with a single Weyl spinor) where parity cannot even be reasonably defined, the free Hamiltonian always commutes with parity. Thus parity can be defined in terms of the free theory, allowing the parities of asymptotic particles to be defined.
- In the above discussion, we have neglected time reversal. When we account for both parity and time reversal and allow for projective representations, we find eight possibilities in total, though the so-called Pin groups are the mathematically nicest.

2.3 Parity

Now we investigate parity more precisely, beginning with the scalar field. As we motivated above, we focus on defining parity on free fields.

- A scalar field has plane wave expansion

$$\phi(x) = \sum_p a(p)e^{-ipx} + c^\dagger(p)e^{ipx}, \quad \sum_p = \int \frac{đ\mathbf{p}}{2E_\mathbf{p}}$$

  where $a^\dagger(p)/c^\dagger(p)$ create particles/antiparticles with momentum p. We use the relativistic normalization convention, so the created states have squared norm $2E_\mathbf{p}$.
- Now, parity should preserve the number of particles and flip the momentum, so

$$\widehat{P}a^\dagger(p)|0\rangle = \eta^{a*}a^\dagger(p_P)|0\rangle$$

  where p_P is the parity-flipped four-momentum and η^{a*} is a phase, by unitarity.
- Inserting $\widehat{P}^{-1}\widehat{P} = 1$ above and assuming the vacuum is parity invariant, $\widehat{P}|0\rangle = |0\rangle$, we find

$$\widehat{P}a^\dagger(p)\widehat{P}^{-1} = \eta^{a*}a^\dagger(p_P), \quad \widehat{P}c^\dagger(p)\widehat{P}^{-1} = \eta^{c*}c^\dagger(p_P).$$

  Taking the adjoint, we find

$$\widehat{P}a(p)\widehat{P}^{-1} = \eta^{a}a(p_P), \quad \widehat{P}c(p)\widehat{P}^{-1} = \eta^{c}c(p_P).$$

- Now, the parity conjugate of the scalar field is defined as

$$\phi^P(x) \equiv \widehat{P}\phi(x)\widehat{P}^{-1} = \sum_p \eta^a a(p_P)e^{-ipx} + \eta^{c*}c^\dagger(p_P)e^{ipx} = \sum_p \eta^a a(p)e^{-ip_Px} + \eta^{c*}c^\dagger(p)e^{ip_Px}$$

  where we reindexed the sum. This looks rather different from our previous expression; moreover, $[\phi(x), {\phi^\dagger}^P(y)]$ does not necessarily vanish for spacelike x and y.
- These problems are solved if $\eta^a = \eta^{c*} \equiv \eta_P$, so

$$\phi^P(x) = \eta_P\phi(x_P).$$

  The phase η_P is called the intrinsic parity of ϕ.
- If ϕ is a real field, then $c(p) = a(p)$, so $\eta^c = \eta^a$, which implies that η_P is real, and hence it is ± 1. The case $+1$ is a scalar, and the case -1 is a pseudoscalar.

- On the other hand, for a complex field η_P can be an arbitrary phase, but there is a $U(1)$ internal symmetry which may yield a conserved charge Q. In this case, we can always replace $\widehat{P}$ with $\widehat{P}e^{-i\alpha Q}$, where α may be chosen so that $\widehat{P}^2 = 1$, so that $\eta_P = \pm 1$.
- Another way of saying this is that the complex scalar isn't really a different case than a real scalar. Everything that can be expressed in terms of complex scalars can be expressed in terms of pairs of real scalars with appropriate $U(1)$ symmetries. Choosing a description in terms of complex scalars is purely a matter of convention and convenience, which pays off when the $U(1)$ symmetries at least approximately hold in the interacting theory.
- In the case of vector fields, we have

  $$V^\mu(x) = \sum_{p,\lambda} \epsilon^{\lambda\mu}(p) a^\lambda(p) e^{-ipx} + \epsilon^{\lambda\mu *}(p) c^{\lambda *}(p) e^{ipx}$$

  where the $\epsilon^{\lambda\mu}$ are polarization vectors. It can be shown, using the desired properties of parity defined in the previous section, that they transform as

  $$\epsilon^{\lambda\mu}(p_P) = -\mathcal{P}^\mu{}_\nu \epsilon^{\lambda\nu}(p).$$

- The rest of the argument goes as before, so for a real vector field

  $$\widehat{P} V^\mu(x) \widehat{P}^{-1} = -\eta_P \mathcal{P}^\mu{}_\nu V^\nu(x_P)$$

  where $\eta_P = -1$ for a polar vector and $\eta_P = 1$ for an axial vector.

We now review conventions for the Dirac field, which is more subtle.

- A solution to the free Dirac equation can be expanded as

  $$\psi(x) = \sum_{p,s} b^s(p) u^s(p) e^{-ipx} + d^{s\dagger}(p) v^s(p) e^{ipx}$$

  where $b^\dagger$ and $d^\dagger$ create particles and antiparticles of momentum p, and the spinors satisfy

  $$(\not{p} - m)u(p) = 0, \quad (\not{p} + m)v(p) = 0$$

  for components $s = \pm 1/2$. In the chiral representation their components are

  $$u^s(p) = \begin{pmatrix} \sqrt{p\cdot\sigma}\xi^s \\ \sqrt{p\cdot\overline{\sigma}}\xi^s \end{pmatrix}, \quad v^s(p) = \begin{pmatrix} \sqrt{p\cdot\sigma}\zeta^s \\ -\sqrt{p\cdot\overline{\sigma}}\zeta^s \end{pmatrix}, \quad \sigma = (1, \boldsymbol{\sigma}), \quad \overline{\sigma} = (1, -\boldsymbol{\sigma})$$

  and a useful basis of two-component spinors is $\xi^{1/2} = (1,0)^T$ and $\xi^{-1/2} = (0,1)^T$, which have spin up and spin down along $\hat{z}$ for both the positive and negative frequency solutions. We'll use a different basis for the negative frequency solutions for convenience, as explained below.
- The spin angular momentum operator can be found by taking the conserved quantity due to rotations and subtracting off the orbital contribution, giving

  $$S_i = \frac{i}{4}\epsilon_{ijk}\gamma^j\gamma^k = \frac{1}{2}\begin{pmatrix} \sigma^i & 0 \\ 0 & \sigma^i \end{pmatrix}, \quad \gamma^5 S^i = S^i\gamma^5 = \frac{1}{2}\gamma^0\gamma^i.$$

  Multiplying the massless Dirac equation $\not{p}u = \not{p}v = 0$ by γ^0/p^0 then gives

  $$(1 - 2\mathbf{S}\cdot\hat{\mathbf{p}}\gamma^5)u^s(p) = (1 - 2\mathbf{S}\cdot\hat{\mathbf{p}}\gamma^5)v^s(p) = 0.$$

- For a classical solution to the Dirac equation, define $h = \mathbf{S} \cdot \hat{\mathbf{p}}$. Inserting a factor of $P_L + P_R$,

$$hu^s_{L,R} = \mp\frac{1}{2}u^s_{L,R}, \quad hv^s_{L,R} = \mp\frac{1}{2}v^s_{L,R}$$

  where the L/R subscripts indicate left-chiral or right-chiral Weyl fields.

- The physical interpretation is a bit tricky. For positive frequency solutions, h is equal to the helicity λ of the corresponding particle. For negative frequency solutions, the parameter p is the opposite of the physical momentum, as they are proportional to e^{ipx} rather than e^{-ipx}.

- Upon quantization negative frequency solutions become holes, which flips p, $\mathbf{S}$, and all other quantum numbers. The fact that p is already flipped once in the definition of $v^s(p)$ means that the particle corresponding to $v^s(p)$ indeed has momentum p, with no sign. But since the spin is flipped, we have $h = -\lambda$ for negative frequency solutions. Since the charge is flipped, these particles are called antiparticles.

- Thus, a left-chiral Weyl field annihilates a left-helicity (negative helicity) particle and creates a right-helicity (positive helicity) antiparticle. Similarly, a right-chiral Weyl field annihilates a right-helicity particle and creates a left-helicity antiparticle. We see that each of these Lorentz irrep fields gives rise to two Poincare particle irreps.

- For example, a "left-chiral antiquark field" is one which annihilates a left-helicity antiquark. It would be the charge conjugate of a left-chiral quark field, and the parity conjugate of a right-chiral antiquark field, assuming these fields exist at all in the theory; if they are not, parity and charge conjugation aren't defined.

- For reference, for a massless particle moving in the $+\hat{\mathbf{z}}$ direction, we have

$$\text{spin up: } u(p) = \begin{pmatrix}0\\0\\1\\0\end{pmatrix}, v(p) = \begin{pmatrix}1\\0\\0\\0\end{pmatrix}, \quad \text{spin down: } u(p) = \begin{pmatrix}0\\1\\0\\0\end{pmatrix}, v(p) = \begin{pmatrix}0\\0\\0\\1\end{pmatrix}$$

  where spin up indicates positive helicity λ.

Next, we turn to the parity transformation of the Dirac field.

- By the same reasoning as for the scalar field, we should have

$$\widehat{P}b^s(p)\widehat{P}^{-1} = \eta^b b^s(p_P), \quad \widehat{P}d^{s\dagger}\widehat{P}^{-1} = \eta^{d*}d^{s\dagger}(p_P).$$

  Therefore, the transformation of the Dirac field is

$$\psi^P(x) \equiv \widehat{P}\psi(x)\widehat{P}^{-1} = \sum_{p,s} \eta^b b^s(p)u^s(p_P)e^{-ipx_P} + \eta^{d*}d^{s\dagger}(p)v^s(p_P)e^{ipx_P}$$

  where we reindexed the sum as for the scalar field.

- One can show that the spinors transform as

$$u^s(p_P) = \gamma^0 u^s(p), \quad v^s(p_P) = -\gamma^0 v^s(p)$$

as can be verified in the chiral basis using $p \cdot \sigma = p_P \cdot \overline{\sigma}$. Requiring the transformed field to take the same form as the original field, we must have

$$\psi^P(x) = \eta_P \gamma^0 \psi(x_P), \quad \eta_P = \eta^b = -\eta^{d*}.$$

Similarly, for the adjoint field we have

$$\overline{\psi}^P(x) = \eta_P^* \overline{\psi}(x_P)\gamma^0.$$

- By applying projectors, we find that parity flips the chirality,

$$\widehat{P}\psi_L(x)\widehat{P}^{-1} = \eta_P \gamma^0 \psi_R(x_P), \quad \widehat{P}\overline{\psi}_L\widehat{P}^{-1} = \eta_P^* \overline{\psi}_R(x_P)\gamma^0.$$

  This is a special case of the fact that parity maps the (s_1, s_2) Lorentz irrep to (s_2, s_1). We can then straightforwardly check that ψ^P satisfies the Dirac equation, that $\overline{\psi}\psi$ is a scalar and $\overline{\psi}\gamma^5\psi$ is a pseudoscalar, and so on.

- We have freedom in choosing the phase η_P as described above, using global $U(1)$ symmetries, and in the SM this freedom is used to set the intrinsic parities of the proton, neutron, and charged leptons to $+1$. Note that this point is unrelated to the transformations of Dirac bilinears, where η_P cancels out.

- We also note that, regardless of phase adjustments, we have $\eta^b \eta^d = -1$, which means that a two-particle state containing a fermion and its antiparticle has an extra factor of -1 in its intrinsic parity, as we previously noted in our qualitative overview. This logic holds unchanged for Majorana fermions, where the fermion and its antiparticle coincide. The same result holds for the charge conjugate of a fermion and its antiparticle.

2.4 Charge Conjugation

Charge conjugation is different from the other discrete symmetries, since it does not arise from the structure of $O(3,1)$. Instead, it arises from the generic prediction of antiparticles in quantum field theory. It is especially confusing because there are two related notions of it.

- Consider a set of classical fields ψ_i that transform under some representation R. Then the complex conjugate fields ψ_i^* transform under the conjugate representation R^*, though they generally won't be in the "standard" basis. We return to the standard basis using a "charge conjugation matrix" C, and call the operation $\psi \to \psi^{(c)} = C\psi^*$ charge conjugation.

- Since the $(1/2, 0)$ and $(0, 1/2)$ Lorentz representations are conjugate, this notion of charge conjugation flips the chirality. This is the notion of charge conjugation we used when studying group theory. It comes from the classical theory, and useful mainly for constructing real, singlet Lagrangians. It is not the same as the $\widehat{C}$ we study below, which instead corresponds to the intuitive idea of "exchanging matter and antimatter".

- To translate this idea to particles, let ψ transform under a representation R of an internal symmetry group. Then ψ annihilates particles which transform under R and creates particles which transform under R^*, and are hence called antiparticles.

- The field $\psi^{(c)}$ simply does the reverse: it annihilates what ψ creates, and vice versa. In particular, classical charge conjugation doesn't modify the particle content at all; a Lagrangian written in terms of only ψ is equivalent to one written in terms of only $\psi^{(c)}$.
- Note that if R is complex, the particles definitely cannot be identified with their antiparticles, while if R is real they might or might not be.
- The situation is more complicated when we are talking about spacetime symmetries, since fields have Lorentz symmetry and particles have Poincare symmetry; we've seen how chirality for fields corresponds to helicity for particles above.
- A rough heuristic is that classical charge conjugation conjugates both internal and spacetime representations, while, in a $\widehat{C}$-symmetric theory, $\widehat{C}$ conjugates exactly the internal representations (when acting on the free "in/out" states), and in a $\widehat{C}$-asymmetric theory, $\widehat{C}$ might not even defined on those states at all. In terms of representations, the two notion of charge conjugation differ essentially by a parity transformation, leading to confusion when people use different versions of it.

Now we define $\widehat{C}$, starting with the scalar field.

- We begin by demanding that the particle and antiparticle operators should be exchanged,

$$\widehat{C}a(p)\widehat{C}^{-1} = \eta_C c(p), \quad \widehat{C}c(p)\widehat{C}^{-1} = \eta_C^* a(p)$$

  where we used Lorentz invariance as before to constrain the phases. Then we have, for instance

$$\widehat{C}|p, \text{particle}\rangle = \widehat{C}a^\dagger(p)|0\rangle = \eta_C^* c^\dagger(p)|0\rangle = \eta_C^*|p, \text{antiparticle}\rangle.$$

- In terms of the fields, we have

$$\widehat{C}\phi(x)\widehat{C}^{-1} = \eta_C \phi^\dagger(x), \quad \widehat{C}\phi^\dagger(x)\widehat{C}^{-1} = \eta_C^* \phi(x).$$

  For a real scalar field, this implies $\eta_C = \pm 1$, while for a complex scalar field we can perform a rotation so that $\eta_C = 1$. In the former case, this means that particles are eigenstates of $\widehat{C}$, so the symmetry can provide selection rules.

- The photon field must obey

$$\widehat{C}A_\mu(x)\widehat{C}^{-1} = -A_\mu(x)$$

  for $\widehat{C}$ to be a symmetry of QED. That is, photons have intrinsic charge eigenvalue $\widehat{C} = -1$. Physically, this is because the coupling to matter is in the form $A^\mu J_\mu$, where the current J_μ certainly flips sign under charge conjugation.

Next, we proceed to the Dirac field.

- We define the positive frequency and negative frequency basis spinors to be related by

$$\zeta^s = i\sigma^2 \xi^{s*}.$$

  This gives an extra sign flip at the classical level, which ensures that the particles created by $b^{s\dagger}(p)$ and $d^{s\dagger}(p)$ have the same physical spin, just as they have the same physical momentum.

- Next, we define a charge conjugation matrix C that acts on spinors by

$$\gamma^{\mu T} = -C^{-1}\gamma^\mu C.$$

  One can show that C is real, anti-symmetric, and unitary, ${\gamma^5}^T = C^{-1}\gamma^5 C$, and the $\gamma^\mu C$ are symmetric, using only the properties of the Clifford algebra. For the chiral representation,

$$C = -i\gamma^0\gamma^2 = \begin{pmatrix} i\sigma_2 & 0 \\ 0 & -i\sigma_2 \end{pmatrix}.$$

- Under these definitions, we have the simple relationships

$$v^s(p) = C\overline{u}^{sT}(p), \quad u^s(p) = C\overline{v}^{sT}(p).$$

- Now, for the Dirac field, we have

$$\widehat{C}b^s(p)\widehat{C}^{-1} = \eta_C d^s(p), \quad \widehat{C}d^{s\dagger}(p)\widehat{C}^{-1} = \eta_C b^{s\dagger}(p)$$

  where the phases are equated as usual, and we used the fact that $\widehat{C}$ doesn't change spacetime quantum numbers such as momentum and spin. Thus $\widehat{C}$ preserves helicity.

- Since a right-chiral field is defined by annihilating positive helicity, $\widehat{C}$ preserves chirality for quantum fields. Note this is the opposite of the result for classical charge conjugation.

- The charge conjugated field ψ^c, not to be confused with $\psi^{(c)}$, is

$$\psi^c(x) \equiv \widehat{C}\psi(x)\widehat{C}^{-1} = \eta_C \sum_{p,s} d^s(p)u^s(p)e^{-ipx} + b^{s\dagger}(p)v^s(p)e^{ipx}.$$

  On the other hand, the adjoint field transposed to a column vector is

$$\overline{\psi}^T(x) = \sum_{p,s} b^{s\dagger}(p)\overline{u}^{sT}(p)e^{ipx} + d^s(P)\overline{v}^{sT}(p)e^{-ipx}.$$

  Therefore, by our spinor identities we have

$$\psi^c(x) = \eta_C C\overline{\psi}^T(x), \quad \overline{\psi}^c(x) = -\eta_C^*\psi^T(x)C^{-1}.$$

  These equations can also be used (sometimes unwittingly) to define $\widehat{C}$ on *classical* fields, with the caveat that this differs from classical charge conjugation by a parity transformation.

Note. In practice, the simple definition of $\widehat{C}$ above might not work, while a slightly different definition which lacks some of the usual properties of $\widehat{C}$ (such as flipping all internal quantum numbers) may be more useful. For example, in the Standard Model with a sterile neutrino, charge conjugation must exchange the active and sterile neutrinos, if it is to keep the spacetime quantum numbers the same. But the active and sterile neutrinos don't have opposite internal quantum numbers, e.g. the active neutrinos have hypercharge and the sterile neutrinos don't. A strict interpretation would lead to the conclusion that $\widehat{C}$ can't be defined in such a theory. However, it is more common to loosen the criteria and allow $\widehat{C}$ to be defined this way anyway. This is useful because it leads to an approximate symmetry, which is only broken by weak interactions.

This illustrates an important point when discussing discrete symmetries. The point of symmetries is precisely to be able to use them to understand the dynamics. It doesn't make sense to worry about whether some operator is "the true $\widehat{C}$" in some metaphysical sense. Nature doesn't care: the theory described above will still have a $\widehat{C}$-like symmetry constraining it, whether we call it that or not. As another example, in some "left-right symmetric theories", it is conventional to allow parity to switch the internal representations of $SU(2)_L$ and $SU(2)_R$, which is again useful precisely because it leads to an approximate symmetry. (However, to give credit to the mathematicians, the definition of $\widehat{C}\widehat{P}\widehat{T}$ is more "canonical", because it is the conserved quantity guaranteed to us by the CPT theorem. This operator always flips all internal quantum numbers and the helicity.)

Finally, we can check on a few applications of charge conjugation.

- The charge conjugate spinor ψ^c satisfies the Dirac equation. To see this, take the transpose of the Dirac equation for $\overline{\psi}$ for
$$(-i\gamma^{\mu T}\partial_\mu - m)\overline{\psi}^T(x) = 0.$$
Inserting factors of $C^{-1}C$ and using $\gamma^{\mu T} = -C^{-1}\gamma^\mu C$ gives the result.

- A Majorana fermion has $b^s(p) = d^s(p)$. That is, they are Dirac fermions that are their own antiparticles, $\psi^c = \psi$. They arise from quantizing solutions to the Dirac equation obeying a reality condition. Then a spin up Majorana fermion can be described by either the spinor $\zeta^{1/2}$ or $\xi^{1/2}$, where the spinors are related by $\zeta^s = i\sigma^2\xi^{s*}$. Note that a Majorana field doesn't have a definite chirality, just like a Dirac field.

- The vector current is odd under $\widehat{C}$. To see this cleanly, write
$$j^\mu(x) = \frac{1}{2}(\overline{\psi}\gamma^\mu\psi - \psi^T\gamma^{\mu T}\overline{\psi}^T) = \frac{1}{2}(\gamma^\mu)_{ij}[\overline{\psi}_i(x), \psi_j(x)]$$
where the sign flip from the transpose is explained in the notes on Quantum Field Theory. Applying charge conjugation, we have
$$\widehat{C}j^\mu\widehat{C}^{-1} = \frac{1}{2}(\gamma^\mu)_{ij}[\widehat{C}\overline{\psi}_i\widehat{C}^{-1}, \widehat{C}\psi_j\widehat{C}^{-1}] = -\frac{1}{2}(\gamma^\mu)_{ij}[(\psi^T C^{-1})_i, (C\overline{\psi}^T)_j] = \frac{1}{2}(\gamma^\mu)_{\ell k}[\psi_k, \overline{\psi}_\ell] = -j^\mu.$$
On the other hand, we know that the electromagnetic field is coupled as $A^\mu j_\mu$, so for QED to be charge conjugation invariant, we must define $\widehat{C}A_\mu\widehat{C}^{-1} = -A_\mu$.

- Similarly, one can show that the axial current is even under $\widehat{C}$. This implies that it is impossible to couple a linear combination of the vector and axial currents to a single field without violating $\widehat{C}$, and this is exactly what happens in the weak interactions.

Majorana spinors can be a bit confusing, because people use the term in many distinct ways, so we treat them carefully.

- To avoid confusion, we start with two-component Weyl fields, since Dirac and Majorana fields are built out of them. Suppose we have a left-chiral Weyl spinor field ψ which transforms under a representation R of an internal symmetry group. It annihilates a particle with negative helicity in the representation R, and creates a particle with positive helicity in the representation $\overline{R}$.

- In general, complex conjugating a quantum field just reverses which particles it creates and annihilates. The conjugate field $\psi^\dagger$ is a right-chiral Weyl spinor with internal symmetry representation $\overline{R}$. It annihilates a particle with positive helicity in the representation $\overline{R}$, and creates a particle with negative helicity in the representation R.

- Therefore, to describe a set of particles with $|h| = 1/2$, we can use only left-chiral Weyl fields, or only right-chiral Weyl fields, or a mixture of both. The field content of a theory is somewhat arbitrary. Note that the framework of fields can only describe particles which come in matter-antimatter pairs: for every particle species transforming in a given internal representation, there *must* be another particle species with opposite helicity and the same mass, transforming in the conjugate internal representation. This is a consequence of CPT symmetry.

- With a single left-chiral Weyl field ψ, there are only two ways to produce quadratic Lorentz-invariant terms in the Lagrangian. We know ψ transforms in $(1/2, 0)$, and its conjugate transforms in $(0, 1/2)$. Since $(1/2, 0) \times (1/2, 0) = (1, 0) + (0, 0)$, contracting the field with itself can yield a scalar. Since $(1/2, 0) \times (0, 1/2) = (1/2, 1/2)$, contracting the field with its conjugate yields a four-vector, which can yield a scalar upon contraction with ∂^μ.

- Therefore the two possible Lagrangian terms are

$$\mathcal{L} \supset i\psi^\dagger \overline{\sigma}^\mu \partial_\mu \psi + m\psi\psi$$

  where the second term is a two-component spinor contraction, defined in the notes on Supersymmetry, and the $\overline{\sigma}^\mu$ are just coefficients that isolate the appropriate scalar contraction.

- Now suppose ψ transforms in a representation R of an internal symmetry group. The first term is automatically invariant, but the second term transforms as $R \times R$, so it can only be invariant if R is a real representation. The logic is precisely the same if the symmetry group is a gauge group, except that ∂_μ must be replaced with an appropriate covariant derivative D_μ.

- For a right-chiral Weyl field χ, the logic is the same, but the terms are written as

$$\mathcal{L} \supset i\chi^\dagger \sigma^\mu \partial_\mu \chi + m\chi\chi.$$

  Again, with a gauge field, ∂_μ is replaced with a covariant derivative.

- Now, we can always stack a Weyl field and its conjugate into a four-component spinor field,

$$\Psi = \begin{pmatrix} \psi \\ \psi^{(c)} \end{pmatrix}.$$

  This is just a change of notation. There are still two possible terms in the Lagrangian,

$$\mathcal{L} \supset \frac{1}{2}\overline{\Psi}(i\partial\!\!\!/ - m)\Psi.$$

  which are just the same as the original ones, up to conventions for factors of 2, once one expands out the products. As before, the mass term is only allowed if R is real.

- Here's the tricky part: if there's a gauge field, the kinetic term should become

$$\mathcal{L} \supset \frac{1}{2}\overline{\Psi}(i(\partial\!\!\!/ + ie\gamma^5 A\!\!\!/) - m)\Psi.$$

That is, we do *not* use the minimal coupling prescription. Minimal coupling is a procedure for generating a scalar Lagrangian given fields which transform in known representations. But in general, Ψ does *not* transform in a well-defined representation of the internal symmetry group, because the top half transforms in R and the bottom half transforms in $\overline{R}$. Again, we can confirm the γ^5 has to be there by expanding everything in components. (A more common way to do this would be to add a chiral projector P_L. It leads to the same result when expanded in components, but our way treats the two halves of Ψ symmetrically.)

- Calculations with Standard Model fermions can be done with either two-component or four-component spinor fields. In both cases, explicit mass terms are forbidden, but masses are permitted by the Higgs mechanism. The advantage of four-component notation is that one can use familiar techniques for the traces of gamma matrices; the disadvantage is that γ^5 appears.

- Now we're ready to answer the key question: what is a Dirac spinor? Often, particles transforming in a representation R can be paired with other particles, of the same mass and *same* helicity, transforming in the representation $\overline{R}$. For instance, this can be done for all particles if the theory is symmetric under charge conjugation. We can describe a pair of such particle species using a pair of left-chiral and right-chiral Weyl spinor fields, ψ and χ, which transform in the same representation R.

- This allows a new term in the Lagrangian: we can contract one with the conjugate of the other to get a scalar, no matter what R is. This is called a Dirac mass term, and it is most easily written in four-component notation. Stacking these fields into a four-component spinor,

$$\Psi = \begin{pmatrix} \psi \\ \chi \end{pmatrix}$$

the Lagrangian is

$$\mathcal{L} \supset \overline{\Psi}(i(\not{\partial} + ie\not{A}) - m)\Psi$$

where there's no factor of $1/2$, since the two halves of Ψ are distinct particles, and we simply have a covariant derivative with no need for γ^5, since both halves of Ψ transform in R. At the level of two-component spinors the Dirac mass term looks like $\psi\chi + \overline{\psi}\overline{\chi}$.

- There are at least two distinct ways to define Majorana spinors.
  - Starting with a Dirac spinor transforming in a representation R, one can define a Majorana spinor by additionally imposing a reality condition $\Psi^{(c)} = \Psi$. In our language, this is equivalent to setting $\psi = \chi$, and demanding invariance of the kinetic term implies R must be real. This is the source of the claim that Majorana spinors can't be charged.
  - Starting with a Weyl spinor transforming in a representation R, one can define a Majorana spinor by stacking it on its conjugate. Demanding invariance of the explicit mass term implies R must be real, but if there is no such term, R can be arbitrary. This is not contradictory with the previous point, because in this case Ψ does not transform in a well-defined representation of the internal symmetry group.

Note. Imposing a reality condition might seem a bit artificial; alternatively, it's simple to produce Majorana spinors starting from only Dirac spinors. For example, suppose there is a global $U(1)$ symmetry, a Dirac field with charge 1, and a scalar field H with charge -2. We can then write

down terms like $\psi\psi H$, which turns into a Majorana mass for ψ when H gets a vev. This doesn't contradict the statement that massive Majorana spinors can't be charged, because the $U(1)$ is spontaneously broken by H.

This simple mechanism won't show up in typical textbooks, because they often only consider the $U(1)$ of electromagnetism, which we know holds to extreme precision. However, it's a common tool in model building for dark matter, where we might have a "dark" $U(1)$ separate from the Standard Model gauge groups. When the Majorana mass terms are much larger than the Dirac mass term, we get two distinct Majorana spinors, while if they're smaller, then we have the "pseudo-Dirac" case where the Majoranas ψ and χ have only a small splitting. The same idea can be applied to separate the components of a complex scalar, giving the "inelastic scalar" case. The latter two are concrete examples of "inelastic dark matter", where collisions can excite or de-excite the dark matter by interconverting the two particle species, leading to distinctive experimental signatures.

Note. Chiral gauge theories. Consider the fermions in a gauge theory. If, for every positive helicity particle in a representation R of the gauge group G, there is a negative helicity particle in the same representation R, the theory is not chiral; it doesn't distinguish between the two helicities.

Suppose we write all spinor fields in a theory as left-chiral Weyl spinors. They are collectively in a large representation S, and if S is not complex, the theory is not chiral. This remains true if spontaneous symmetry breaking reduces G to $H \subseteq G$, because S will still remain real; it will split into real representations plus pairs of conjugate representations of H.

This places a strong constraint on GUTs, because the SM is a chiral gauge theory. If S were not complex in a GUT, then it would yield unwanted extra "mirror matter" transforming in the conjugates of the SM particle representations. The mirror matter would have to be made very heavy while keeping ordinary matter light, and it is unclear how to achieve this naturally.

2.5 Time Reversal

We conclude with time reversal symmetry.

- We recall that time reversal symmetry takes $x \to x_T$, $p \to p_T$ where

  $$x_T^\mu = (-x^0, \mathbf{x}), \quad p_T^\mu = (p^0, -\mathbf{p}).$$

  In addition, time reversal flips the sign of the angular momentum. Note that $p_T x = -p x_T$.

- For the scalar field, we have

  $$\widehat{T} a(p) \widehat{T}^{-1} = \eta_T a(p_T), \quad \widehat{T} c^\dagger(p) \widehat{T}^{-1} = \eta_T c^\dagger(p_T)$$

  where the phases are equal as usual. Then

  $$\widehat{T}\phi(x)\widehat{T}^{-1} = \sum_p \widehat{T}a(p)\widehat{T}^{-1}e^{ipx} + \widehat{T}c^\dagger(p)\widehat{T}^{-1}e^{-ipx} = \eta_T \sum_p a(p)e^{-ipx_T} + c^\dagger(p)e^{ipx_T} = \eta_T \phi(x_T)$$

  where we used the antilinearity of $\widehat{T}$ in the first step, then reindexed the sum.

- For the Dirac field, we define

  $$\widehat{T} b^s(p) \widehat{T}^{-1} = \eta_T (-1)^{1/2-s} b^{-s}(p_T), \quad \widehat{T} d^{s\dagger} \widehat{T}^{-1} = \eta_T (-1)^{1/2-s} d^{-s\dagger}(p_T), \quad s = \pm 1/2$$

  where b maps to b because both the momentum and spin are flipped, keeping the helicity the same, and the extra phase factors are again constrained by Lorentz invariance.

- Given this definition, one can show the spinors satisfy

$$(-1)^{1/2-s}u^{-s*}(p_T) = -C^{-1}\gamma^5 u^s(p), \quad (-1)^{1/2-s}v^{-s*}(p_T) = -C^{-1}\gamma^5 v^s(p)$$

  and we define

$$B = C^{-1}\gamma^5 = -\gamma^5 C = \gamma^1\gamma^3 = \begin{pmatrix} i\sigma^2 & 0 \\ 0 & i\sigma^2 \end{pmatrix}.$$

- It is then straightforward to show that the Dirac field transforms as

$$\widehat{T}\psi(x)\widehat{T}^{-1} = \eta_T B\psi(x_T), \quad \widehat{T}\overline{\psi}(x)\widehat{T}^{-1} = \eta_T^*\overline{\psi}(x_T)B^{-1}.$$

  Then $\overline{\psi}(x)\psi(x) \to \overline{\psi}(x_T)\psi(x_T)$, which makes sense since charge density is T-even classically.

- To check the transformation properties of other bilinears, we use

$$B^{-1}\gamma^{5*}B = \gamma^5, \quad B^{-1}\gamma^{0*}B = \gamma^0, \quad B^{-1}\gamma^{i*}B = -\gamma^i.$$

  Then we have

$$\widehat{T}\overline{\psi}(x)\gamma^\mu\psi(x)\widehat{T}^{-1} = \overline{\psi}(x_T)B^{-1}\gamma^{\mu*}B\psi(x_T)$$

  so that $\overline{\psi}\gamma^\mu\psi$ has its spatial parts flipped. The axial current $\overline{\psi}\gamma^5\gamma^\mu\psi$ transforms the same way, essentially because $\widehat{T}$ is blind to chirality, and the currents only differ by chirality.

- We may also explicitly check that chirality is preserved, as

$$\widehat{T}\psi_L\widehat{T}^{-1} = \eta_T B\psi_L(x_T), \quad \widehat{T}\overline{\psi}_L\widehat{T}^{-1} = \eta_T^*\overline{\psi}_L(x_T)B^{-1}.$$

We now apply time reversal symmetry to S-matrix elements.

- The definition of the S-matrix in the interaction picture is

$$S = T\exp\left(-i\int dt\, V(t)\right), \quad V(t) = -\int d\mathbf{x}\,\mathcal{L}_I(x).$$

  For example, in QED, the interaction term is $\mathcal{L}_I = -e\overline{\psi}\gamma^\mu A_\mu\psi$.

- In a theory with C, P, and T symmetry, the Lagrangian is C, P, and T-even. The first two imply that the S-matrix satisfies

$$\widehat{P}S\widehat{P}^{-1} = S, \quad \widehat{C}S\widehat{C}^{-1} = S.$$

  Then the amplitude for $|i\rangle \to |f\rangle$ is the same as the amplitude for $\widehat{P}|i\rangle \to \widehat{P}|f\rangle$ or $\widehat{C}|i\rangle \to \widehat{C}|f\rangle$.

- Time reversal is more complicated. Note that $V(t)$ is real, and the time ordering puts later times to the left. Under conjugation by $\widehat{T}$, the factors of $-i$ are conjugated, and the time ordering is now in reverse. This is equivalent to an overall complex conjugation, so

$$\widehat{T}S\widehat{T}^{-1} = S^\dagger.$$

  Now we have

$$\langle i_T|S|f_T\rangle = \langle i|\widehat{T}^\dagger|S\widehat{T}|f\rangle = \langle i|\widehat{T}^\dagger S\widehat{T}|f\rangle^* = \langle f|S|i\rangle$$

  where the bar indicates the direction the antilinear operators act; swapping the direction picks up a complex conjugation. Then the amplitude for $|i\rangle \to |f\rangle$ equals that for $\widehat{T}|f\rangle \to \widehat{T}|i\rangle$.

Note. A summary table for gamma matrices. The fourth column is representation-independent, the first three are highly representation-dependent, and the last two are by definition.

	γ^*	γ^T	$\gamma^\dagger$	γ^{-1}	$C^{-1}\gamma C$	$B^{-1}\gamma B$
0	$+$	$+$	$+$	$+$	$(-)^T$	$(+)^*$
1	$+$	$-$	$-$	$-$	$(-)^T$	$(-)^*$
2	$-$	$+$	$-$	$-$	$(-)^T$	$(-)^*$
3	$+$	$-$	$-$	$-$	$(-)^T$	$(-)^*$
5	$+$	$+$	$+$	$+$	$(+)^T$	$(+)^*$

Note. A summary table for discrete symmetries, for a real scalar ϕ, a real vector V^μ, the special case A^μ, and Dirac bilinears. For objects with vector indices, we define $\mathcal{P} = \text{diag}(1,-1,-1,-1)$ and $\mathcal{T} = -\mathcal{P} = \text{diag}(-1,1,1,1)$.

	ϕ	V^μ	A^μ	$\overline{\psi}\psi$	$i\overline{\psi}\gamma^5\psi$	$\overline{\psi}\gamma^\mu\psi$	$\overline{\psi}\gamma^\mu\gamma^5\psi$	$\overline{\psi}\sigma^{\mu\nu}\psi$	∂_μ
$\widehat{C}$	$\eta_c = \pm 1$	η_c	-1	1	1	-1	1	-1	1
$\widehat{P}$	$\eta_p = \pm 1$	$\eta_p\mathcal{P}$	$\mathcal{P}$	1	-1	$\mathcal{P}$	$-\mathcal{P}$	$\mathcal{P}^\mu\mathcal{P}^\nu$	$\mathcal{P}$
$\widehat{T}$	$\eta_t = \pm 1$	$\eta_t\mathcal{T}$	$-\mathcal{T}$	1	-1	$-\mathcal{T}$	$-\mathcal{T}$	$-\mathcal{T}^\mu\mathcal{T}^\nu$	$\mathcal{T}$
$\widehat{C}\widehat{P}\widehat{T}$	1	-1	-1	1	1	-1	-1	1	-1

where the last line requires choosing $\eta_c\eta_p\eta_t = 1$, which can always be arranged. Note that CPT just gives a factor of -1 for each Lorentz index, so any Lorentz invariant Lagrangian is automatically CPT invariant. In addition, in non-abelian gauge theories, the potential, field strength, and current transform under C and P with extra matrix transposes.

3 Spontaneous Symmetry Breaking

3.1 Classical Fields

We know that statistical fields can experience spontaneous symmetry breaking, so similarly quantum fields can as well. We begin with the case of classical field theory. Note that despite the formal similarity, nothing we say will have anything to do with phase transitions; all of our quantum field theory is at zero temperature.

Example. The linear sigma model. Consider N real scalar fields with Lagrangian

$$\mathcal{L}(\phi, \partial_\mu \phi) = \frac{1}{2}(\partial_\mu \phi)^2 + \frac{1}{2}\mu^2 \phi^2 - \frac{\lambda}{4}\phi^4$$

where ϕ has N components, and we've suppressed dot products. Then the Lagrangian has an $O(N)$ symmetry. The dispersion relation about $\phi = 0$ contains excitations with negative mass squared, indicating a potential maximum rather than a minimum. The lowest-energy classical field configuration is a constant field ϕ_0. The potential is minimized for

$$\phi_0^2 = \frac{\mu^2}{\lambda}.$$

Since ϕ_0 can only take a single value, choosing it breaks the $O(N)$ symmetry down to $O(N-1)$, since we are still free to rotate in the directions orthogonal to ϕ_0. Suppose we pick

$$\phi_0 = (0, 0, \ldots, 0, v), \quad v = \frac{\mu}{\sqrt{\lambda}}.$$

We can expand the Lagrangian about the minimum by defining

$$\phi(x) = (\pi^1(x), \ldots, \pi^{N-1}(x), v + \sigma(x)).$$

Then we have

$$\mathcal{L} = \frac{1}{2}(\partial_\mu \pi^k)^2 + \frac{1}{2}(\partial_\mu \sigma)^2 - \frac{1}{2}(2\mu^2)\sigma^2 + \text{cubic and quartic interactions.}$$

That is, we find one massive field and $N-1$ massless fields. In the case $N = 2$, this reduces to the usual picture of a "Mexican hat potential".

Note. The crucial step that breaks the symmetry is selecting a specific vacuum state, not rewriting the Lagrangian. The new Lagrangian still has an $O(N)$ symmetry, though it's harder to see as it's nonlinearly realized; we couldn't have broken any symmetry because we merely redefined variables.

Note. Consider $N = 1$, where the broken symmetry is a discrete symmetry, $\mathbb{Z}_2$. In this case the experimental signature is not a Goldstone boson, but a domain wall. Without symmetry breaking, the $\mathbb{Z}_2$ symmetry means that the parity of the number of particles is conserved. With spontaneous symmetry breaking, we see cubic interaction terms, but the symmetry is still there, in a sense, because we can think of the fourth particle as coming from our vacuum, which acts as a source.

Example. We don't need to begin with negative squared masses. In the case of a potential $V(\phi) \sim -|\phi|^4 + |\phi|^6$ for a complex scalar ϕ, we start with two massless particles and end up with one massive particle and one massless Goldstone boson after symmetry breaking.

Note. To prove Goldstone's theorem classically, consider a Lagrangian of the form

$$\mathcal{L} = \text{kinetic} - V(\phi)$$

and let ϕ_0 be a constant field that minimizes V. Then we have

$$\frac{\partial V}{\partial \phi^a} = 0, \quad \frac{\partial^2 V}{\partial \phi^a \partial \phi^b} \equiv m_{ab}^2$$

where the derivatives are evaluated at ϕ_0. The number of Goldstone bosons is equal to the number of zero eigenvalues of the symmetric mass matrix m_{ab}^2. Now, a general continuous symmetry transformation has the form

$$\phi \to \phi + \alpha \Delta(\phi)$$

where α is infinitesimal. The condition for this to leave the Lagrangian invariant is

$$\Delta^a(\phi) \frac{\partial}{\partial \phi^a} V(\phi) = 0.$$

Differentiating with respect to ϕ^b and evaluating at $\phi = \phi_0$ gives

$$\Delta^a(\phi_0) \left(\frac{\partial^2 V}{\partial \phi^a \partial \phi^b} \right) = 0.$$

Now, the symmetry is spontaneously broken if $\Delta(\phi_0) \neq 0$, since it changes the vacuum. Then $\Delta(\phi_0)$ is an eigenvector of the mass matrix with zero eigenvalue, and hence a Goldstone boson.

Note. To count the Goldstone bosons, let G act on the fields leaving the Lagrangian invariant and let $H \subset G$ leave the vacuum ϕ_0 invariant. Let M be the set of vacua. If the vacuum degeneracy is entirely due to the group G, then the action of G on M must be transitive; that is, if it were not, we should really be working with a larger initial symmetry group. Then by the orbit-stabilizer theorem, $M \cong G/H$. The number of Goldstone bosons is simply the number of independent directions we can travel along M,

$$\dim M = \dim G - \dim H$$

so there are $\dim G - \dim H$ Goldstone bosons as desired. For example, in our model above, the initial symmetry group had dimension $N(N-1)/2$ and the new symmetry group had dimension $(N-1)(N-2)/2$, a decrease of $N-1$, and there were indeed $N-1$ Goldstone bosons. In words, the vacua live in the coset space, while the Goldstone bosons live in the tangent space to this space.

Note. The Goldstone boson counting doesn't depend on how we describe the symmetries. Consider the linear sigma model with N complex scalar fields; then a similar analysis to the above shows that the $U(N)$ symmetry is broken to $U(N-1)$, giving $2N-1$ Goldstone bosons. But the system actually has the larger symmetry group $O(2N)$ when written in terms of $2N$ real scalar fields, which is broken to $O(2N-1)$. This also gives $2N-1$ Goldstone bosons. These results match, even though $O(2N)$ is larger than $U(N)$, because the action of $U(N)$ alone on the vacua is still transitive.

Note. Here's another way of thinking about Goldstone bosons. When there's spontaneous continuous symmetry breaking, we can define our fields about the symmetry broken vacuum, around which there will be a symmetry which locally looks like a shift. That implies a Noether current of the form $J^\mu \sim \partial^\mu \phi + (\text{nonlinear terms})$. Conservation of that current gives $\partial^2 \phi = (\text{nonlinear terms})$, implying that ϕ is massless. Furthermore, the existence of the shift symmetry means couplings of ϕ, at least at leading order, should only involve derivative terms like $\partial^\mu \phi$.

3.2 Quantum Fields

In the quantum case, proving Goldstone's theorem is significantly trickier.

Note. The first puzzle is why spontaneous symmetry breaking should even be possible. For a $\mathbb{Z}_2$ symmetry, there are two candidate vacua $|\pm\rangle$. This is analogous to the case of a double well potential in quantum mechanics. In that case, quantum tunneling between them splits the degeneracy, and the true ground state is the symmetric combination; both energy eigenstates have zero vev.

This reasoning does not apply to quantum field theory because there are many more degrees of freedom. Then the amplitude for tunneling between the $|\pm\rangle$ vacua is exponentially suppressed; it is analogous to the amplitude for an infinite set of double well oscillators to all tunnel at once. Hence the symmetry may be broken by *any* effect that does not treat the $|\pm\rangle$ states symmetrically, such as a tiny external field. This is the same reason that all macroscopic objects have a definite orientation, even though their quantum ground state is spherically symmetric.

Note. Spontaneous symmetry breaking can be justified without needing any external field. Suppose we have a set of orthogonal degenerate vacuum states $|n\rangle$, where a vacuum state is defined as a state with zero momentum not part of a continuum of states, with no amplitude for tunneling between them by the arguments above. Consider local operators $A(x)$ and $B(y)$. Inserting the identity,

$$\langle n|A(x)B(y)|n'\rangle = \sum_m \langle n|A(x)|m\rangle\langle m|B(y)|n'\rangle + \sum_N \int d\mathbf{p}\, \langle n|A(x)|N_{\mathbf{p}}\rangle\langle N_{\mathbf{p}}|B(y)|n'\rangle$$

where N indexes over non-vacuum states. By the translational invariance of the vacuum, we have

$$\langle n|A(x)B(y)|n'\rangle = \sum_m \langle n|A(0)|m\rangle\langle m|B(0)|n'\rangle + \sum_N \int d\mathbf{p}\, e^{-i\mathbf{p}\cdot(\mathbf{x}-\mathbf{y})}\langle n|A(0)|N_{\mathbf{p}}\rangle\langle N_{\mathbf{p}}|B(0)|n'\rangle.$$

Then in the limit $|\mathbf{x}-\mathbf{y}| \to \infty$, the integral term goes to zero. Moreover, at spacelike separation $A(x)$ and $B(y)$ commute, so the matrices $\langle n|A(x)|m\rangle$ and $\langle m|B(y)|n'\rangle$ commute and can be simultaneously diagonalized.

Now, in a theory with one vacuum $|0\rangle$, we take as an axiom the cluster decomposition principle, which states that in the limit of large separation,

$$\langle 0|A(x)B(y)|0\rangle \to \langle 0|A(x)|0\rangle\langle 0|B(y)|0\rangle.$$

In this case, cluster decomposition only holds if we work in a basis of vacua where A and B are diagonal. In our examples above, the quantum field ϕ itself is a local operator, picking out the $|\pm\rangle$ states as valid vacua.

Note. Spontaneous symmetry breaking appears in the path integral through the choice of boundary condition. In the case of quantum mechanics, these boundary conditions don't matter for long times because of quantum tunneling, but for quantum field theory they do.

Note. We can easily prove Goldstone's theorem with the effective action. Considering a scalar field theory for simplicity, we have an effective potential $V_{\text{eff}}(\phi)$ whose minima give the allowed vevs. Assuming the symmetry is linearly realized on the fields, $V_{\text{eff}}(\phi)$ shares the same symmetries, as shown in the notes on Quantum Field Theory. Then our classical argument goes through, showing that broken symmetries give zero eigenvalues of the matrix of second derivatives of $V_{\text{eff}}(\phi)$. On the other hand, this matrix is related to the reciprocal of the momentum-space propagator by

$$\frac{\partial^2 V_{\text{eff}}(\phi)}{\partial\phi_n \partial\phi_m} = \Delta^{-1}_{nm}(p=0).$$

Then a zero eigenvalue of V_{eff} corresponds to a zero eigenvalue of the exact mass matrix, and hence a massless particle.

We now show Goldstone's theorem without using the effective action, in scalar field theory.

- Let ϕ be a vector of scalar fields and consider a continuous symmetry group G with generators indexed by a, with corresponding conserved currents $j^{\mu a}(x)$ and conserved charges Q^a. For the infinitesimal symmetry $\phi \to \phi + \epsilon t^a \phi$, Noether's theorem gives

$$Q^a = \int d\mathbf{x}\, J_0^a(\mathbf{x}) = \int d\mathbf{x}\, \pi_i(\mathbf{x}) t^a \phi^i$$

  where we work in Schrodinger picture, and $[H, Q^a] = 0$.

- The charge Q^a generates the symmetry transformation just as it does classically: using the equal-time commutation relations, we have

$$[Q^a, \phi(0)] = -it^a \phi(0).$$

  Then intuitively, the current $J_0^a(\mathbf{x})$ generates the symmetry "localized" near $\mathbf{x}$. For example, if the symmetry rotates ϕ_1 into ϕ_2, then $J_0(\mathbf{x})$ creates a "pion" $\phi_1 \overline{\phi}_2$ localized at $\mathbf{x}$.

- By definition, spontaneous symmetry breaking exists if the vacuum $|0\rangle$ is charged, $Q^a|0\rangle \neq 0$. Since H and Q^a commute, $Q^a|0\rangle$ is degenerate with $|0\rangle$.

- Next, we construct the states

$$|\pi^a(\mathbf{p})\rangle \sim \int d\mathbf{x}\, e^{i\mathbf{p}\cdot\mathbf{x}} J_0^a(\mathbf{x})|0\rangle.$$

  If E_0 is the vacuum energy, then these states have energy $E_0 + E(\mathbf{p})$. But when $\mathbf{p}$ is zero, the state is proportional to $Q|0\rangle$ and hence has energy E_0, so $E(0) = 0$. Then the states must satisfy a massless dispersion relation; they are the desired Goldstone bosons.

- Note that ϕ need not be a fundamental field. For example, in a theory with Dirac spinors, we could take $\phi = \overline{\psi}\psi$.

We can also show this result more formally.

- By inserting copies of the identity $\sum_n |n\rangle\langle n| = 1$ and using the translational invariance of the vacuum, we find

$$\langle 0|[j^{a\mu}(x), \phi(0)]|0\rangle = i \int d^4k\, \rho^{a\mu}(x) e^{-ikx} - \tilde{\rho}^{a\mu}(x) e^{ikx}$$

  where we define the spectral densities

$$i\rho^{a\mu}(k) = \sum_n \delta(k - p_n)\langle 0|j^{a\mu}(0)|n\rangle\langle n|\phi(0)|0\rangle, \quad i\tilde{\rho}^{a\mu}(k) = \sum_n \delta(k - p_n)\langle 0|\phi(0)|n\rangle\langle n|j^{a\mu}(0)|0\rangle.$$

  These are analogous to the spectral densities we found for the exact propagator, except that instead of $\phi \to \phi$ we are describing the amplitude for $\phi \to B$ where B is the particle created by the current. We will see that B can be interpreted as a Goldstone boson.

- Since ρ and $\tilde{\rho}$ are Lorentz vectors that only depend on k, they must be proportional to k^μ. They must also be zero for negative energy since the states all have positive energy. Then

$$\rho^{a\mu}(k) = k^\mu \theta(k^0)\rho^a(k^2), \quad \tilde{\rho}^{a\mu}(k) = k^\mu\theta(k^0)\tilde{\rho}^a(k^2).$$

  Substituting this in, we have

$$\langle 0|[j^{a\mu}(x), \phi(0)]|0\rangle = -\partial^\mu \int d^4k\, \theta(k^0)(\rho^a(k^2)e^{-ikx} + \tilde{\rho}^a(k^2)e^{ikx}).$$

- At this point, the result looks similar to the free propagator,

$$\langle 0|\phi(z)\phi(y)|0\rangle = \int \frac{d^4p}{(2\pi)^3}\theta(p^0)\delta(p^2-\sigma)e^{-ip(z-y)} = D(z-y,\sigma)$$

  where σ is the mass squared, but with weighting factors ρ^a and $\tilde{\rho}^a$. We write

$$\rho(k^2) = \int d\sigma\, \rho(\sigma)\delta(k^2-\sigma)$$

  and plug this in to find

$$\langle 0|[j^{a\mu}(x), \phi(0)]|0\rangle = -(2\pi)^3\partial^\mu \int d\sigma\, \rho^a(\sigma)D(x,\sigma) + \tilde{\rho}^a(\sigma)D(-x,\sigma).$$

- Now, the left-hand side must vanish at spacelike separation by causality. We know that for spacelike x, $D(x,\sigma) = D(-x,\sigma)$, so $\rho^a(\sigma) = -\tilde{\rho}^a(\sigma)$. Then we have

$$\langle 0|[j^{a\mu}(x), \phi(0)]|0\rangle = -\partial^\mu \int d\sigma\, \rho^a(\sigma) i\Delta(x,\sigma)$$

  where

$$i\Delta(x,\sigma) = (2\pi)^3(D(x,\sigma) - D(-x,\sigma)) = \int d^4k\, \delta(k^2-\sigma)\,\mathrm{sign}(k^0)e^{-ikx}$$

- Next, apply ∂_μ to both sides. By current conservation, the left-hand side vanishes, and we can simplify the right-hand side with the Klein–Gordan equation, for

$$0 = \int d\sigma\, \sigma\rho^a(\sigma) i\Delta(x,\sigma).$$

  For this to hold for all x, we require $\sigma\rho^a(\sigma) = 0$, since σ and ρ are positive.

- In the case where $\rho^a(\sigma) = 0$, we have $\langle 0|[Q^a, \phi(0)]|0\rangle \propto \langle 0|t^a\phi(0)|0\rangle = t^a\phi_0 = 0$, so the symmetry is unbroken. Otherwise, we have $\rho^a(\sigma) = N^a\delta(\sigma)$ and we can explicitly calculate

$$t^a\phi_0 = i\langle 0|[Q^a, \phi(0)]|0\rangle = N^a \int d\mathbf{x}\, \partial^0\Delta(x,0) = -(2\pi)^3N^a \neq 0$$

  so the symmetry is indeed broken. Returning to the definition of the spectral density, there must be families of massless states $|B(p)\rangle$ where

$$\langle 0|j^{\mu a}(0)|B(p)\rangle = iF_B^a p^\mu, \quad \langle B(p)|\phi(0)|0\rangle = Z^B$$

  by dimensional analysis and Lorentz invariance. The states $|B(p)\rangle$ are spinless since $\phi(0)|0\rangle$ is rotationally invariant and carry the same quantum numbers as j^0. They are the desired Goldstone bosons.

Note. There is another simple proof that Goldstone bosons remain massless, though it is only valid perturbatively. If the exact propagator $\Delta(k^2)$ of a Goldstone bosons is to retain a pole at $k^2 = 0$, then the self-energy should satisfy $\Pi(k^2 = 0) = 0$. However, Goldstone bosons are derivatively coupled, so all diagrams one can draw come with powers of the external momentum k. Since $\Pi(k^2 = 0)$ can be evaluated at $k = 0$, these diagrams vanish.

Note. There are several exceptions to Goldstone's theorem. In $d \leq 2$, the Mermin–Wagner theorem ensures that spontaneous continuous symmetry breaking can never occur in the first place; concretely, the effective potential will never develop an instability at $\phi = 0$.

Gauge symmetry also complicates the picture. If a global symmetry is gauged then its current cannot create Goldstone bosons, because it merely takes a state to the very same physical state. If we try to work only with physical states, we either lose manifest Lorentz invariance, which invalidates the formal proof above, or we have states with negative norm (e.g. in Lorenz gauge). This also invalidates the proof above since ρ may be negative. Instead, we would see that the would-be Goldstone bosons are "eaten" to produce massive gauge bosons.

3.3 Gauge Theories

We now consider the case where spontaneous symmetry breaking occurs to a global symmetry, whose corresponding local symmetry is gauged.

Note. This situation is also called "spontaneous breaking of a gauge symmetry", but this is a misnomer. The local gauge symmetry remains a gauge symmetry; the choice of vacuum doesn't affect the fact that states related by a gauge transformation are physically the same. Actually breaking gauge symmetry would be disastrous; it occurs in the case of a gauge anomaly and destroys the Ward identities, making the theory inconsistent. Breaking the global symmetry does not violate gauge symmetry since we require gauge transformations to vanish at infinity.

Incidentally, it is also possible to have local symmetries that are not gauge symmetries, e.g. in lattice spin systems considered in the notes on Statistical Field Theory. However, Elitzur's theorem states that such a local symmetry can never be broken. There are two ways to argue this. We can imagine introducing a symmetry breaking field h and taking the limits $N \to \infty$ followed by $h \to 0$, in which case a global symmetry is broken because the energy cost Nh goes to infinity, while a local symmetry isn't because the energy cost is $O(h)$ which goes to zero.

Alternatively, suppose there is no external field; then we care about tunneling between ground states related by the symmetry by local thermal fluctuations. Suppose the symmetry is discrete. In the case of a global symmetry, there is an extensive energy cost since we must form a domain wall, and hence cannot happen for $d > 1$. But in the case of a local symmetry, we can always relate two such ground states by a series of local transformations which each cost no energy.

Thus, a local gauge symmetry can't be broken in a consistent theory, while a local non-gauge symmetry can't be broken at all.

Example. The abelian Higgs model. Consider the theory of scalar QED with a potential,

$$\mathcal{L} = -\frac{1}{4}(F_{\mu\nu})^2 + |D_\mu \phi|^2 - V(\phi), \quad V(\phi) = -\mu^2 \phi^* \phi + \frac{\lambda}{2}(\phi^* \phi)^2, \quad \lambda > 0$$

with $D_\mu = \partial_\mu + ieA_\mu$. This Lagrangian has the $U(1)$ gauge symmetry

$$\phi(x) \to e^{i\alpha(x)}\phi(x), \quad A_\mu(x) \to A_\mu(x) - \frac{1}{e}\partial_\mu \alpha(x).$$

As usual, the field ϕ acquires a vev

$$\langle\phi\rangle = \phi_0 = \left(\frac{\mu^2}{\lambda}\right)^{1/2}$$

which we have chosen to be real, breaking the global $U(1)$ symmetry. Expanding about the vev,

$$\phi(x) = \phi_0 + \frac{1}{\sqrt{2}}(\phi_1(x) + i\phi_2(x))$$

for real scalar fields ϕ_1 and ϕ_2, where we've chosen the constant so that the ϕ_i have canonical kinetic terms. Now the potential becomes

$$V(\phi) = -\frac{1}{2\lambda}\mu^4 + \mu^2\phi_1^2 + \text{interactions}$$

so ϕ_1 has a mass $m = \sqrt{2}\,\mu$ and it appears that ϕ_2 is a massless Goldstone boson. However, there are new terms in the kinetic term of ϕ,

$$|D_\mu\phi|^2 = \frac{1}{2}(\partial_\mu\phi_1)^2 + \frac{1}{2}(\partial_\mu\phi_2)^2 + \sqrt{2}\,e\phi_0 A_\mu\partial^\mu\phi_2 + e^2\phi_0^2 A_\mu A^\mu + \text{interactions}.$$

The first new term allows the photon and ϕ_2 to mix, while the second term is a photon mass term. In particular, it turns out that the ϕ_2 becomes the third, longitudinal mode of the now massive photon. The easiest way to see this is to go to unitary gauge, where ϕ is real. Then

$$|D_\mu\phi|^2 = (\partial_\mu\phi)^2 + e^2\phi^2 A_\mu A^\mu$$

which provides a photon mass term, $m_A = \sqrt{2}\,e\phi_0$, with no massless Goldstone bosons. We say the Goldstone boson ϕ_2 has been "eaten" by the gauge boson to gain mass, keeping the total number of degrees of freedom the same.

Note. To see how the Goldstone boson is eaten diagrammatically, note that the Goldstone-photon vertex has a factor of $m_A k^\mu$. Resumming the propagator using the photon mass vertices gives a pole in the appropriate place, and including the Goldstone-photon vertex provides the right numerator. This can be seen at lowest order, where the correction is

$$im_A^2\left(\eta^{\mu\nu} - \frac{k^\mu k^\nu}{k^2}\right)$$

which is exactly the right numerator polarization structure for a massive particle.

Note. One has to be careful with unitary gauge. Consider scalar QED without symmetry breaking. If we switch to unitary gauge, we apparently find a real scalar and a massless vector, without gauge symmetry; the number of degrees of freedom appears to have dropped. The problem is that now the vector has a third, massive, spin zero degree of freedom.

Next, we consider a generic non-abelian example.

- Consider a theory of scalar fields ϕ_i invariant under a compact group G, represented by

  $$\phi_i \to (1 + i\alpha^a t^a)_{ij}\phi_j = (1 - \alpha^a T^a)_{ij}\phi_j.$$

  Without loss of generality we can work with only real fields. Since the representation is finite-dimensional, it is unitary, so the t^a are imaginary and Hermitian, and the T^a are real and antisymmetric.

- Promoting the symmetry to a local gauge symmetry, the covariant derivative is

$$D_\mu \phi = (\partial_\mu + igA^a_\mu t^a)\phi = (\partial_\mu - gA^a_\mu T^a)\phi.$$

  Then the kinetic energy term for the ϕ_i is

$$\frac{1}{2}(D_\mu \phi_i)^2 = \frac{1}{2}(\partial_\mu \phi_i)^2 - gA^a_\mu(\partial_\mu \phi_i T^a_{ij}\phi_j) + \frac{1}{2}g^2 A^a_\mu A^{b\mu}(T^a\phi)_i(T^b\phi)_i.$$

- Now let the field acquire a vev, $\langle\phi_i\rangle = \phi_{0i}$. The last term yields a gauge boson mass term,

$$\Delta\mathcal{L} = \frac{1}{2}m^2_{ab}A^a_\mu A^{b\mu}, \quad m^2_{ab} = g^2(T^a\phi_0)_i(T^b\phi_0)_i.$$

  The mass matrix is positive semidefinite, so the gauge bosons receive nonnegative masses.

- Suppose a generator T^a leaves the vacuum invariant, $T^a\phi_0 = 0$. Then the corresponding gauge boson is massless, as expected.

- The interaction between the Goldstone bosons and the gauge bosons is

$$\Delta\mathcal{L} = -gA^a_\mu \partial_\mu \phi_i (T^a\phi_0)_i.$$

  As expected, the only components of ϕ that mix are those parallel to $T^a\phi_0$ for some transformation T^a, which is precisely the set of Goldstone bosons. In other words, each massive gauge boson eats the Goldstone corresponding to the broken symmetry it generates. Just as in the abelian case, this provides the desired third polarization in the gauge boson propagator.

- The mass eigenstates in a gauge theory with gauge group G are multiplets of G. For example, quarks are color triplets and gluons form a color octet. In the case of symmetry breaking, one can show that mass eigenstates are instead multiplets of the unbroken gauge group H, by checking that the mass matrices all commute with the generators of H in the appropriate representation. For example, the particles in the SM have definite electric charge.

We now consider a series of non-abelian examples.

Example. Consider an $SU(2)$ gauge theory where ϕ transforms in the spinor representation. Then

$$D_\mu \phi = (\partial_\mu + igA^a_\mu \tau^a)\phi, \quad \tau^a = \frac{\sigma^a}{2}.$$

If ϕ acquires a vev, then by the $SU(2)$ symmetry we can let it be

$$\langle\phi\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ v\end{pmatrix}.$$

The mass term for the gauge bosons has the form

$$\Delta\mathcal{L} = \frac{1}{2}g^2 \begin{pmatrix}0 & v\end{pmatrix}\tau^a\tau^b\begin{pmatrix}0\\ v\end{pmatrix}A^a_\mu A^{b\mu} = \frac{1}{4}g^2\begin{pmatrix}0 & v\end{pmatrix}\{\tau^a,\tau^b\}\begin{pmatrix}0\\ v\end{pmatrix}A^a_\mu A^{b\mu} = \frac{g^2v^2}{8}A^a_\mu A^{a\mu}$$

where we used $\{\tau^a,\tau^b\} = \delta^{ab}/2$. Therefore all three gauge bosons receive the mass $m_A = gv/2$.

Example. Consider the same example, but let ϕ transform in the vector representation. The covariant derivative is, in components,

$$(D_\mu\phi)_a = \partial_\mu\phi_a - g\epsilon_{abc}A^b_\mu\phi_c$$

and we choose the vev to be $\langle\phi_a\rangle = v\delta_{a3}$. Then the mass term is

$$\Delta\mathcal{L} = \frac{g^2v^2}{2}(\epsilon_{abc}A^b_\mu\delta_{c3})^2 = \frac{g^2v^2}{2}((A^1_\mu)^2 + (A^2_\mu)^2)$$

so we only get two massive gauge bosons. This makes sense, since the symmetry of rotations about the ϕ_3 axis is preserved. Since the model contains both massive and massless gauge bosons, it was once proposed as a candidate theory of the weak interaction, but it's not quite right: we require two massive bosons that only couple to left-handed fields (the $W^\pm$ bosons), and one massive boson that couples to both handednesses (the Z). This can't be achieved by breaking $SU(2)$ in any way.

Example. Consider an $SU(3)$ gauge theory where ϕ transforms in the adjoint representation. Then

$$(D_\mu\phi)_a = \partial_\mu\phi_a - gf_{abc}A^b_\mu\phi_c, \quad \Delta\mathcal{L} = \frac{1}{2}(D_\mu\phi)_a(D^\mu\phi)_a \supset \frac{g^2}{2}(f_{abc}A^b_\mu\phi_c)^2.$$

In this case, it's more convenient to work without components. We let $\Phi = \phi_a t^a$, so

$$D_\mu\Phi = \partial_\mu\Phi + ig[A_\mu, \Phi], \quad \Delta\mathcal{L} = \operatorname{tr}(D_\mu\Phi D^\mu\Phi) \supset -g^2\operatorname{tr}([t^a, \Phi][t^b, \Phi])A^a_\mu A^{b\mu}.$$

where the normalization is fixed by using the Gell-Mann matrices. Now, we can always rotate Φ so that it is diagonal, but there are still several distinct possibilities. If $\Phi_0 = |\phi|\operatorname{diag}(1, 1, -2)$, then the masses of the A^a_μ are

$$a \in \{1, 2, 3, 8\}\colon 0, \quad a \in \{4, 5, 6, 7\}\colon 3g|\phi|$$

so the symmetry is broken to $SU(2) \times U(1)$. If $\Phi_0 = |\phi|\operatorname{diag}(1, -1, 0)$, then the masses are

$$a \in \{3, 8\}\colon 0, \quad a \in \{1, 2\}\colon 2g|\phi|, \quad a \in \{4, 5, 6, 7\}\colon g|\phi|$$

and the symmetry is broken to $U(1) \times U(1)$. We see that matter fields in the adjoint can't break the symmetry corresponding to the Cartan subalgebra. To break these symmetries, we would have to add another matter field transforming in a different representation.

Now we give a more formal analysis of the Higgs mechanism.

- So far we have considered symmetry breaking by scalar fields picking up vevs, but other mechanisms could be possible; we will see such a mechanism for chiral symmetry in QCD. Hence we would like an analysis that is independent of how the symmetry is broken.
- Consider a theory with a global symmetry G and let α parametrize the global symmetry. By the usual Noether trick, if we promote α to $\alpha(x)$, then

$$\delta\mathcal{L} = -(\partial_\mu\alpha^a)J^{\mu a}, \quad \partial_\mu J^{\mu a} = 0.$$

- We may couple this globally symmetric theory to a non-abelian gauge field by

$$\mathcal{L}' = \mathcal{L} + gA^a_\mu J^{\mu a} + O(A^2)$$

which has the effect of gauging the symmetry. By directly plugging in $\delta\mathcal{L}$, we see that $\mathcal{L}'$ is indeed gauge invariant, up to unspecified $O(A^2)$ terms. However, we will only need the linear term to compute matrix elements involving only one insertion of the gauge field.

- As mentioned above, if the global symmetry G is spontaneously broken, the currents $J^{\mu a}$ generate the Goldstone bosons,

$$\langle 0|J^{\mu a}(x)|\pi_k(p)\rangle = -ip^\mu F^a{}_k\, e^{-ipx}.$$

Here, the $F^a{}_k$ are only nonzero if the symmetry a is broken.

- As an example, in the cases we studied earlier, we have

$$J^{\mu a} = \partial_\mu \phi_i T^a_{ij} \phi_j.$$

The Goldstone bosons are the ϕ_i which are shifted by the global transformation, and indeed,

$$\langle 0|J^{\mu a}(x)|\phi_i(p)\rangle = (T^a\phi_0)_i \langle 0|\partial^\mu\phi_i|\phi_i(p)\rangle = -ip^\mu (T^a\phi_0)_i e^{-ipx}.$$

This is just of the form above, with the identification

$$F^a{}_i = T^a_{ij}\phi_{0j}.$$

- Now, the vacuum polarization amplitude for the gauge bosons is

$$\Pi^{\mu\nu}_{ab}(k^2) = i\left(\eta^{\mu\nu} - \frac{k^\mu k^\nu}{k^2}\right)(m^2_{ab} + O(k^2))$$

where m_{ab} is the gauge boson mass matrix. To compute this, note that the pole at $k^2 = 0$ comes from the diagram with an intermediate Goldstone boson. Using our two equations involving $J^{\mu a}$ above, the $A^a_\mu\phi_j$ vertex factor is $-gk^\mu F^a{}_j$, giving

$$\Pi^{\mu\nu}_{ab}(k^2) \supset (gk^\mu F^a{}_j)\frac{i}{k^2}(-gk^\nu F^b{}_j)$$

so the gauge boson mass matrix is

$$m^2_{ab} = g^2 F^a{}_j F^b{}_j.$$

The mass matrix is again manifestly positive semidefinite.

Note. If we wanted to analyze how electroweak symmetry breaking occurred earlier in our universe, we would have to understand how thermal effects change the Higgs potential. This can be done using standard techniques from thermal field theory, which are sketched in my dissertation.

One should also account for quantum corrections to the potential, which can be quite important. For instance, in massless ϕ^4 theory coupled to a $U(1)$ gauge field, quantum effects cause spontaneous symmetry breaking even though it doesn't happen classically; the effective potential in this case is called the Coleman-Weinberg potential. We managed to ignore this above by either implicitly assuming weak coupling or working with the effective potential.

3.4 Quantization

In this section we consider the quantization of theories with spontaneous symmetry breaking of a gauged global symmetry.

- First, we focus on the abelian case. To establish conventions, we define

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + (D_\mu\phi)^*(D^\mu\phi) - \frac{g}{2}\left(\phi^*\phi - \frac{1}{2}v^2\right)^2, \quad D_\mu\phi = \partial_\mu\phi - ieA_\mu\phi$$

  where the gauge transformations are

$$\phi \to e^{ie\alpha}\phi, \quad A_\mu \to A_\mu - \partial_\mu\alpha.$$

  The potential ensures $|\phi| = v/\sqrt{2}$. In unitary gauge, we can see that the gauge field gains a mass $M_A^2 = e^2v^2$. There is also a remaining massive scalar field corresponding to the radial part of ϕ. with mass $m^2 = gv^2$.

- As shown in the notes on Quantum Field Theory, the propagator for a massive vector boson is

$$D_{\mu\nu}(k^2) = -\frac{i}{k^2 - M_A^2 + i\epsilon}\left(\eta_{\mu\nu} - \frac{k_\mu k_\nu}{M_A^2}\right).$$

  However, this makes renormalizability unclear, because the propagator does not fall off at high k. Since we are dealing with a gauge theory, we should also be more careful to account for Faddeev–Popov ghosts.

- To be more explicit, we show what happens before gauge fixing. We parametrize

$$\phi(x) = \frac{1}{\sqrt{2}}\left(v + f(x) + i\varphi(x)\right)$$

  where we use a global symmetry to set v real. Then the full kinetic term is

$$D_\mu\phi^* D^\mu\phi = \frac{1}{2}\partial_\mu f\partial^\mu f + \frac{1}{2}\partial_\mu\varphi\partial^\mu\varphi + \frac{1}{2}e^2v^2A_\mu A^\mu + evA^\mu\partial_\mu\varphi$$
$$- eA^\mu(\varphi\partial_\mu f - f\partial_\mu\varphi) + e^2vfA^\mu A_\mu + \frac{1}{2}A^\mu A_\mu(f^2 + \varphi^2).$$

  This is quite complicated, but the terms are mostly interaction terms; the most problematic term is the mixing term,

$$\mathcal{L} \supset evA^\mu\partial_\mu\varphi = -M_A(\partial_\mu A^\mu)\varphi$$

  which, as we've seen, makes it difficult to interpret A^μ or φ alone.

- As such, any convenient gauge fixing must suppress this term. In path integral quantization, we may choose the gauge fixing function

$$F(A, \varphi) = \partial_\mu A^\mu - \xi M_A\varphi.$$

  Furthermore, we integrate over gauge fixings with a Gaussian weight of width ξ. As shown in the notes on Quantum Field Theory, the Lagrangian picks up the terms

$$\mathcal{L} \supset -\frac{1}{2\xi}F(A)^2 + \bar{c}\Delta_{\text{FP}}c.$$

  This is a generalization of R_ξ gauge.

- Unlike previous examples, the gauge fixing function F now depends on matter fields as well as the gauge field. To evaluate the Faddeev–Popov determinant, we go back to the definition,

$$\Delta_{\rm FP} = \frac{\partial F}{\partial \alpha} = -\frac{\partial F}{\partial A_\mu} D_\mu + \frac{\partial F}{\partial \varphi} e(v+f) = -\partial^2 - \xi e M_A (v+f).$$

  Since this is an abelian gauge theory, the ghosts do not couple to the gauge field directly, but have indirect effects by their coupling to f.

- Expanding the extra $F^2/2\xi$ terms, the mixing term is cancelled, leaving quadratic terms

$$\begin{aligned}\mathcal{L}_{\rm quad} = \frac{1}{2}(\partial_\mu f)(\partial^\mu f) - \frac{1}{2}m^2 f^2 + \frac{1}{2}\partial_\mu\varphi\partial^\mu\varphi - \frac{\xi}{2}M_A^2\varphi^2 \\ - \frac{1}{2}A_\mu\left(-\eta^{\mu\nu}\partial^2 + \left(1-\frac{1}{\xi}\right)\partial^\mu\partial^\nu - M_A^2\eta^{\mu\nu}\right)A_\nu - \bar{c}\partial^2 c - \xi M_A^2 v\bar{c}c\end{aligned}$$

  where the f mass term is from the potential for ϕ. The propagators are now

$$D_{\mu\nu}(k^2) = -\frac{i}{k^2 - M_A^2 + i\epsilon}\left(\eta_{\mu\nu} - (1-\xi)\frac{k_\mu k_\nu}{k^2 - \xi M_A^2}\right), \quad D_\varphi(k^2) = \frac{i}{k^2 - \xi M_A^2 + i\epsilon}$$

  and

$$D_c(k^2) = \frac{i}{k^2 - \xi M_A^2}, \quad D_f(k^2) = \frac{i}{k^2 - m_f^2}.$$

  Renormalizability is now easier to show, as all propagators fall off as $1/k^2$, but not all the fields are physical, as signaled by the ξ-dependent masses of φ and the ghost field. The f field is physical, and in the Standard Model will correspond to the Higgs boson.

- The physical results should not depend on ξ, which can be a useful cross-check in computations, just as it was in QED. The ξ-independence may be proven generally using the BRST symmetry of the gauge-fixed Lagrangian.

- One useful special case is $\xi = 0$, where the Goldstone boson φ is massless and the gauge field is fully transverse,

$$D_{\mu\nu}(k^2) = -\frac{i}{k^2 - M_A^2 + i\epsilon}\left(\eta_{\mu\nu} - \frac{k_\mu k_\nu}{k^2}\right), \quad D_\varphi(k^2) = \frac{i}{k^2 + i\epsilon}.$$

  Both propagators have poles at $k^2 = 0$, but they don't correspond to physical particles.

- Another useful special case is $\xi = 1$, where

$$D_{\mu\nu}(k^2) = -\frac{i\eta_{\mu\nu}}{k^2 - M_A^2 + i\epsilon}, \quad D_\varphi(k^2) = \frac{i}{k^2 - M_A^2 + i\epsilon}.$$

  This gauge, called the Feynman–'t Hooft gauge, is the most convenient for general higher-order computations.

- We recover unitary gauge in the limit $\xi \to \infty$, where the unphysical fields decouple; the unphysical poles in k^2 go to infinity. This is called unitary gauge because every pole found by evaluating Feynman diagrams corresponds to the propagation of physical intermediate states, consistent with the Cutkosky rules, so unitarity is manifest.

- In 1972, 't Hooft and Veltman used R_ξ gauge to prove the renormalizability of the Standard Model at all orders in perturbation theory. For any finite ξ, it is easy to show that the divergences can be cancelled by a finite number of counterterms, since the usual power counting arguments will work. 't Hooft and Veltman additionally showed that the counterterms preserved local gauge invariance, and the ξ-independence of S-matrix elements.

Though we have now set up the trickiest Feynman rules, loop computations in the Standard Model are quite complicated, and we will not perform any in these notes.

4 Electroweak Theory

4.1 Gauge Theory

We now describe the Weinberg–Salam theory of the electroweak interaction.

- We postulate a gauge group $SU(2)_L \times U(1)_Y$, where the factors are called weak isospin and hypercharge, and a complex scalar field ϕ, called the Higgs field. The Higgs transforms as a weak isospin doublet with a $U(1)$ hypercharge $Y = 1/2$,

$$\phi(x) \to e^{i\alpha^a(x)\tau^a} e^{i\beta(x)/2}\phi(x), \quad \tau^a = \frac{\sigma^a}{2}.$$

- We suppose the Higgs acquires a vev, through the same potential as in the abelian Higgs model. Using the $SU(2)_L \times U(1)_Y$ global symmetry, without loss of generality we can pick

$$\phi_0 = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ v\end{pmatrix}.$$

  This breaks the symmetry to $U(1)_{\text{EM}}$, generated by gauge transformations with $\alpha^3(x) = \beta(x)$.

- The covariant derivative for the Higgs is

$$D_\mu\phi = (\partial_\mu + igA^a_\mu\tau^a + \frac{i}{2}g'B_\mu)\phi$$

  and the Lagrangian is

$$\mathcal{L} = -\frac{1}{2}\operatorname{tr} F^A_{\mu\nu}F_{A\mu\nu} - \frac{1}{4}F^B_{\mu\nu}F^{B\mu\nu} + (D_\mu\phi)^\dagger(D^\mu\phi) - \mu^2|\phi|^2 - \lambda|\phi|^4.$$

  Here, the field strengths are

$$F^{Aa}_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g\epsilon^{abc}A^b_\mu A^c_\nu, \quad F^B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu.$$

- Expanding out the gauge boson mass term, we have

$$\begin{aligned}\mathcal{L} &\supset \frac{1}{2}\begin{pmatrix}0 & v\end{pmatrix}(gA^a_\mu\tau^a + \frac{1}{2}g'B_\mu)(gA^{b\mu}\tau^b + \frac{1}{2}g'B^\mu)\begin{pmatrix}0\\ v\end{pmatrix}\\ &= \frac{1}{2}\frac{v^2}{4}\left(g^2(A^1_\mu)^2 + g^2(A^2_\mu)^2 + (-gA^3_\mu + g'B_\mu)^2\right).\end{aligned}$$

  Therefore, we find three massive vector bosons, which we write as

$$W^\pm_\mu = \frac{1}{\sqrt{2}}(A^1_\mu \mp iA^2_\mu), \quad Z^0_\mu = \frac{1}{\sqrt{g^2+g'^2}}(gA^3_\mu - g'B_\mu), \quad m_W = g\frac{v}{2}, \quad m_Z = \sqrt{g^2+g'^2}\,\frac{v}{2}.$$

  The fourth vector field, which is orthogonal to Z^0_μ, remains massless,

$$A_\mu = \frac{1}{\sqrt{g^2+g'^2}}(g'A^3_\mu + gB_\mu).$$

  This field is identified with the QED vector potential.

- The general covariant derivative may be rewritten in terms of the mass eigenstates as

$$\begin{aligned} D_\mu &= \partial_\mu + igA_\mu^a T^a + iYg'B_\mu \\ &= \partial_\mu + \frac{ig}{\sqrt{2}}(W_\mu^+ T^+ + W_\mu^- T^-) + \frac{i}{\sqrt{g^2+g'^2}} Z_\mu (g^2 T^3 - g'^2 Y) + i\frac{gg'}{\sqrt{g^2+g'^2}} A_\mu (T^3 + Y) \end{aligned}$$

  where the T^a are $SU(2)$ generators in the appropriate representation and $T^\pm = T^1 \pm iT^2$. To simplify this result, we define the electron charge e and electric charge Q as

$$e = \frac{gg'}{\sqrt{g^2+g'^2}}, \quad Q = T^3 + Y.$$

  Note that Y is sometimes defined with an extra factor of 2.

- Note that the Z also couples directly to anything with hypercharge, so it can couple to particles that are $SU(2)_L$ singlets. This isn't a phenomenological problem, because the photon can couple to anything the Z can. That is, the Z boson doesn't produce any new decays; at low energies its effect is totally washed out by that of the photon.

- Next, we define the weak mixing angle or Weinberg angle θ_w so that

$$\begin{pmatrix} Z^0 \\ A \end{pmatrix} = \begin{pmatrix} \cos\theta_w & -\sin\theta_w \\ \sin\theta_w & \cos\theta_w \end{pmatrix} \begin{pmatrix} A^3 \\ B \end{pmatrix}, \quad \cos\theta_w = \frac{g}{\sqrt{g^2+g'^2}}.$$

  Then the covariant derivative simplifies to

$$D_\mu = \partial_\mu + \frac{ig}{\sqrt{2}}(W_\mu^+ T^+ + W_\mu^- T^-) + \frac{ig}{\cos\theta_w} Z_\mu (T^3 - \sin^2\theta_w Q) + ieA_\mu Q$$

  and we have

$$e = g\sin\theta_w, \quad m_W = m_Z \cos\theta_w.$$

- The theory is predictive: with just four parameters (g, g', μ^2, and λ), all of the masses and (self-)interactions of the electroweak bosons and the Higgs are fixed. For example, the Higgs trilinear and quartic couplings are predicted but currently not measured to any precision; they will be targeted by future colliders.

Note. In the Standard Model, the mechanism of spontaneous symmetry breaking is much less constrained than the other parts of it. It is therefore interesting to consider which of the above predictions follow solely from the spontaneous symmetry breaking pattern $SU(2)_L \times U(1)_Y \to U(1)_{\text{EM}}$, and which additionally rely on there being a single, $SU(2)_L$ doublet Higgs field.

We focus on the mass matrix of the four gauge bosons. First, note that $U(1)_{\text{EM}}$ transformations don't commute with two of the $SU(2)_L$ transformations; this implies that a pair of $SU(2)_L$ bosons pick up opposite electric charges. Thus, the mass matrix must take the form

$$\begin{pmatrix} m_1^2 & & & \\ & m_2^2 & & \\ & & m_3^2 & m^2 \\ & & m^2 & m_0^2 \end{pmatrix}$$

where additional off-diagonal terms are forbidden by $U(1)_{\text{EM}}$ symmetry, which additionally forces $m_1 = m_2 \equiv m_W$. Because $U(1)_{\text{EM}}$ is unbroken, there must be a massless gauge boson A_μ,

which implies $m^2 = \pm|m_0 m_3|$. Requiring that A_μ take the same form as found above implies $\tan\theta_W = |m_0/m_3|$. Therefore, the only thing not determined is the mass of the Z-boson,

$$m_Z^2 = m_0^2 + m_3^2 = m_3^2/\cos^2\theta_W.$$

This matches the Standard Model prediction precisely when $m_3 = m_W$. We can thus search for deviations from the Standard Model by measuring $\rho = m_W^2/(m_Z^2\cos^2\theta_W)$, which at tree level is $\rho_0 = 1$. Loop corrections contribute $\Delta\rho \approx 0.008$.

4.2 Coupling to Matter

Next, we couple fermions to the gauge bosons and Higgs. We begin with leptons.

- It suffices to find the $SU(2)_L$ and $U(1)_Y$ representations the fermions transform in; we can then read the interaction terms off the covariant derivative. We are guided by the experimental fact that the weak interactions only affect left-helicity particles and right-helicity antiparticles.
- We postulate the left-handed electron and electron neutrino fit in an isospin doublet,

  $$L(x) = \begin{pmatrix}\nu_e(x)\\ e_L(x)\end{pmatrix}, \quad e_L(x) = \frac{1}{2}(1-\gamma^5)e(x).$$

  We define $R(x) = e_R(x)$. To get the observed electric charges, we need

  $$Y = -\frac{1}{2} \text{ for } L(x), \quad Y = -1 \text{ for } R(x).$$

  The lepton-gauge boson part of the electroweak Lagrangian is thus

  $$\mathcal{L} \supset \overline{L} i\not{D} L + \overline{R} i\not{D} R$$

  which is just the Weyl Lagrangian with covariant derivatives.
- Expanding out the covariant derivatives, we can show the interaction terms are

  $$\mathcal{L} \supset -\frac{g}{2\sqrt{2}}(J^\mu W_\mu^+ + J^{\mu\dagger} W_\mu^-) - eJ_{\text{EM}}^\mu A_\mu - \frac{g}{2\cos\theta_W} J_n^\mu Z_\mu$$

  where we have defined the leptonic charged weak, neutral weak, and electromagnetic currents

  $$J^\mu = \overline{\nu}_e\gamma^\mu(1-\gamma^5)e, \quad J_n^\mu = \frac{1}{2}(\overline{\nu}_e\gamma^\mu(1-\gamma^5)\nu_e - \overline{e}\gamma^\mu(1-\gamma^5-4\sin^2\theta_w)e), \quad J_{\text{EM}}^\mu = -\overline{e}\gamma^\mu e.$$

  Note that right-handed electrons can couple to the Z, as mentioned earlier. Also note that because $\sin^2\theta_w$ is close to $1/4$, the coupling of the charged leptons to the Z boson is almost purely axial, i.e. proportional to γ^5.

Note. One might wonder if the weak force can form bound states; for instance, the Z boson mediates an attractive interaction between neutrinos. However, while all potential wells in 1D and 2D have bound states, sufficiently weak potential wells in 3D don't, and indeed there are no weak bound states in the SM. However, certain models of WIMP dark matter could have them.

Next, we consider lepton couplings to the Higgs.

- Next, we wish to write down a mass term for the electron, but the Dirac mass term $m_e(\overline{e}_L e_R + \overline{e}_R e_L)$ is not gauge invariant. Instead, all mass terms in the SM come from Yukawa couplings to the Higgs. We work in unitary gauge where

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}$$

and choose $Y = 1/2$ for the Higgs doublet so that the Higgs boson h is electrically neutral. Note that the components of ϕ with charge are exactly the ones 'eaten' by the charged $W_\mu^\pm$.

- Then the coupling to the Higgs is

$$\mathcal{L} \supset -\sqrt{2}\lambda_e(\overline{L}\phi R + \overline{R}\phi^\dagger L) = -\lambda_e(v+h)(\overline{e}_L e_R + \overline{e}_R e_L) = -m_e \overline{e}e - \lambda_e h \overline{e}e$$

giving a mass $m_e = \lambda_e v$ and a Yukawa coupling λ_e to the Higgs, proportional to m_e. For now, we'll take the neutrino to be massless.

- In reality, there are three generations of leptons, so we write

$$L^1 = \begin{pmatrix} \nu_e \\ e_L \end{pmatrix}, \quad L^2 = \begin{pmatrix} \nu_\mu \\ \mu_L \end{pmatrix}, \quad L^3 = \begin{pmatrix} \nu_\tau \\ \tau_L \end{pmatrix}, \quad R^1 = e_R, \quad R^2 = \mu_R, \quad R^3 = \tau_R.$$

Then the Higgs-lepton coupling has the generic form

$$\mathcal{L} \supset -\sqrt{2}(\lambda^{ij}\overline{L}^i \phi R^j + \lambda^{\dagger ij}\overline{R}^i \phi^\dagger L^j).$$

Here the generation indices are kept explicit, the spinor indices are contracted between $\overline{L}$ and R, and the weak isospin indices are contracted between L and ϕ. Note that the adjoint/dagger acts on all spaces, so $\overline{L}$ is a row vector in weak isospin space with $Y = 1/2$. Similarly, $\phi^\dagger$ is a row vector in weak isospin space with $Y = -1/2$.

- In general, the weak interactions and the Higgs interactions will pick out two different bases, the flavor basis and the mass basis. In the SM, neutrinos have no mass, so this problem doesn't arise. It also doesn't occur for the charged leptons, as we now show.

- Now λ is an arbitrary complex matrix, so it can't be diagonalized in the usual way. But since $\lambda\lambda^\dagger$ is Hermitian and positive, we have

$$\lambda\lambda^\dagger = U\Lambda^2 U^\dagger$$

where Λ^2 is diagonal and positive, and U is unitary. Taking Λ to also be diagonal and positive, we define $S = \lambda^\dagger U \Lambda^{-1}$, so S is unitary as well, and

$$\lambda^\dagger\lambda = S\Lambda^2 S^\dagger, \quad \lambda = U\Lambda S^{-1}.$$

Hence we may diagonalize λ if we use different unitaries on both ends.

- We now redefine the lepton fields by

$$L^i \to U^{ij}L^j, \quad R^i \to S^{ij}R^j, \quad \overline{L}^i \to U^{ij*}\overline{L}^j = \overline{L}^j(U^\dagger)^{ji}, \quad \overline{R}^i \to \overline{R}^j(S^\dagger)^{ji}.$$

Here, the transformations for the barred quantities follow because taking the Dirac adjoint performs a complex conjugation. The covariant derivative terms aren't affected, while the mass matrix λ is diagonalized to Λ, as desired.

Note. More about weak isospin. Given two objects A^α and B^α in the fundamental representation of $SU(2)_L$, their inner product $A^{\alpha *}B^\alpha$ is invariant. To unpack this, we note that $A^{\alpha *} \equiv A_\alpha$ transforms in the antifundamental representation. The contraction $A_\alpha B^\alpha$ is then invariant; forming invariants is just a matter of matching up the indices, as for the Lorentz group. In the examples above, everything starts with an upper index, and taking the adjoint lowers the index. Note that the fundamental representation of $SU(2)$ is pseudoreal and hence similar to the antifundamental, i.e. we may raise and lower indices with $\epsilon^{\alpha\beta}$.

Next, we perform the same procedure for the quark fields.

- The left-handed quarks fit into $SU(2)_L$ doublets,

$$Q_L^i = \begin{pmatrix} u^i \\ d^i \end{pmatrix}_L = \left(\begin{pmatrix} u \\ d \end{pmatrix}_L \quad \begin{pmatrix} c \\ s \end{pmatrix}_L \quad \begin{pmatrix} t \\ b \end{pmatrix}_L \right).$$

  To get the right electric charges, we take $Y = 1/6$.

- The right-handed quarks fit into $SU(2)_L$ singlets, which we write as

$$u_R^i = (u_R, c_R, t_R), \quad d_R^i = (d_R, s_R, b_R)$$

  with hypercharge $Y = 2/3$ and $Y = -1/3$ respectively.

- The quarks couple to the gauge bosons by

$$\mathcal{L} \supset \overline{Q}_L i \not{D} Q_L + \overline{u}_R i \not{D} u_R + \overline{d}_R i \not{D} d_r$$

  as usual. These terms violate C and P, but obey CP and T symmetry; note that here we are referring to the quantum $\widehat{C}$, so the CP symmetry acts on fields like classical C symmetry, conjugating them.

- The most general renormalizable gauge invariant quark-Higgs couplings are

$$\mathcal{L} \supset -\sqrt{2} \left(\lambda_d^{ij} \overline{Q}_L^i \phi d_R^j + \lambda_u^{ij} \overline{Q}_L^i \phi^c u_R^j + \text{h.c.} \right), \quad \phi^{c\alpha} \equiv \epsilon^{\alpha\beta} \phi_\beta^\dagger.$$

  In the second term, we need to use $\phi^\dagger$ to get hypercharge invariance, and a Levi–Civita to get a weak isospin invariant contracting with $\overline{Q}_\alpha$. Also, by hypercharge, there's no term involving u_R and d_R. Since CP conjugates the fields, the coupling is CP invariant if and only if λ_d^{ij} and λ_u^{ij} are real. Roughly speaking, complex physical parameters indicate CP violation.

- Next, we switch to the mass basis, as we did for the leptons. As before, we let

$$\lambda_u = K_u \Lambda_u S_u^\dagger, \quad \lambda_d = K_d \Lambda_d S_d^\dagger$$

  and redefine the quark fields by

$$u_L \to K_u u_L, \quad d_L \to K_d d_L, \quad u_R \to S_u u_R, \quad d_R \to S_d d_R.$$

  Then we have, for example, in unitary gauge

$$\sqrt{2} \lambda_d^{ij} \overline{Q}_L^i \phi d_R^j \supset v\, \overline{d}_L^i \lambda_d^{ij} d_R^j \to v\, \overline{d}_L K_d^\dagger K_d \Lambda_d S_d^\dagger S_d d_R = v\, \overline{d}_L \Lambda_d d_R$$

  and the quark mass term becomes

$$\mathcal{L} \supset -v \left(\Lambda_d^{ij} \overline{d}_L^i d_R^j + \Lambda_u^{ij} \overline{u}_L^i u_R^j + \text{h.c.} \right) = -v \sum_i m_d^i \overline{d}_L^i d_R^i + m_u^i \overline{u}_L^i u_R^i + \text{h.c.}$$

- Now, this redefinition affects the gauge couplings. The terms $\overline{u}_R i\not{D} u_R + \overline{d}_R i\not{D} d_r$ are not affected, but $\overline{Q}_L i\not{D} Q_L$ is because the covariant derivative mixes u_L and d_L. In particular, the charged weak current transforms as

$$J^\mu = \overline{u}^i \gamma^\mu (1-\gamma^5) d^i = 2\overline{u}^i_L \gamma^\mu d^i_L \to 2\overline{u}^i_L \gamma^\mu (K_u^\dagger K_d)^{ij} d^j_L.$$

  However, the neutral current remains diagonal, because it does not convert up-type quarks to down-type quarks. Thus the SM, at tree level, has no flavor-changing neutral currents.

- We define the CKM matrix by

$$K_u^\dagger K_d = V_{\text{CKM}} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}.$$

  The off-diagonal elements quantify the mismatch between the mass basis and the flavor basis. When we talk about an "up quark", we conventionally mean the mass basis.

Note. We began this discussion in the flavor basis, where we noted that gauge boson couplings are CP invariant, but introducing the mass terms broke CP symmetry. But usually we work in the mass basis, where we say that the mass terms have CP symmetry, while the gauge boson couplings break CP, as discussed below. So which term "really" breaks CP? Neither. The point is that there is no single "objective" definition of CP, because discrete symmetries are only defined up to an arbitrary linear transformation on the fields. This extra transformation can be chosen to leave the gauge boson couplings invariant, or the mass terms invariant, but not both at once.

Next, we'll introduce a useful way to count degrees of freedom, and apply it to the CKM matrix.

- Consider an atom in an external electric or magnetic field. Naively, the field has three degrees of freedom, but we can always take it to be along the z-axis, giving only one degree of freedom. The reason is that the atom and field still have $SO(3)$ symmetry provided we rotate them together, so we can align the field with the z-axis without loss of generality. To count the number of degrees of freedom of the perturbation, we note that the atom alone has only $SO(2)$ symmetry, by rotations orthogonal to the field direction. We have lost $3-1=2$ symmetry generators, which were the ones used to align the field with the z-axis, so the field is described by only $3-2=1$ parameter.

- In a more general situation, suppose that some couplings break a symmetry. We can formally think of the couplings as spurions (i.e. effectively as external fields) which transform under that symmetry, reducing the reasoning to the previous case. The number of parameters needed to break the couplings is the naive number, minus the number of broken symmetry generators.

- Now consider a general $n \times n$ complex matrix. Each entry is a complex number with a magnitude and phase, so there are n^2 real parameters and n^2 phase parameters.

- An $n \times n$ orthogonal matrix has $n(n-1)/2$ real parameters. However, an orthogonal matrix can just be thought of as a unitary matrix with the phases removed, and a unitary matrix has n^2 parameters, so a unitary matrix has $n(n-1)/2$ real parameters and $n(n+1)/2$ phases.

- This reasoning can also be understood by realizing that unitary matrices correspond with ordered bases of $\mathbb{C}^n$. The first basis vector is described by $n-1$ real parameters, for the magnitudes of the components, and n phases, for the phases of the components. Restricting to the orthogonal subspace, the second basis vector is described by $n-2$ real parameters and $n-1$ phases, and so on.
- Now we apply these results to the CKM matrix. Without the Yukawa couplings, the quark sector has a $U(3)^3$ symmetry, by unitary transformations of u_R, d_R, and Q_L individually; this corresponds to 9 real parameters and 18 phases. Adding the Yukawa couplings breaks this to a $U(1)_B$ symmetry, which is 1 phase.
- The Yukawa couplings take the form of two 3×3 complex matrices, with 18 real parameters and 18 phases. Hence the quark sector has 9 real parameters (6 quark masses and 3 CKM angles) and 1 phase.
- One might worry that anomalies upset this parameter counting, once we account for quantum effects. Indeed, the $U(3)^3$ symmetry includes the axial $U(1)_A$ symmetry, which is anomalous. The corresponding new term is the QCD θ-term, which has no effect classically.
- In the above derivation, we derived the CKM matrix rather differently, as we used a $U(3)^4$ quark field redefinition, which is not a symmetry of the Lagrangian even for $\lambda_u = \lambda_d = 0$. This is precisely why the form of the rest of the Lagrangian changed, i.e. why we picked up the CKM matrix in the first place. We will find the physical parameters of the CKM matrix below explicitly, but this heuristic analysis using $U(3)^3$ symmetry tells us what to expect.

Example. Spurions are ubiquitous. Consider a theory of a complex scalar field and a Weyl fermion,

$$\mathcal{L} = (\partial_\mu \phi)^2 + i\psi^\dagger \bar{\partial} \psi - m_\phi^2 |\phi|^2 - \frac{1}{2} m_\psi \psi^2 + \mathcal{L}_{\text{int}}.$$

In the limit $m_\phi \to 0$, we recover a shift symmetry for the scalar, so m_ϕ is a spurion for this symmetry. Assuming the interaction obeys this theory, the mass of the scalar can't become much larger than m_ϕ, at least perturbatively. Similarly, m_ψ is a spurion for chiral symmetry. Scale invariance is restored when both mass terms go to zero, and supersymmetry is restored when the mass terms become equal. Supersymmetry effectively transfers the chiral symmetry of the spinor to the scalar.

Next, we investigate the degrees of freedom in the CKM matrix.

- In the case of two generations, unitarity implies that V_{CKM} has four parameters, which can be expressed as an angle and three phases,

  $$V_{\text{CKM}} = \begin{pmatrix} \cos\theta_c e^{i\alpha} & \sin\theta_c e^{i\beta} \\ -\sin\theta_c e^{i(\alpha+\gamma)} & \cos\theta_c e^{i(\beta+\gamma)} \end{pmatrix}.$$

  However, in the absence of the CKM matrix, the Lagrangian would be invariant under a global phase rotation of any quark field,

  $$q_L^i \to e^{i\alpha^i} q_L^i, \quad q^i \in \{u, d, s, c\}.$$

  On the other hand, a rotation of all four quark simultaneously doesn't change the CKM matrix, because it is the $U(1)_B$ symmetry. Since the CKM matrix breaks three $U(1)$ symmetries, we can use them to remove all phases in the CKM matrix; then there is no CP violation. The remaining angle is called the Cabibbo angle.

- In this case, the charged weak current is

$$\frac{1}{2}J^\mu = \cos\theta_c \bar{u}_L\gamma^\mu d_L + \sin\theta_c \bar{u}_L\gamma^\mu s_L - \sin\theta_c \bar{c}_L\gamma^\mu d_L + \cos\theta_c \bar{c}_L\gamma^\mu s_L.$$

- Kobayashi and Maskawa proposed a third generation of quarks, which would allow for CP violation. In this case, there are 3 angles and 6 phases, but only 5 quark phases available. Thus the CKM matrix can be parametrized in terms of three angles and one phase.

- The CKM matrix is conventionally written in terms of the Wolfenstein parameters,

$$V_{\text{CKM}} = \begin{pmatrix} 1-\lambda^2/2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\lambda^2 & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{pmatrix} + O(\lambda^4)$$

  where $\lambda \approx 0.22$, $A \approx 0.81$, $\rho \approx 0.12$, and $\eta \approx 0.36$. This is useful because it parametrizes generation-mixing effects as powers of λ. For example, crossing from the first to the third generation is penalized by a factor λ^3. The CP violating phase is parametrized by η. Note that the top-left block is simply the 2×2 CKM matrix with Cabibbo angle λ.

- The unitarity of the CKM matrix is often tested by plotting unitarity triangles. We know that the inner product of any two distinct columns, or any two distinct rows, must vanish, and each inner product is the sum of three complex numbers, so there are six 'unitarity triangles' in the complex plane that must close. In most cases, the triangle is very flat because some terms are much bigger than others, so we usually plot

$$\sum_i V_{id}V_{ib}^* = 0$$

  because every term is $O(\lambda^3)$.

Note. Summarizing the matter content, we have the isospin doublets

$$\phi = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v+h \end{pmatrix}, Y = 1/2, \quad L = \begin{pmatrix} \nu_e \\ e_L \end{pmatrix}, Y = -1/2, \quad Q_L = \begin{pmatrix} u_L \\ d_L \end{pmatrix}, Y = 1/6$$

and the isospin singlets

$$R = e_R, Y = -1, \quad u_R, Y = 2/3, \quad d_R, Y = -1/3.$$

Note that the hypercharge is always the average electric charge of a weak isospin multiplet.

Note. It's important to avoid thinking of the Higgs sector of the SM as obvious. Over the 50 years between its proposal and discovery, many influential physicists expressed skepticism, as described in the historical review The Theoretical Physics Ecosystem Behind the Discovery of the Higgs Boson.

At the time of its proposal, it was not clear that the Higgs mechanism (i.e. the pattern of electroweak symmetry breaking which gives mass to the W and Z bosons) was even necessary. Glashow had proposed a model in 1961 where these masses were simply put in by hand, as an explicit symmetry breaking, and viewed it as no less legitimate than breaking flavor symmetry by hand. The Higgs mechanism, proposed in 1964 and used to by Weinberg and Salam to complete the SM in 1967 and 1968, gained greater acceptance in 1971 when 't Hooft showed it to be renormalizable,

in contrast to Glashow's setup. (But in the 1970s we also learned that renormalizability was less important of a criterion than had been thought, due to the rise of Wilsonian ideas.)

Even when the Higgs mechanism became more accepted, the Higgs *boson* was not. The Higgs field was simply the analogue of the Ginzburg–Landau order parameter field in superconductivity. In that case, the field was meant to measure some aspect of the collective behavior of the electrons, so the natural analogue would have been to view the Higgs field as representing a condensate of other particles. (Examples of such theories included top quark condensates, where top quarks play the role of electrons, and technicolor, which breaks electroweak symmetry by strong gauge interactions, and has no discernible Higgs excitation at all.) Many physicists, especially condensed matter physicists, thought that postulating an elementary Higgs was naive, the result of taking an order parameter field too literally. A further issue, realized throughout the 1970s, is that an elementary Higgs boson requires fine tuning. As a result, thousands of papers have been written on alternatives to an elementary Higgs.

The current experimental results have mostly wiped out Higgsless theories, because we now know there is a new scalar with a mass of about 125 GeV. Currently, it is known that this scalar has the same quantum numbers as the Higgs, and its direct Yukawa couplings to bottom and top quarks have been measured, assuming the Higgs vev is as expected. However, we have measured none of the other Yukawa couplings, or any features of the Higgs potential. For example, it is possible that there is "induced electroweak symmetry breaking",

$$V \supset \mu^2 H^\dagger \tilde{H} + m^2 |H|^2 + V(\tilde{H})$$

where a second Higgs doublet $\tilde{H}$ acquires a vev, creating a linear term in the Higgs potential and leading to the observed Higgs mass and vev. In this case, there could be no Higgs quartic term. In addition, some models with composite Higgs bosons remain viable. Distinguishing between these options would be a task for a post-LHC collider. Of course, if the Higgs continues to appear fundamental, and nothing else shows up, the fine-tuning problems pointed out 50 years ago would become even more severe.

4.3 Symmetries of the Standard Model

Now we step back and examine the symmetries of the SM, neglecting neutrino masses.

- All quarks couple to gluons with the same strength, because they all transform in the fundamental of $SU(3)_C$. In addition, all leptons couple to W bosons with equal strength, because all the L^i transform in the fundamental of $SU(2)_L$. This result is known as lepton universality.
- Lepton universality doesn't apply to quarks, because of the CKM matrix, but one can still get nice results upon summing over quarks. For example, for W^+ decay at tree level, we have

  $$\Gamma = \Gamma(W^+ \to e^+ \nu_e) \left(3 + 3 \sum_{n=1}^{2} \sum_{m=1}^{3} |V_{nm}|^2 \right)$$

  where the first factor of 3 comes from the three generations of leptons, and the next factor of 3 comes from the three quark colors. Since the CKM matrix unitary, the sum is equal to 2, giving the simple result that the W^+ decays to hadrons 2/3 of the time.
- All CP violation is due to the complex phase in the CKM matrix. Thus, any CP violating process must involve all three generations, giving a suppression of $\lambda^6 \sim 10^{-3}$. Measurements of CP violation are therefore sensitive probes of new physics.

- The only particles in the SM that connects fermions with different flavors are the W bosons, through the off-diagonal elements of the CKM matrix.

Next, we consider flavor changing neutral currents (FCNC).

- In QED, matter is coupled to photons through the interaction $A_\mu J^\mu_{\text{EM}}$. Therefore, if we integrate the photon out, we get an effective interaction $J^\mu_{\text{EM}} J_{\mu,\text{EM}}$. A similar structure appears when we integrate out all the other SM bosons, yielding a "charged current" interaction $J^\mu J_\mu$ and a "neutral current" interaction $J^\mu_n J_{\mu,n}$, distinguished by the electric charges of J^μ and J^μ_n.

- At tree level, charged current interactions are mediated by W bosons, while neutral current interactions (if we define the term rather inclusively) are mediated by the Z boson, gluons, photons, and the Higgs. However, a loop of W bosons could also contribute to the neutral current interaction.

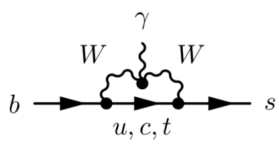
- The Standard Model turns out to have no tree-level FCNC, as we have already seen above. This is obvious for gluons and photons, whose interactions are flavor diagonal. For the Higgs, it occurs because the Yukawa couplings to the Higgs are proportional to the masses, but it wouldn't be true for a more complicated Higgs sector, such as a two Higgs doublet model.

- Tracing back, we might wonder why there was no CKM matrix for Z bosons, which would have led to tree-level FCNC. This occurred because all the up-type quarks (and down-type quarks) coupled to the Z boson identically. Thus, the matrix of couplings was proportional to the identity, and the matrices from changing to mass basis cancelled out, $K_\mu^\dagger K_\mu = K_d^\dagger K_d = I$.

- Now consider a general matter sector. The mass terms can connect fields with the same $SU(3)_C \times U(1)_{\text{EM}}$ irrep, so each type of irrep corresponds to a mass matrix. Meanwhile, fields in the same $SU(3)_C \times SU(2)_L \times U(1)_Y$ irrep couple the same way to the Z. Therefore, the logic above goes through as long as all fields with the same $SU(3)_C \times U(1)_{\text{EM}}$ irrep automatically have the same $SU(3)_C \times SU(2)_L \times U(1)_Y$ irrep.

- This holds in the SM, but before the discovery of the charm quark, the strange quark was proposed to be in an $SU(2)_L$ singlet, which would have led to tree-level FCNC with the down quark. This would have produced a sizable rate for the neutral kaon decay $K^0 \to \mu^+\mu^-$, but it was measured to be very rare, with a branching fraction of about 10^{-9}. That result led to the prediction of the charm quark.

- The GIM mechanism further suppresses FCNC, through a cancellation at loop level. Consider the process $b \to s\gamma$ by a W loop which emits a photon, as shown below.



  Concretely, this might be part of a B meson decay process. The amplitude is proportional to

$$\sum_{i \in \{u,c,t\}} V_{ib} V_{is}^* f(m_i^2/m_W^2).$$

  Now consider Taylor expanding the function f. At zeroth order, the result vanishes by unitarity of the CKM matrix. Beyond zeroth order, we can't have a bare factor of $\log(m_i^2/m_W^2)$, since this

would blow up as $m_i \to 0$, which means the leading correction is at least suppressed by m_i^2/m_W^2 (possibly multiplied by a logarithm). This is small for everything but the very massive top quark, which is the reason flavor physics can be used to "measure" the top quark mass. However, the top quark amplitude is suppressed by several powers of the Wolfenstein parameter λ. Thus, either the top or charm quark loops could be dominant, depending on the circumstances.

- To estimate the contribution from the charm quark loop, note that

$$\int đp\, f(p^2) = \frac{1}{(4\pi)^2} \int dx\, x f(x)$$

so in general a loop gives a numerical factor of roughly $1/(4\pi)^2$. Then the amplitude scales as

$$\frac{1}{(4\pi)^2} \frac{m_c^2}{m_W^2} \frac{1}{m_W^2}$$

compared with a generic $1/\Lambda^2$ for new physics. Thus, in general, loop suppression of a process in the SM allows us to probe new physics at scales up to $\Lambda \sim 10 m_W \sim$ TeV, while the GIM mechanism gives an additional factor of m_W/m_c, reaching an incredible $\sim 100\,\text{TeV}$.

- However, new physics can still exist below this scale. For instance, a new particle's couplings could have a trivial flavor structure, coupling identically to each up-type and down-type quark, in which case FCNC is not modified at tree level. (One example is the minimal dark photon.)

- Alternatively, the new couplings could be proportional to (powers of) the existing Yukawa couplings. This is the paradigm of "minimal flavor violation", which makes the SM Yukawa couplings the only source of flavor violation. It also suppresses new FCNC, and is commonly used in SUSY model building. A final possibility is "flavor alignment", where the new couplings and the Yukawa couplings can be simultaneously diagonalized. So flavor constraints don't totally rule out new physics, but rather place strong constraints on how it can look.

Note. The general rule of thumb for loops given above is useful in many contexts. For example, corrections due to a gluon loop are of order $g_3^2/(4\pi)^2 \sim 10^{-2}$, corrections due to a photon loop are of order $e^2/(4\pi)^2 \sim 10^{-3}$, and weak loops are intermediate. On the other hand, depending on the process, loop corrections may also come with logarithmic factors, which could be as large as $\log(m_W^2/\Lambda_{\text{QCD}}^2) \sim 10$ for gluon loops and $\log(m_W^2/m_e^2) \sim 20$ for photon loops.

Processes involving charged particles can also proceed with an extra photon in the final state. The rule of thumb is that the rate comes with a factor of $e^2(2\pi)/(2\pi)^3 = \alpha/\pi \sim 2 \times 10^{-3}$, where the numerator comes from the angular integration and the denominator comes from the momentum integration measure. (The rest of the phase space integral is not substantially affected, as long as the photon is soft.) As a concrete example,

$$\frac{\text{Br}(\mu^- \to e^- \overline{\nu} \nu \gamma)}{\text{Br}(\mu^- \to e^- \overline{\nu} \nu)} = (1.4 \pm 0.4)\%$$

where the enhancement is due to a large logarithm $\log(m_\mu^2/m_e^2)$.

The SM also has a number of accidental global symmetries.

- An accidental symmetry is a symmetry that arises from the field content, renormalizability, and other symmetries, but is not put in by hand. For example, in QED, the most general renormalizable Lagrangian is

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + iaF_{\mu\nu}\widetilde{F}^{\mu\nu} + i\overline{\psi}\not{D}\psi + \overline{\psi}(m + i\gamma^5 m_5)\psi.$$

  However, the second term is a total derivative and the final term may be removed by a chiral rotation $\psi \to e^{i\alpha\gamma^5}\psi$. Then we have accidental C, P, and T symmetry.

- The SM also contains all possible renormalizable terms, and has several accidental symmetries: the baryon number $U(1)_B$ and the individual lepton numbers $U(1)_{L_e}$, $U(1)_{L_\mu}$, and $U(1)_{L_\tau}$. The dimension 5 neutrino mass violates both individual and total lepton number, while dimension 6 operators can violate baryon number.

- Note that either $U(1)_B$ or $U(1)_L$ alone is sufficient to prevent the proton from decaying. Also, if the proton decays, it must decay into an odd number of fermions by Lorentz invariance, which requires the parity of the fermion number to be conserved. The only fermions lighter than the proton are leptons, so lepton number must be violated in the decay.

- Proton decay has been tested stringently, placing a high bound on Λ. For a rough estimate, we have $\tau > 10^{33}$ years while the decay rate should be m_p^5/Λ^4 by dimensional analysis, where the numerator accounts for the phase space; then $\Lambda > 10^{15}$ GeV, a result similar to the bound from neutrino masses. Thus new physics is either very far away, or respects baryon number.

- Violations of the individual $U(1)_{L_i}$ have also been searched for, most stringently through the unobserved decay $\mu \to e\gamma$, which will be probed further by the upcoming MEG II experiment. There is also the upcoming Mu2e experiment, which will search for $\mu \to e$ conversion in nuclei.

- It turns out that anomalies violate B and L conservation, as discussed further in the notes on Quantum Field Theory, but $L_i - L_j$ remains exactly conserved, as does $B - L$ if there is a sterile neutrino. Neutrino masses break $L_i - L_j$, while Majorana neutrino masses also break L.

Note. In the Standard Model, the result $\rho_0 = 1$ follows from an approximate accidental symmetry. Since the Higgs is a complex doublet, thereby containing 4 real fields, its most general possible global symmetry is $O(4)$. This symmetry is preserved by the Higgs potential, as it only depends on the combination $\phi^\dagger\phi$. When the Higgs field develops a vev, it is broken to $O(3)$, and since $\mathfrak{o}(3) \cong \mathfrak{su}(2)$, this residual symmetry is called custodial $SU(2)$.

Writing it in terms of real fields is clunky, but we can work with complex fields by defining the Higgs matrix $\Phi = (\phi, \epsilon\phi^*)$. The Higgs potential is a function of $\mathrm{tr}(\Phi^\dagger\Phi)$, which preserves the $SU(2)_L \times SU(2)_R$ symmetry

$$\Phi \to L\Phi R^\dagger$$

where $L, R \in SU(2)$. The $SU(2)_L$ factor acts just like electroweak $SU(2)_L$, while $U(1)_Y \subseteq SU(2)_R$. When Φ acquires a vev, which we can take to be proportional to the identity, the diagonal subgroup corresponding to $L = R$ is preserved, and this is the custodial $SU(2)$.

The coupling to gauge fields can be written as

$$\mathcal{L} \supset \mathrm{tr}(D_\mu\Phi)^\dagger D^\mu\Phi, \quad D_\mu\Phi = \partial_\mu\Phi + igA_\mu^a\tau^a\Phi + \frac{i}{2}g'B_\mu\Phi\sigma_3$$

where the σ_3 is on the right, because ϕ and $\epsilon\phi^*$ have opposite hypercharge. The coupling to the $SU(2)_L$ gauge bosons is invariant under $SU(2)_L$ by construction, and under $SU(2)_R$ by the cyclic property of the trace. Therefore, if we set $g' = 0$, the custodial $SU(2)$ survives, and implies the three massive gauge bosons must be degenerate. For $g' \neq 0$, the same logic implies that a 3×3 block of the mass matrix must be proportional to the identity, which implies $\rho_0 = 1$.

This logic would not have applied if, e.g. the Higgs field had been an $SU(2)_L$ triplet, so measurements of ρ provide information about the mechanism of electroweak symmetry breaking. On the other hand, it is not necessary to have one $SU(2)_L$ doublet. There can be multiple doublets, or more radically, if electroweak symmetry had been broken solely by the QCD condensate, there would be the custodial symmetry $SU(2)_V$, which implies the same result.

The coupling to $U(1)_Y$ breaks custodial $SU(2)$, which leads to the radiative correction

$$\rho - 1 \supset -\frac{11 G_F m_Z^2 \sin^2\theta_W}{24\sqrt{2}\,\pi^2} \log\frac{m_h^2}{m_Z^2}$$

in the $\overline{\text{MS}}$ scheme. In addition, the Yukawa couplings generally break custodial symmetry. Within each generation of quarks, we have custodial symmetry if the up-type and down-type quarks have the same mass, i.e. if there is isospin symmetry. Therefore, the high mass of the top quark produces a significant loop correction to ρ,

$$\rho - 1 \supset \frac{3G_F}{8\sqrt{2}\,\pi^2}\left(m_t^2 + m_b^2 - 2\frac{m_t^2 m_b^2}{m_t^2 - m_b^2}\log\frac{m_t^2}{m_b^2}\right).$$

Measurements of ρ therefore allowed the huge top quark mass to be predicted before it was discovered. More generally, new physics corrections to the WW, ZZ, and $\gamma\gamma$, and γZ two-point functions (called "oblique" corrections, in contrast to direct modifications of the fermion-boson couplings) are commonly parametrized by the so-called Peskin–Takeuchi parameters S, T, and U.

Note. Suppose that $B - L$ was gauged, and that the corresponding gauge boson was massless. The result is a long-range force which makes baryons repel, leptons repel, and baryons and leptons attract each other. It turns out that the constraints on the gauge coupling g are extremely strong. First, the energy levels of the deuteron would be shifted relative to hydrogen's by order g^2/e^2, placing a constraint $g^2 \lesssim 10^{-7}$ from spectroscopy. Next, for $g^2 \ll e^2$ there is a strong constraint from stellar physics, as otherwise $B - L$ gauge boson emission would dramatically accelerate stellar evolution since such particles could escape more readily than photons. This rules out couplings stronger than $g^2 \lesssim 10^{-20}$.

Even strong constraints come from terrestrial physics. Since the Earth is charge neutral, its $B - L$ charge is roughly its neutron number. This leads to a repulsion which would have destroyed the Earth unless the $B - L$ force is weaker than gravitational, $g^2 \lesssim G_N m_n^2 \sim (m_n/M_{\text{pl}})^2 \sim 10^{-36}$. (We can try to avoid this constraint by supposing the Earth has trapped a compensating number of neutrinos, but it doesn't work; if the residual charge is strong enough to keep neutrinos trapped, it is also strong enough to destroy the Earth.) But even if the Earth is stable, the presence of a $B - L$ force would cause different materials to feel different effective values of $\mathbf{g}$. Precision tests of the equivalence principle therefore bound $g^2 \lesssim 10^{-48}$. On the other hand, the constraints are significantly weaker if the $B - L$ gauge boson has a mass, and hence a finite range.

Note. In the limit of massless neutrinos, the SM has another long-range force: leptons can interact with each other through a loop of neutrinos. In fact, Feynman briefly speculated that this could

explain the gravitational force! However, there is a major immediate obstacle. When a massless gauge boson is exchanged with momentum k, the matrix element is $\mathcal{M} \sim 1/k^2$, which implies $V(r) \sim 1/r$ in the Born approximation. For a neutrino loop, the matrix element is proportional to G_F^2, which implies $\mathcal{M} \sim k^2 G_F^2$ by dimensional analysis; Fourier transforming this gives $V(r) \sim 1/r^5$, which is dramatically different but still technically long-ranged. A more precise calculation gives

$$V(r) = \frac{G_F^2}{4\pi^3 r^5}$$

though at higher orders there is also nontrivial velocity dependence. Unfortunately, the force is so weak that it has never been observed.

Note. All dimension 6 operators, which number over 2,000, are considered in "Standard Model EFT" (SMEFT) analyses. These terms have various signatures, such as CP violation, baryon number violation, and changing the overall rates and high momentum tails of various processes. Choosing to express experimental results as constraints on SMEFT coefficients has some advantages: it is quite general and unambiguous, and can easily be combined between different searches. But it's hard to interpret the results, in terms of specific models of UV physics.

The SMEFT takes the Higgs doublet as a field, while the less popular HEFT uses the physical Higgs, i.e. the SMEFT expands about the electroweak symmetry preserving vacuum, while the HEFT expands about the physical vacuum. The SMEFT is more "straightforward" to work with, but the HEFT is more general, e.g. it can accommodate other Higgs sectors.

Finally, we take a look at some of the unsolved problems of the SM.

- The first problem is that the SM does not account for neutrino masses and mixings, which we've covered above. On astronomical and cosmological scales, the SM does not account for dark matter, which is neutral, colorless, cold, non-baryonic, and massive. It also does not contain enough CP violation to account for the matter/anti-matter asymmetry in our universe.
- The SM has three naturalness problems: the Higgs hierarchy problem, the cosmological constant problem, and the strong CP problem. The first two are more urgent, in the sense that they are also fine-tuning problems; on the other hand, the strong CP problem can't be solved by anthropics, making it arguably more robust.
- There are also a number of problems which might or might not have an explanation.
  - Why is the amount of matter, radiation, and vacuum energy in the universe roughly equal today? These quantities varied by many orders of magnitude in the universe's history.
  - Why are there three fermion families, and why do they display a hierarchical structure in their masses and mixings? There are many candidate theories, but none are compelling enough to earn widespread acceptance.
  - Why are the three gauge couplings all relatively close in size?
  - Why are there four spacetime dimensions, and one time dimension?
  - Why is electric charge quantized? This is not explained by $U(1)_Y$, because we must allow for projective representations, and the universal cover of $U(1)$ is $\mathbb{R}$. It could be explained by a grand unified theory where $U(1)_Y$ is embedded in a larger group.

4.4 Electroweak Decays

Next, we look at some concrete electroweak decay processes. First, we set up the effective field theory for the weak interaction.

- We recall the weak part of the Lagrangian has the form

$$\mathcal{L}_W = -\frac{g}{2\sqrt{2}}(J^\mu W_\mu^+ + J^{\mu\dagger} W_\mu^-) - \frac{g}{2\cos\theta_W} J_n^\mu Z_\mu.$$

  Therefore, expanding the S matrix, we have

$$\begin{aligned} S &= T\exp\left(i\int dx\,\mathcal{L}_W(x)\right) \\ &= 1 - \frac{g^2}{8}\int dx dx'\, J^{\mu\dagger}(x) D^W_{\mu\nu}(x-x') J^\nu(x') + \frac{1}{\cos^2\theta_W} J_n^{\mu\dagger}(x) D^Z_{\mu\nu}(x-x') J_n^\nu(x') + O(g^4). \end{aligned}$$

  where $D^W_{\mu\nu}$ and $D^Z_{\mu\nu}$ are massive vector propagators.

- For the Z boson, the Euler–Lagrange equation is

$$\partial^2 Z_\mu - \partial_\mu \partial_\nu Z^\nu + m_Z^2 Z_\mu = -j_\mu$$

  and taking the divergence of each side gives $m_Z^2 \partial_\mu Z^\mu = -\partial_\mu j^\mu$. Substituting this back into the equation of motion gives

$$(\partial^2 + m_Z^2) Z_\mu = -\left(\eta_{\mu\nu} - \frac{\partial_\mu \partial_\nu}{m_Z^2}\right) j^\nu$$

- The Green's function/propagator satisfies

$$Z_\mu(x) = i\int dy\, D^Z_{\mu\nu}(x-y) j^\nu(y)$$

  and taking a Fourier transform yields the familiar massive vector boson propagator,

$$D^Z_{\mu\nu}(x-y) = \int đp\, e^{-ip(x-y)} \widetilde{D}^Z_{\mu\nu}(p), \quad \widetilde{D}^Z_{\mu\nu}(p) = \frac{i}{p^2 - m_z^2 + i\epsilon}\left(-\eta_{\mu\nu} + \frac{p_\mu p_\nu}{m_Z^2}\right).$$

  The Green's function for the W boson is similar. At low energies, we can approximate

$$\widetilde{D}^{W/Z}_{\mu\nu}(p) \approx \frac{i\eta_{\mu\nu}}{m^2_{W/Z}}, \quad D^{W/Z}_{\mu\nu}(x-y) \approx \frac{i\eta_{\mu\nu}}{m^2_{W/Z}}\delta(x-y)$$

  so that the weak interactions can be described by a four-fermion contact interactions.

- Therefore we get the same S-matrix using the effective weak Lagrangian

$$\mathcal{L}_W^{\text{eff}}(x) = -\frac{G_F}{2}\left(J^{\mu\dagger}(x) J_\mu(x) + \rho J_n^{\mu\dagger}(x) J_{n\mu}(x)\right), \quad \frac{G_F}{\sqrt{2}} = \frac{g^2}{8m_W^2}.$$

  This is indeed an effective theory since the four-fermion operator has dimension 6. Higher-order diagrams would give further contributions, but they are suppressed by more powers of large masses; this is the reason we don't have to include the top quark, as it only appears internally in diagrams where there is already a W boson.

- Why do we include the effects of the W and Z boson, but not that of the Higgs boson? Integrating out the Higgs yields interactions of the form $\overline{f}f\overline{f}'f'$, but they are further suppressed by small Yukawa couplings $m_f m_f'/v^2$. In addition, these terms don't break symmetries the same way the W and Z-mediated interactions do, so when they do contribute to processes, they tend to be swamped by the larger strong or electromagnetic interactions.

Example. The muon's "Michel" decay, $\mu \to e\overline{\nu}_e\nu_\mu$. It occurs via the leptonic charged weak current,

$$J^\rho = \overline{\nu}_e\gamma^\rho(1-\gamma^5)e + \overline{\nu}_\mu\gamma^\rho(1-\gamma^5)\mu + \overline{\nu}_\tau\gamma^\rho(1-\gamma^5)\tau.$$

Since the muon mass $m_\mu = 105\,\mathrm{MeV}$ is much less than $m_W = 80\,\mathrm{GeV}$, we can use the effective theory above, where

$$S - 1 = \int dx\, \mathcal{L}_W^{\mathrm{eff}}(x).$$

Here, the position integration enforces momentum conservation. We will compute the amplitudes $\mathcal{M}$ which have $i\not{\delta}(\sum_i p_i)$ factored out of the matrix element. Factoring out the delta function is equivalent to dropping the position integration, so

$$\begin{aligned}\mathcal{M} &= \langle e^-(k)\overline{\nu}_e(q)\nu_\mu(q')|\mathcal{L}_W^{\mathrm{eff}}(0)|\mu^-(p)\rangle \\ &= -\frac{G_F}{\sqrt{2}}\langle e^-(k)\overline{\nu}_e(q)|\overline{e}\gamma^\rho(1-\gamma^5)\nu_e|0\rangle\langle\nu_\mu(q')|\overline{\nu}_\mu\gamma_\rho(1-\gamma^5)\mu|\mu^-(p)\rangle \\ &= -\frac{G_F}{\sqrt{2}}\overline{u}_e(k)\gamma^\rho(1-\gamma^5)v_{\nu_e}(q)\overline{u}_{\nu_\mu}(q')\gamma_\rho(1-\gamma^5)u_\mu(p)\end{aligned}$$

where we picked up on-shell spinors, with no sign flips, and all phases canceled. We then sum over final spins and average over the initial spin, using $\gamma^{5\dagger} = \gamma^5$, for

$$\frac{1}{2}\sum_{\text{spins}}|\mathcal{M}|^2 = \frac{G_F^2}{4}S_1^{\rho\sigma}S_{2\rho\sigma}$$

where since the neutrinos are massless, the spinor traces are

$$S_1^{\rho\sigma} = \mathrm{tr}\left[(\not{k}+m_e)\gamma^\rho(1-\gamma^5)\not{q}\gamma^\sigma(1-\gamma^5)\right], \quad S_{2\rho\sigma} = \mathrm{tr}\left[\not{q}'\gamma_\rho(1-\gamma^5)(\not{p}+m_\mu)\gamma_\sigma(1-\gamma^5)\right].$$

We simplify the spinor traces using the usual identities, noting that $(1-\gamma^5)^2 = 2(1-\gamma^5)$, for

$$S_1^{\rho\sigma} = 8(k^\rho q^\sigma + k^\sigma q^\rho - (k\cdot q)\eta^{\rho\sigma} - i\epsilon^{\rho\sigma\mu\nu}k_\mu q_\nu), \quad S_{2\rho\sigma} = 8(p_\rho q'_\sigma + p_\sigma q'_\rho - (p\cdot q')\eta_{\rho\sigma} - i\epsilon_{\rho\sigma\mu\nu}q'^\mu p^\nu)$$

and contracting the Levi–Civitas with the identity

$$\epsilon^{\mu\nu\rho\sigma}\epsilon_{\mu\nu\lambda\tau} = -2(\delta^\rho_\lambda\delta^\sigma_\tau - \delta^\rho_\tau\delta^\sigma_\lambda)$$

where the minus sign comes from the determinant of the metric, we find

$$\frac{1}{2}\sum_{\text{spins}}|\mathcal{M}|^2 = 64G_F^2(p\cdot q)(k\cdot q').$$

Finally, we must perform the integral over final state momenta. We have

$$\Gamma = \frac{1}{2m_\mu}\int \frac{đ\mathbf{k}đ\mathbf{q}đ\mathbf{q}'}{8k^0q^0q'^0}\not{\delta}(p-k-q-q')\frac{1}{2}\sum_{\text{spins}}|\mathcal{M}|^2 = \frac{G_F^2}{8\pi^5 m_\mu}\int\frac{d\mathbf{k}d\mathbf{q}d\mathbf{q}'}{k^0q^0q'^0}\delta(p-k-q-q')(p\cdot q)(k\cdot q').$$

To perform this tricky three-body integral it's best to separate out the massless neutrinos,

$$I_{\mu\nu}(Q) = \int \frac{d\mathbf{q}d\mathbf{q}'}{|\mathbf{q}||\mathbf{q}'|}\,\delta(Q-q-q')q_\mu q'_\nu, \quad Q = p-k.$$

Then by Lorentz invariance we must have $I_{\mu\nu}(Q) = aQ_\mu Q_\nu + b\eta_{\mu\nu}Q^2$. Contracting with $\eta^{\mu\nu}$ and $Q^\mu Q^\nu$ and using the delta function to simplify, we find

$$a+4b = \frac{I}{2}, \quad a+b = \frac{I}{4}, \quad I = \int \frac{d\mathbf{q}d\mathbf{q}'}{|\mathbf{q}||\mathbf{q}'|}\,\delta(Q-q-q').$$

The integral I is Lorentz invariant, so we work in the center-of-mass frame $Q = (\sigma, \mathbf{0})$,

$$I = \int \frac{d\mathbf{q}}{|\mathbf{q}|^2}\delta(\sigma - 2|\mathbf{q}|) = 4\pi\int_0^\infty d|\mathbf{q}|\,\delta(\sigma-2|\mathbf{q}|) = 2\pi$$

from which we conclude $a = \pi/3$ and $b = \pi/6$. Then we find

$$\Gamma = \frac{G_F^2}{3m_\mu(2\pi)^4}\int \frac{d\mathbf{k}}{k^0}\left(2p\cdot(p-k)\,k\cdot(p-k) + (p\cdot k)(p-k)^2\right).$$

We work in the frame of the muon and approximate the electron as massless with energy E,

$$p = (m_\mu, 0, 0, 0), \quad k = (E, E, 0, 0),$$

which yields the final expression

$$\Gamma = \frac{2G_F^2 m_\mu}{3(2\pi)^3}\int_0^{m_\mu/2} dE\; E^2(3m_\mu - 4E) = \frac{G_F^2 m_\mu^5}{192\pi^3}$$

where the upper bound is attained when the neutrinos exit in the same direction. The size of this result is substantially smaller than one would get by counting 2π factors, mostly because the final phase space integral happens to give a numeric factor of $1/16$. However, a similar suppression often occurs whenever there is a three-body decay to light particles. Also note that the energy distribution for the electron is monotonic: it is most likely to emerge with the maximum possible energy $m_\mu/2$, while the probability for lower energy is suppressed as E^2.

Note. Helicity suppression. Consider the case where the electron and muon neutrino exit in the z-direction and the electron antineutrino exits in the $-z$-direction. Then

$$|\mathcal{M}|^2 \propto k\cdot q' = \sqrt{m_e^2 + k_z^2}\,q'_z - k_z q'_z$$

which vanishes in the limit $m_e \to 0$. This is because in this limit, chirality coincides with helicity. Since the electron and muon neutrino are left-handed and the electron antineutrino is right-handed, the z components of the spin would sum to $-3/2$, so the decay is forbidden.

There are two ways to think about the effect of an electron mass. We can think of the electron as a Dirac spinor, in which case a left-handed electron does not have definite helicity, so the process is allowed. Alternatively, we can think of the electron as made of two massless Weyl spinors, where chirality and helicity match, and treat the mass as an interaction term that flips the chirality.

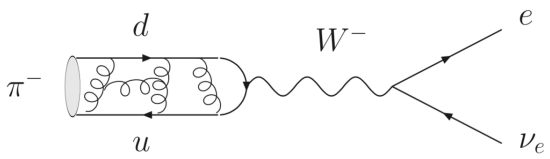
Note. The above decay channel is the only one allowed for the muon, so it provides a precise way to measure the Fermi constant,

$$\tau = \frac{1}{\Gamma} = 2.1970 \times 10^{-6}\,\text{s}, \quad G_F = 1.164 \times 10^{-5}\,\text{GeV}^2.$$

One-loop corrections only affect G_F at the per-million level. A similar calculation can be performed for the τ, which has the two leptonic decay channels $\tau \to e\overline{\nu}_e\nu_\tau$ and $\tau \to \mu\overline{\nu}_\mu\nu_\tau$, as well as decays into hadrons. We can estimate the decay rate in each leptonic channel by simply replacing m_μ by m_τ. Thus, one can measure G_F from these decays, and the results match that found for muons due to lepton universality.

Note. In the 1950s, it was thought that there was only one neutrino, a conclusion supported by lepton universality. However, this would imply the decay $\mu \to e\gamma$ was possible through a loop of a W boson and neutrino, with a branching ratio of order α. The nonobservation of this decay led to the conclusion that there was a separate neutrino for each generation.

Example. Pion decay, $\pi^- \to e\overline{\nu}_e$, has the Feynman diagram shown below.



The d and $\overline{u}$ quarks do not propagate freely, but rather are bound together by nonperturbative dynamics; thus we'll have to parametrize our ignorance using form factors. The decay is again solely through the charged weak current, where the hadronic weak current is

$$J^\mu = V^\mu - A^\mu, \quad V^\mu = \overline{u}\gamma^\mu(V_{ud}d + V_{us}s + V_{ub}b) + \ldots, \quad A^\mu = \overline{u}\gamma^\mu\gamma^5(V_{ud}d + V_{us}s + V_{ub}b) + \ldots$$

where we've defined vector and axial components with definite parity, and the overall "vector minus axial" form is because the charged current only couples to left-handed quark fields. Then

$$\mathcal{M} = \langle e^-(k)\overline{\nu}_e(q)|\mathcal{L}_W^{\text{eff}}(0)|\pi^-(p)\rangle = -\frac{G_F}{\sqrt{2}}\overline{u}_e(k)\gamma_\mu(1-\gamma^5)v_{\nu_e}(q)\langle 0|J^\mu_{\text{had}}|\pi^-(p)\rangle.$$

The QCD vacuum is parity even and the pion is parity odd. Then $\langle 0|V^\mu_{\text{had}}|\pi^-(p)\rangle$ must be an axial vector, but there are no axial vectors it could be equal to, so it must simply be zero. On the other hand, $\langle 0|A^\mu_{\text{had}}|\pi^-(p)\rangle$ must be a vector, so it has to be proportional to p^μ. We cannot compute the matrix element perturbatively, so we absorb it into a single dimensionful parameter called the pion decay constant F_π, so that

$$\langle 0|\overline{u}\gamma^\mu\gamma^5 d|\pi^-(p)\rangle = i\sqrt{2}F_\pi p^\mu.$$

By momentum conservation we have $p = k + q$ and the on-shell spinor identities

$$\overline{u}_e(k)\not{k} = \overline{u}_e(k)m_e, \quad \not{q}v_{\nu_e}(q) = 0$$

so the amplitude simplifies to

$$\mathcal{M} = iG_F F_\pi m_e V_{ud}\overline{u}_e(k)(1-\gamma^5)v_{\nu_e}(q).$$

We expect helicity suppression, since the pion has spin zero and, in the pion's rest frame, the two particles come out back-to-back, giving a total of spin one in the massless limit. This is reflected in the fact that $\mathcal{M} \propto m_e$.

Next, summing over the final spins we have

$$\sum_{\text{spins}} |\mathcal{M}|^2 = 2|G_F F_\pi m_e V_{ud}|^2 \operatorname{tr}\left((\not{k} + m_e)(1-\gamma^5)\not{q}\right) = 8|G_F F_\pi m_e V_{ud}|^2 (k \cdot q).$$

Abbreviating the squared quantity as C, the decay rate in the pion rest frame is

$$\Gamma = \frac{1}{m_\pi} \int \frac{đ\mathbf{k} đ\mathbf{q}}{4k^0 q^0} \, đ(p-k-q) \sum_{\text{spins}} |\mathcal{M}|^2 = \frac{C}{4\pi^2 m_\pi} \int \frac{d\mathbf{k}}{E|\mathbf{k}|} \, \delta(m_\pi - E - |\mathbf{k}|)(E + |\mathbf{k}|)|\mathbf{k}|$$

where we defined $E = k^0$ and integrated over $\mathbf{q}$ so that $q = (|\mathbf{k}|, -\mathbf{k})$. The angular integral is 4π, and integrating the delta function yields

$$\Gamma = \frac{C}{4\pi} m_\pi \left(1 - \frac{m_e^2}{m_\pi^2}\right)^2.$$

We still don't know what F_π is, but we can compute branching ratios, such as

$$\frac{\Gamma(\pi \to e\overline{\nu}_e)}{\Gamma(\pi \to \mu\overline{\nu}_\mu)} = \frac{m_e^2}{m_\mu^2} \left(\frac{m_\pi^2 - m_e^2}{m_\pi^2 - m_\mu^2}\right)^2 = 1.28 \times 10^{-4}.$$

The experimental result is 1.230×10^{-4}, with the difference accounted for by loop diagrams.

Note. Above, we saw another example of helicity suppression, which is a rather common effect in the ultrarelativistic limit. Yet another example occurs in the scattering of spin-polarized electrons and positrons. If we neglect their masses, scattering via a photon is forbidden if the particles have the same helicities (and hence opposite angular momenta), because the resulting product of Poincare irreps has helicity zero, while photons have helicity ± 1. This is an example of how the chiral components decouple in massless QED.

For this argument to work, it's essential that we think in terms of massless Poincare irreps with helicity, rather than massive Poincare irreps with spin, since the combination of two antiparallel spin 1/2 particles does have a spin 1 component (with $L_z = 0$). For this reason, when we account for a nonzero mass, the scattering can happen, but it's helicity suppressed.

An objection one could make for the muon and pion decays is: why can't the decay products come out with orbital angular momentum? Up to a basis change, orbital angular momentum just corresponds to a particular pattern of superposition of directions of the outgoing particles, with a state that looks like $\int d\Omega \, f(\hat{\mathbf{n}})|k\hat{\mathbf{n}}, -k\hat{\mathbf{n}}\rangle$. (Instead of an integral over angles, one could also express this state as a sum over l and m involving spherical harmonics, giving a partial wave expansion. This is more useful for low-energy scattering, where the s-wave typically dominates, and in this case orbital angular momentum expresses itself as a non-s-wave component. But the point is that one doesn't need to do, and indeed can't do, both expansions at once; both bases here are complete.)

Now consider a symmetry argument involving only rotations about the z-axis. Such rotations don't rotate the $|k\hat{\mathbf{z}}, -k\hat{\mathbf{z}}\rangle$ with others, so the argument can be used to show that $f(\hat{\mathbf{z}}) = 0$ without caring about what the other values of $f(\hat{\mathbf{n}})$ are. And then, since $\hat{\mathbf{z}}$ was arbitrary, this shows that $f(\hat{\mathbf{n}}) = 0$ in general. This is the proper way to phrase the arguments we made above. (A slick, but somewhat mysterious way of summarizing this is that "the orbital angular momentum is perpendicular to the linear momentum, so it doesn't affect helicity".)

4.5 CP Violation

Finally, we investigate neutral kaon mixing, which demonstrates CP violation.

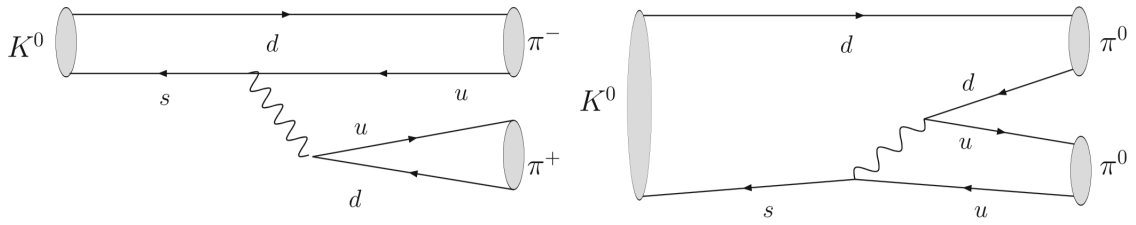
- Kaons are pseudoscalar mesons containing either a strange quark or a strange antiquark. The neutral kaons K^0 and $\overline{K}^0$ have quark content $\overline{s}d$ and $\overline{d}s$ respectively.
- Under $\widehat{C}$, the neutral kaons are mapped to each other. As discussed earlier, $\widehat{C}$ and $\widehat{P}$ have some freedom in phase redefinition, and we may choose these phases so that

$$\widehat{C}\widehat{P}|K^0\rangle = -|\overline{K}^0\rangle, \quad \widehat{C}\widehat{P}|\overline{K}^0\rangle = -|K^0\rangle.$$

  We thus have the CP eigenstates

$$|K^0_\pm\rangle = \frac{|K^0\rangle \mp |\overline{K}^0\rangle}{\sqrt{2}}$$

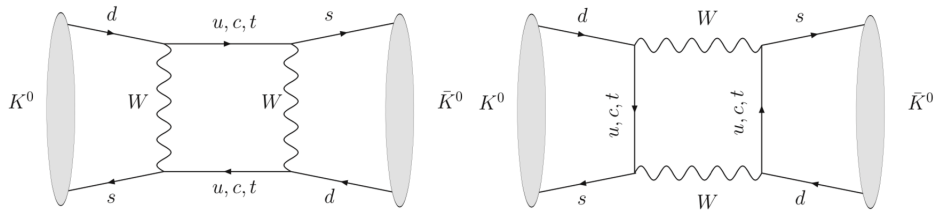
  so that K^0_+ is CP even and K^0_- is CP odd.
- We consider the decays of neutral kaons to two pions, either $\pi^+\pi^-$ or $\pi^0\pi^0$. Since this is a flavor-changing interaction, it is mediated by a weak current, as shown below.



  The pions are all pseudoscalars, and their total angular momentum must be zero since the kaon has spin zero, so parity simply exchanges the pions without any signs, and charge conjugation simply flips the charges. Thus both possible final states $|\pi^+\pi^-\rangle$ and $|\pi^0\pi^0\rangle$ are CP even, and only $|K^0_+\rangle$ can decay to two pions if CP is conserved. The $|K^0_-\rangle$ should have a longer lifetime, being only able to decay to three pions or other final states.
- Experimentally, it is indeed observed that there are two neutral kaons, K^0_S and K^0_L, with a short and long lifetime respectively. We can create a pure sample of K^0_L by waiting for a time much longer than the lifetime of the K^0_S. However, we occasionally observe the K^0_L decay into two pions. Specifically, we have

$$\frac{\langle\pi^+\pi^-|H_W|K^0_L\rangle}{\langle\pi^+\pi^-|H_W|K^0_S\rangle} \approx \frac{\langle\pi^0\pi^0|H_W|K^0_L\rangle}{\langle\pi^0\pi^0|H_W|K^0_S\rangle} \approx 2.2\times10^{-3} \neq 0$$

  indicating that CP is violated.
- To understand this physically, we consider how the K^0 and $\overline{K}^0$ mix. Since the strangeness changes by 2, the mixing must involve two W bosons and hence involves a loop. The most important contributions are from the six box diagrams shown below.



If CP symmetry holds, the amplitude for K^0 to transition to $\overline{K}^0$ is the same as the amplitude to go the other way, and the eigenstates $K^0_{S/L}$ coincide with the CP eigenstates.

- On the other hand, if we have a CP violating phase in the CKM matrix, the amplitude for $K^0 \to \overline{K}^0$ is not the same as that for $\overline{K}^0 \to K^0$. Thus the 'mass basis' is not the same as the 'CP basis', so the K^0_L can decay to two pions. Another way of phrasing this is that CP violating effects produce oscillations between the CP states $|K^0_\pm\rangle$.

Next, we investigate the oscillation quantitatively with a simple phenomenological model.

- We write the mass eigenstates as combinations of the CP eigenstates,

$$|K^0_S\rangle = \frac{|K^0_+\rangle + \epsilon_1 |K^0_-\rangle}{\sqrt{1+|\epsilon_1|^2}}, \quad |K^0_L\rangle = \frac{|K^0_-\rangle + \epsilon_2 |K^0_+\rangle}{\sqrt{1+|\epsilon_2|^2}}.$$

  Since the kaons decay, the mass eigenstates have complex energies. Here, we're making the 'Wigner–Weisskopf' assumption, i.e. we aren't keeping track of the 'environment' state at all, so we guarantee an exponential decay.

- The Hamiltonian is the weak Hamiltonian at next-to-leading order in perturbation theory,

$$H = H_W - \sum_n \frac{H_W |n\rangle\langle n| H_W}{E_n - m_0 - i\epsilon}$$

  and we write its matrix elements as

$$\begin{pmatrix} \langle K^0|H|K^0\rangle & \langle K^0|H|\overline{K}^0\rangle \\ \langle \overline{K}^0|H|K^0\rangle & \langle \overline{K}^0|H|\overline{K}^0\rangle \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}.$$

- Now, we write the CPT operator as $\widehat{\Theta}$. We may choose the phases so that

$$\widehat{T}|K^0\rangle = |K^0\rangle, \quad \widehat{T}|\overline{K}^0\rangle = |\overline{K}^0\rangle$$

  which implies that

$$\widehat{\Theta}|K^0\rangle = -|\overline{K}^0\rangle, \quad \widehat{\Theta}|\overline{K}^0\rangle = -|K^0\rangle.$$

  Next, we note that under a CPT transformation,

$$\widehat{\Theta} H \widehat{\Theta}^{-1} = H^\dagger$$

- The only nontrivial constraint this yields is

$$R_{11} = \langle K^0|(\widehat{\Theta}^{-1}\widehat{\Theta})|H(\widehat{\Theta}^{-1}\widehat{\Theta})|K^0\rangle = \langle \overline{K}^0|H^\dagger|\overline{K}^0\rangle^* = R_{22}$$

  where we picked up a complex conjugation by flipping the direction of action of $\widehat{\Theta}$.

- If we further had T invariance, which would imply CP invariance, then $\widehat{T}H\widehat{T}^{-1} = H^\dagger$, so

$$R_{12} = \langle K^0|\widehat{T}^{-1}\widehat{T}|H\widehat{T}^{-1}\widehat{T}|\overline{K}^0\rangle = \langle \overline{K}^0|H|K^0\rangle = R_{21}$$

  so a difference of R_{12} and R_{21} implies CP violation.

- Finally, assuming for simplicity that $\epsilon_1 = \epsilon_2 = \epsilon$, which turns out to be correct, one can straightforwardly calculate

$$\epsilon = \frac{\sqrt{R_{12}} - \sqrt{R_{21}}}{\sqrt{R_{12}} + \sqrt{R_{21}}}.$$

  The R_{ij} can be computed in perturbation theory, and then ϵ can be related to the branching ratios for $K^0_{S/L}$ decay, giving a quantitative calculation of a CP violating effect.

5 Neutrinos

5.1 Historical Review

Next, we turn to neutrino masses, the leading correction of the SM. We begin with a history of neutrino physics.

- 1914: Chadwick demonstrates the energy of the outgoing electron in β decay has a continuous spectrum, which seems to contradict energy-momentum conservation. (This took almost two decades from the discovery of β radiation, since such measurements were difficult.)
- A gap in progress occurs because of World War I.
- 1920s: there was much confusion around this time. Nuclei were thought to be made of protons and electrons, but this gave the wrong statistics and a much too large magnetic moment. Ignoring these issues, the continuous spectrum could then be explained by assuming violation of energy-momentum conservation, which was justified in a 1931 textbook by Gamow by saying that we already knew such electrons had to behave strangely because of all the *other* problems.
- 1930: Pauli postulates an additional, nearly undetectable light neutral fermion contained in the nucleus, called the neutron ν, that solves all the problems above. This is first presented in absentia by his "dear radioactive ladies and gentlemen" letter.
- 1932: Chadwick discovers the neutron. This is too heavy to be Pauli's postulated particle, so Fermi renames it to the neutrino, because that means "little neutron" in Italian. (The -ino ending was then hijacked for the rest of particle physics to mean a generic fermion, even if they aren't "little".)
- 1934: Fermi introduces a four-fermion theory of weak interactions, allowing calculations. This accounts for beta decay as the process $n \to p + e^- + \nu$. This is actually quite a theoretical advance, because it is the first example of fermion production not in particle-antiparticle pairs.
- 1935: the nucleus is understood as being composed of protons and neutrons, with the neutrinos and electrons being newly created upon decay. Yukawa postulates a nuclear strong force mediated by a "meson" (i.e. pion) to hold the nucleus together, with a Yukawa potential.
- 1937: the meson is "discovered" in cosmic rays, which has the right mass but seems to interact far too weakly with nuclei. A long confusion ensues, until people eventually realize it is a new particle, the muon, which is like a heavy electron. It is initially thought to be an excited state of the electron, but the expected decay $\mu^- \to e^-\gamma$ is not observed through many experiments.
- It turns out that cosmic rays are actually high-energy protons, which produce pions upon impact with atoms in the atmosphere. These pions decay into the muons that we call cosmic rays above; this is the most common decay because of helicity suppression.
- A gap in progress occurs because of World War II.
- 1947: Marshak and Bethe propose the "two-meson hypothesis", where π is produced in cosmic rays but quickly decays to μ. This ridiculous ad-hoc idea is confirmed to be correct; pions are observed in cosmic rays high in the atmosphere.

- 1956: Reines and Cowan observe the neutrino-induced reaction $\overline{\nu} + p \to n + e^+$ in "project poltergeist", with a nuclear reactor as a neutrino source, directly benefiting from World War II technology. (Note that this gap between theory and experiment is already several decades!) Specifically, they observe the gamma rays due to the annihilation of the e^+ and the absorption of the n, and require these two to be roughly coincident to reduce backgrounds.

- 1958: shortly after the Wu experiment (1956), neutrinos are observed by Goldhaber et al. to always have left-handed helicity, which would make sense if they were massless.

- 1962: at this point, the muon neutrino has been theorized, and electron/muon number conservation has been postulated to explain the absence of the decay $\mu^- \to e^-\gamma$. (Actually, this decay can occur due to neutrino masses, but is exceptionally rare in the SM because of the GIM mechanism.) This means that pion decay is actually $\pi^- \to \mu^-\overline{\nu}_\mu$. A beam of muon neutrinos fired at nuclei is then expected to produce muons and not electrons, which is confirmed at Brookhaven in this year.

- 1968: the Homestake experiment detects solar electron neutrinos by the reaction

$$\nu_e + {}^{37}\mathrm{Cl} \to e^- + {}^{37}\mathrm{Ar}$$

  and then filtering out the argon and measuring its decay. It finds 1/3 as much compared to detailed astrophysical calculations based on the proton-proton chain. The discrepancy is called the solar neutrino problem. Later, other experiments, such as Kamiokande, support this result. However, this evidence is not definitive, since solar physics is rather complicated.

- Note that nearly all of the neutrinos produced in the Sun are expected to be electron neutrinos. This is because the Sun is "low-energy" by the standards of particle physics. Neutrinos are hence only produced by charged current interactions, and there is not enough energy to form muons or taus. We also do not expect electron antineutrinos. The electron neutrinos produced have MeV scale energies. By comparison, atmospheric neutrinos from cosmic rays go into the GeV scale.

- The Homestake and related experiments are not sensitive to muon or tau neutrinos, because absorption by a nucleus would have to produce a muon or tau, and there is not enough energy to do so.

- 1957: Pontecorvo and Gribov formulate the theory of neutrino flavor oscillations, which violate electron/muon/tau number. The oscillations require neutrino masses, since massless particles "do not experience time" and hence can't oscillate. Later, Mikheyev, Smirnov, and Wolfenstein refine this into a solution for the solar neutrino problem, which we cover below.

- 1975: the tau is discovered at SLAC, leading to the prediction of the tau neutrino.

- 1970s to 1990s: followups on the Homestake experiment are done. SAGE and GALLEX/GNO use gallium (lower threshold energy) while SNO, Kamioka, and SuperK use oxygen nuclei in water (higher threshold energy, but cheaper), confirming the puzzling result. Some of these are repurposed proton decay experiments motivated by GUTs. Throughout this time, many aren't convinced the solar neutrino problem is a real one, since the experiments are difficult and the nuclear physics of the Sun is complicated.

- 1987: neutrinos from a supernova, SN1987A, are detected. The neutrinos arrive at about the same time as light (actually earlier, since the light is delayed during core collapse), providing a strong upper bound on the neutrino mass.
- 1998: Super-Kamiokande provides definitive evidence for atmospheric neutrino oscillations. These neutrinos are created from cosmic rays by reactions like

$$\pi^+ \to \mu^+ + \nu_\mu \to e^+ + \nu_e + \overline{\nu}_\mu + \nu_\mu.$$

  For a detector on the ground, one expects an equal rate of muon neutrinos coming down and up from the other side of the Earth, by a shell-theorem like argument, assuming the isotropy of high-energy cosmic rays. But Super-Kamiokande found almost exactly half as much going up, which is explained by their oscillation into tau neutrinos.
- 2000: the ν_τ is directly observed by the DONUT experiment at Fermilab with the same strategy as for muon neutrinos, using a tau neutrino beam. This is a very difficult experiment. The discovery paper itself had only 4 events, and to date only about 10 tau events have been directly seen by all experiments combined! However, note that the ν_τ had earlier been observed indirectly from the Z decay width at LEP.
- 2001: the SNO experiment becomes sensitive to all three flavors of solar neutrinos. The experiment uses heavy water, containing deuterons (loosely bound pn bound states). Neutrinos can scatter off the deuteron by a neutral current interaction (same for all three flavors), breaking it apart, and one then measures the produced neutron. SNO finds a total flux in accordance with expectation, decisively confirming that solar neutrinos oscillate.
- 2005: KamLAND uses reactor neutrinos to directly observe neutrino oscillations for anti-electron neutrinos. Varying the distance can be achieved because Japan has over 50 existing nuclear reactors at varying distances from the (stationary) detector. Sociologically, this is because Japan is an island and hence has plentiful water for reactor cooling.
- 2010s: Double Chooz (France), Daya Bay (China), and RENO (South Korea) all find that the parameter θ_{13} in the PMNS matrix is nonzero, using reactor neutrinos. NOνa (Fermilab) and T2K (Japan) do the same with accelerator neutrinos. These experiments do not have the luxury of KamLAND's multiple sources; instead they generally use two detectors, a "near" one and a "far" one, to see how much the neutrino flux decreases.
- Reactor neutrino experiments find an unexpectedly large number of neutrinos at around 5 MeV, which has not been resolved. There is also an outstanding accelerator neutrino anomaly from LSND, which was checked by MiniBooNE. MiniBooNE in turn found yet another anomaly, which has been checked by MicroBooNE, but the interpretation of all three experiments remains unclear. (Historically, neutrino physics has generated a very large number of anomalies.)

5.2 Neutrino Oscillations

Next, we take a closer look at neutrino oscillations.

- For the moment, we assume the neutrinos have Majorana masses and ignore issues of gauge invariance. We write the mass terms as

$$\mathcal{L} \supset -\frac{1}{2}\left(m_{ab}\overline{\nu}_a P_L \nu_b + \text{h.c.}\right)$$

where the ν_a are the neutrino fields. We define the PMNS matrix V to map from the mass basis to the flavor basis. (Note that when one refers to just "the neutrinos", one means the flavor basis. This is in contrast to quarks, where one means the mass basis.)

- Now we count the degrees of freedom, for Majorana masses.
  - The lepton sector has a $U(3)^2$ symmetry, with 6 real parameters and 12 phases.
  - The lepton Yukawa coupling is a complex matrix with 9 real parameters and 9 phases.
  - This leads a theory with 3 real parameters (the charged lepton masses) and the three $U(1)$ lepton number symmetries.
  - When we introduce Majorana masses, these symmetries are broken completely.
  - The matrix m above is complex symmetric, and has 6 real parameters and 6 phases.
  - Hence the masses can be described in terms of 6 real parameters and 3 phases.

  Of these parameters, 3 of the real parameters are just the neutrino masses.

- The rest are in the PMNS matrix, which can be written as

$$V = UK$$

  where U has the same parametrization as the CKM matrix, and K can be chosen to be, e.g., $\mathrm{diag}(e^{i\alpha_1}, e^{i\alpha_2}, 1)$. The matrix K can't be measured by neutrino oscillation experiments.

- Now consider the case where neutrinos have Dirac masses.
  - This requires introducing a set of right-handed neutrino fields, which gives another $U(3)$ symmetry to use.
  - The $U(1)^3 \times U(3)$ symmetry is broken to $U(1)_L$, allowing us to absorb 3 real parameters and 8 phases.
  - The Yukawa coupling between the left-handed and right-handed neutrino fields is again a complex matrix with 9 real parameters and 9 phases.
  - Hence the masses can be described in terms of 6 real parameters and 1 phase.

  Again, 3 of the real parameters are neutrino masses, while the rest are in the PMNS matrix, which in this case can be written in the same form as the CKM matrix.

- While this parameter counting is comprehensive and reliable, it can be simplified if we only care about the PMNS matrix.
  - This is naively a general unitary matrix with 3 real parameters and 6 phases.
  - In the Majorana case, phases can only be removed by rephasing the charged lepton fields (since this rephases the flavor basis), giving 3 remaining phases.
  - In the Dirac case, both the charged lepton and neutrino fields can be rephased, but a uniform phase shift does nothing to the PMNS matrix because of the $U(1)_L$ symmetry. This leaves $6 - (6 - 1) = 1$ phase.

- The matrix U is parametrized by three mixing angles θ_{12}, θ_{23}, and θ_{13}, and a CP-violating phase δ. Currently, all of the mixing angles have been found to be nonzero, though $\theta_{13} \approx 8^\circ$ is smaller than the rest and took much longer to measure, while δ is only nonzero at 2σ.

Next, we turn to neutrino oscillations.

- Neutrinos are typically produced and absorbed in charged-current weak interactions, i.e. in flavor eigenstates $|\nu_a\rangle$. We consider the amplitude

$$\langle \nu_b(\mathbf{x},t)|\nu_a(0,0)\rangle = \langle \nu_b|e^{-iHt+i\mathbf{P}\cdot\mathbf{x}}|\nu_a\rangle$$

  and insert a complete basis of mass eigenstates $|\nu_i\rangle$, for

$$\langle \nu_b(\mathbf{x},t)|\nu_a(0,0)\rangle = \sum_{i,\sigma}\int d\mathbf{k}\, e^{-iE_i(k)t+i\mathbf{k}\cdot\mathbf{x}}\langle\nu_b|\nu_i(\mathbf{k},\sigma)\rangle\langle\nu_i(\mathbf{k},\sigma)|\nu_a\rangle.$$

  The spin part is flavor-independent; for simplicity we take the initial and final states to have spin up. But we also find flavor-dependent phases since the dispersion relations $E_i(k)$ differ.

- We apply the ultrarelativistic approximation,

$$|\mathbf{x}| \approx t, \quad |\mathbf{k}| \approx E - m_i^2/E$$

  which yields the simplification

$$\langle \nu_b(\mathbf{x},t)|\nu_a(0,0)\rangle = e^{i\xi}\sum_i e^{im_i^2L/2E}\langle\nu_b|\nu_i\rangle\langle\nu_i|\nu_a\rangle = e^{i\xi}\sum_i e^{-im_i^2L/2E}V_{bi}V_{ai}^*$$

  where ξ is an unimportant global phase.

- Squaring, we find the probability is

$$P_{\nu_a\to\nu_b}(E,L) = \sum_{ij} e^{-i(m_i^2-m_j^2)L/2E}V_{bi}V_{bj}^*V_{aj}V_{ai}^*$$

  where the overall phases in K have canceled out; neutrino oscillations cannot measure them.

- For concreteness, we can focus on the case of two neutrinos, where

$$P_{\nu_a\to\nu_b}(E,L) \approx \sin^2(2\theta)\sin^2\left(\frac{\Delta m^2 L}{4E}\right).$$

  The length scale of oscillations is

$$\lambda = \frac{2E}{\Delta m^2} = 500\,\mathrm{m}\left(\frac{E}{1\,\mathrm{GeV}}\right)\left(\frac{1\,\mathrm{eV}^2}{\Delta m^2}\right), \quad \Delta m^2 \lesssim 3\times10^{-3}\,\mathrm{eV}^2.$$

  Here, typical atmospheric and accelerator neutrinos have an energy of 1 GeV, so λ is much greater than the thickness of the atmosphere, but much less than the size of the Earth. Reactor neutrinos have lower energies, and can be used to probe smaller mass splittings.

- In the limit of small L, we of course have

$$P_{\nu_a\to\nu_b}(E,L) \approx \sin^2(2\theta)\left(\frac{\Delta m^2 L}{4E}\right)^2.$$

  Accelerator neutrino experiments are in this regime, and there is a tradeoff between having large L and higher probability of oscillation, and smaller L with higher flux. The distances are of order 100 km.

- In the limit of large L, we note that E typically has some range, so the rapidly oscillating random phase averages the second factor to $1/2$, giving

$$P_{\nu_a \to \nu_b}(E, L) \approx \frac{1}{2} \sin^2(2\theta).$$

  Atmospheric neutrino experiments are in this regime for up-going neutrinos. In between, the probability can oscillate.

Next, we consider some further subtleties of neutrino mixing.

Note. Just like the CKM matrix, the PMNS matrix breaks CP and hence T. Indeed, we have

$$P(\nu_a \to \nu_b) \neq P(\nu_b \to \nu_a)$$

and

$$P(\nu_a \to \nu_b) \neq P(\overline{\nu}_a \to \overline{\nu}_b)$$

where antineutrinos have mixing matrix V^*. (Also note that V^* is the matrix that appears in the Lagrangian, because neutrino fields create antineutrinos.) The differences of these probabilities is proportional to the Jarlskog invariant for the PMNS matrix. **(cover in more detail)** However, CPT implies that antineutrinos have the same masses as the corresponding neutrinos, which gives

$$P(\nu_a \to \nu_b) = P(\overline{\nu}_b \to \overline{\nu}_a)$$

in general.

Note. To define the PMNS matrix, we must fix a convention for the mass eigenstates. We let $m_1^2 < m_2^2$ and let m_3^2 be the one far from the other two. However, we don't know if m_3^2 is larger ("normal" hierarchy) or smaller ("inverted" hierarchy). Under this convention, the elements of the PMNS matrix are

$$V = \begin{pmatrix} V_{e1} & V_{e2} & V_{e3} \\ V_{\mu 1} & V_{\mu 2} & V_{\mu 3} \\ V_{\tau 1} & V_{\tau 2} & V_{\tau 3} \end{pmatrix} \sim \begin{pmatrix} 0.8 & 0.4 & 0.1 \\ 0.4 & 0.5 & 0.7 \\ 0.4 & 0.6 & 0.7 \end{pmatrix}$$

where the numbers above are extremely approximate. Assuming the neutrino mass is Dirac, the PMNS matrix has three physical angles and one physical phase, which we defined to as

$$\tan^2 \theta_{12} = \frac{|V_{e2}|^2}{|V_{e1}|^2}, \quad \tan^2 \theta_{23} = \frac{|V_{\mu 3}|^2}{|V_{\tau 3}|^2}, \quad V_{e3} = \sin \theta_{13}\, e^{-i\delta}.$$

Current measurements of δ are still consistent with zero within a few sigma. Our knowledge of θ_{12} comes from solar neutrinos, θ_{23} from atmospheric neutrinos, and θ_{13} from reactor neutrinos.

From this matrix, we see that in the large L limit, we lose roughly half of both initial electron neutrinos and initial muon neutrinos, while "2" neutrinos are composed of each flavor equally. A nice way to remember this is to use the "tribimaximal" form,

$$|V_{ai}|^2 = \begin{pmatrix} 2/3 & 1/3 & 0 \\ 1/6 & 1/3 & 1/2 \\ 1/6 & 1/3 & 1/2 \end{pmatrix}$$

which was used in many earlier models, but is now ruled out, e.g. since θ_{13} is nonzero. This is important, as if it were zero, there would be no CP violation at all.

Note. Why are we allowed to restrict to two neutrino flavors sometimes? First, one oscillation frequency is several times smaller than the others, so for experiments with small lengths (e.g. reactor neutrinos) we can ignore the slow frequency. Second, for other ones (e.g. solar neutrinos) we often only measure electron neutrinos, and here $|V_{e3}|^2$ is quite small.

Note. Neutrino oscillations are a bit puzzling, because if the mass-basis neutrinos have different dispersion relations, then it is impossible for a flavor-basis neutrino to have a definite four-momentum. But the electroweak Feynman diagrams that produce flavor-basis neutrinos impose momentum conservation at every vertex, so the final state in the reaction $e^- + X \to X' + \nu_e$ looks like

$$|\text{definite flavor } X', \nu_e\rangle = \sum_i |\text{definite momentum } X', \nu_i\rangle$$

where each of the states on the right has a different momentum for X'. Tracing out the X', it would appear that we cannot have interference between the neutrino states. But this is no problem, for the same reason that a Stern–Gerlach apparatus doesn't destroy superpositions: the momenta of the X was not well-defined to begin with! In the case of solar neutrinos, even demanding that the X lie in the Sun requires a large enough spread in momentum that the X' states of different momentum almost completely overlap.

However, this raises the possibility that neutrino oscillations can decohere. For instance, this occurs if the distance traveled is great enough that the different components of the neutrino wavepacket stop overlapping. An exhaustive review of the subtleties of neutrino oscillations is given in *Paradoxes of neutrino oscillations*.

Note. If we used the formulas above, we would expect that the Homestake experiment saw 1/2 as many neutrinos as expected, rather than the actual 1/3. (Sometimes the 1/3 is naively explained by saying that there are 3 neutrino flavors, but this is a drastic oversimplification.) The 1/3 results from the MSW effect: while electron neutrinos are created at the center of the Sun, they will be affected by the electrons in the Sun, so that the neutrinos exiting the Sun are *not* electron neutrinos.

The mass eigenstates satisfy a Schrodinger equation in space,

$$i\frac{d}{dL}|\nu_i\rangle = \frac{m_i^2}{2E}|\nu_i\rangle$$

In terms of flavor eigenstates, we have

$$i\frac{d}{dL}|\nu_\beta\rangle = V_{\beta i}\frac{m_i^2}{2E}V_{i\alpha}^\dagger|\nu_\alpha\rangle.$$

Tau neutrinos are not important here, so we restrict to two flavors,

$$i\frac{d}{dL}\begin{pmatrix}|\nu_e\rangle\\|\nu_\mu\rangle\end{pmatrix} = \frac{\Delta m^2}{2E}\begin{pmatrix}\sin^2\theta & \cos\theta\sin\theta\\ \cos\theta\sin\theta & \cos^2\theta\end{pmatrix}\begin{pmatrix}|\nu_e\rangle\\|\nu_\mu\rangle\end{pmatrix}.$$

Now consider the interaction of electrons with electron-neutrinos, which is the leading (i.e. tree-level) interaction in this context. The effective four-fermion interaction is

$$\mathcal{L} \supset 2\sqrt{2}G_F(\overline{\nu}_{eL}\gamma_\mu e_L)(\overline{e}_L\gamma^\mu\nu_{eL}) = -2\sqrt{2}G_F(\overline{\nu}_{eL}\gamma_\mu\nu_{eL})(\overline{e}_L\gamma^\mu e_L)$$

where we used a Fierz identity. In a matter background with electron number density N_e, in the matter rest frame, we may set

$$\langle\overline{e}_L\gamma_\mu e_L\rangle = \delta_{\mu 0}\frac{N_e}{2}$$

where the 1/2 is from the two possible helicities. The effective Lagrangian for electron neutrinos is

$$\mathcal{L} \supset \overline{\nu}_{eL}\partial\!\!\!/\nu_{eL} - iA(\overline{\nu}_{eL}\gamma_0\nu_{eL}), \quad A = \sqrt{2}G_F N_e.$$

The matter term produces an effective potential. To see this, consider the equation of motion

$$(\partial\!\!\!/ - iA\gamma_0)|\nu_e\rangle = 0.$$

Multiplying by $\partial\!\!\!/ - iA\gamma_0$, we have

$$0 = (\partial^2 - 2iA\partial_0 + A^2)|\nu_e\rangle = (E^2 - |\vec{p}|^2 \mp 2AE + A^2)|\nu_e\rangle$$

where the other sign arises for antineutrinos. This gives the dispersion relation

$$E = |\vec{p}| \pm A$$

which shows the matter-induced potential. Hence the Schrodinger equation becomes

$$i\frac{d}{dL}\begin{pmatrix}|\nu_e\rangle \\ |\nu_\mu\rangle\end{pmatrix} = \left[\frac{\Delta m^2}{2E}\begin{pmatrix}\sin^2\theta & \cos\theta\sin\theta \\ \cos\theta\sin\theta & \cos^2\theta\end{pmatrix} + \begin{pmatrix}A & 0 \\ 0 & 0\end{pmatrix}\right]\begin{pmatrix}|\nu_e\rangle \\ |\nu_\mu\rangle\end{pmatrix}.$$

One way to parametrize this matrix is to subtract off a multiple of the identity, giving

$$\begin{pmatrix}A & (\Delta/2)\sin 2\theta \\ (\Delta/2)\sin 2\theta & \Delta\cos 2\theta\end{pmatrix}, \quad \Delta = \frac{\Delta m^2}{2E}$$

which can be written in the original form with a different Δ and θ,

$$P(e\to\mu) = \sin^2(2\theta_M)\sin^2\frac{\Delta_M L}{2}, \quad \Delta_M = \sqrt{(A - \Delta\cos 2\theta)^2 + \Delta^2\sin^2 2\theta}, \quad \Delta_M \sin 2\theta_M = \Delta\sin 2\theta.$$

Note that neutrinos and antineutrinos oscillate differently; this is compatible with CPT because the matter background spontaneously breaks it. Also note that the MSW effect depends on the sign of Δ, and hence can in principle tell between the normal and inverted mass hierarchy.

In the case of the Sun, A is high in the core, so that a produced $|\nu_e\rangle$ is approximately a mass eigenstate. As the neutrino exits, A adiabatically transitions to zero, so the neutrino exits in a mass eigenstate, namely the heavier one $|\nu_2\rangle$ because of avoided level crossing, and afterward do not oscillate. The fraction of electron-neutrinos we see is

$$P_{ee} = |\langle\nu_e|\nu_2\rangle|^2 = \sin^2\theta \approx \frac{1}{3}$$

which is the correct result. So ironically, the first evidence for neutrino oscillations doesn't even involve neutrino oscillations. (Though strictly speaking, the Sun produces neutrinos with a wide range of energies, and this argument only applies to the high-energy ones. For lower-energy neutrinos, measured in experiments after Homestake, this effect is less important.) The MSW effect also causes a "day-night" effect for solar neutrinos, which have to pass through the Earth at night.

Note. Various signs get flipped for antineutrinos, which raises the question: does these results change if neutrinos are Majorana? The answer is actually no, because if so, then what we call "neutrino" and "antineutrino" just stands for left-helicity neutrino and right-helicity neutrino in the lab frame, since that determines how we can detect them. (Strictly speaking, the weak force couples to definite chirality; the mismatch between chirality and helicity gives errors, but they are suppressed by powers of m_ν/E, which is tiny for all neutrinos ever detected.)

5.3 Neutrino Masses

Next, we consider the possibility of sterile neutrinos, first ignoring any gauge structure.

- In general, any fermion that mixes with ordinary neutrinos but does not couple to anything else in the SM is called a sterile neutrino. By definition, sterile neutrinos must have no charge under any gauge group.
- Concretely, suppose we introduce N sterile Majorana neutrino fields. (As mentioned above, this does not lose any generality. Only Lagrangian terms break symmetries, not how we package the fields in them.) The most general Majorana mass terms include mixing terms between the sterile and ordinary neutrinos, with mass matrix

$$\begin{pmatrix} m & \mu \\ \mu^T & M \end{pmatrix}$$

  where m is the mass matrix introduced earlier.
- First, consider the case $\mu \ll m, M$. Then there is negligible mixing between sterile and ordinary neutrinos, and the sterile neutrinos don't do anything at all, though there may be constraints on them from cosmology.
- Next, consider the case of Dirac neutrinos, $m = M = 0$. In this case, we can write the fields as $N - 3$ massless decoupled sterile neutrinos and 3 massive Dirac neutrinos. The result is exactly analogous to the quark fields. Lepton number is conserved, and the phases α_i are all zero.
- It seems that it would be easy to rule out Dirac neutrinos, because ordinary neutrinos would quickly oscillate into sterile neutrinos, which have the opposite chirality, leading to an easily measurable missing probability. Chirality oscillations indeed occur for other fermions, but the the mixing angle between the chirality and mass basis in the Dirac equation goes to zero as the neutrino becomes ultrarelativistic. Since all neutrinos available are ultrarelativistic, the oscillation amplitude is extremely small. This effect is known as helicity suppression.
- On the other hand, if we have light sterile neutrinos, $m \sim \mu \sim M$, the previous argument doesn't apply. These models are indeed tightly constrained by "missing probability".
- Finally, seesaw neutrinos are the case $m \ll \mu \ll M$. The eigenvectors are almost purely sterile and normal, with masses on the order of M and $m + \mu^2/M$. These models are experimentally acceptable, since the sterile neutrinos are too heavy to be produced, and give the right neutrino mass naturally, as we'll see below.
- When it is asked whether neutrinos are Dirac or Majorana, a Dirac neutrino would simply correspond to $m = M = 0$. If there are Majorana mass terms, $U(1)_L$ is violated and neutrinos can annihilate themselves. This isn't forbidden, since $U(1)_L$ is merely an accidental global symmetry of the SM anyway, and is even anomalous.
- However, note that $U(1)_{B-L}$ is also an accidental global symmetry of the SM, which isn't anomalous. In extensions of the SM where $U(1)_{B-L}$ is gauged and not spontaneously broken, we must have $m = M = 0$. Alternatively, we could rule out these terms by just postulating that $U(1)_L$ or $U(1)_{B-L}$ are exact global symmetries.

- It would also be easy to tell the difference between Majorana or Dirac neutrinos if we could detect nonrelativistic neutrinos, which appear in the cosmic neutrino background. This is extremely challenging, since cross sections scale with the neutrino energy; we have never seen any nonrelativistic neutrinos.

Next, we embed the neutrino models above in a gauge invariant formalism.

- The simplest possible sterile neutrinos are a set of three right-handed neutrinos

$$N^i = \nu_R^i = (\nu_{eR}, \nu_{\mu R}, \nu_{\tau R})$$

  which are gauge singlets. Then we can include a Yukawa mass term,

$$\mathcal{L} \supset -\sqrt{2}(\lambda_\nu^{ij} \overline{L}^i \phi^c N^j + \text{h.c.})$$

  where we use ϕ^c to make the hypercharge work out. Integrating out the Higgs, this corresponds to the case of Dirac neutrinos. Such a term is not $SU(2)_L \times U(1)_Y$ invariant, but this is acceptable since the only residual symmetry below the Higgs scale is $U(1)_A$.

- We can also write down a gauge invariant Majorana mass term for the sterile neutrinos. The natural scales for the mass matrices are then

$$m \sim 0 \text{ (gauge invariance)}, \quad \mu \sim M_{\text{EW}}, \quad M \sim \Lambda$$

  where Λ is the SM cutoff. We hence get a seesaw mechanism with neutrino masses Λ and M_{EW}^2/Λ, and the latter is the right mass if Λ is about the GUT scale, a compelling coincidence.

- To have Dirac neutrinos, we need to force M to be small somehow. For example, we could use an exact $B - L$ symmetry. However, μ must also be much smaller than M_{EW}. This is technically natural in the same way that the lightness of the up and down quarks relative to M_{EW} is, but it's a bit unsatisfying because it adds more unexplained flavor structure.

- If we assume that whatever physics produces the neutrino masses is heavy, then we can simply use effective field theory. Here, neutrinos receive mass by the dimension 5 "Weinberg operator",

$$\mathcal{L} \supset -\frac{Y^{ij}}{\Lambda}(L^{iT}\phi^c)C(\phi^{cT}L^j) + \text{h.c.}$$

  where the conjugates ensure gauge invariance. The mass is therefore about M_{EW}^2/Λ, giving a simple reason that the seesaw mechanism, and related mechanisms, work.

- From the effective field theory point of view, neutrino masses are the leading correction to the SM, because they are the only dimension 5 operator we can write down. If we take Λ to be near the GUT scale as inferred from the neutrino masses, then dimension 6 operators are very hard to measure.

- There are many more ways to UV complete the Weinberg operator. Above, we have only considered the "type I seesaw", which introduces a right-handed fermionic singlet. But one can also introduce a scalar weak triplet (type II seesaw) or a fermionic weak triplet (type III seesaw), as any of these can play the role of the intermediate heavy particle.

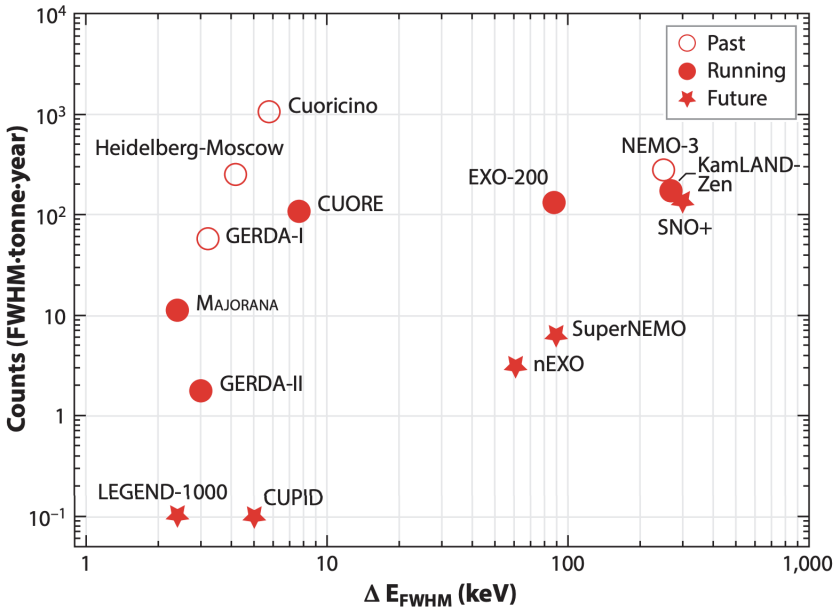
- There are also "radiative" mass generation models where the neutrino mass is only generated at loop level, such as the Zee model and the Ma model. The Ma model is called "scotogenic" ("from darkness") since the mass comes from neutrino interactions with a dark matter candidate.

Note. Why did we discuss neutrino masses in a set of notes on the Standard Model? After all, doesn't the Standard Model require neutrino masses to be zero? This is debatable, because if one reads the term literally, as the *standard* model one uses to describe nature, then it has changed significantly since the advent of "the" Standard Model, and now includes neutrino masses. For some historical discussion, see *The Once and Present Standard Model of Elementary Particle Physics*.

Note. The clearest experimental signature for a Majorana mass term would be neutrinoless double beta decay. Some nuclei cannot decay by beta decay, because the resulting product is heavier, but can decay if two beta decays occur at once, a rare process. In neutrinoless double beta decay, the two neutrinos produced annihilate, which is even rarer due to helicity suppression,

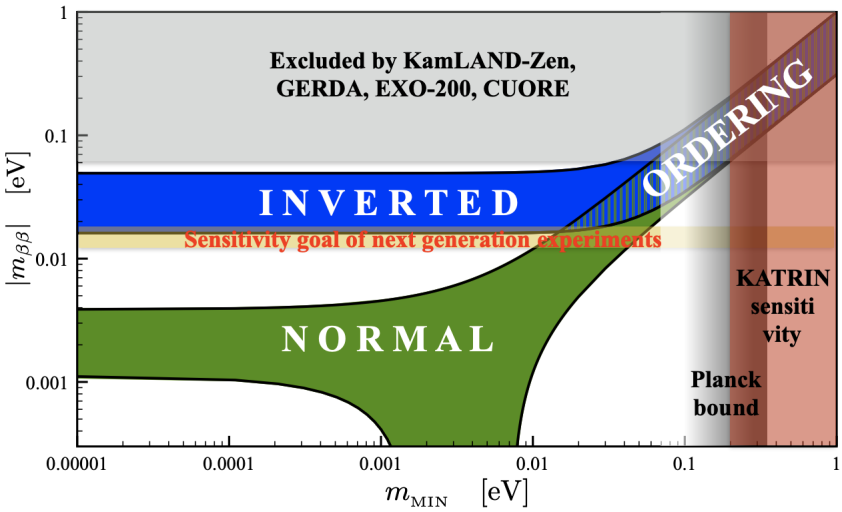
$$\Gamma \sim G_F^4 \frac{m_\nu^2}{E^2} E^9 \sim 10^{-31}\,\text{years}^{-1}$$

assuming $m_\nu \sim 0.01\,\text{eV}$, $E \sim 1\,\text{MeV}$. The process can be identified by an incredibly sharp peak in the energy spectrum of the resulting electrons, which requires very sensitive energy measurements. Relevant experiments are reviewed here and here. Some experiments, with background count rate plotted against energy resolution, are shown below.



There is a tradeoff between large size, at the right of the plot, and good energy resolution, at the bottom. For example, CUORE uses precise bolometers (i.e. calorimeters) in a dilution fridge; it will upgrade to CUPID by adding a light detector to veto most of the background, in the form of degraded alpha particles. On the other extreme, KamLAND-Zen uses about 800 kg of liquid ^{136}Xe dissolved in liquid scintillator, and operates much like direct dark matter detection experiments, though its threshold is at MeV, while WIMP recoils are keV and lower.

Current experiments probe down to $\Gamma \sim 10^{-28}\,\text{years}^{-1}$, while future experiments have the concrete goal of probing the inverted neutrino hierarchy.



Past experiments had the potential to probe heavier, quasi-degenerate neutrinos, but these are in tension with cosmology, so one needs to add epicycles to fix this. Thus, as experiments get more precise, we actually move towards testing the simplest models. Unfortunately, the possibility remains that a cancellation occurs for the normal hierarchy, making the rate very small.

6 Quantum Chromodynamics

6.1 Hadron Production

Before beginning, we consider the running coupling.

- We take the QCD Lagrangian to be

$$\mathcal{L} = -\frac{1}{4}F^{a\mu\nu}F_{a\mu\nu} + \sum_f \overline{q}_f(i\not{D} - m_f)q_f, \quad D_\mu = \partial_\mu + igA^a_\mu T^a$$

  where $T^a = \lambda^a/2$ as usual, and the field is

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - gf^{abc}A^b_\mu A^c_\nu$$

- In general, the one-loop beta function for the coupling is

$$\beta(g) = -\beta_0\frac{g^3}{16\pi^2}, \quad \beta_0 = \frac{11}{3}C_A - \frac{4}{3}\sum_f T_f$$

  where C_A is the quadratic Casimir of the adjoint representation, and T_f is the Dynkin index of the representation for quark flavor f. We see that fermions provide screening, while nontrivial gluon-gluon interactions provide 'antiscreening', which favor asymptotic freedom.

- In the case of QCD, we have the group $SU(3)$, so $C_A = 3$, and all the quarks transform in the fundamental representation where $T_F = 1/2$, so

$$\beta(g) = -\beta_0\frac{g^3}{16\pi^2}, \quad \beta_0 = 11 - \frac{2}{3}N_f$$

  where N_f is the number of flavors. Then the beta function is negative if $N_f < 33/2$. This also holds for QED, where $C_A = 0$ and the beta function is positive for any nonzero N_f.

- For high energies, we have $N_f = 6$, so the beta function is negative. Defining $\alpha_s = g^2/4\pi$,

$$\frac{d\alpha_s}{d\log\mu} = -\frac{\beta_0}{2\pi}\alpha_s^2.$$

  Integrating, we have

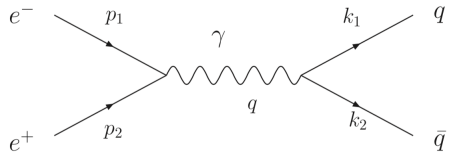
$$\alpha_s(\mu) = \frac{2\pi}{\beta_0}\frac{1}{\log(\mu/\mu_0) + 2\pi/\beta_0\alpha_s(\mu_0)} = \frac{2\pi}{\beta_0\log(\mu/\Lambda_{\text{QCD}})}$$

  where we defined Λ_{QCD} as the scale where the coupling diverges.

- We're implicitly using a mass-independent scheme, so each quark continues to contribute even when μ is much less than its mass. In practice, when we drop below the top quark mass we 'manually' stop its running, matching the coupling and then setting $N_f = 5$, and so on. Doing this yields $\Lambda_{\text{QCD}} \approx 200 - 500\,\text{MeV}$, though the answer depends on the subtraction scheme.

Next, we consider the cross section for $e^+e^- \to$ hadrons.

- First, we consider the process $e^+e^- \to q\bar{q}$, which we treat perturbatively by asymptotic freedom. By simplicity we consider only the tree-level process where the intermediate particle is a virtual photon, as shown below.



- If the quarks have electric charge eQ, the matrix element is

$$\mathcal{M} = (-ie)^2 Q \overline{u}_q(k_1)\gamma^\mu v_q(k_2)\frac{-i\eta_{\mu\nu}}{q^2}\overline{v}_e(p_2)\gamma^\nu u_e(p_1).$$

  Neglecting the quark and electron masses and summing/averaging over spins,

$$\frac{1}{4}\sum_{\text{spins}}|\mathcal{M}|^2 = \frac{e^4Q^2}{4q^4}\operatorname{tr}(\not{k}_1\gamma^\mu\not{k}_2\gamma^\nu)\operatorname{tr}(\not{p}_1\gamma_\mu\not{p}_2\gamma_\nu) = \frac{8e^4Q^2}{q^4}[(p_1\cdot k_1)(p_2\cdot k_2)+(p_2\cdot k_1)(p_1\cdot k_2)].$$

- Next, we work in the center of mass frame,

$$p_1 = (|\mathbf{p}|,\mathbf{p}), \quad k_1 = (|\mathbf{k}|,\mathbf{k}), \quad \mathbf{p}\cdot\mathbf{q} = |\mathbf{p}||\mathbf{q}|\cos\theta, \quad q = (2|\mathbf{p}|,0)$$

  where we have

$$\frac{1}{4}\sum_{\text{spins}}|\mathcal{M}|^2 = e^4Q^2(1+\cos^2\theta).$$

- Next, the basic formula for the differential cross section is

$$d\sigma = \frac{1}{|\mathbf{v}_1-\mathbf{v}_2|}\frac{1}{4p_1^0p_2^0}\frac{d\mathbf{k}_1}{2k_1^0}\frac{d\mathbf{k}_2}{2k_2^0}\not{\delta}(q-k_1-k_2)\,\frac{1}{4}\sum_{\text{spins}}|\mathcal{M}|^2.$$

  Since the particles are massless, $|\mathbf{v}_1-\mathbf{v}_2| = 2$, giving

$$d\sigma = \frac{e^4Q^2}{8\pi^2q^2}\frac{d\mathbf{k}_1}{4|\mathbf{k}_1|^2}\delta(\sqrt{q^2}-2|\mathbf{k}_1|)\,(1+\cos^2\theta).$$

- Writing $d\mathbf{k}_1 = |\mathbf{k}_1|^2 d|\mathbf{k}_1|d\Omega$ and performing the delta function gives

$$\frac{d\sigma}{d\Omega} = \frac{\alpha^2Q^2}{4q^2}(1+\cos^2\theta)$$

  using $\alpha = e^2/4\pi$, and performing the angular integration gives

$$\sigma = \frac{4\pi\alpha^2}{3q^2}Q^2.$$

  This matches the result from our more specific formulas for the cross section.

- Note that the cross section only depends on the identity of the final particles through Q. Then to reduce experimental and theoretical uncertainties, we can test this result by comparing it to the cross section for $e^+e^- \to \mu^+\mu^-$.

Next, we account for the hadronic final states.

- Let $|X\rangle$ denote a generic hadronic final state and let $|0\rangle$ denote the QCD vacuum. Then the amplitude to produce $|X\rangle$ is approximately

$$\mathcal{M}_X = \frac{e^2}{q^2}\langle X|J_h^\mu|0\rangle \overline{v}_e(p_2)\gamma_\mu u_e(p_1), \quad J_h^\mu = \sum_f Q_f \overline{q}_f \gamma^\mu q_f$$

  where J_h^μ is the hadronic electric current, and we are essentially assuming that the process goes as $e^+e^- \to \gamma^* \to q\overline{q} \to$ hadrons with a clean separation between the steps. Summing over all final states gives

$$\sigma = \frac{1}{8p_1^0 p_2^0}\sum_X \frac{1}{4}\sum_{\text{spins},\, p_X} \not{\delta}(q - p_X)|\mathcal{M}_X|^2$$

  where the sum over p_X includes the appropriate Lorentz invariant phase space factors.

- To simplify this, we introduce the hadronic spectral density as we did for a scalar field,

$$\rho_h^{\mu\nu}(q) = (2\pi)^3 \sum_{X,p_X} \delta(q - p_X)\langle 0|J_h^\mu|X\rangle\langle X|J_h^\nu|0\rangle.$$

  By Lorentz invariance, it is proportional to a linear combination of $g^{\mu\nu}$ and $q^\mu q^\nu$. The Ward identity gives $q_\mu \rho^{\mu\nu} = q_\nu \rho^{\mu\nu} = 0$, so

$$\rho_h^{\mu\nu}(q) = (-\eta^{\mu\nu}q^2 + q^\mu q^\nu)\theta(q^0)\rho_h(q^2)$$

  where the theta function exists because the $|X\rangle$ states have positive energy.

- The cross section in the center of mass frame thus simplifies to

$$\sigma = \frac{16\pi^3\alpha^2}{q^2}\rho_h(q^2).$$

  In general, $\rho_h(q^2)$ is a complicated nonperturbative function.

- In order to make progress, we essentially neglect hadronization entirely, writing

$$\sum_{X \in \text{hadrons}} |X\rangle\langle X| = \sum_{Y \in q,\overline{q},g\,\text{states}} |Y\rangle\langle Y|.$$

  Switching from a hadron-level to quark-level description of the process is called quark-hadron duality. Using this assumption, the computation is essentially identical to our earlier computation for $q\overline{q}$ final states.

- Concretely, the spectral density is now

$$\rho_h^{\mu\nu}(q^2) = N_c \sum_f Q_f^2 \int \frac{đ\mathbf{k}_1 đ\mathbf{k}_2}{4k_1^0 k_2^0}(2\pi)^3 \delta(q - k_1 - k_2)\,\text{tr}((\not{k}_1 + m_f)\gamma^\mu(\not{k}_2 - m_f)\gamma^\nu)$$

  where the final state are on-shell, $k_1^2 = k_2^2 = m_f^2$. Unlike our previous computation, we maintain the masses of the quarks. We know the integral must take the form

$$I^{\mu\nu} = Aq^\mu q^\nu + B\eta^{\mu\nu}$$

where A and B are found by contracting both sides with $\eta_{\mu\nu}$ and $q_\mu q_\nu$. We thus find

$$\rho_h(q^2) = \frac{N_c}{12\pi^2}\sum_f Q_f^2\theta(q^2-4m_f^2)\left(1-\frac{4m_f^2}{q^2}\right)^{1/2}\frac{q^2+2m_f^2}{q^2}.$$

- Neglecting the specific dependence on the quark masses, we have

$$\sigma = N_c\frac{4\pi\alpha^2}{3q^2}\sum_f Q_f^2$$

where the sum is over all quarks light enough to be produced; we thus expect a series of plateaus between jumps. Experimental results confirm that $N_c = 3$ and display the same plateaus, with extra resonances throughout. There is a resonance between each plateau, corresponding to the lightest meson containing the new quark that can be produced, e.g. the J/ψ for the charm quark. The result is good for $\sqrt{s} \in [2, 20]\,\text{GeV}$. At high energies, we run into the broad Z pole, while at low energies, α_s is large.

Next, we discuss jets and higher-order corrections.

- At next-to-leading order, we must account for a gluon loop on the $q\overline{q}\gamma$ vertex. The loop is UV finite after renormalization but IR divergent, giving a divergent negative contribution to the cross section. This cancels with the IR divergences in the tree-level cross section for $e^+e^- \to q\overline{q}g$. Thus the total cross section for

$$e^+e^- \to q\overline{q} + \text{possible soft gluons}$$

is finite, and the result is that the total cross-section is multiplied by $1 + \alpha/\pi$. Physically, the q and $\overline{q}$ are seen as jets, so we've computed the differential cross-section for jet production.

- Intuitively, the intermediate photon is very far off-shell, decaying in time $1/\sqrt{q^2}$ by the energy-time uncertainty principle. The emission of soft gluons takes place over a much longer timescale, so it can't retroactively change how the photon decayed. Thus the IR divergences simply account for how the hard quarks are "dressed" after their production and cannot affect the total rate, so they must cancel. Indeed, the corrections in the next-to-leading order cross section come from kinematic regions where the virtual gluon is hard.

- The formal proof that IR divergences cancel is rather difficult. For QED, the result is the Bloch–Nordsieck theorem, while for the general non-abelian case it is the KLN theorem.

- Similarly, consider the tree-level differential cross-section for $e^+e^- \to q\overline{q}g$. By the above considerations, when the gluon is soft, we see two jets, not three. If we restrict to regions where gluon is sufficiently hard, we can trust the result, giving a QCD prediction for the distribution of three-jet events.

- A related question is what scale to choose for the running coupling α_s. Intuitively, it should be set at the scale of the momenta in the question, i.e. $\sqrt{s}$ for the dijets. However, for more complicated processes there will be multiple invariant momenta; for the three-jet event, one might choose the transverse momentum of the gluon.

- Since the running coupling takes the form

$$\alpha_s(\mu) \sim \frac{1}{\log(\mu/\Lambda)}$$

modifying the scale μ by a factor of 2 changes α_s by $O(\alpha_s^2)$. Thus ambiguities in the scale can be resolved by computing to the next order.

Finally, we take a closer look at the spectral density.

- We define the two-point function

$$\Pi_h^{\mu\nu}(x,y) = i\langle 0|TJ^\mu(x)J^\mu(y)|0\rangle.$$

Since J^μ is the hadronic electric current, we have the QED Ward identity is $q_\mu \Pi_h^{\mu\nu} = 0$, so

$$\Pi_h^{\mu\nu}(q) = (-\eta^{\mu\nu}q^2 + q^\mu q^\nu)\Pi_h(q^2).$$

Intuitively, since the hadronic current couples as $J^\mu A_\mu$, the two-point function is essentially the set of hadronic loop corrections to the amputated photon propagator.

- The analogue of the Kallen–Lehmann spectral representation for vectors gives

$$\Pi_h(q^2) = \int_0^\infty ds\, \frac{\rho_h(s)}{s - q^2 - i\epsilon}.$$

Then $\Pi_h(q^2)$, as a function of complex q^2, gets a branch cut starting at the masses of the lightest hadrons.

- Note that we can compute $\Pi_h(q^2)$ at large spacelike momenta, $-q^2 \gg 1$, where perturbation theory holds and we are far from the nonanalyticities. We can then analytically continue to large timelike momenta, which are relevant for hard scatterings.

- We can then use this information to compute $\rho_h(q^2)$ and thereby make experimental predictions. Note that we may invert the formula above for

$$\rho_h(q^2) = \lim_{\delta\to 0} \frac{\Pi_h(q^2 + i\delta) - \Pi_h(q^2 - i\delta)}{2\pi i}.$$

This can be computed by integrating the derivative of $\Pi_h(z)$ along any contour connecting the two points. In particular, we can take a large circle of radius q^2. We won't go into much more detail here, but the idea of performing QCD computations by taking advantage of analyticity is related to S-matrix theory and dispersion relations, and leads to "sum rules".

6.2 Deep Inelastic Scattering

First, we review the basics of the deep inelastic scattering process.

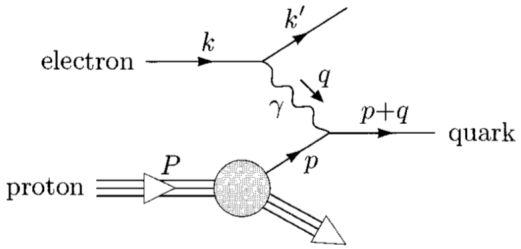
- Historically, the first hint that the strong interaction was asymptotically free came from hadron-hadron scattering experiments. In these experiments, the hadrons were shattered into many constituents, but most of them had low transverse momentum, indicating that the components of the hadrons were loosely bound and could not absorb a large momentum.

- We need to talk about transverse momentum because, while the lab frame coincides with the CM frame of the two protons, it generally doesn't coincide with the CM frame of two proton constituents colliding, which may have a longitudinal boost. Another way to quantify the momentum transfer q is via its square q^2, as if q^2 is large and spacelike, the components of $\mathbf{q}$ must be large in any frame.

- In the 1960s, deep inelastic scattering experiments involving electrons and protons indicated that the proton was made of a small number of pointlike constituents, called partons. The physical picture is that the hard scattering involved a photon exchange between one electron and one parton, while subsequent small-q^2 exchanges between the struck parton and the others produced jets, as we've seen above.

- Just as in Newtonian mechanics, elastic scattering refers to a scattering event where kinetic energy is conserved. In deep inelastic scattering, the proton instead absorbs energy, shattering into many pieces.

- Specifically, consider the following process.



From our earlier work, we know that

$$\sum_{\text{spins}} |\mathcal{M}|^2 \sim e^4 Q_i^2 \frac{s^2+u^2}{t^2}$$

where s, t, and u are the Mandelstam variables for the electron-quark collision.

- Approximating the electron and quark as massless, we find

$$\frac{d\sigma}{dt} \sim \frac{\alpha^2 Q_i^2}{s^2} \frac{s^2+(s+t)^2}{t^2}.$$

Note that $t = q^2$. Since the momentum transfer is spacelike, we define $Q^2 = -q^2$ for convenience.

- Suppose the parton carries a fraction ξ of the proton's momentum, $p = \xi P$. Then

$$s = (p+k)^2 = 2p \cdot k = 2\xi P \cdot k = \xi s'$$

where s' is a Mandelstam variable for the electron-proton collision. Since the electron-parton scattering is elastic,

$$0 \approx (p+q)^2 = 2p \cdot q + q^2 = 2\xi P \cdot q - Q^2$$

so we may measure ξ from observations of the electron alone,

$$\xi = x \equiv \frac{Q^2}{2P \cdot q}.$$

- We let $f_i(x)$ be a parton distribution function, denoting the probability that the constituent i carries longitudinal momentum fraction x. Combining our results,

$$\frac{d^2\sigma}{dxdQ^2} \sim \sum_i f_i(x) Q_i^2 \frac{\alpha^2}{Q^4} \left(1 + \left(1 - \frac{Q^2}{xs'}\right)^2\right).$$

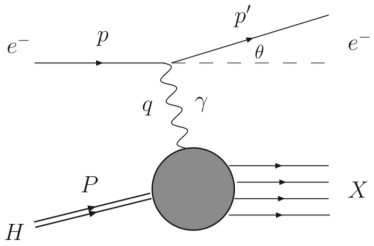
  The only part of this cross section that depends on the strong interaction is $f_i(x)$, while everything else is just from the QED amplitude and the phase space kinematics. Dividing the cross section by these extra factors gives a cross section independent of Q^2, a prediction known as Bjorken scaling, validated to 10% accuracy for $Q \gtrsim 1\,\mathrm{GeV}$.

- Physically, Bjorken scaling means that the proton appears the same to an electromagnetic probe, no matter how hard the proton is struck. This is sensible, because for high Q, the scattering process is much faster than the internal dynamics of the proton.

- On the other hand, Bjorken scaling should be corrected by emission of high-momentum partons; this remains possible at arbitrarily high energies as the strong coupling only decays to zero logarithmically. Thus the parton distribution functions depend logarithmically on Q^2 and their RG evolution equations are called the Altarelli–Parisi or DGLAP equations.

Next, we turn to a quantitative analysis of the cross section for deep inelastic scattering.

- We consider a scattering process $H + e^- \to X + e^-$, where X stands for a hadronic final state.



  Applying ordinary QED to the interaction, we have

$$\mathcal{M} = (-ie)^2 \overline{u}_e(p') \gamma^\mu u_e(p) \frac{-i\eta_{\mu\nu}}{q^2} \langle X | J_h^\nu | H(P) \rangle.$$

- Working in the hadron rest frame, we have

$$d\sigma = \frac{1}{4EM} \frac{đ\mathbf{p}'}{2p'^0} \sum_{X,p_X} \text{ð}(q + P - p_X) \frac{1}{2} \sum_{\text{spins}} |\mathcal{M}|^2$$

  where the sum over p_X includes the Lorentz invariant phase space factors for a variable number of final state particles, we average over the initial spin of the electron, assuming the hadron H is spinless, and M is the mass of H.

- The squared matrix element takes the form

$$\frac{1}{2} \sum_{\text{spins}} |\mathcal{M}|^2 = \frac{e^4}{2q^4} L_{\mu\nu} \langle H(P) | J_h^\mu | X \rangle \langle X | J_h^\nu | H(P) \rangle.$$

Treating the electron as massless, we have

$$L_{\mu\nu} = \sum_{\text{spins}} \overline{u}(p)\gamma_\mu u(p')\overline{u}(p')\gamma_\nu u(p) = \text{tr}(\not{p}\gamma_\mu \not{p}'\gamma_\nu) = 4(p_\mu p'_\nu + p'_\mu p_\nu - \eta_{\mu\nu} p\cdot p').$$

- Next, we define

$$W_H^{\mu\nu}(q,P) = \frac{1}{4\pi}\sum_X \not{\phi}(q+P-p_X)\langle H(P)|J_h^\mu|X\rangle\langle X|J_h^\nu|H(P)\rangle$$

so the differential cross section is

$$E'\frac{d\sigma}{d\mathbf{p}'} = \frac{1}{32\pi^2 EM}\frac{e^4}{q^4}L_{\mu\nu}W_H^{\mu\nu}.$$

- Since $W_H^{\mu\nu}$ is contracted with $L_{\mu\nu}$, we can take it to be symmetric. By current conservation, $q^\mu L_{\mu\nu} = 0$, which means we can choose $q_\mu W_H^{\mu\nu} = 0$. Then we have two form factors,

$$W_H^{\mu\nu} = \left(-\eta^{\mu\nu} + \frac{q^\mu q^\nu}{q^2}\right)W_1 + \left(P^\mu - \frac{P\cdot q}{q^2}q^\mu\right)\left(P^\nu - \frac{P\cdot q}{q^2}q^\nu\right)W_2$$

where W_1 and W_2 are Lorentz scalars that only depend on

$$Q^2 \equiv -q^2 = 2p\cdot p', \quad \nu = P\cdot q.$$

Using $q^\mu L_{\mu\nu} = 0$ we find

$$L_{\mu\nu}W_H^{\mu\nu} = 8p\cdot p'\,W_1 + 4(2(p\cdot P)(p'\cdot P) - M^2\,p\cdot p')W_2 = 4Q^2W_1 + 2M^2(4EE' - Q^2)W_2$$

where E and E' are the initial and final electron energies.

- Introducing the dimensionless variables

$$x = \frac{Q^2}{2\nu}, \quad y = \frac{P\cdot q}{P\cdot p} = \frac{\nu}{ME}$$

where $x, y \in [0,1]$ and taking the high-energy limit,

$$L_{\mu\nu}W_H^{\mu\nu} \approx 8EM\left(xyW_1 + \frac{1-y}{y}\nu W_2\right)$$

where we neglected the Q^2W_2 term since it is a factor of M/E smaller.

- Next, the momentum differential may be rewritten as

$$d\mathbf{p}' = 2\pi E'^2\,d(\cos\theta)dE' = \pi E'\,dQ^2dy = 2\pi E'\,dxdy.$$

Therefore, the differential cross section is

$$\frac{d\sigma}{dx\,dy} = \frac{4\pi\alpha^2}{Q^4}2ME\left(xy^2F_1(x,Q^2) + (1-y)F_2(x,Q^2)\right)$$

where we have defined the dimensionless structure functions $F_1 = W_1$ and $F_2 = \nu W_2$.

We can further parametrize the structure functions using light-cone variables.

- Given an arbitrary four-vector V^μ we define

$$V^\pm = V^0 \pm V^3, \quad \mathbf{V}_\perp = (V^1, V^2)$$

so the inner product is

$$V \cdot U = \frac{1}{2}(V^+U^- + V^-U^+) - \mathbf{V}_\perp \cdot \mathbf{U}_\perp.$$

- Again working in the rest frame of the hadron, we choose the photon momentum to be along $\hat{e}_3$, so $\mathbf{P}_\perp = \mathbf{q}_\perp = 0$. Then

$$Q^2 = -q^+q^-, \quad \nu = \frac{1}{2}(q^+P^- + q^-P^+).$$

In the deep inelastic limit, we take $q^- \to \infty$ with $q^+ \sim P^+$, giving

$$x \sim -\frac{q^+}{P^+}, \quad \nu \sim \frac{q^-P^+}{2}.$$

- Also, in this frame we have

$$W_H^{+-}(q,P) = -W_1 + \left(P - \frac{P\cdot q}{q^2}q\right)^2 W_2 = -W_1 + \left(M^2 + \frac{\nu^2}{Q^2}\right)W_2 \equiv F_L(x,Q^2).$$

In the deep inelastic limit,

$$F_L(x,Q^2) \approx -F_1(x,Q^2) + \frac{1}{2x}F_2(x,Q^2).$$

We also have the identities

$$W_H^{++}(q,P) = \frac{(q^+)^2}{Q^2}F_L(x,Q^2), \quad W_H^{--}(q,P) = \frac{(q^-)^2}{Q^2}F_L(x,Q^2).$$

Thus everything can be parametrized in terms of the longitudinal structure function F_L.

We can further simplify by applying the parton model, writing the structure functions in terms of parton distribution functions, but we won't go into the details here.

6.3 Chiral Symmetry

We now introduce the chiral symmetry of QCD, leading up to chiral perturbation theory.

- We consider QCD with N flavors of quarks,

$$\mathcal{L} = -\frac{1}{4}G_a^{\mu\nu}G_{a\mu\nu} + \sum_i \overline{q}_i(i\not{D} - m_f)q_i.$$

We suppress color indices for the quarks throughout, contracting them implicitly.

- For massless quarks, we have the symmetry group

$$G = U(1)_L \times U(1)_R \times SU(N)_L \times SU(N)_R$$

from unitary rotations of the left-chiral and right-chiral quarks. This group may be generated by a combination of vector symmetries, which rotate them the same way, and axial symmetries, which rotate them oppositely.

- The $U(1)_V$ and $U(1)_A$ symmetries are

$$q \to \exp(-i\theta)q, \quad q \to \exp(-i\theta\gamma_5)q$$

where q is a vector of the quark fields, with conserved currents

$$V_\mu = \overline{q}\gamma_\mu q, \quad A_\mu = \overline{q}\gamma_\mu\gamma_5 Q.$$

The $U(1)_V$ symmetry corresponds to baryon number, while the $U(1)_A$ symmetry is anomalous. Since $U(1)$ is abelian, these symmetries do not yield any degeneracies.

- The vector and axial $SU(N)$ symmetries are

$$q \to \exp(-i\alpha^a T^a)q, \quad q \to \exp(-i\alpha^a T^a\gamma_5)q$$

respectively. The corresponding conserved currents are

$$V^a_\mu = \overline{q}\gamma_\mu T^a q, \quad A^a_\mu = \overline{q}\gamma_\mu\gamma_5 T^a q.$$

The total axial charge is parity-odd, since γ_0 and γ_5 anticommute, and thus it should yield degeneracies between hadrons with opposite parities. But no such degeneracies are observed.

- Sometimes we will be sloppy and write $SU(N)_L \times SU(N)_R \cong SU(N)_V \times SU(N)_A$. This isn't correct, because composing axial transformations instead yields all of $SU(N)_L \times SU(N)_R$. We can define $SU(N)_A$ as the coset space $SU(N)_L \times SU(N)_R / SU(N)_V$, which is a manifold that is topologically $SU(N)$, but it does not have a group structure. Similarly, the set of axial generators $\mathfrak{su}(N)_A$ is well-defined, but it is not an algebra, because the bracket isn't closed. (However, this doesn't change the fact that the generators correspond to conserved currents, as long as each individual one indeed generates a symmetry.)

- Also note that a set of N uncharged quarks has a much larger symmetry group, since we can separate the two chiral components of each Dirac field,

$$\mathcal{L} = i\overline{q_L}\partial\!\!\!/ q_L + i\overline{q_R}\partial\!\!\!/ q_R.$$

Using charge conjugation, we can rewrite this in terms of $2N$ Weyl fields,

$$\mathcal{L} = i\overline{\Psi}\partial\!\!\!/\Psi, \quad \Psi = \begin{pmatrix} \psi_L \\ (\psi_c)_L \end{pmatrix}, \quad \psi_c = C\psi^*,$$

which has an $SU(2N) \times U(1)$ symmetry. This is a perfectly legitimate symmetry, but it is hidden when working with Dirac fields because in that case it would be partly antilinear. It is not useful in QCD, because the extra symmetries are broken by the coupling to the gauge field.

- Now we consider quark masses, which break further symmetries. Mass terms have the form

$$\mathcal{L} \supset \overline{q_L} M q_R + \text{h.c.}$$

so that under a chiral transformation,

$$M \to L^\dagger M R.$$

Hence a chiral field redefinition can be used to make M real and diagonal.

- Mass terms always break axial symmetry. If all the masses are different, the symmetry is broken down to $U(1)^N$, the individual quark numbers. If the masses are the same, the full vector symmetry is preserved. (The different quark electric charges also break symmetries, but we ignore these because electromagnetism is weak.)

- In QCD, only the up, down, and strange quarks are reasonably light. If we consider only the up and down, the small mass difference means $SU(2)_V$ is a very good symmetry; it is isospin. Since their absolute masses are small compared to Λ_{QCD}, $SU(2)_A$ should be almost as good, but this is not observed.

To fix this problem, we introduce the quark condensate, starting with two quark flavors.

- We postulate the QCD vacuum has a "quark condensate" analogous with the condensate of Cooper pairs in a superconductor,

$$\langle \Omega | \overline{q}_{Ri} q_{Lj} | \Omega \rangle = -v^3 \delta_{ij}, \quad v \approx 250\,\text{GeV}$$

where the minus sign ensures the vacuum energy is lowered. It isn't known exactly how to show the right-hand side is proportional to δ_{ij}, but we know this must be the case, or else $SU(2)_V$ would be badly broken.

- The vacuum is only invariant under $U(1)_V$ and $SU(2)_V$, so we expect pseudo-Nambu–Goldstone bosons (PNGBs) from the broken symmetries, $U(1)_A$ and "$SU(2)_A$". This hypothesis was once called "partially conserved axial current" (PCAC).

- The formation of the quark condensate occurs in a phase transition, which occurs approximately concurrently with the confinement transition for QCD, since Λ_{QCD} is the only relevant mass scale in the theory. However, in supersymmetric gauge theories only confinement occurs.

- The PNGBs corresponding to $SU(2)_A$ are the pions, which are indeed far lighter than any other hadron. They are generated by the axial current,

$$\langle \Omega | A^a_\mu(x) | \pi^b(q) \rangle = i f_\pi q_\mu \delta^{ab} e^{-iqx}, \quad f_\pi \approx 92\,\text{MeV}.$$

However, there is no PNGB observed for $U(1)_A$, and this was called the $U(1)_A$ problem.

- Because of the axial anomaly, the $U(1)_A$ current obeys

$$\partial_\mu J^\mu_5 = \frac{\alpha_s}{4\pi} G^{\mu\nu}_a \tilde{G}_{a\mu\nu}, \quad J^\mu_5 = \frac{1}{2} \overline{Q} \gamma^\mu \gamma_5 Q.$$

However, this doesn't solve the $U(1)_A$ problem alone because $G\tilde{G}$ is a total derivative, so one can define a modified (approximately) conserved axial current, which again requires a PNGB. The real resolution requires accounting for QCD instantons, as covered in the notes on Quantum Field Theory.

- More generally, we may extend to three flavors, where the PNGBs corresponding to $SU(3)_A$ are the three pions, the four kaons, and the η. The particle that should be the PNGB for $U(1)_A$ is the η', which is much heavier; this is the $U(1)_A$ problem again. Explicitly, the masses are

$$\pi^\pm, \pi^0 \colon 140\,\text{MeV}, \quad K^\pm, K^0, \overline{K}^0 \colon 500\,\text{MeV}, \quad \eta \colon 550\,\text{MeV}, \quad \eta' \colon 950\,\text{MeV}.$$

  These mesons are pseudoscalars because they are the Goldstone bosons related to the breaking of axial symmetry, and axial symmetries pick up an extra sign under parity.

- Note that for two quark flavors, the $SU(2)_L$ symmetry is simply weak isospin, which uncoincidentally is also called $SU(2)_L$. Hence the quark condensate would break electroweak symmetry if the electroweak phase transition hadn't done it already.

Example. Alternative scenarios of symmetry breaking. First suppose the color group was $SO(3)$ and the quarks transformed in the 3. Since this representation is real, we may write the theory in terms of $2N$ Weyl spinors $\chi_{\alpha i}$ transforming in the 3. The symmetry group is $SU(2N)$, and the $U(1)$ factor is anomalous as before. The condensate that preserves the most symmetry is

$$\langle \Omega | \chi_{\alpha i} \chi_{\alpha j} | \Omega \rangle = -v^3 \delta_{ij}$$

and the remaining symmetry group is $SO(2N)$, by redefinitions of the form

$$\chi_{\alpha i} \to O_{ij} \chi_{\alpha j}.$$

Next, suppose that the color group was $SU(2)$ and the quarks transformed in the 2. By the same reasoning, the symmetry group is $SU(2N)$. The general form of the condensate is

$$\langle \Omega | \epsilon^{\alpha\beta} \chi_{\alpha i} \chi_{\alpha j} | \Omega \rangle = -v^3 \eta_{ij}$$

where the $\epsilon^{\alpha\beta}$ factor is necessary to get a color singlet, which forces η_{ij} to be antisymmetric. Then the most symmetric condensate is one where $\eta^2 = -I$, preserving the symmetry $Sp(2N)$.

Example. Consider N real scalar fields that transform in a representation R. If R is real, the flavor symmetry is clearly $SO(N)$. If R were complex, there would have to be another N real scalar fields transforming in $\overline{R}$, because the overall representation must be real; they are collectively equivalent to N complex scalar fields transforming in R with flavor symmetry $SU(N)$. If R is pseudoreal, it turns out the flavor symmetry is not $SO(N)$, but $Sp(2N)$.

6.4 Chiral Perturbation Theory

Putting aside the $U(1)_A$ problem, we can be more quantitative using chiral perturbation theory, historically one of the most sophisticated examples of effective field theory, which describes the low-energy dynamics of QCD. We begin by ignoring quark masses.

- In chiral perturbation theory, we allow the quark condensate to vary,

$$\delta_{ij} \to U_{ij}(x), \quad U(x) = \exp\left(\frac{2i\pi^a(x) T^a}{f_\pi}\right)$$

  so the fields $\pi^a(x)$ correspond to the pions and other light mesons. Since we are ignoring $U(1)_A$, we have $U(x) \in SU(3)$, and the T^a are the Gell-Mann matrices. If we only consider the up and down quarks, then $U(x) \in SU(2)$ and the T^a are the Pauli matrices.

- The field $U(x)$ transforms in the bifundamental of $SU(3)_L \times SU(3)_R$,

$$U(x) \to LU(x)R^\dagger.$$

The pion fields $\pi^a(x)$ parametrize the vacuum manifold. For a vector transformation we have

$$U(x) \to LU(x)L^\dagger, \quad \pi(x) \to L\pi(x)L^\dagger$$

while for an axial transformation,

$$U(x) \to LU(x)L, \quad \pi(x) \to \pi^a(x) + f_\pi \alpha^a + \ldots, \quad L = e^{i\alpha^a T^a}.$$

- We see the unbroken symmetries act on the $\pi^a(x)$ linearly, while the broken symmetries act infinitesimally by shifts, ensuring the Goldstone bosons are massless. The general procedure for writing broken and unbroken symmetries in this form is called the CCWZ construction. At the quantum level, we build our Hilbert space on the physical vacuum. The unbroken symmetries act on this Hilbert space, while the broken symmetries relate these states to excitations about other vacuums, and hence do not yield degeneracies.

- Next, we write down every term consistent with the symmetries. The Lagrangian must be invariant under $SU(3)_L \times SU(3)_R$, which means that a U must always be next to a $U^\dagger$. But we always need derivatives, because $U^\dagger U = 1$. Hence we have

$$\mathcal{L} = \frac{f_\pi^2}{4} \operatorname{tr}(\partial_\mu U^\dagger \partial^\mu U) + \ldots.$$

Since U is a nonlinear function of the pion fields, the quadratic term alone is enough to do nontrivial computations. Expanding and neglecting $O(1)$ factors and $SU(3)$ indices,

$$\mathcal{L} = \frac{1}{2}\partial_\mu \pi \partial^\mu \pi + \frac{1}{f_\pi^2}\pi^2 \partial_\mu \pi \partial^\mu \pi + \ldots.$$

In particular, the symmetry ensures that there are no relevant interaction terms at all!

- There are relations between the infinitely many coefficients in the Lagrangian in terms of π. These relations were originally understood by current algebra, but in chiral perturbation theory, we get them easily because we know that U transforms linearly.

- The second term above is the leading contribution to $\pi\pi \to \pi\pi$ scattering. The cross section can be estimated as follows.
  - If the pions are relativistic, we may ignore their masses, so the only mass scales are f_π and the Mandelstam variables.
  - The f_π can only enter through the matrix element, $\mathcal{M} \propto 1/f_\pi^2$, so $\sigma \propto 1/f_\pi^4$.
  - By dimensional analysis, σ has degree 1 in the Mandelstam variables. There is no way to get a Mandelstam variable in the denominator, because we aren't doing any kind of particle exchange (compare the $1/t$ in $e^+e^- \to e^+e^-$ scattering).
  - In general, for two particles in the final state, the phase space integrals will contribute a factor like $1/8\pi$, while for three particles we get $1/64\pi^3$.

Therefore, we conclude $\sigma \sim (s+t)/8\pi f_\pi^4$.

- Contributions also come from higher order terms, such as

$$\mathcal{L} \supset \operatorname{tr}(\partial_\mu U^\dagger \partial^\mu U)^2 \supset \frac{1}{f_\pi^4}(\partial_\mu \pi \partial^\mu \pi)^2.$$

  The matrix element here is smaller by a factor of p^2/f_π^2. Hence the cross section, which is the square of the sum of the matrix elements, is modified by $O(p^2/f_\pi^2)$. Hence chiral perturbation theory is a perturbation series in p^2/f_π^2.

- For pion-pion scattering at tree level, only the terms quadratic and quartic in U can contribute, since all others have too many powers of π. Hence tree level results are fully parametrized by only a few parameters, and the real test is controlling the loop corrections. The logarithms in cross sections due to loops are called chiral logs and are a signature of quantum effects.

- Loop corrections are suppressed by a factor of $1/(4\pi)^2$, so the real expansion parameter is $p/4\pi f_\pi$. Hence we were justified in treating the pions relativistically, because

$$4\pi f_\pi \approx 1\,\text{GeV} \gg m_\pi.$$

  Moreover, we may estimate error of any calculation to n loops as $(p/4\pi f_\pi)^{n+1}$.

Note. Chiral perturbation theory is closely related to sigma models. In the original linear sigma model, one takes a set of two complex scalar fields and performs spontaneous symmetry breaking, yielding a massive field σ and massless Goldstone bosons associated with the pions. At low energies, the σ decouples, yielding a "nonlinear sigma model", which now is a generic term for any field theory whose fields take values on a manifold. Chiral perturbation theory is simply a nonlinear sigma model where the σ would have corresponded to a variation of v.

Next, we account for the quark masses and other couplings.

- We are justified in treating the strange quark mass perturbatively because $m_s \ll 4\pi f_\pi$. In general, a mass will contribute to the Lagrangian as

$$\mathcal{L} \supset -\overline{q}_L M q_R + \text{h.c.}$$

  where M is a complex matrix. If we simply replace $\overline{q_L} q_R$ with its vev, we get the leading term,

$$\mathcal{L} = v^3 \operatorname{tr}(MU + M^\dagger U^\dagger).$$

  We can find higher-order terms by treating M as a spurion field that transforms as $M \to LMR^\dagger$.

- Expanding this leading term, we have

$$\mathcal{L} \supset -\frac{4v^3}{f_\pi^2} \operatorname{tr}(MT^aT^b)\pi^a\pi^b = -\frac{2v^3}{4_\pi^2} \operatorname{tr}(M\{T^a, T^b\})\pi^a\pi^b.$$

  In the case of two quark flavors, this may be simplified using $\{T^a, T^b\} = \delta^{ab}/2$ for $SU(2)$ generators. We find that all three pions satisfy the Gell-Mann–Oakes–Renner equation

$$m_\pi^2 = 2(m_u + m_d)v^3/f_\pi^2.$$

  In reality, the pions are split due to electromagnetic effects.

- For three quark flavors, a longer version of the same computation relates the pion, kaon, and η masses, giving the Gell-Mann–Okubo formula.
- To establish the parameter f_π is really the pion decay constant, we can couple the pion fields to leptons using the four-Fermi interaction and calculate the decay rate.
- To compute nucleon-pion scattering, for two quark flavors, we have a Dirac nucleon field

$$N = \begin{pmatrix} p \\ n \end{pmatrix}$$

that transforms as

$$P_L N \to L P_L N, \quad P_R N \to R P_R N.$$

The only possible mass term for the nucleons is

$$\mathcal{L} \supset -m_N \overline{N}(U^\dagger P_L + U P_R) N$$

which yields a pion-nucleon-nucleon coupling.

- Alternatively, we may couple the pion fields to quark fields, which transform similarly. This is valid for energies above Λ_{QCD} and below $4\pi f_\pi$. In both cases we write down all terms consistent with chiral symmetry.
- To account for the different electric charges of the quarks, pions, or nuclei, we simply promote derivatives in the chiral Lagrangian to covariant derivatives.

Further refinements, including a solution to the $U(1)_A$ problem, require an understanding of anomalies, which are covered in the notes on Quantum Field Theory.

6.5 The Strong CP Problem

First, we introduce the strong CP problem.

- The θ parameter is modified by chiral rotations of the quark fields by the chiral anomaly. Specifically, after electroweak symmetry breaking the quark mass matrices are neither Hermitian nor diagonal; integrating out the Higgs we have

$$\mathcal{L} \supset \overline{q}_{R_i} M_{ij} q_{L_j} + \text{h.c.}$$

and the matrix may be diagonalized by $SU(N_f)_A \times SU(N_f)_V$ transformations on the quark fields. To remove the overall phase, we require $U(1)_A$ transformations,

$$q_R \to e^{i\alpha/2} q_R, \quad q_L \to e^{-i\alpha/2} q_L$$

which shift the vacuum angle by an amount proportional to α, as can be understood by considering the $SU(3)^2 U(1)_A$ anomaly. The invariant quantity is

$$\overline{\theta} = \theta - \arg\det M$$

where M is the quark mass matrix.

- The value of $\overline{\theta}$ may be measured from the neutron electric dipole moment, yielding
$$\overline{\theta} \lesssim 10^{-10}.$$
The strong CP problem asks why this holds. It is especially puzzling because the two contributions to $\overline{\theta}$ appear to be completely unrelated.

- The chiral anomaly explains why all the θ-vacua are equivalent if any quark, such as the up quark, is exactly massless. In that case we can perform arbitrary chiral rotations on that quark field without physical effect, rotating θ to zero. This was once proposed as a solution to the strong CP problem, though it is disfavored by lattice computations.

- For the purposes of calculation, it's most useful to rotate so that $\theta = 0$, and work with the complex masses in chiral perturbation theory. The presence of the masses yields a $\overline{\theta}$-dependent QCD vacuum energy, which will provide the axion potential. This setup is also used to compute the neutron electric dipole moment.

We now compute the QCD vacuum energy.

- By rotating $\overline{\theta}$ into the quark mass matrix, we can take without loss of generality,
$$\theta = 0, \quad M = \begin{pmatrix} m_u & & \\ & m_d & \\ & & m_s e^{-i\overline{\theta}} \end{pmatrix}$$
where the m_i are all real. As a result, to minimize the vacuum energy, the pion fields pick up vevs. We can neglect all non-diagonal fields, because if they picked up a vev, they would break $U(1)_A$ symmetry. The general intuition here is that a minimum is often a point of enhanced symmetry (the exception being spontaneous symmetry breaking), which is essentially the reason that the vacuum energy is minimized at the CP preserving point $\theta = 0$.

- Plugging this expression in, we find
$$\mathcal{L} \supset 2v^3(m_u \cos\varphi_u + m_d \cos\varphi_d + m_s \cos(\varphi_s + \overline{\theta}))$$
where the pion fields are
$$U = \operatorname{diag}(e^{i\varphi_u}, e^{i\varphi_d}, e^{i\varphi_s}), \quad \varphi_u + \varphi_d + \varphi_s = 0.$$
We wish to maximize this expression to minimize the potential energy.

- Since $m_s \gg m_u, m_d$, we will have $\varphi_u + \varphi_d \approx \overline{\theta}$. Deviations from this equality can lower the potential energy by terms higher order in m_u/m_s, which we neglect. Differentiating, we have
$$m_u \sin\varphi_u = m_d \sin\varphi_d.$$
By applying the law of sines and law of cosines to an appropriately chosen triangle,
$$\frac{\sin\varphi_u}{m_d} = \frac{\sin\varphi_d}{m_u} = \frac{\sin\overline{\theta}}{\sqrt{m_u^2 + m_d^2 + 2m_u m_d \cos\overline{\theta}}}.$$
Plugging this in and using the Gell-Mann–Oakes–Renner equation, we find the vacuum energy
$$V(\overline{\theta}) = -m_\pi^2 f_\pi^2 \sqrt{1 - \frac{4m_u m_d}{(m_u + m_d)^2} \sin^2\left(\frac{\overline{\theta}}{2}\right)} \approx \frac{1}{2} m_\pi^2 f_\pi^2 \frac{m_u m_d}{(m_u + m_d)^2} \overline{\theta}^2.$$

- Note that a semiclassical one-instanton approximation would give $V(\overline{\theta}) \sim -\cos\overline{\theta}$, which is quite different. In general little is rigorously known about the function $V(\overline{\theta})$, except that it is periodic with a minimum at $\overline{\theta} = 0$. (This is established rigorously by the Vafa–Witten theorem.) We may also compute the curvature at the minimum. There could be conceivably be other local minima, or even discontinuities.

- It is difficult to compute $V(\overline{\theta})$ in lattice QCD, because lattice QCD must be done in Euclidean signature to make the path integral converge numerically, but the theta term has a single time derivative and hence appears as an imaginary part in the Euclidean action.

- The neutron EDM can also be computed in chiral perturbation theory. In this case the neutron EDM comes from pion loops.

Next, we describe how the axion solves the strong CP problem. There exists an enormous literature on axions; here we are just giving the very basics. Further discussion is given in the dissertation on Cosmological Relaxation.

- For motivation, we could try to solve the strong CP problem in the same way the "electroweak" CP problem is solved. That is, suppose there were a new chiral global symmetry $U(1)_{\text{PQ}}$, called Peccei–Quinn symmetry, where the quarks transform as

$$q_{Li} \to e^{ie_i\alpha/2} q_{Li}, \quad q_{Ri} \to e^{-ie_i\alpha/2} q_{Ri}$$

  where e_i is the $U(1)_{PQ}$ charge of the i^{th} quark flavor. If $\sum_i e_i$ is nonzero, there would be a $U(1)_{\text{PQ}}SU(3)_C^2$ anomaly, and the θ term could be rotated away.

- However, this global symmetry cannot exist in the SM. Conventionally normalizing $\sum_i e_i = 1$, the quark Yukawa coupling $\overline{Q}_L H D_R$ cannot be invariant unless all quarks have $e_i = 1/6$, and the Higgs transforms as $H \to e^{i\alpha/6} H$. However, the other quark Yukawa coupling $\overline{Q}_L H^c U_R$ requires the Higgs transforms as $H \to e^{-i\alpha/6} H$. (Note that in multi-axion theories one typically does not normalize $\sum_i e_i = 1$, which leads to extra coefficients in the results below.)

- This objection does not hold if $U(1)_{\text{PQ}}$ is spontaneously broken at a high scale f_a. Surprisingly, the strong CP problem is still solved if this happens, because the spontaneous symmetry breaking yields a Goldstone field, the axion $a(x)$, which transforms by a shift,

$$a(x) \to a(x) + \alpha f_a.$$

  The $U(1)_{\text{PQ}}SU(3)_C^2$ anomaly manifests by breaking the shift symmetry of the axion,

$$\mathcal{L} \supset \frac{g^2}{32\pi^2} \frac{a}{f_a} G^{\mu\nu a} \tilde{G}^a_{\mu\nu}$$

  so that the observed θ parameter is

$$\theta_{\text{obs}} = \overline{\theta} + \frac{a(x)}{f_a}.$$

  Hence the strong CP problem is solved if the axion vev adjusts so that $\theta_{\text{obs}} = 0$.

- By moving θ_{obs} into complex quark masses, we find a vacuum energy that depends on $a(x)$, and hence an axion mass term,

$$m_a \approx 0.5 \frac{m_\pi f_\pi}{f_a} \sim 10^{-3}\,\text{eV} \left(\frac{10^{10}\,\text{GeV}}{f_a}\right)$$

  by our work above. More roughly one could find this by dimensional analysis, $m_a \sim \Lambda_{\text{QCD}}^2/f_a$.

- Note that since the $U(1)_{\text{PQ}}$ symmetry is axial, the axion must be a pseudoscalar. Since $G\tilde{G}$ is P odd and C even, $aG\tilde{G}$ is both P and C even and hence obeys CP.

- The strong CP problem differs from other "naturalness" problems, as there seems to be no anthropic explanation. Also, our discussion above introduces a new high scale f_a, which potentially makes the hierarchy problem worse.

6.6 Axion Phenomenology

Next, we consider explicit axion models. We begin with the Weinberg–Wilczek axion.

- The simplest way to implement PQ symmetry is to add a second Higgs doublet and let the Yukawa interactions be

$$\mathcal{L} \supset \overline{Q}_L H_1 D_R + \overline{Q}_L H_2^c U_R$$

  with $U(1)_{\text{PQ}}$ charges $e_i = 1/6$ for each quark field, and the transformations

$$H_1 \to e^{i\alpha/6} H_1, \quad H_2 \to e^{-i\alpha/6} H_2.$$

  This is similar to how a second Higgs doublet is introduced in the MSSM.

- Spontaneous breaking of $U(1)_{\text{PQ}}$ occurs along with electroweak symmetry breaking, where

$$H_1 = e^{i\beta(x)/6} \begin{pmatrix} 0 \\ v_1/\sqrt{2} \end{pmatrix}, \quad H_2 = e^{-i\beta(x)/6} \begin{pmatrix} 0 \\ v_2/\sqrt{2} \end{pmatrix}.$$

  The W and Z boson masses are determined by

$$\sqrt{v_1^2 + v_2^2} \equiv v = 246\,\text{GeV}.$$

- The Goldstone boson is the relative phase of the two fields, and

$$\mathcal{L} \supset |\partial_\mu H_1|^2 + |\partial_\mu H_2|^2 \supset \frac{f_a^2}{2} \partial_\mu \beta \partial^\mu \beta, \quad f_a = \frac{v}{6}.$$

  Hence the axion field in this model is $a(x) = f_a \beta(x)$, and our formula above gives $m_a \sim 150\,\text{keV}$. Since f_a is low in this model, the axion interacts strongly with other particles, and hence this scenario is experimentally ruled out.

In "invisible axion" models, the scale f_a is much higher than the electroweak scale. The most well-known models are the DFSZ and KSVZ axions, though many models are possible.

- In the DFSZ model, one takes the previous model and adds a complex scalar Φ which is a singlet under the SM gauge group, with $\Phi \to e^{i\alpha/6}\Phi$, and adds the terms $\Phi^\dagger\Phi$ and $H_1^\dagger H_2\Phi^2$. The axion field is now a linear combination of the phases of the fields H_1, H_2, and Φ, but now
$$f_a = \frac{\sqrt{v_1^2 + v_2^2 + v_\Phi^2}}{6}.$$
- In the KSVZ model, the ordinary quarks have zero $U(1)_{\text{PQ}}$ charge, but we add an additional heavy quark Dirac field Ψ which is a triplet under $SU(3)_C$ and an electroweak singlet, with unit $U(1)_{\text{PQ}}$ charge. We also introduce a singlet field Φ as above with $\Phi \to e^{i\beta}\Phi$. Note that adding only Dirac fields avoids more chiral fermions, so there are no issues with gauge anomalies.
- Giving Ψ a $U(1)_{\text{PQ}}$ charge of $1/2$, we now have a $U(1)_{\text{PQ}}$ invariant Yukawa interaction
$$\mathcal{L} \supset \lambda\Phi\overline{\Psi}_L\Psi_R + \text{h.c.}$$
and $U(1)_{\text{PQ}}$ is spontaneously broken by the vev $|\langle\Phi\rangle| = v_\Phi/\sqrt{2}$. We then have
$$\Phi = (f_a + \sigma(x))e^{ia(x)/\sqrt{2}f_a}$$
and the vev provides the quarks with a mass on the order of λf_a. The singlet σ, called the "saxion", also has a mass on the order of f_a.
- Note that one may either normalize the charges so that $a \in [0, 2\pi f_a]$, or normalize them so that $\theta_{\text{obs}} \supset a/f_a$. We have chosen the latter option, though both are common.
- There are many variants. For example, in the original KSVZ model, Ψ has a hypercharge $-1/3$, motivated by embedding the theory in a GUT.
- There is a model-dependent coupling to photons by the $U(1)_{\text{PQ}}U(1)_{\text{EM}}^2$ anomaly, proportional to $(a/f_a)F_{\mu\nu}\tilde{F}^{\mu\nu}$. In "photophobic" models this coupling is exactly zero in the UV. Note that this term doesn't contribute to the axion potential because $U(1)_{\text{EM}}$ has no instantons.
- As for interactions with matter, symmetries allow couplings of the form
$$\mathcal{L} \supset \frac{1}{f_a}(\partial_\mu a)j^\mu_{\text{PQ}}, \quad j^\mu_{\text{PQ}} = \sum_i \frac{e_i}{2}\overline{f}_i\gamma^\mu\gamma^5 f_i$$
since the axion is a Nambu–Goldstone boson, where the f_i stand for all fermions. The important point is that the interaction is suppressed by f_a and depends only on $\partial_\mu a$.
- These interactions are modified by loop effects; equivalently one must RG flow down to a low scale $\mu \ll 1\,\text{GeV}$. In particular, the axion mixes with the π^0, η, and η', which are also SM singlet neutral pseudoscalar mesons, generating a substantial axion-photon coupling, even if none exists in the UV. This mixing can be computed explicitly in chiral perturbation theory.
- Both the KSVZ and DFSZ axions interact with nucleons by
$$\mathcal{L} \supset ig_{aNN}\frac{\partial_\mu a}{f_a}(\overline{N}\gamma^\mu\gamma^5 N)$$
where g_{aNN} is an $O(1)$ parameter. In the case of the KSVZ model, this interaction is induced by gluon loops, which don't suppress the coupling since QCD is strongly coupled in a nucleon. The KSVZ axion is called a hadronic axion because interactions with leptons only occur through photon loops, and are hence suppressed by a factor of $\alpha_e^2/4\pi$.

Note. Putting aside these models, how would we define an axion from the bottom up? First off, axions are pseudoscalar Nambu–Goldstone bosons, which explains their light mass. But that isn't specific enough, because it also applies to many other things, like pions. Axions additionally have an *exact* abelian discrete shift symmetry $a \to a + 2\pi f_a$, which forbids many potential couplings, but allows terms like $aF^{\mu\nu}\tilde{F}_{\mu\nu}$ and $aG^{\mu\nu}\tilde{G}_{\mu\nu}$ with discrete coefficients. These are allowed because the spacetime integrals of $F^{\mu\nu}\tilde{F}_{\mu\nu}$ and $G^{\mu\nu}\tilde{G}_{\mu\nu}$ are quantized for topological reasons, as explained in the notes on Geometry, so that the discrete shift won't change e^{iS}. Some people will say that an axion *must* have a $aG^{\mu\nu}\tilde{G}_{\mu\nu}$ coupling, while others would say this defines a QCD axion, while axions without it are called "axion-like particles" (ALPs).

Next, we consider axion production.

- Following the WIMP paradigm, one might think about a thermal population of axions. It turns out this would require heavy axions with $m_a \sim 100\,\text{eV}$. This dark matter would be heavy, and it is also ruled out by the constraints below. Instead, axions are produced by the nonthermal "misalignment mechanism".
- The axion only exists when Peccei–Quinn (PQ) symmetry breaks, so there are two scenarios. If PQ symmetry breaks for the last time before inflation, then we expect the axion field to have a uniform value in our Hubble patch. If it does so afterward, because either $H_I \gtrsim f_a$ or $T_{\text{reheat}} \gtrsim f_a$, then the axion field will have different values on each Hubble patch.
- In either case, right after PQ symmetry breaking the axion field will have some random misalignment angle θ_0 because its potential is flat. As the temperature lowers, the axion potential turns on, giving the required DM energy density.
- The situation is a bit subtle, because the potential gradually turns on, and also isn't perfectly harmonic. A numeric calculation gives

$$\Omega_a h^2 \approx 0.35 \left(\frac{\theta_0}{0.001}\right)^2 \left(\frac{f_a}{3\times 10^{17}\,\text{GeV}}\right)^n, \quad n = \begin{cases} 1.17 & f_a \lesssim 3\times 10^{17}\,\text{GeV} \\ 1.54 & f_a \gtrsim 3\times 10^{17}\,\text{GeV} \end{cases}$$

  The two cases depend on whether we have $H \sim m_a$ during radiation domination or matter domination, and the exact solution involves Bessel functions. Note that a more naive estimate would simply be $V \sim \Lambda_{\text{QCD}}^4 \theta_0^2$, with no dependence on f_a. The effects accounted for above mean that the axion density actually decreases with the axion mass, which is generic for coherent production mechanisms (i.e. ones where we can treat the DM as a field rather than particles).
- If PQ symmetry breaks after inflation, we must average over Hubble patches, giving

$$\langle \theta_0^2 \rangle = \frac{\pi^2}{3}$$

  in the case where the potential is sinusoidal; in general there are anharmonic correction factors. To achieve closure density, $\Omega_a h^2 = 0.12$, we have $m_a \sim 10^{-5}\,\text{eV}$. This is the classical axion window, which has been investigated by ADMX and other experiments. It also allows high f_a, all the way up to the Planck scale, which is favored by some string theory models.
- On the other hand, if PQ symmetry breaks before inflation, then θ_0 can be arbitrary, so we can have a smaller axion mass if θ_0 is small. This is the anthropic axion window. DM radio, CASPEr, and ABRACADABRA are more sensitive to the lower axion masses in this window.

- Note that there will be some energy in non-zero momentum modes of the axion field, but they dilute away rapidly since they behave like radiation. The energy density from the zero momentum mode, on the other hand, dilutes like matter. Hence the axion field can behave as cold dark matter, despite being extremely light.
- Also note that in both cases, the axion potential is a little more complicated because it isn't simple harmonic. In general, it's easy to get reasonable estimates, but almost all precision statements about axions must be found numerically.

We now consider issues with the classical axion window.

- Cosmic strings generically appear when a $U(1)$ symmetry is broken. In the anthropic window, the cosmic strings are diluted away. In the classical window, we expect one per Hubble volume.
- The axionic strings decay to axions, producing another relic population of axions. This is a complex process that is hard to compute precisely and whose details are still under dispute. However, under one calculation, it can provide closure density if $m_a = 26.2 \pm 3.4\,\mu\text{eV}$.
- Practical computations involving axionic string decay usually treat the string with an effective description such as the Nambu–Goto action, in which case string decay and axion production must be added in as additional effects.
- In models where instantons preserve a $\mathbb{Z}_N$ symmetry, domain walls are produced. This is unacceptable in the classical axion window, as there are strong constraints on domain walls.
- The initial Hubble-scale perturbations can form gravitationally bound clumps of axions on small scales called "miniclusters", with many astrophysical consequences. Generally, axion searches don't take into account this non-homogeneous phase space distribution, but some features can make detection easier. For example, rather than a general spread $\Delta v \sim 10^{-3}$, there can be low dispersion streams. And miniclusters can be quite light, passing through the Earth regularly.

Next, we turn to the anthropic axion window.

- One big simplification is that topological defects such as strings and domain walls are diluted away. Also, inflation makes the field extremely smooth, preventing the formation of structures like miniclusters. However, the axion density depends on an arbitrary parameter, the misalignment angle, so the scenario is less predictive. One can take the axion mass to be much lower if θ_0 is also assumed lower.
- The main issue is the production of isocurvature fluctuations, which place an upper bound on H_I. In particular, an observably large value of r_T, as was claimed by BICEP2, would rule out much of the parameter space, though this measurement did not pan out.
- The anthropic tuning required for very light axions is not as ill-defined as the usual anthropics, because there is a clear measure to use, i.e. uniform over θ.
- Very low axion masses are disfavored by black hole superradiance. That is, they would cause rotating BHs to "spin down", so observing such BHs constraints the axions. The region ruled out is $6 \times 10^{-13}\,\text{eV} \lesssim m_a \lesssim 10^{-11}\,\text{eV}$.

Finally, we consider constraints on axions.

- In invisible axion models the parameter f_a can exist in a wide range,

$$f_a \in [10^2, 10^{19}]\,\text{GeV}, \quad m_a \in [10^{-12}, 10^6]\,\text{eV}.$$

  String theory also motivates axions with $f_a \sim m_{\text{str}} \sim 10^{18}\,\text{GeV}$. The so-called classical axion window has $m_a \sim \mu\text{eV}$, with an order of magnitude uncertainty in $g_{a\gamma\gamma}$ due to model dependence. (Smaller values would be possible, but would require tuned cancellations between the UV coupling and the RG flow.)

- The axion-photon coupling is an important interaction experimentally. If one turns on a background electric/magnetic field, then the axion will mix with the magnetic/electric field.

- One detection method would be by polarization effects. Turning on a background static magnetic field, photons with polarization along $\mathbf{B}$ will be preferentially converted to axions, and moreover will experience a different index of refraction than photons with polarization perpendicular to $\mathbf{B}$. These effects, known as vacuum magnetic birefringence and dichroism, also exist in QED, and were probed by PVLAS.

- Axions could be thermally produced in stars and quickly escape. In "helioscope" experiments such as CAST, we use a magnetic field to convert them back to photons using a background electromagnetic field, which are expected to have a thermal spectrum with the core temperature of the Sun. The IAXO experiment will set a stronger bound; it will not be sensitive at the level of standard QCD axion DM, but could detect other axion-like particles.

- Another approach, developed at DESY, is "light shining through walls". A laser is shined at an opaque wall in a background magnetic field. Light will go through if it turns into an axion, goes through the wall, and turns back into a photon. However, the bounds from these experiments are not strong because they require two axion-photon conversion events in the lab.

- The experiments ADMX, HAYSTAC, DM Radio, ABRACADABRA, SHAFT, ORGAN, TASEH, ALPHA, MADMAX, TOORAD, and BREAD are instead sensitive to existing axions, not those created by lasers or stars, by assuming they constitute DM and converting them to photons in a static magnetic field. Hence these experiments are "haloscopes".

- The resulting line width is just $O(mv^2/mc^2) \sim 10^{-6}$ using the DM virial velocity, so we can get an $O(10^6)$ boost in sensitivity using a high-Q resonant cavity of size m_a^{-1}, which is a fraction of a meter. Note that this is very different from helioscopes, where the axions are relativistic.

- In this case, parametrically the power absorbed in the cavity is

$$P \sim (ga_0B_0)^2 V \omega_a \min(Q, 10^6) \lesssim \left(\frac{B_0}{1\,\text{T}}\right)^2 \left(\frac{V}{1\,\text{m}^3}\right) \frac{10^{11}\,\text{GeV}}{f_a} \left(10^{-22}\,\text{W}\right)$$

  for a typical QCD axion, which corresponds to $O(10^3)$ photons per second. This is the approximate sensitivity of ADMX.

- Why is it possible to detect axion dark matter at all, for such high f_a? Roughly, it's because

$$a \sim \frac{\sqrt{\rho_{\text{DM}}}}{m_a} \sim \frac{\sqrt{\rho_{\text{DM}}} f_a}{\Lambda_{\text{QCD}}^2}$$

which means that holding $\rho_{\rm DM}$ constant, the effect of the axion a/f_a is independent of f_a. That is, all QCD axions are coupled the "same" amount; the challenge to probing other m_a and f_a is really about devising a system with the appropriate resonant frequency.

- Incidentally, some string compactifications motivate an "axiverse" of many light pseudoscalar particles. The linear combination of them that couples to the QCD θ term, $\mathcal{L} \supset \sum_i (a_i/f_i) G_{\mu\nu}\tilde{G}^{\mu\nu}$, becomes the QCD axion. The rest are "axion-like particles", which don't couple to QCD directly, but can couple to the photon or fermions, whose mass is not related to their f_a.

There are also relevant astrophysical and cosmological constraints.

- Axions can generically decay to two photons by the same coupling. Hence one basic constraint is that they must be stable on cosmological timescales if they are to be the DM, giving $m_a \lesssim 20\,\text{eV}$.
- A more stringent constraint comes from the cooling of SN 1987A, which gives $m_a \lesssim 10^{-2}\,\text{eV}$, which is the current best astrophysical constraint. Accelerated cooling of the Sun, which would shorten its lifetime, gives a weaker constraint $m_a \lesssim 1\,\text{eV}$. (However, note that faster cooling makes the core heat up, because of higher gravitational contraction.)
- Why should stellar bounds be competitive with precision laboratory experiments? The point is that the pace of stellar evolution is determined by the long time required for photons to diffuse from the core of a star to the surface. Thus, producing weakly coupled axions, which quickly exit the star, would accelerate stellar cooling. (This also means that stellar bounds stop working at *high* axion couplings, in which case the axions get trapped, though such couplings are already ruled out anyway.)
- Let's do a very rough estimate to justify this. Such cooling occurs through the axion-photon coupling via the Primakoff process $\gamma + e^- \to a + e^-$, which scales as

$$\Gamma \sim \frac{g_{a\gamma\gamma}^2 T_{\rm core}^2}{16\pi^2 f_a^2}.$$

  Therefore, the total cooling rate is

$$\frac{dQ_a}{dt} \sim V_{\rm core} \frac{g_{a\gamma\gamma}^2}{16\pi^2 f_a^2} T_{\rm core}^7$$

  where we multiplied by the photon density T^4 and the typical axion energy T.
- This is to be compared with the ordinary cooling rate by radiation,

$$\frac{dQ_{\rm rad}}{dt} \sim \sigma A_{\rm surf} T_{\rm surf}^4$$

  where the Stefan–Boltzmann constant is $\sigma = \pi^2/60$ in natural units. Dropping all $O(1)$ factors,

$$\frac{dQ_a/dt}{dQ_{\rm rad}/dt} \sim \frac{R T_{\rm core}^7}{f_a^2 T_{\rm surf}^4}$$

  and for this to be less than one, we require $f_a \gtrsim 10^6\,\text{GeV}$ which corresponds to $m_a \lesssim 100\,\text{eV}$.

7 Effective Field Theory

7.1 Introduction

So far, we have touched on the principles of effective field theory above, as well as in the notes on Statistical Field Theory and the notes on Quantum Field Theory. In this section, we dive deeper into the subject, beginning with a heuristic review.

- The "zeroth step" of working with an EFT is to identify the relevant degrees of freedom. Next, we write down a general action for those degrees of freedom, including all terms allowed by the symmetries. The action will involve undetermined constants, which we call couplings, reflecting our ignorance of where the theory comes from.
- If there is a separation of scales in the problem, we will use it to write the action as a series in their small ratio ("power counting"). A typical example is the ratio of the energy scales of the processes considered, to a high cutoff energy. For example, the Euler–Heisenberg Lagrangian applies to processes below m_e, while Fermi theory applies to processes below m_W. Or, in cases where there are nontrivial backgrounds, like a macroscopically occupied bosonic field, one might expand in a vev divided by a mass scale.
- The power counting expansion will allow us to compute physical quantities to any desired precision, by working up to the required order. At a fixed order, there will be a finite number of unknown couplings, and the theory starts to be predictive when we measure more physical quantities than unknowns.
- If the power counting expansion parameter stops being small, the predictive power of the EFT breaks down, because infinitely many unknown couplings contribute. In these situations, the EFT could be completed into a "full theory" with fewer relevant parameters. (And every "full" theory is itself an EFT approximation to a deeper full theory!)
- However, EFTs can be useful even when we already know the full theory, because the EFT can be more physically transparent and amenable to calculation. For instance, in the Standard Model, hadronic decays of B mesons involve nonperturbative matrix elements. But in heavy quark effective theory (HQET), these matrix elements just fix coupling coefficients, and there is a perturbative expansion in $\Lambda_{\text{QCD}}^2/m_b^2$. We have already seen a similar situation for chiral perturbation theory (χPT), which can compute light meson scattering perturbatively. Furthermore, passing to an EFT can yield new symmetries at leading order, such as spontaneously broken chiral symmetry in χPT and a "spin-flavor" symmetry in HQET, simplifying calculations.
- In these cases, the full theory yields information about the couplings in the EFT, which we can incorporate by "matching". That is, we compute some physical quantity in both the EFT and the full theory (which only involves external degrees of freedom that are present in the EFT), and demand the answers match.
- Of course, if this is intractable, or the full theory is unknown, we can also match to experimental data. Alternatively, if we are particularly lucky, the EFT may exhibit "universality", where the symmetries are so restrictive that, at some order in the power counting expansion, the interesting physical observables in the EFT are *independent* of the full theory.

- The same full theory can have multiple EFTs, depending on the process and regime of interest. For example, QCD has χPT at low energy, HQET for heavy mesons, and SCET for jet formation. In fact, we might use more than one theory within the same process, thanks to factorization. In a hadronic process at a particle collider, we might compute the hard process using QCD and parton distribution functions, then switch to SCET for the jets.
- Another great feature of EFTs is that they tell us when they break down, even when we don't know the full theory, because we can infer the size of the power counting expansion by measuring individual couplings in experiments.

Example. The hydrogen atom. In introductory quantum mechanics, we consider the Hamiltonian

$$H = \frac{p^2}{2m_e} + \frac{e^2}{r}$$

which can be regarded as the leading term in an EFT. Corrections to atomic energy levels due to a finite nuclear mass enter at order m_e/m_p, by changing m_e to the reduced mass, and fine structure enters at higher order in α. Nonperturbative Standard Model effects also appear, as hyperfine structure enters at order m_e/m_p and requires nonperturbative QCD to find the proton magnetic moment. Continuing to higher accuracy, we run into the Lamb shift and even electroweak effects.

As another example, we can focus on the multipole expansion of the proton's electromagnetic field. Even this by itself yields an infinite series of couplings, suppressed by powers of r_p/r, where r_p is the proton radius, which includes the electric charge, electric and magnetic dipole moments, quadrupole moments, anapole moments, and charge radii. All of these parameters are determined by the full theory (i.e. QCD, or the entire Standard Model), and some are simply zero due to symmetries. The matching could be done, for example, by computing the scattering amplitude for the proton in the presence of a classical current. Conversely, measuring these moments gives us information about the proton's radius, even if we don't know about QCD.

Example. For a theory with scalars and fermions, we can enumerate operators by dimension.

- At $D = 0$ we have 1, which represents a cosmological constant.
- At $D = 1$ we have ϕ, which can usually be removed by shifting ϕ.
- At $D = 2$ we have ϕ^2, which is a mass term. (We neglect $\partial^\mu\phi$, which breaks Lorentz invariance.)
- At $D = 3$, we have the fermion mass term $\overline{\psi}\psi$ and the self-interaction ϕ^3.
- At $D = 4$, we have the self-interaction ϕ^4, the Yukawa interaction $\overline{\psi}\psi\phi$, the scalar kinetic term $\partial_\mu\phi\,\partial^\mu\phi$, and the fermion kinetic term $\overline{\psi}\not{\partial}\psi$. Note that we are neglecting total derivatives, so $\phi\,\partial^2\phi$ is equivalent to $\partial_\mu\phi\,\partial^\mu\phi$.

Many of the above remarks are generalities, which hold just as well in classical physics. Now we add a bit more detail, in the context of the quantum field theories considered in these notes, assuming a cutoff scale M and physics at $m \ll M$.

- Even when the full theory is weakly coupled, calculations can break down because of the phenomenon of "large logarithms". Generically, the polynomially divergent parts of loop amplitudes will give contributions scaling as $(M/m)^n$, which are absorbed by counterterms. What is left over is the part of the loop integral scaling as $\int d^dk/k^d \sim \int d(\log k)$, which receives contributions at all scales, and gives a factor of $\log(M/m)$. More generally, we can get a logarithm of the power counting parameter.

- These logarithmic contributions are "nonlocal in energy" and are much subtler to deal with. In our previous examples of effective theories, we mostly avoided them, either by working at tree level, or by working in 0 dimensions, but we focus on them here.

- The perturbation series in the full theory thus includes factors of $\log(M/m)$, and if M/m is sufficiently large, this can significantly affect its accuracy. The solution, as noted in the notes on Quantum Field Theory, is to use a "running coupling" $g(\mu)$ which depends on a renormalization scale μ. When we bring μ down to $\mu \sim m$, we have "resummed the logarithms". Hence the same theory at $\mu \sim m$ and $\mu \sim M$ can be thought of as a very simple example of an EFT and a full theory.

- One might thus ask, in the context of weakly-coupled full theories, why an EFT is necessary at all if we can just use running couplings. In more subtle multi-scale processes, such as those in SCET, the large logarithms cannot be removed by running couplings alone. (Also, more conceptually, the physical content might be clearer in terms of the EFT degrees of freedom.)

- As a simple example, consider a divergent loop integral with a hard UV cutoff $\Lambda_{\text{UV}} \gg m$,

$$I = i \int \frac{d^4\ell}{\ell^2 - m^2} = \frac{1}{8\pi^2} \int_0^{\Lambda_{\text{UV}}} d\ell \, \frac{\ell^3}{\ell^2 + m^2} = \frac{1}{16\pi^2} \left(\Lambda_{\text{UV}}^2 - m^2 \log \frac{m^2 + \Lambda_{\text{UV}}^2}{m^2} \right).$$

  This gives the expected quadratic and logarithmic divergences.

- Now, a very naive approach to get a power counting expansion would be to "Taylor expand before integrating", which would give

$$I = \frac{1}{8\pi^2} \int_0^{\Lambda_{\text{UV}}} d\ell \, \ell^3 \left(\frac{1}{\ell^2} - \frac{m^2}{\ell^4} + \frac{m^4}{\ell^6} + \dots \right).$$

  However, this is too crude of an approximation. We no longer have the logarithmic dependence on m at all, and meanwhile a spurious IR divergence has been introduced! Regulating it with an IR cutoff $\Lambda_{\text{IR}} \ll m$, we get

$$I = \frac{1}{16\pi^2} \left(\Lambda_{\text{UV}}^2 + m^2 \log \frac{\Lambda_{\text{UV}}^2}{\Lambda_{\text{IR}}^2} + m^4 \left(\frac{1}{\Lambda_{\text{IR}}^2} - \frac{1}{\Lambda_{\text{UV}}^2} \right) + \dots \right)$$

  which has little resemblance to the true answer.

- The point of this example is that simply Taylor expanding everything in sight is not a valid approach, though one can get this idea to work if one uses the "method of regions".

- We will use DR rather than a hard cutoff, as the latter tends to become messy for nontrivial calculations, and also breaks a vast array of symmetries. (Other advantages of mass-dependent versus mass-independent regulators have been discussed in the notes on Quantum Field Theory.) In particular, both the EFT and the full theory will depend on the renormalization scale μ. We will constrain the EFT by matching physical quantities at scale $\mu \sim M$, then RG flow down to $\mu \sim m$ to do perturbation theory in the EFT without large logarithms.

Note. In EFT, people often use the equations of motion to simplify the Lagrangian. For example, the operators $\phi^3 \partial^2 \phi$ and $m^2 \phi^4$ are supposedly equivalent because $(\partial^2 - m^2)\phi = 0$. Thus, an important part of setting up an EFT is using identities such as integration by parts, and the

equations of motion, to eliminate unwanted terms (often those with more derivatives). This is done for dimension-6 operators in the SM here.

A dumb way to justify this procedure is to say "the equations of motion are true, so why not use them?" Of course, this doesn't make sense even classically. As noted in the notes on Undergraduate Physics, plugging the equations of motion back into the Lagrangian isn't valid even in point particle mechanics, and applying it to a field theory like the one above clearly changes the solutions at the classical level. And in a quantum theory, it is even less clear what it means for the equations of motion to be "true" (e.g. the path integral integrates over off-shell configurations).

The reason this procedure works is that the two subtleties above cancel each other out. In the QFT calculations we typically use EFTs for, we have little interest in the field itself. As discussed in the notes on Group Theory, the field just serves as a tool for creating and annihilating particles, and we are really interested in scattering amplitudes for the particles, which are the true gauge invariant physical observables. As such, it's clear that one can use the equations of motion at lowest order. For instance, the operators $\phi^3\partial^2\phi$ and $m^2\phi^4$ both mediate $2 \to 2$ scattering at tree level, but all particles in the diagram are external. The former gives factors of p^2 for each external leg, and the latter gives factors of m^2, but these are the same thing because external legs are on-shell.

That above argument is only valid at lowest order, but the underlying reason works at all orders. Correlation functions of fields are related to S-matrix elements by the LSZ reduction formula. We will find the same scattering amplitudes as long as we use field operators that have nonzero overlap between the vacuum and the desired one-particle states we are scattering. As such, we are free to redefine the fields to some degree, and it turns out this freedom is equivalent to using the equations of motion in the Lagrangian. The reason that this is not emphasized in standard field theory textbooks is that the procedure only works order by order in the power counting expansion, i.e. one can eliminate a term at the cost of introducing infinitely many higher-order ones. However, this is perfectly acceptable in an EFT where we already had those terms to start with, and were planning on neglecting them anyway.

In order to show this properly, we will follow the paper *Reduced Effective Lagrangians* (also see this paper). For concreteness, we consider an EFT including a scalar field ϕ and work to first order in the power counting parameter η,

$$\mathcal{L} = \mathcal{L}^{(0)} + \eta\mathcal{L}^{(1)} + O(\eta^2).$$

Now consider working in terms of the field ϕ', where

$$\phi = \phi' + \eta T[\Phi]$$

where T is any local function of the fields Φ, carrying no powers of η. The generating functional

$$Z[J] = \int \mathcal{D}\Phi \, \exp\left(i\int dx\, \mathcal{L}(\phi) + \sum_i J_i\Phi_i\right)$$

can be written in terms of this new field,

$$Z[J] = \int \mathcal{D}\Phi' \, \det\left(\frac{\delta\phi}{\delta\phi'}\right) \exp\left(i\int dx\, \mathcal{L}^{(0)}(\phi') + \eta\mathcal{L}^{(1)}(\phi') + \eta T[\Phi']\frac{\delta\mathcal{L}^{(0)}}{\delta\phi} + J_\phi\phi' + J_\phi\eta T + \dots\right)$$

where for brevity we are abusing notation by defining the "same spacetime" functional derivative,

$$\frac{\delta\mathcal{L}^{(0)}}{\delta\phi} \equiv \frac{\partial\mathcal{L}^{(0)}}{\partial\phi} - \partial_\mu\frac{\partial\mathcal{L}^{(0)}}{\partial(\partial_\mu\phi)}.$$

At this order, we pick up a Jacobian from the change of variables, a shift to the $O(\eta)$ part of the action proportional to the $O(\eta^0)$ equations of motion, and a change in the coupling to currents. The second effect is precisely what we want, because $\delta\mathcal{L}^{(0)}/\delta\phi = 0$ is simply the zeroth order equation of motion. Thus, by a field redefinition, we can effectively use it to simplify the Lagrangian at first order, at the cost of shifting the couplings in higher order terms which we are neglecting anyway. It is clear that this generalizes, e.g. to subsequently eliminate $O(\eta^2)$ terms we would shift the field again by an amount proportional to η^2. Also, it is necessary that the field redefinition preserves the symmetries of the EFT, so no new terms are introduced.

The Jacobian factor is nontrivial, since we are going beyond linear field redefinitions. As in Yang–Mills, it can be handled by introducing a ghost field, with

$$\mathcal{L}_{\text{ghost}} = -\left(\bar{c}c + \eta\bar{c}\frac{\delta T}{\delta\phi}c\right).$$

However, the kinetic term for the ghosts must appear in the second term, and hence is $O(\eta)$. Thus, for a canonically normalized ghost field, the first term gives a mass $1/\eta$, which is at the cutoff; hence the ghosts can be integrated out just like the heavy fields. Note that for this part of the argument to work, it was essential that the field redefinition be of the form $\phi = \phi' + O(\eta)$. Intuitively, this condition means that the redefinition "preserves one-particle states".

Finally, consider the effect of changing the couplings to currents. This is important because it changes Green's functions,

$$G^{(n)} = \langle(\phi + \eta T)_1 \dots (\phi + \eta T)_n\rangle$$

where the right-hand side is a time-ordered vev, and the subscripts i indicate x_i arguments. To see why this does not affect S-matrix elements, it suffices to consider a few examples. When $T = \phi$, the field redefinition just scales ϕ. This changes the field renormalization Z_ϕ in a compensating way, and so drops out of the LSZ reduction formula. When $T = \phi^2$, the Green's function essentially picks up contributions that would have been part of $G^{(n+1)}$, and hence don't have the right pole structure to contribute to n-point scattering amplitudes. In the case of derivatives, $T = \partial^2\phi$, we could simply write $T = (\partial^2 - m^2)\phi + m^2\phi$. The first term gives contributions that cancel the corresponding pole in the Green's function, while the second has already been taken care of.

Thus, the only physical effect is that of shifting the Lagrangian, so the (zeroth order) equations of motion can be used freely to simplify it. Another useful fact is that, at lowest order only, integrating out a heavy field is equivalent to simply solving its equations of motion and plugging the solution back into the Lagrangian. Note that at higher orders, the equations of motion themselves are changed already at lower order, so using them to simplify the Lagrangian is tricky; it is better to just think in terms of field redefinitions.

7.2 Scalar Example

As a concrete example of matching at loop level, we consider a simple scalar field theory. We will use the same conventions as in the notes on Quantum Field Theory, with the exception that DR is defined with $d = 4 - 2\epsilon$.

- The full theory has fields ϕ of mass m, and Φ of mass M, with $M \gg m$. It is valid up to scales $\mu > M$, and has Lagrangian

$$\mathcal{L}^{\text{Full}}_{\text{kin}} = \frac{1}{2}(\partial_\mu\phi)^2 - \frac{1}{2}m^2\phi^2 + \frac{1}{2}(\partial_\mu\Phi)^2 - \frac{1}{2}M^2\Phi^2$$

and

$$\mathcal{L}_{\text{kin}}^{\text{Full}} = -\frac{1}{3!}a\phi^3 - \frac{1}{2}b\phi^2\Phi - \frac{1}{4}\kappa\phi^2\Phi^2 - \frac{1}{3!}\rho\phi^3\Phi - \frac{1}{4!}\eta\phi^4.$$

To keep things simple, we will only turn some of these couplings on at once.

- In this simple case, the power counting parameter is m/M, and we can do power counting in the EFT by ordinary dimensional analysis. However, in a more general situation, this would be more subtle. For example, a derivative ∂_μ has mass dimension 1 and hence power counting dimension 1 here, but in a nonrelativistic EFT where the power counting parameter is v/c, such as NRQED, we have $\partial_0 \sim (v/c)^2$ and $\partial_i \sim (v/c)$. Similar phenomena would happen in an EFT defined in the soft or collinear limit.

- In these more nontrivial cases, we simply recall that the scaling of the field itself is always fixed by making the kinetic term marginal, and coordinates scale inversely to derivatives.

-