MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Physics Department

Physics 8.286: The Early Universe November 29, 2018
Prof. Alan Guth

Lecture Notes 9
THE MAGNETIC MONOPOLE PROBLEM

INTRODUCTION:

In addition to the horizon and flatness problems discussed in Lecture Notes 8, the conventional (non-inflationary) hot big bang model potentially suffers from another problem, known as the magnetic monopole problem. If one accepts the basic ideas of grand unified theories (GUT's) in addition to those of the conventional cosmological model, then one is led to the conclusion that there is a serious problem with the overproduction of particles called "magnetic monopoles". While a full understanding of the particle physics of grand unified theories is obviously much more than can be accomplished in a single set of lecture notes, the goal here will be to give you a qualitative understanding of what a grand unified theory is and how magnetic monopoles arise in such theories. We will not try to give a solid justification for all the steps along the way, but we will get far enough so that you will be able to estimate for yourself the magnetic monopole production in the early universe, verifying that far too many monopoles are predicted in the context of the conventional hot big bang cosmology.

THE STANDARD MODEL OF PARTICLE PHYSICS:

Before discussing grand unified theories, there are a few things that should be said about the "standard model of particle physics," which is the bedrock of our understanding of particle physics. The standard model, which was developed in the early 1970s, has enjoyed enormous success, giving predictions in agreement with all reliable particle physics experiments so far. The model has been enlarged since its initial discovery, adding a third generation of fundamental fermions, but the form of the standard model has remained unchanged. The original formulation described massless neutrinos, but the model can easily be modified to include neutrino masses (which are now known to be nonzero, due to neutrino oscillations). There is more than one way to add neutrino masses, however, and we are still not sure what is the correct way to do it.

Physicists divide the known interactions in nature into four classes: (1) the strong interactions, which bind quarks together inside protons, neutrons, and other strongly interacting particles, and also provide the residual force responsible for the interactions between these particles; (2) the weak interactions, responsible for example for beta decay ($n \to p + e^- + \bar{\nu}_e$; e.g. neutron $\to$ proton + electron + anti-electron-neutrino); (3) electromagnetic interactions; and (4) gravity. The standard model of particle physics describes the first three of these, omitting gravity. In practice there is no problem ignoring gravity at the level of elementary particle interactions, as the gravitational force between two elementary particles is so weak that it has never been detected. The gravitational force between two protons, for example, is 10^{36} times weaker than the electrostatic force between the same two particles.

The elementary particle content of the standard model of particle physics is shown in Fig. 9.1,* at the right. All particles are classified as either fermions or bosons. Fermions are particles with spins that in units of $\hbar$ are equal to $\frac{1}{2}$, $\frac{3}{2}$, etc., where for the fundamental particles the spin is always $\frac{1}{2}\hbar$. Fermions obey the Pauli exclusion principle. Bosons are particles with spins that are integer multiples of $\hbar$. They obey quantum mechanical rules which are the opposite of the Pauli exclusion principle, so that bosons

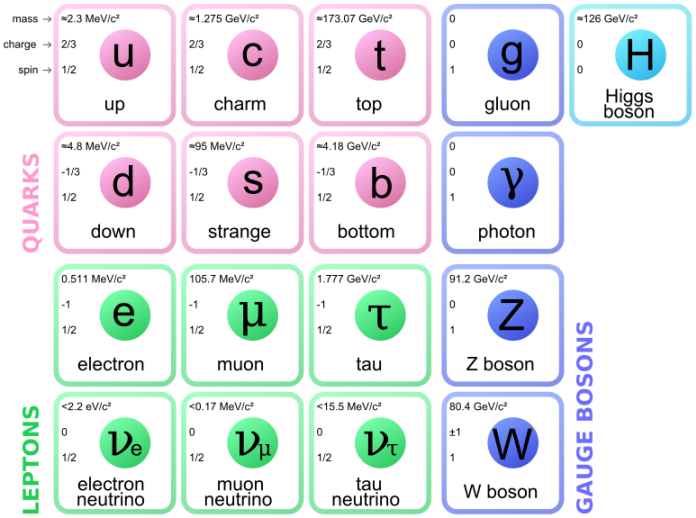

Figure 9.1: The particles of the standard model of particle physics.

have an enhanced tendency to fall into the same quantum state — that is the underlying principle behind the workings of a laser. The fermions of the standard model belong to three "generations," where the second and third generations are essentially copies of the first, except they are more massive. (The neutrinos are possibly an exception to this, as we do not know either the values or the ordering of the neutrino masses.) Each generation contains two quarks, with charges 2/3 and -1/3 in units of the magnitude of the electron charge, and also a neutrino and a lepton. Each quark comes in three different "color" states, and all the fermions have associated antiparticles. The color of a quark of course has nothing to do with its visual appearance, but is simply a label which was dubbed "color" because there are three possible values, like the three primary colors.

The interactions of the standard model are mainly provided by the "gauge bosons" shown in the fourth column of the diagram. These are spin-1 particles, and hence bosons, of which the most familiar is the photon γ. The photon is its own antiparticle, and is said to be the "carrier" of the electromagnetic interactions. The gluons g are the carriers of the strong interaction, and there are eight of them, including antiparticles. The Z^0, W^+, and W^- particles are the carriers of the weak interactions. The Z^0 is its own antiparticle, while the W^+ and W^- are antiparticles of each other.

At the right of the diagram is the Higgs particle (named for Peter W. Higgs of the University of Edinburgh), the existence of which was established in July 2012 at the Large Hadron Collider (LHC) at CERN. The Higgs particle can be seen only in very high energy processes, by modern accelerator standards, and even then only rarely. For example, in

* From the Wikimedia Commons. Source: PBS NOVA, Fermilab, Office of Science, United States Department of Energy, Particle Data Group.

the collision of two high energy protons at the LHC, it is possible for two gluons inside the protons to fuse into a Higgs particle, which can then decay to two photons. In addition to its role in describing Higgs particles, the Higgs field is responsible for the masses of the $W^{\pm}$, the Z^0, the quarks, and the $e^{\pm}$, the $\mu^{\pm}$, and the $\tau^{\pm}$. I will come back to the question of what exactly we mean when we say that the Higgs field is responsible for these masses. It may also be responsible for the neutrino masses, but it is not responsible for the mass of the proton; even without the Higgs field, it would be possible for massless quarks to form a massive bound state, such as the proton.

The spin-1 particles are called "gauge" particles because the standard model is an example of what is called a gauge theory. We will not have time to describe in detail what this means, but I will attempt to convey some partial understanding. You already know about one gauge theory — electromagnetism — but electromagnetism is a little too simple to make it obvious how to generalize the idea. The gauge theory aspect of electromagnetism can be seen only if it is written in terms of its potentials: the vector potential $\vec{A}$ and the scalar potential ϕ, which are related to the electric field $\vec{E}$ and the magnetic field $\vec{B}$ by

$$\begin{aligned} \vec{E} &= -\vec{\nabla}\phi - \frac{1}{c}\frac{\partial \vec{A}}{\partial t} \ , \\ \vec{B} &= \vec{\nabla} \times \vec{A} \ . \end{aligned} \tag{9.1}$$

The quantum theory of electromagnetism is always formulated in terms of these potentials. $\vec{A}$ and ϕ can be put together relativistically to define a four-potential A_μ,

$$A_\mu = (-\phi, A_i) \ . \tag{9.2}$$

As you have probably seen, the potentials themselves cannot be measured, but are in fact subject to the symmetry of gauge transformations. That is, given any scalar function $\Lambda(t, \vec{x})$, one can define new potentials ϕ' and $\vec{A}'$ by

$$\begin{aligned} \phi'(t, \vec{x}) &= \phi - \frac{\partial \Lambda}{\partial t} \ , \\ \vec{A}'(t, \vec{x}) &= \vec{A} + \vec{\nabla}\Lambda \ , \end{aligned} \tag{9.3}$$

which can be written in four-vector notation as

$$A'_\mu(x) = A_\mu(x) + \frac{\partial \Lambda}{\partial x^\mu} \ . \tag{9.4}$$

Gauge transformations are a symmetry in the sense that the new fields $A'_\mu(x)$ describe exactly the same physical situation as the original fields: $\vec{E}(t, \vec{x})$ and $\vec{B}(t, \vec{x})$ are unchanged.

If we consider two successive gauge transformations described by functions $\Lambda_1(t,\vec{x})$ and $\Lambda_2(t,\vec{x})$, the combined transformation can be described by a new function $\Lambda_3(t,\vec{x})$ given by

$$\Lambda_3(t,\vec{x}) = \Lambda_1(t,\vec{x}) + \Lambda_2(t,\vec{x}) \ , \tag{9.5}$$

so the combination of gauge transformations is described mathematically by the addition of real numbers. The combination is abelian — it does not matter in what order the gauge transformations of Λ_1 and Λ_2 are performed. The extension of gauge theories to nonabelian (i.e., non-commutative) transformations was invented in 1954 by Chen Ning (Frank) Yang and Robert Mills, and this idea became a key ingredient of the standard model of particle physics and its extensions. The standard model is based on gauge transformations that follow the form of three mathematical groups: SU(3), SU(2), and U(1). SU(3) is defined as the group of complex 3×3 matrices which are special (S) and unitary (U). The group operation is matrix multiplication. A matrix is special if its determinant is 1. A matrix U is unitary if it obeys the relation $U^\dagger U = I$, where $U^\dagger$ is called the adjoint of the matrix U, which means that $U^\dagger$ is the matrix obtained by transposing U (interchange rows and columns) and then taking its complex conjugate. I denotes the identity matrix. The definition of a unitary matrix is equivalent to saying that U has the property that if it multiplies a complex column vector v of the same size, then the norm of v is unchanged: $|Uv| = |v|$, where $|v| \equiv \sqrt{v^\dagger v}$. The group SU(2) is defined analogously, except that it uses 2×2 matrices. The group SU(2) is in fact essentially the same as the group of rotations in three spatial dimensions, although that fact is not obvious if you have not seen it. (It is actually a 2:1 map, with two matrices in SU(2) corresponding to each rotation matrix.) Finally, $U(1)$ is simply the group of complex phases, in the sense that an element of U(1) can be represented as a complex number $z = e^{i\theta}$, where θ is a real number.

In this language standard electromagnetism is a gauge theory based on U(1). From Eq. (9.5) it looks like we should be talking about the group of real numbers under addition, but in fact both descriptions are okay. If we included Dirac fields in our theory, to describe relativistic electrons, the Dirac field ψ would transform under the gauge transformation as

$$\psi'(x) = \exp\{ie_0\Lambda(x)\}\,\psi(x) \ , \tag{9.6}$$

where e_0 is the charge of the electron. So the transformation is fully described by the complex phase $z = \exp\{ie_0\Lambda(x)\}$. Note that this phase does not give us enough information to find Λ, because Λ can be changed by a multiple of $2\pi/e_0$ without changing the phase. But it nonetheless does give us enough information to find the gauge transformation of $A_\mu(x)$, since

$$\frac{\partial\Lambda}{\partial x^\mu} = \frac{1}{ie_0}\, e^{-ie_0\Lambda(x)}\, \frac{\partial}{\partial x^\mu}\, e^{ie_0\Lambda(x)} \ . \tag{9.7}$$

Yang and Mills invented a procedure to construct a field theory based on any of these gauge groups. A gauge transformation is defined by specifying an element of the

group at each point in spacetime, and two successive transformations combine according to the definition of multiplication in the mathematical group. Since SU(3) and SU(2) are nonabelian, these are called nonabelian gauge theories. These theories require specific spin-1 fields, which are the gauge fields shown in column 4 of Fig. 9.1. The equations of motion of these gauge fields are completely determined by the gauge symmetry, except for one *coupling constant* g for each gauge symmetry, which describes the size of the nonlinear terms in the equations of motion. Note that linear equations of motion allow any two solutions to be superimposed to obtain a third solution, which means that waves of the field do not interact, and the corresponding particles are free (i.e., do not interact). The coupling constant g therefore describes the strength of the interactions of the gauge particles, both with themselves and with other particles. For electromagnetism the gauge coupling constant is e, the magnitude of the electron charge, which in fact describes the strength of the interaction between photons and any charged particle. Photons are atypical among gauge particles, however, in that photons do not interact with other photons, which is a consequence of the fact that photons obey an abelian gauge theory. In addition to the gauge fields, gauge theories also allow other fields to be present, with interactions that are strongly restricted by the gauge symmetry, but not completely determined by it.

The three gauge symmetries of the standard model are usually described together as a product group, SU(3)×SU(2)×U(1). An element of such a product group is simply an ordered triplet (u_3, u_2, u_1), where u_3 is an element of SU(3), u_2 is an element of SU(2), and u_1 is an element of U(1). Thus the product group provides a compact notation, but really has the same information content as you would get by thinking about the three groups individually.

The SU(3) part describes the strong interactions, while the SU(2) and U(1) together describe the electromagnetic and weak interactions. The two are intertwined in their effect, however, so together they describe the *electroweak* interactions. While electromagnetism is a U(1) gauge theory, the U(1) of electromagnetism is actually a combined transformation that involves the U(1) of the standard model and a rotation about one fixed direction within the SU(2) group.

The Higgs field of the standard model is a complex doublet; i.e.,

$$H(x) \equiv \begin{pmatrix} h_1(x) \\ h_2(x) \end{pmatrix} \ , \tag{9.8}$$

where $h_1(x)$ and $h_2(x)$ are each complex numbers defined at each point $x \equiv (t, \vec{x})$ in spacetime. The functions $h_1(x)$ and $h_2(x)$ are called the *components* of the Higgs field $H(x)$. The doublet transforms under SU(2) gauge transformations in what is called the fundamental representation. I.e.,

$$H'(x) = u_2(x)H(x) \ , \tag{9.9}$$

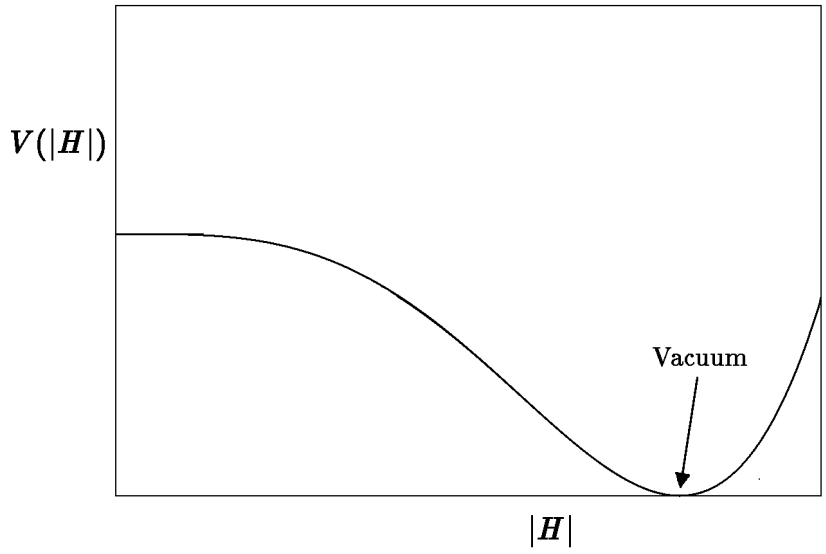

Figure 9.2: The approximate shape of the potential energy function for the Higgs field in the standard model of particle physics.

where u_2 denotes the 2×2 complex unitary matrix that defines the element of the gauge transformation at the spacetime point x. The gauge symmetry implies that the potential energy density of the Higgs field must be gauge invariant, which in turn means that it can depend only on the norm of the field,

$$|H| \equiv \sqrt{|h_1|^2 + |h_2|^2} \ . \tag{9.10}$$

The potential energy function for the Higgs field is assumed to have a peculiar form, as shown in Fig. (9.2). The potential energy function (which actually describes the potential energy density) is chosen to produce a phenomenon called spontaneous symmetry breaking. Spontaneous symmetry breaking is actually a common phenomenon in many branches of physics, including familiar processes such as the freezing of water. In the case of water, the relevant symmetry is rotational invariance. The laws of physics that describe water are completely rotationally invariant, with no direction preferred over any other direction. However, when water freezes, it forms a crystalline lattice which is not rotationally invariant. The crystalline lattice picks out definite directions along which the molecules align. The initial allignment is chosen randomly as the first molecules bind together, and then the rest of the molecules follow the pattern as they join onto the crystal. In general, whenever the ground state of a system has less symmetry than the underlying laws that describe it, it is called spontaneous symmetry breaking.

The equations of the standard model are exactly invariant under the gauge symmetry, but the only value of H that would be invariant under the gauge transformation (9.9)

is $H = 0$. But the potential energy function is designed so that the state $H = 0$ has a high energy, and the vacuum state — the state with the lowest possible energy density — has a nonzero value of $|H|$. This means that in the vacuum H must have a value that breaks the symmetry. There are of course an infinite number of directions in (h_1, h_2) space which would have the same value of $|H|$, and all would minimize the energy just as well. Like the direction of the crystal axes, the direction is picked out at some early time, and after that it is "frozen", becoming constant in time and over large regions of space.

The SU(2)×U(1) part of the standard model gauge symmetry implies that in the fundamental equations there is no distinction between electrons and neutrinos. The distinction arises entirely from the spontaneous symmetry breaking. The lepton fields interact with the Higgs fields, and those which interact with the components of the Higgs fields that have nonzero values will behave differently from the components that remain zero. Thus some components of the lepton fields will describe electrons, and some will describe neutrinos.

Before leaving the standard model, I'd like to try to qualitatively explain the connection between Higgs fields and mass. First, when we say that the Higgs field is responsible for the mass of the quarks, leptons, W^+, W^-, and the Z, we are talking about their rest masses. If the Higgs were not included in the theory, all these particles would have zero rest mass, like the photon. To understand how one field can influence the rest mass of another, remember that particles in a quantum field theory are simply the quantized excitations of fields. The rest mass times c^2 is the least energy that a particle can have, so the rest mass is just $1/c^2$ times the energy of the smallest possible excitation. To find this smallest excitation, we imagine describing the field inside a rectangular box, with for example periodic boundary conditions, and expand the field in terms of its normal modes. For small oscillations each normal mode behaves as a harmonic oscillator. When a harmonic oscillator is described quantum mechanically, the lowest energy level is $E = \frac{1}{2}\hbar\omega$, where ω is the (angular) frequency of the oscillator. The excited energy levels are evenly spaced, with the n'th energy level given by

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega \ . \tag{9.11}$$

Thus the smallest excitation is given by

$$\Delta E = E_1 - E_0 = \hbar\omega \ . \tag{9.12}$$

For any field, the mode with the smallest frequency is the homogeneous mode, the mode where the whole field oscillates uniformly. Thus, the mass of the particle is simply

$$m_0 = \frac{\hbar\omega_0}{c^2} \ , \tag{9.13}$$

where ω_0 is the angular frequency for homogeneous oscillations of the field. This formula implies that the rest mass of the photon is zero, since the frequency of a photon is given by $\omega = 2\pi c/\lambda$, so it approaches zero as λ approaches infinity, which is the limit of homogeneous oscillations. This is also the case for the quarks, leptons, W^+, W^-, and the Z of the standard model, if the Higgs field has zero value. But when some components of the Higgs fields have a nonzero value, the equations of the theory imply that these Higgs fields interact with the other fields, in some cases creating a restoring force, proportional to the value of one of the Higgs components, which pushes the other fields towards zero value. This restoring force results in a nonzero frequency for homogeneous oscillations, and hence a rest mass for the corresponding particles.

While the standard model (with some modification for neutrino masses) has been spectacularly successful in explaining all particle physics experiments, few if any physicists regard it as the final story, for at least two types of reasons. First, the theory is incomplete: it does not include gravity, nor does it contain any particle which can account for the dark matter in the universe. Second, the theory is viewed by physicists as being too inelegant (i.e., too ugly) to be the final theory. Specifically, the theory has many more seemingly arbitrary features and free parameters than one would hope. Why should there be three unconnected gauge symmetries, and why should there be three generations of fermions? The theory in its original form has 19 free parameters, such as the masses of each of the fermions and the strengths of the three fundamental interactions, which have values that must be measured, but cannot be deduced from any known principle. To account for neutrino masses, 7 or 8 new parameters must be added. What determines the values of all these parameters? Thus, while the standard model is certainly very accurate over a huge range of phenomena, the field of "beyond-the-standard-model" (BSM) particle physics is burgeoning.

Grand Unified Theories:

Grand unified theories, which were first proposed in the 1970s, are one promising attempt to go beyond the standard model. Grand unified theories are aimed primarily at unifying the three gauge interactions of the standard model, namely the SU(3), SU(2), and U(1) interactions. This is accomplished by embedding all three symmetry groups into a single, larger group, which becomes the gauge symmetry of the full theory. In the context of the full grand unified symmetry, there is no distinction between a neutrino, an electron, or a quark. The distinction is entirely created by the spontaneous breaking of the symmetry. The breaking of the full gauge symmetry down to SU(3)×SU(2)×U(1) is accomplished by introducing Higgs fields to produce the needed spontaneous symmetry breaking. (These fields are different from the Higgs field of the standard model, but they are also called Higgs fields because they play a completely analogous role.)

We are not attempting a full description of any of these topics, but I will explain how the three groups can be embedded in one larger group. There are many ways to

do it, but the simplest is to embed them in SU(5), the group of 5×5 unitary matrices with determinant one. This was the original grand unified theory, proposed in 1974 by Howard Georgi and Sheldon L. Glashow of Harvard.* To do this, we can let the SU(3) subgroup of SU(5) be the set of matrices of the form

$$g_3 = \begin{pmatrix} x & x & x & 0 & 0 \\ x & x & x & 0 & 0 \\ x & x & x & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} , \tag{9.14}$$

where the 3×3 block of x's represents an arbitrary SU(3) matrix. Similarly the SU(2) subgroup can be described by matrices of the form

$$g_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & x & x \\ 0 & 0 & 0 & x & x \end{pmatrix} , \tag{9.15}$$

where this time the 2×2 block of x's represents an arbitrary SU(2) matrix. Note that these matrices commute with matrices of the form shown above in Eq. (9.14). Finally, we need to find a set of U(1) matrices, complex phases, which commute with both classes of matrices described above. This can be done by setting

$$g_1 = \begin{pmatrix} e^{2i\theta} & 0 & 0 & 0 & 0 \\ 0 & e^{2i\theta} & 0 & 0 & 0 \\ 0 & 0 & e^{2i\theta} & 0 & 0 \\ 0 & 0 & 0 & e^{-3i\theta} & 0 \\ 0 & 0 & 0 & 0 & e^{-3i\theta} \end{pmatrix} , \tag{9.16}$$

where the factors of 2 and 3 in the exponents were chosen so that the determinant — in this case the product of the diagonal entries — is one, as it must be for an SU(5) matrix.

---

* H. Georgi and S. L. Glashow, "Unity of All Elementary-Particle Forces," *Phys. Rev. Letters*, vol. 32, pp. 438-441 (1974). Available from *Phys. Rev. Letters* at http://prl.aps.org/abstract/PRL/v32/i8/p438_1, from *Phys. Rev. Letters* with an MIT certificate as http://prl.aps.org.libproxy.mit.edu/abstract/PRL/v32/i8/p438_1, or it can be found for example at http://puhep1.princeton.edu/~kirkmcd/examples/EP/georgi_prl_32_438_74.pdf. The paper had a one sentence abstract: "Strong, electromagnetic, and weak forces are conjectured to arise from a single fundamental interaction based on the gauge group SU(5)." According to the INSPIRE database, it has been cited over 4,600 times.

In the SU(5) theory there is only one gauge interaction strength, while in the standard model there are three. The trick of relating one SU(5) interaction strength to the three interaction strengths of SU(3)×SU(2)×U(1) was a key step in the development of grand unified theories. The interaction strengths, it turns out, are not fixed constants, but vary with energy in a calculable way. When the three interaction strengths are extrapolated from the measured values to much higher energies, it is found that to a good approximation they meet, as shown in Figure 9.3.*

In Figure 9.3, α_1, α_2, and α_3 are the coupling strengths of the U(1), SU(2), and SU(3) interactions, respectively, as measured at low energies and extended to high energies according to the theory. (The α_i are related to the coupling constants g_i mentioned earlier by $\alpha_i \equiv g_i^2/4\pi$.) The horizontal axis shows the base-10 logarithm of the energy scale in GeV. The top graph shows the calculation for the standard model, while the bottom graph shows the more promising calculation for the Minimal Supersymmetric Standard Model, an extension of the standard model that incorporates supersymmetry. (Supersymmetry is a proposed, approximate symmetry that connects fermions to bosons and vice versa, which would connect each of the known particles to a partner that is slightly too massive to have yet been seen.) This graph is perhaps one of the strongest pieces of evidence for supersymmetry, and for grand unification. It implies a unification scale of about 10^{16} GeV.

The grand unified theory is constructed so that the spontaneous symmetry breaking gives masses of order 10^{16} GeV to those gauge bosons that represent interactions that are part of SU(5), but not part of the SU(3)×SU(2)×U(1) subgroup of the standard model. Energies of order 10^{16} GeV are totally unattainable; the LHC is designed to reach an energy of 7 TeV = 7,000 GeV per beam, or 14 TeV total. Nonetheless, we can speak theoretically about energies high compared to 10^{16} GeV, and then the spontaneous symmetry breaking of the grand unified theory would become unimportant. At such very high energies, the full gauge symmetry would become apparent. But at energies low compared to 10^{16} GeV, these 10^{16} GeV-scale particles would be too heavy to ever produce, and we would expect to see precisely the particle physics of the SU(3)×SU(2)×U(1) standard model.

The Magnetic Monopole Problem:

A magnetic monopole is a particle with a net North or South magnetic charge. The magnetic field of a monopole points radially outward (or inward), with a magnitude proportional to $1/r^2$, just like the Coulomb field of a point electric charge. Such particles

---

* Taken from the Particle Data Group 2016 Review of Particle Physics, C. Patrignani et al. (Particle Data Group), Chin. Phys. C, **40**, 100001 (2016), Chapter 16, *Grand Unified Theories*, Revised January 2016 by A. Hebecker and J. Hisano, http://www-pdg.lbl.gov/2016/reviews/rpp2016-rev-guts.pdf.

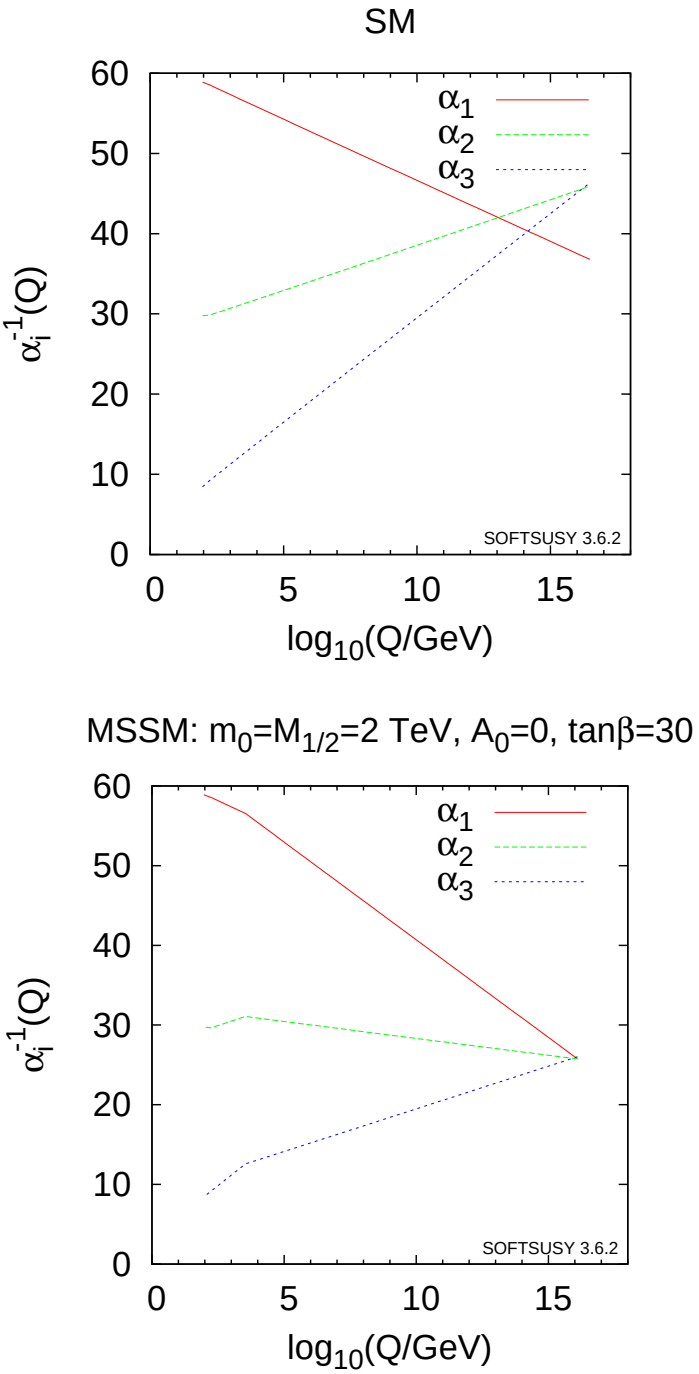

Figure 9.3: Running of the U(1), SU(2), and SU(3) interaction strengths with energy, in the standard model (SM) of particle physics and in the minimal supersymmetric standard model (MSSM). The horizontal axis is $\log_{10} Q$, where Q is the energy in GeV.

do not exist in the usual formulation of electromagnetism, in which all magnetic effects arise from electric currents. An ordinary bar magnet, with internal currents associated with the alignment of electronic orbits, has the form of a dipole, with North and South poles at the two ends. If a bar magnet is cut in half, one obtains two dipoles, each with a North and South pole. Grand unified theories (GUTs), however, imply that magnetic monopoles necessarily exist. They are generally superheavy particles, with mass energies of approximately 10^{18} GeV. That is, they are about two orders of magnitude heavier than the unification scale.

The magnetic monopoles of grand unified theories are constructed as twists, or knots, in the GUT Higgs fields, so the production of magnetic monopoles is closely linked to the behavior of the GUT Higgs fields as the early universe expanded and cooled. More technically, these knots in the GUT Higgs fields are called *topological defects.*

At high temperatures these Higgs fields will undergo large thermal fluctuations. In many grand unified theories the high temperature thermal equilibrium state is one in which the values of the fields average to zero, which means that the GUT symmetry is unbroken. As the system cools a phase transition is encountered. A phase transition is characterized by a specific temperature, called the critical temperature, at which some thermal equilibrium properties of the system change discontinuously. In this case, at temperatures below the critical temperature, some subset of the Higgs fields acquire nonzero mean values in the thermal equilibrium state — the GUT symmetry is thereby spontaneously broken. There may be one or perhaps several such phase transitions before the system reaches the lowest temperature phase — the phase which includes the vacuum. For simplicity, we will discuss the case in which there is only one such phase transition. In any case, the broken symmetry state which exists below the critical temperature is not unique, for precisely the same reason that the vacuum state is not unique.

In the conventional cosmological model, it is assumed that this phase transition occurs quickly once the critical temperature is reached. Thus, in any given region of space the Higgs fields will settle into a broken symmetry state, in which some subset of the Higgs fields acquire nonzero mean values. The choice of this subset is made randomly, just as the orientation of the axes of a crystal are determined randomly when the crystal first starts to condense from a molten liquid. The other particles in the theory, such as the quarks and leptons, are also described by fields, which interact with the Higgs fields in a manner consistent with the GUT symmetry. Through these interactions, the randomly selected combination of nonzero Higgs fields determines what combination of the fields will act like an electron, what combination will act like a u-quark, etc. The same random choice determines what combination of vector boson fields will act like the photon field, and what combinations will act like the W's, Z's, or gluons. In addition, some vector bosons acquire masses of the order of 10^{16} GeV, and these vector bosons are then irrelevant to the low energy physics which we observe in present-day accelerator experiments.

As mentioned above, the magnetic monopoles are examples of defects which form in the phase transition. The defects arise when regions of the high temperature symmetric phase undergo a transition to different broken-symmetry states. In the analogous situation when a liquid crystallizes, different regions may begin to crystallize with different orientations of the crystallographic axes. The domains of different crystal orientation grow and coalesce, and it is energetically favorable for them to smooth the misalignment along their boundaries. The smoothing is often imperfect, however, and localized defects remain.

The detailed nature of these defects is too complicated to explain here, so I will settle for the statement of some general facts. There are three types of defects that can occur. The simplest type is a surfacelike defect called a domain wall. This type of defect arises whenever the broken-symmetry state in one region of space cannot be smoothly deformed into the broken-symmetry state in a neighboring region of space. A domain wall then forms at the interface between the two regions. Some grand unified theories allow for the formation of such domain walls, and others do not. The second type is a linelike defect called a cosmic string. Again, some grand unified theories allow such defects to exist, and others do not. Finally, the third type is a pointlike defect, called a magnetic monopole. In contrast to the first two types of defects, magnetic monopoles exist in ***any*** grand unified theory.

To see how a pointlike defect can arise, let us consider the simplest theory in which they occur. This theory is too simple to describe the real world, but it serves as a "toy" model which is useful to illustrate many features of spontaneously broken gauge theories. The theory has a three-component multiplet of Higgs fields, which I will denote by ϕ_a, where $a = 1, 2,$ or 3. The symmetry which operates on this multiplet is identical in its mathematical form to the transformations that describe how the three components of an ordinary vector are modified by a rotation. The potential energy density associated with the Higgs fields is then a function of the three components ϕ_a. The energy density function, however, is an ingredient of the fundamental theory which must be invariant under the symmetry. Thus, the energy density can depend only on

$$|\phi| \equiv \sqrt{\phi_1^2 + \phi_2^2 + \phi_3^2} \ . \tag{9.17}$$

The potential energy density for this field will be assumed to have the general form shown earlier for $|H|$, so that spontaneous symmetry breaking ensues. The energy density will be minimized when $|\phi|$ has some particular nonzero value, which we will call ϕ_v. Now consider the following static configuration of the Higgs field:

$$\phi_a(\vec{r}) = f(r)\hat{r}_a \ , \tag{9.18}$$

where $r \equiv |\vec{r}|$, $\hat{r}_a$ denotes the a-component of the unit vector $\hat{r} = \vec{r}/r$, and $f(r)$ is a function which vanishes when $r = 0$ and approaches ϕ_v as $r \to \infty$. This configuration

is sketched below in Fig. 9.4. An arrow is drawn at each point in space, and the three vector components of the arrow are used to represent the three components of the Higgs field:

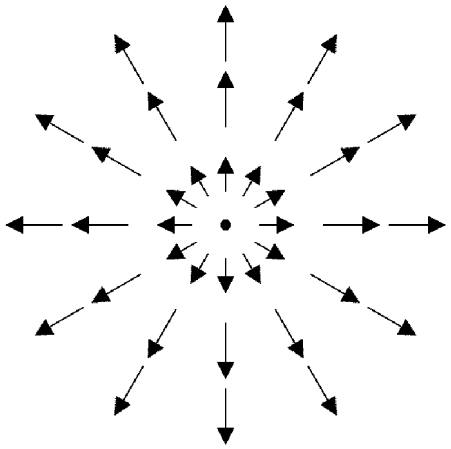

Figure 9.4: Graphical represention of the three-component Higgs field in the vicinity of a magnetic monopole.

If the diagram were constructed as a three-dimensional model, then all of the arrows would point radially outward from the origin. (An antimonopole is described by a similar picture, except that the arrows would point radially inward.) Note that the index a on ϕ_a normally has nothing to do with any direction in physical space — ϕ_1, ϕ_2, and ϕ_3 are just three scalar fields. Their behavior is related by a symmetry of fields — the gauge symmetry — but this symmetry is unrelated to the symmetry of rotations in physical space. Nonetheless, each field ϕ_a is allowed to be an arbitrary function of position, so there is nothing to prevent the fields from assuming the form of Eq. (9.18), as illustrated in Fig. (9.4).

The Higgs fields for the monopole configuration are in a vacuum state at large distances, but the fields differ from their vacuum values in the vicinity of $r = 0$, resulting in a concentration of energy. It can be proven that this configuration is "topologically stable" in the following sense: if the boundary conditions for the fields at infinity are held fixed, and if the fields are required to be continuous functions of position, then there must always be at least one point at which all three components of the Higgs field vanish. I will not attempt to prove this theorem, but I recommend that you stare at the diagram until the theorem becomes believable. Because of this topological property of the magnetic monopole configuration, it is sometimes referred to as a "knot" in the Higgs field. The configuration involves a concentration of energy localized around a point, and it behaves exactly as a particle.

So far I have not mentioned anything about magnetic fields, so the astute reader is no doubt wondering why these particles are called magnetic monopoles. For present

purposes the important property is that these objects are topologically stable knots in the Higgs fields, but in fact they must have a net magnetic charge. The reason comes from energy considerations. In the absence of any other fields, the energy of the magnetic monopole Higgs field configuration would be infinite. To understand this infinity, you must accept without proof the fact that the expression for the energy density of a Higgs field contains a term proportional to the square of the gradient. The form of Eq. (9.18) for large r (with $f(r) \to \phi_v$) then implies that the gradient of ϕ_a falls off as $1/r$ at large distances. The total energy within a large sphere is therefore proportional to

$$4\pi \int r^2 \, \mathrm{d}r \left(\frac{1}{r}\right)^2$$

and therefore diverges linearly with the radius of the sphere. However, the expression for the energy density becomes more complicated when the vector boson fields are included. It is beyond what we have time to discuss here, but it can be shown that the total energy of the Higgs field configuration of Eq. (9.18) can be made finite only if the configuration includes vector boson fields that correspond to a net magnetic charge. (The usual $\vec{\nabla} \cdot \vec{B} = 0$ equation holds far away from the center of the monopole, where $|\phi|$ is very close to ϕ_v, but $\vec{\nabla} \cdot \vec{B}$ can be nonzero near the center of the monopole, where the Higgs fields become small and the other gauge fields become part of the dynamics.) Even the magnitude of this magnetic charge is determined uniquely. The magnetic charge must correspond to a value $1/(2\alpha)$ times the electric charge of an electron. Here α denotes the usual fine structure constant of electrodynamics: $\alpha = e^2/\hbar c$ in cgs units, or $\alpha = e^2/4\pi\epsilon_0\hbar c$ in SI (mks) units. In any case, $\alpha \approx 1/137$. This means that the magnetic charge of a monopole is 68.5 times as large as the electric charge of an electron, and the force between two monopoles is then $(68.5)^2$ times as large as the force between electrons at the same distance.

The mass of a monopole can be estimated in these models, and it turns out to be extraordinary. The mass is approximately $1/\alpha$ times the mass scale at which the unification of forces occurs. Since the unification of forces occurs roughly at 10^{16} GeV, it follows that Mc^2 for a monopole is about 10^{18} GeV.

Having gone through the basic physics, we are now in a position to discuss how one estimates the number of magnetic monopoles that would be produced in the GUT phase transition. I will present a crude argument which is probably accurate to within one or two orders of magnitude. Although the argument will be crude, there is really no need to carry out a more accurate calculation. The magnetic monopole problem is so severe than an ambiguity of two orders of magnitude in the estimate is unimportant to the conclusion.

Recall that the monopoles are really knots in the Higgs field, so their number density is related to the misalignment of the Higgs field in different regions of space. This

misalignment can be characterized by a "correlation length" ξ. We will need only an approximate definition of this correlation length, so it will suffice to say that ξ is the minimum length such that the Higgs field at a given point in space is almost uncorrelated with the Higgs field a distance ξ away. One then estimates that the number density of magnetic monopoles and antimonopoles is given roughly by

$$n_M \approx 1/\zeta^3 \ . \tag{9.19}$$

In words, we are estimating that every cube with a side of length ξ will have, on the average, approximately one magnetic monopole in it. This estimate was first proposed by T.W.B. Kibble of Imperial College (London).

The remaining problem is to estimate ξ. Here we will be working in the context of conventional cosmology, in which it is assumed that the phase transition occurs quickly once the critical temperature is reached. Under these assumptions the phase transition has no significant effect on the evolution of the early universe. When the universe cools below the critical temperature T_c of the GUT phase transition (with $kT_c \approx 10^{16}$ GeV), it becomes thermodynamically probable for the Higgs field to align uniformly over reasonably large distances. If the system were allowed time to reach thermal equilibrium, then very few monopoles would be present — their abundance would be suppressed by the usual Boltzmann factor

$$e^{-Mc^2/kT}$$

from statistical mechanics. For this case the factor is roughly $e^{-100} \approx 10^{-43}$. However, if the whole process must happen on the time scales at which the early universe evolves, then there is not enough time for this long range correlation of the Higgs field to become established. While we are not prepared to calculate the correlation length in these circumstances, we can safely say that the correlation length must be less than the horizon distance — this statement assumes only that the correlation of the Higgs field requires the transmission of information, and special relativity implies that information cannot propagate faster than the speed of light. It is then a straightforward calculation, which you will do in Problem Set 9, to find a lower bound on the number of monopoles that would have been produced at the GUT phase transition under these assumptions.

If you do the problem right, you should find that the contribution to Ω today, from monopoles, would be bigger than 10^{20}, according to this calculation.

This number is obviously unacceptable, but one way to drive this point home is to consider the age of the universe. As you recall, a large value of Ω implies that the universe slowed down rapidly to its present expansion rate, giving a low predicted age for the universe. The formula for the age of the universe was derived in Lecture Notes 4, assuming that it can be approximated as being matter-dominated throughout its evolution. Since the monopoles would behave as nonrelativistic matter, this would be an

excellent approximation here. For $\Omega > 2$, during the expanding phase, the age is given by

$$t = \frac{\Omega}{2H(\Omega-1)^{3/2}} \left\{ \sin^{-1}\left(\frac{2\sqrt{\Omega-1}}{\Omega}\right) - \frac{2\sqrt{\Omega-1}}{\Omega} \right\} ,$$

where the inverse sine function is to be evaluated in the range $\frac{\pi}{2}$ to π. For very large Ω the inverse sine function approaches π, and the age is approximated by

$$t = \frac{\pi}{2H\sqrt{\Omega}} . \tag{9.20}$$

Taking Ω as 10^{20}, the age turns out to be 2.2 years. The prediction that you will find will be even smaller than that, since you should find a value of Ω bigger than 10^{20}.

Thus if grand unified theories are correct — which is plausible but not necessarily true — then we have another serious problem for the conventional hot big bang model. The universe, after all, is certainly more than 2.2 years old!