Physics 408: RELATIVITY

2004 Spring MWF 10:40-11:35 P128

Warren Siegel Math Tower 6-103 632-7978
http://insti.physics.sunysb.edu/~siegel/408.html
mailto:siegel@insti.physics.sunysb.edu

Special relativity is a symmetry of nature. In this course we will examine its implications for both particles and waves, and their dynamics. It is a modern course, including things we have learned about relativity after 1916. If time permits, we'll consider its generalization to curved space (general relativity), as applied to gravity or strings.

Outline

Spacetime: nonrelativistic, Minkowski, examples, experiments, conformal

Spin: nonrelativistic tensors, SU(2), nonrelativistic spinors, relativistic spinors

Actions: equations of motion, conservation

Particles: momentum, antiparticles, action, interactions, pair creation

Waves: correspondence, external fields, electromagnetism

Cosmology: dilaton, expansion, red shift

Schwarzschild solution: metric, gravitational redshift, geodesics, black holes

Strings: geometry, classical mechanics, gauges, quantum mechanics

Yang-Mills: Lie algebra, nonabelian gauge theory, lightcone gauge

General Relativity: coordinate tensors, gauge invariance, covariant derivatives

Grading

Grading will be based entirely on homework. Problems will be taken from those in these lecture notes. You may discuss problems with classmates, but the write-up must be your own. Homework is due one week after assignment, at the beginning of class. (Put it on my desk when you enter.) No late homework is accepted; it may be handed in early, but only to me in person.

1. Spacetime

1.1. Nonrelativistic

One of the most important principles in physics is symmetry. A symmetry is a transformation (a change of variables) under which the laws of nature do not change. It places strong restrictions on what kinds of objects can exist, and how they can interact. When dynamics are described by an action principle (Lagrangian, Hamiltonian, etc.; see below), as required by quantum mechanics, symmetries are directly related to conservation laws, which are the sole content of Newton's laws.

Certain symmetries apply to space and time. (The rest are called "internal symmetries".) Writing the (Cartesian) coordinates of space as

$$x^i = (x^1, x^2, x^3) = (x, y, z)$$

(where the coordinates are labeled by the superscript $i = (1, 2, 3)$), and including time as an additional coordinate

$$x^m = (x^0, x^i) = (t, x^i)$$

($m = (0, 1, 2, 3)$) we can write a coordinate transformation as

$$x^m \to x'^m(x)$$

for some functions x'^m of x^m. Some transformations can be written in terms of parameters that take continuous values down to zero, where by definition they are just the identity transformation. Such transformations can be built up from infinitesimal transformations

$$x'^m(x) = x^m + \delta x^m(x)$$

where δx^m is infinitesimal.

In particular, the simplest and most useful ones are the "affine" transformations

$$x'^m = x^n A_n{}^m + B^m$$

or

$$\delta x^m = x^n a_n{}^m + b^m$$

for some constants $A_n{}^m$ and B^m (or infinitesimal ones $a_n{}^m$ and b^m), where we have used the "Einstein summation convention": We implicitly sum over ("contract") the repeated index n in the above equation, where usually one of the 2 indices is written as a superscript and one as a subscript. Affine transformations are exactly those that can be written conveniently in matrix notation:

$$x' = xA + B, \qquad \delta x = xa + b$$

for row vectors x and B (b), and matrix A (a).

For field theory (waves) we also need to consider transformations of partial derivatives: In the case of affine transformations,

$$\frac{\partial}{\partial x'^m} = \frac{\partial x^n}{\partial x'^m}\frac{\partial}{\partial x^n} = (A^{-1})_m{}^n \frac{\partial}{\partial x^n}$$

We usually abbreviate $\partial/\partial x^m \equiv \partial_m$. Then in matrix notation

$$\partial' = A^{-1}\partial, \qquad \delta\partial = -a\partial$$

Of course, not all coordinate transformations are symmetries of physics (except in general relativity). The symmetries of nonrelativistic physics ("Galilean group") include translations in space and time:

$$\delta x^m = b^m$$

rotations in space

$$\delta t = 0; \qquad \delta x^i = x^j a_j{}^i, \quad a_{ij} = -a_{ji}$$

and "Galilean boosts"

$$\delta t = 0; \qquad \delta x^i = t a_0{}^i$$

where $a_0{}^i$ is the constant velocity of the boost. Translations and rotations preserve the infinitesimal length ds,

$$ds^2 = (dx^i)^2 = dx^2 + dy^2 + dz^2$$

as long as the matrix a is antisymmetric. However, length is not preserved by boosts, unless the curve along which this length is measured is instantaneous (all points at the same time).

In the finite case, the matrix A must be "orthogonal",

$$AA^T = I, \qquad A_i{}^k A_j{}^k = \delta_{ij}$$

in terms of the transpose A^T (since in matrix notation $ds^2 = (dx)(dx)^T$), and "Kronecker δ" ($\delta_{ij} = 1$ for $i = j$, 0 otherwise). Rotations in 3 dimensions are thus called "O(3)" ("O" for orthogonal). Those obtained from infinitesimal ones are "SO(3)"

("S" for special, meaning determinant 1), and exclude reflections (e.g., "parity": $A = -I$). They are also called "proper rotations". All the "improper" ones can be written as proper ones times $-I$, since

$$det(AA^T) = det(A)det(A^T) = [det(A)]^2 = det(I) = 1 \quad \Rightarrow \quad det(A) = \pm 1$$

$$det(-A) = det(-I)det(A) = -det(A)$$

It's convenient to express these transformations as exponentials:

$$A = e^a, \quad a = -a^T \quad \Rightarrow \quad AA^T = e^a e^{-a} = I, \quad det(A) = e^{tr(a)} = 1$$

We can then identify the matrix in the exponent as that appearing in the infinitesimal transformation, since when it is small

$$A = e^a \approx I + a \quad \Rightarrow \quad x' \approx x + \delta x$$

It's easy to get any SO(3) transformation continuously from the identity, simply by multiplying the exponent by a parameter (since that doesn't violate its antisymmetry), and varying it from 0 to 1.

Spacetime symmetries are useful in that they allow us to choose different reference frames: We can choose the origin of our coordinate system anywhere, and rotate the axes in any direction, and the laws of physics will look the same. We can also choose the origin to be either at rest or moving at constant velocity. (It is also possible to write the laws of physics, even nonrelativistic ones, in an arbitrary coordinate system, such as spherical coordinates, but their form is more complicated. However, in general relativity, where spacetime is curved, this turns out to be necessary, as no coordinate system is preferred in general.) In practice, one never explicitly performs such transformations. Rather than transforming from one frame to another, one just examines all relevant quantities directly in the frames of interest. It is enough to know that such transformations exist, i.e., that physical laws are invariant under such symmetries, so that one can simply choose any frame one likes, and the equations will be the same.

Exercise 1.1.1

Write an arbitrary two-dimensional vector in terms of a complex number as $V = \frac{1}{\sqrt{2}}(v_x - iv_y)$.

a Show that the phase (U(1)) transformation $V' = Ve^{i\theta}$ generates the usual rotation. Show that for any two vectors V_1 and V_2, $V_1{}^*V_2$ is invariant, and identify its real and imaginary parts in terms of well known vector products. What kind of transformation is $V \to V^*$, and how does it affect these products?

b Consider two-dimensional functions in terms of $z = \frac{1}{\sqrt{2}}(x+iy)$ and $z^* = \frac{1}{\sqrt{2}}(x-iy)$. Show by the chain rule that $\partial_z = \frac{1}{\sqrt{2}}(\partial_x - i\partial_y)$. Write the real and imaginary parts of the equation $\partial_{z^*}V = 0$ in terms of the divergence and curl. (Then V is a function of just z.)

c Consider the complex integral

$$\oint \frac{dz}{2\pi i}\, V$$

where "$\oint$" is a "contour integral": an integral over a closed path in the complex plane defined by parametrizing $dz = du(dz/du)$ in terms of some real parameter u. This is useful if V can be Laurent expanded as $V(z) = \sum_{n=-\infty}^{\infty} c_n(z-z_0)^n$ inside the contour about a point z_0 there, since by considering circles $z = z_0 + re^{i\theta}$ we find only the $1/(z-z_0)$ term contributes. Show that this integral contains as its real and imaginary parts the usual line integral and "surface" integral. (In two dimensions a surface element differs from a line element only by its direction.) Use this fact to solve Gauss' law in two dimensions for a unit point charge as $E = 1/4\pi z$.

Exercise 1.1.2

Again consider 2 dimensions:

a Write an arbitrary rotation in two dimensions in terms of the *slope* (dy/dx) of the rotation (the slope to which the x-axis is rotated) rather than the angle. (This is actually more convenient to measure if you happen to have a ruler, which you need to measure lengths anyway, but not a protractor.) This avoids trigonometry, but introduces ugly square roots. Also note that this square root form covers only half of the available angles.

b Show that the square roots can be eliminated by using the slope of *half* the angle of transformation as the variable.

1.2. Minkowski

Special relativity is actually simpler than Galilean symmetry, since space and time are treated on more-equal footing. Instead of length along a curve in space, we define "proper time" s along a curve in spacetime by

$$-ds^2 = (dx)^2 \equiv dx^m dx^n \eta_{mn} = -dt^2 + (dx^i)^2$$

where η_{mn} is the flat-space metric tensor,

$$\eta_{mn} = \begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ \hline 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 \\ 3 & 0 & 0 & 0 & 1 \end{array}$$

The 4-dimensional space defined by this metric is "Minkowski space". Here and from now on, unless otherwise specified, we use units $c = 1$, where c is the speed of light in a vacuum. For example, in astronomical units, c=1 light year/year. In fact, the speed of light is no longer measured, but used to define the meter (since 1986) in terms of the second (itself defined by an atomic clock), as the distance light travels in a vacuum in exactly 1/299,792,458th of a second. So, using metric system units for c is no different than measuring land distance in miles and altitude in feet and writing $ds^2 = dx^2 + dy^2 + b^2 dz^2$, where b=(1/5280)miles/foot is the slope of a line raised up 45°.

One way to interpret the funny minus sign, as compared to an ordinary ("Euclidean") 4-dimensional space with all "+"'s in its metric, is by "Wick rotation": Start with Euclidean space, and redefine the 4th space coordinate by a factor of i. In many cases one can relate the two directly that way, although sometimes one has to do a continuous rotation in the complex plane.

The symmetry group of special relativity is the "Poincaré group", and includes all transformations that leave this new "length", the proper time, invariant. This includes not only spacetime translations, but the generalization of rotations, the "Lorentz transformations": They are still affine transformations as before, but now

$$A\eta A^T = \eta, \qquad a\eta = -\eta a^T$$

The Lorentz transformations are thus a generalization of an orthogonal group, namely O(3,1), indicating 3 +'s and 1 $-$ in the metric. The Poincaré group is a further generalization, IO(3,1), where the "I" (for "inhomogeneous") indicates affine transformations. A "(4-)vector" is defined as anything that transforms as $V' = VA$ under Lorentz transformations, but is invariant under translations: Thus dx^m is a vector, although x^m itself isn't.

Clearly Lorentz transformations include ordinary rotations, as $A_i{}^j$. In terms of the infinitesimal transformations, we see $a_i{}^j$ satisfies the usual antisymmetry (so 3 rotations), while $a_0{}^i = a_i{}^0$ mix space and time: They are the 3 "Lorentz boosts". To relate to the Galilean boosts, we need to take a nonrelativistic limit. For this purpose it's useful to re-introduce c, by replacing $t \to ct$, and take the limit $c \to \infty$. We then find

$$\delta x^i = cta_0{}^i, \qquad \delta ct = x^i a_i{}^0 = x^i a_0{}^i$$

To get a nontrivial limit, we need to absorb a c into $a_0{}^i$, as

$$ca_0{}^i = u^i \quad \Rightarrow \quad \delta x^i \to tu^i, \quad \delta t \to 0$$

which are exactly the Galilean boosts.

We now distinguish superscripts and subscripts. This allows us to absorb factors of the metric by using it to raise and lower indices: For any vector V,

$$V_m \equiv V^n \eta_{nm}, \quad V^m = \eta^{mn} V_n; \qquad \eta^{mn} = \eta_{mn}$$

V^m is called a "contravariant vector" and V_m is called a "covariant vector"; they differ only in that $V_0 = -V^0$. Thus dx^m is a contravariant vector, while ∂_m is a covariant vector. From the definition of O(3,1), we see

$$\delta V^m = V^n a_n{}^m, \qquad \delta V_m = V_n a^n{}_m = -a_m{}^n V_n, \qquad a_{mn} = -a_{nm}$$

The square ("norm") of any vector is then defined as

$$V^2 \equiv V^m V^n \eta_{mn} = V^m V_m$$

For any 2 vectors V and W, we can always define an inner product from a square

$$(V + W)^2 \equiv V^2 + W^2 + 2V \cdot W$$

$$\Rightarrow \quad V \cdot W = V^m W_m = V_m W^m = W \cdot V$$

which is invariant (a "scalar") under Lorentz transformations. An analogy in quantum mechanics is bras and kets, which are different kinds of vectors in Hilbert space, but have a natural inner product, and are related by complex conjugation, which changes the sign of their imaginary parts.

One significant difference caused by the minus sign is that now the "square" of a vector can have either sign:

$$V^2 \left\{ \begin{matrix} < \\ = \\ > \end{matrix} \right\} 0 : \quad \left\{ \begin{matrix} \textit{timelike} \\ \textit{lightlike/null} \\ \textit{spacelike} \end{matrix} \right.$$

These 3 types of vectors are unrelated by Lorentz transformations, since their square is invariant. In particular, we see that $ds^2 = -(dx)^2$ is positive when dt^2 is bigger than $(dx^i)^2$, i.e., along a curve in Minkowski space describing motion with speed less than that of light. So, proper time is real in such realistic situations: From experiment, and from quantum field theory, we know that nothing can travel faster than light. However, proper time is constant along the path of light itself ($ds^2 = 0$): Although proper time is a useful variable for describing motion "Lorentz covariantly", an alternative will be needed for describing the motion of light.

Again there are Lorentz transformations that can't be obtained continuously from the identity, for example: (1) time reversal, changing the sign of just the time, and (2) "parity", changing the sign of all 3 spatial coordinates. Neither one is in SO(3,1), although their product is. However, we will usually be sloppy, and use SO(3,1) to refer to just the proper Lorentz group. All improper Lorentz transformations can be obtained from proper ones by multiplying with time reversal and/or parity. (Time reversal and parity are not related by proper Lorentz transformations, since one involves reflections in a timelike direction and the other in spacelike directions. Reflection in lightlike directions is not a Lorentz transformation.) With respect to proper Lorentz transformations, we can further classify timelike and lightlike vectors as "forward" and "backward", since there is no way to continuously "rotate" a vector from forward to backward without it being spacelike ("sideways"), so only spacelike vectors can have their time component change sign continuously. Time reversal and parity are not symmetries of nature, except in certain approximations; however, their product is.

For many applications it is more convenient to use a different set (or basis) of coordinates: Rather than the above "orthonormal" basis, one can use a "lightcone" basis:

$$x^{\pm} = \tfrac{1}{\sqrt{2}}(x^0 \pm x^1) \quad \Rightarrow \quad -ds^2 = -2dx^+dx^- + dy^2 + dz^2$$

$$\eta_{mn} = \begin{array}{c} \\ + \\ - \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} + & - & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

("Lightcone" is an unfortunate but standard misnomer, having nothing to do with cones in most usages.) A special lightcone basis is the "null basis",

$$x^t = \tfrac{1}{\sqrt{2}}(y - iz), \quad \bar{x}^t = \tfrac{1}{\sqrt{2}}(y + iz) \quad \Rightarrow \quad -ds^2 = -2dx^+dx^- + 2dx^t\overline{dx}^t$$

$$\eta_{mn} = \begin{array}{c} \\ + \\ - \\ t \\ \bar{t} \end{array} \begin{array}{c} \begin{array}{cccc} + & - & t & \bar{t} \end{array} \\ \begin{pmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

where x^t is complex and the "—" means complex conjugate.

Exercise 1.2.1

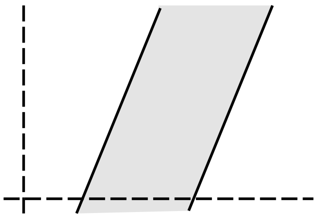
Consider relativity in two dimensions (one space, one time): Show that SO(1,1) is represented in lightcone coordinates by

$$x'^{+} = \Lambda x^{+}, \qquad x'^{-} = \Lambda^{-1} x^{-}$$

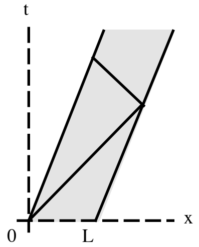
for some (nonvanishing) real number Λ. Write this one Lorentz transformation, in analogy to exercise 1.1.1a on rotations in two space dimensions, in terms of an analog of the angle ("rapidity") for those transformations that can be obtained continuously from the identity. Do the relativistic analog of exercise 1.1.2.

1.3. Examples

There are two standard examples of relativistic effects on geometry. Without loss of generality we can consider 2 dimensions, by considering motion in just 1 spatial direction. One example is called "Lorentz-Fitzgerald contraction": Consider a finite-sized object moving with constant velocity. In our 2D space, this looks like 2 parallel lines, representing the endpoints:



(In higher dimensions, this represents a one-spatial-dimensional object, like a thin ruler, moving in the direction of its length.) If we were in the "rest frame" of this object, the lines would be vertical. In that frame, there is a simple physical way to measure the length of the object: Send light from a clock sitting at one end to a mirror sitting at the other end, and time how long it takes to make the round trip. A clock measures something physical, namely the proper time $T \equiv \int \sqrt{ds^2}$ along its "world-line" (the curve describing its history in spacetime). Since ds^2 is by definition the same in any frame, we can calculate this quantity in our frame. The picture is



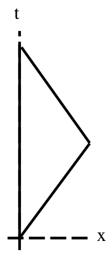
In this 2D picture lightlike lines are always slanted at ±45°. The 2 lines representing the ends of the object are (in this frame) $x = vt$ and $x = L + vt$. Some simple geometry then gives

$$T = \frac{2L}{\sqrt{1-v^2}} \qquad \Rightarrow \qquad L = \sqrt{1-v^2}\, T/2$$

This means that the length L we measure for the object is *shorter* than the length $T/2$ measured in the object's rest frame by a factor $\sqrt{1-v^2} < 1$. Unlike T, the L we have defined is not a physical property of the object: It depends on both the object and our velocity with respect to it. There is a direct analogy for rotations: We can easily define an infinite strip of constant width in terms of 2 parallel lines (the ends), where the width is *defined* by measuring along a line perpendicular to the ends. If we instead measure at an arbitrary angle to the ends, we won't find the width, but the width times a factor depending on that angle.

The most common point of confusion about relativity is that events that are simultaneous in one reference frame are not simultaneous in another (unless they are at the same place, in which case they are the same event). A frequent example is of this sort: You have too much junk in your garage, so your car won't fit anymore. So your spouse/roommate/whatever says, "No problem, just drive it near the speed of light, and it will Lorentz contract to fit." So you try it, but in your frame inside the car you find it is the garage that has contracted, so your car fits even worse. The real question is, "What happens to the car when it stops?" The answer is, "It depends on when the front end stops, and when the back end stops." You might expect that they stop at the same time. That's probably wrong, but assuming it's true, we have (at least) two possibilities: (1) They stop at the same time as measured in the garage's reference frame. Then the car fits. However, in the car's frame (its initial fast frame), the front end has stopped first, and the back end keeps going until it smashes into the front enough to make it fit. (2) They stop at the same time in the car's frame. In the garage's frame, the back end of the car stops first, and the front end keeps going until it smashes out the back of the garage.

The other standard example is "time dilation": Consider two clocks. One moves with constant velocity, so we choose the frame where it is at rest. The other moves at constant *speed* in this frame, but it starts at the position of the first clock, moves away, and then returns. (It is usually convenient to compare two clocks when they are at the same point in space, since that makes it unambiguous that one is reading the two clocks at the same time.) The picture is



A simple calculation shows that when the moving clock returns it measures a time that is *shorter* by a factor of $\sqrt{1-v^2}$. Of course, this also has a Newtonian analog: Curves between two given points are longer than straight lines. For relativity, straight lines are always the *longest* timelike curves because of the funny minus sign in the metric.

Exercise 1.3.1

You are standing in the road, and a police car comes toward you, flashing its lights at regular intervals. It runs you down and keeps right on going, as you watch it continue to flash its lights at you at the same intervals (as measured by the clock in the car). Treat this as a two-dimensional problem (one space, one time), and approximate the car's velocity as constant. Draw the Minkowski-space picture (including you, the car, and the light rays). If the car moves at speed v and flashes its lights at intervals t_0 (as measured by the clock's car), at what intervals (according to your watch) do you see the lights flashing when it is approaching, and at what intervals as it is leaving?

1.4. Experiments

The Galilean group is a symmetry of particles moving at speeds small compared to light, but electromagnetism is symmetric under the Poincaré group (actually the conformal group). This caused some confusion historically: Since the two groups have only translations and rotations in common, it was assumed that nature was invariant under no velocity transformation (neither Galilean nor Lorentz boost). In particular, the speed of light itself would seem to depend on the reference frame, since the laws of nature would be correct only in a “rest frame”. To explain “at rest with respect to what,” physicists invented something that is invariant under rotations and space and time translations, but not velocity transformations, and called this “medium” for wave propagation the “ether,” probably because they were only semiconscious at the time. (The idea was supposed to be like sound traveling through the air, although nobody had ever felt an ethereal wind.)

Many experiments were performed to test the existence of the ether, or at least to show that the wave equation for light was correct only in references frames at rest. So as not to keep you in suspense, we first tell you the general result was that the ether theory was wrong. On the contrary, one finds that the speed of light in a vacuum is measured as c in both of two reference frames that are moving at constant velocity with respect to each other. This means that electromagnetism is right and Newtonian mechanics is wrong (or at least inaccurate), since Maxwell’s equations are consistent with the speed of light being the same in all frames, while Newtonian mechanics is not consistent with any speed being the same in all frames.

The first such experiment was performed by A.A. Michelson and E.W. Morley in 1887. They measured the speed of light in various directions at various times of year to try to detect the effect of the Earth’s motion around the sun. They detected no differences, to an accuracy of 1/6th the Earth’s speed around the sun ($\approx 10^{-4}c$). (The method was interferometry: seeing if a light beam split into perpendicular paths of equal length interfered with itself.)

Another interesting experiment was performed in 1971 by J.C. Hafele and R. Keating, who compared synchronized atomic clocks, one at rest with respect to the Earth’s surface, one carried by plane (a commercial airliner) west around the world, one east. Afterwards the clocks disagreed in a way predicted by the relativistic effect of time dilation.

Probably the most convincing evidence of special relativity comes from experiments related to atomic, nuclear, and particle physics. In atoms the speed of the electrons is of the order of the fine structure constant ($\approx$1/137) times c, and the corresponding effects on atomic energy levels and such is typically of the order of the square of that ($\approx 10^{-4}$), well within the accuracy of such experiments. In particle accelerators (and also cosmic rays), various particles are accelerated to over 99% c, so relativistic effects are exaggerated to the point where particles act more like light waves than Newtonian particles. In nuclear physics the relativistic relation between mass and energy is demonstrated by nuclear decay where, unlike Newtonian mechanics, the sum of the (rest) masses is not conserved; thus the atomic bomb provides a strong proof of special relativity (although it seems like a rather extreme way to prove a point).

Special relativity is so fundamental a part of physics that in some areas of physics *every* experiment is more evidence for it, so that the many early experimental tests of it are more of historical interest than scientific. Therefore, in the rest of this course we will concentrate on the concepts of the theory, and the many differences from Newtonian physics. We will discuss mostly classical physics (mechanics and field theory), but many properties necessarily relate to quantum theory in that one of the basic ideas of quantum mechanics is the relation between particles and waves (fields).

1.5. Conformal

The conformal group is the enlargement of the Poincaré group that preserves only angles, not distances. It consists of the coordinate transformations that scale the proper time by some function:

$$x^m \to x'^m(x) \quad \Rightarrow \quad dx^2 \to dx'^2 = \xi(x'(x))dx^2$$

and thus have a similar effect on the inner product of any 2 vectors. In Euclidean space, we can define angles in terms of inner products: The cosine of the angle between 2 vectors is given by $V \cdot W/\sqrt{V^2W^2}$. We can use a similar definition in Minkowski space, if we don't worry about *cos* vs. *cosh*, etc. Another way to describe conformal symmetry in Minkowski space is as the symmetry that leaves invariant lightlike lines, $dx^2 = 0$. In fact, they are a symmetry of Maxwell's equations in 4 dimensions.

Exercise 1.5.1

Find the conformal group explicitly in two dimensions, and show it's infinite dimensional (not just the SO(2,2) described below). (Hint: Use lightcone coordinates.)

Although conformal symmetry is not observed in nature, it is important for many reasons: For example,

(1) It is useful in the construction of all free field theories. The equations satisfied by conformal theories are easier to find; more general lightlike theories are then found by relaxing some of the conditions. Furthermore, all waves that propagate at less than the speed of light can be formulated in terms of interacting theories of lightlike waves.

(2) It is often convenient to treat arbitrary theories as conformal theories with the conformal invariance broken in some simple way. This is particularly true for the case of gravity. (We'll apply this method to cosmology later.)

(3) Conformal symmetry is also important in finding and classifying solutions to interacting field theories, since at least some terms in the field equations are conformally covariant, so corresponding solutions are related by conformal transformations.

(4) The only physical quantum field theories are ones that are conformal at high energies. The quantum corrections to conformal invariance at lower energies are relatively simple.

This symmetry can be made manifest by starting with a space with one extra space and time dimension:

$$y^A = (y^a, y^+, y^-) \quad \Rightarrow \quad y^2 = y^A y^B \eta_{AB} = (y^a)^2 - 2y^+y^-$$

where $(y^a)^2 = y^a y^b \eta_{ab}$ uses the usual D-dimensional Minkowski-space metric η_{ab}, and the two additional dimensions have been written in a lightcone basis (not to be confused with the similar basis that can be used for the Minkowski metric itself). With respect to this metric, the original SO(D−1,1) Lorentz symmetry has been enlarged to SO(D,2). This is the conformal group in D dimensions. However, rather than also preserving (D+2)-dimensional translation invariance, we instead impose the constraint and invariance

$$y^2 = 0, \quad \delta y^A = \zeta(y) y^A$$

This reduces the original space to the "projective" (invariant under the ζ scaling) lightcone (which in this case really is a cone).

These two conditions can be solved by

$$y^A = e w^A, \quad w^A = (x^a, 1, \tfrac{1}{2} x^a x_a)$$

Projective invariance then means independence from e (y^+), while the lightcone condition has determined y^-. $y^2 = 0$ implies $y \cdot dy = 0$, so the simplest conformal invariant is

$$dy^2 = (e dw + w de)^2 = e^2 dw^2 = e^2 dx^2$$

where we have used $w^2 = 0 \Rightarrow w \cdot dw = 0$. This means any SO(D,2) transformation on y^A will simply scale dx^2, and scale e^2 in the opposite way:

$$dx'^2 = \left(\frac{e^2}{e'^2} \right) dx^2$$

in agreement with the previous definition of the conformal group.

The explicit form of conformal transformations on $x^a = y^a/y^+$ now follows from their linear form on y^A:

$$y'^A = y^B A_B{}^A, \qquad \delta y^A = y^B a_B{}^A$$

For example, a_{+-} just scales x^a. (Scale transformations are also known as "dilatations".) We can also recognize a_{+a} as generating translations on x^a. The only complicated transformations are generated by a_{-a}, known as "conformal boosts" (acceleration transformations). Because they contribute to just one side of the matrix a (like translations), it's easy to exponentiate to find the finite transformations:

$$y' = ye^a, \qquad only\ a_{-a} = -a_{a-} \neq 0$$

Since the conformal boosts act as "lowering operators" for scale weight ($+ \to a \to -$), only the first three terms in the exponential survive:

$$(ya)^- = 0, \qquad (ya)^a = y^- a_-{}^a, \qquad (ya)^+ = y^a a_a{}^+ = y_a a_-{}^a$$

$$\Rightarrow \quad y'^- = y^-,\ y'^a = y^a + a_-{}^a y^-,\ y'^+ = y^+ + a_-{}^a y_a + \tfrac{1}{2}(a_-{}^a)^2 y^-$$

$$\Rightarrow \quad x'^a = \frac{x^a + \frac{1}{2}a_-{}^a x^2}{1 + a_-{}^a x_a + \frac{1}{4}(a_-{}^a)^2 x^2}$$

using $x^a = y^a/y^+$, $y^-/y^+ = \frac{1}{2}x^2$ (and $\eta_{+-} = -1$).

We actually have the full O(D,2) symmetry: Besides the continuous symmetries, and the discrete ones of SO(D−1,1), we have a second "time" reversal (from our second time dimension):

$$y^+ \leftrightarrow -y^- \quad \Rightarrow \quad x^a \leftrightarrow -\frac{x^a}{\frac{1}{2}x^2}$$

This transformation is called an "inversion".

Exercise 1.5.2

Show that a finite conformal boost can be obtained by performing a translation sandwiched between two inversions.

Exercise 1.5.3

The conformal group for Euclidean space (or any spacetime signature) can be obtained by the same construction. Consider the special case of D=2 for these SO(D+1,1) transformations. (This is a subgroup of the 2D conformal group: See exercise 1.5.1.) Use complex coordinates for the two "physical" dimensions:

$$z = \tfrac{1}{\sqrt{2}}(x^1 + ix^2)$$

a Show that the inversion is

$$z \leftrightarrow -\frac{1}{z^*}$$

b Show that the conformal boost is (using a complex number $v = \frac{1}{\sqrt{2}}(v^1 + iv^2)$ also for the boost vector)

$$z \to \frac{z}{1 + v^* z}$$

2. Spin

2.1. Nonrelativistic tensors

Usually nonrelativistic physics is written in matrix or Gibbs' notation. This is insufficient even for 19th century physics: We can write a column or row vector p for momentum, and a matrix T for moment of inertia, but how do we write in that notation more general objects? These are different representations of the rotation group: We can write how each transforms under rotations:

$$p' = pA, \qquad T' = A^T T A$$

The problem is to write *all* representations.

One alternative is used frequently in quantum mechanics: A scalar is "spin 0", a vector is "spin 1", etc. Spin s has $2s+1$ components, so we can write a column "vector" with that many components. For example, moment of inertia is a symmetric 3×3 matrix, and so has 6 components. It can be separated into its trace S and traceless pieces R, which don't mix under rotations:

$$T = R + \tfrac{1}{3}SI, \quad tr(T) = S, \quad tr(R) = 0$$

$$\Rightarrow \quad tr(T') = tr(A^T T A) = tr(AA^T T) = tr(T) \quad \Rightarrow \quad tr(R') = 0, \quad S' = S$$

using the cyclicity of the trace. Thus the "irreducible" parts of T are the scalar s and the spin-2 (5 components) R. But if we were to write R as a 5-vector, it would be a mess to relate the 5×5 matrix that rotates it to the 3×3 matrix A, and even worse to write a scalar like pRp^T in terms of 2 3-vectors and 1 5-vector. (In quantum mechanics, this is done with "Clebsch-Gordan-Wigner coefficients".)

The simplest solution is to use indices. Then it's easy to write an object of arbitrary integer spin s as a generalization of what we just did for spins 0,1,2: It has s 3-vector indices, in which it is totally (for any 2 of its indices) symmetric and traceless:

$$T^{i_1 ... i_s}: \qquad T^{...i...j...} = T^{...j...i...}, \quad T^{...i...j...}\delta_{ij} = 0$$

and it transforms as the product of vectors:

$$T'^{i_1 ... i_s} = T^{j_1 ... j_s} A_{j_1}{}^{i_1} ... A_{j_s}{}^{i_s}$$

In this "tensor notation" there are 2 special tensors, directly related to the "S" and "O" of SO(3): We have the identities

$$\delta'^{ij} = \delta^{kl} A_k{}^i A_l{}^j = \delta^{ij}, \qquad \epsilon'^{ijk} = \epsilon^{lmn} A_l{}^i A_m{}^j A_n{}^k = \epsilon^{ijk} det(A) = \epsilon^{ijk}$$

(and similarly for the tensors with indices down). Here ϵ^{ijk} is the (3D) "Levi-Civita tensor", which is totally antisymmetric. (In D dimensions, it has D indices.) This means its only nonvanishing components are when all the indices take different values; we then normalize by choosing $\epsilon^{123} = 1$. The expression for its transformation law is one of the definitions of the determinant, and thus its invariance under proper rotations is the statement $det(A) = 1$; under improper rotations it changes sign. (The invariance of the Kronecker δ is the statement of orthogonality of A.) As a consequence, we can always replace any pair of antisymmetric indices on a tensor with a single index: e.g.,

$$a^{ij} = -a^{ji} \quad \Rightarrow \quad a_i \equiv \tfrac{1}{2}\epsilon_{ijk}a^{jk}, \quad a^{ij} = \epsilon^{ijk}a_k$$

where we have used the identity

$$\epsilon^{ijk}\epsilon_{lmn} = \delta^i_{[l}\delta^j_m\delta^k_{n]}$$

where "[ ]" means to antisymmetrize in the indices lmn, i.e., sum over the 6 permutations of the indices with minus signs for odd permutations. (This is easy to derive by noting that anything antisymmetric in 3 3-valued indices must be proportional to the Levi-Civita tensor in those indices.) This is often done with the angular momentum

$$J^{ij} \equiv x^{[i}p^{j]} = x^i p^j - x^j p^i \quad \Rightarrow \quad J_i \equiv \epsilon_{ijk}x^j p^k$$

(We can also recognize this as the definition of the cross product.) Since J_i changes sign under parity, it is called an "axial vector" (unlike the usual "polar vectors"). Similarly, since δ^{ij} is a tensor, we can always use it to separate the trace and traceless part of a tensor with respect to any 2 symmetric indices. The end result is that the only irreducible tensors can be written as totally symmetric traceless tensors.

2.2. SU(2)

Since the earliest days of quantum mechanics, we know that half-integer spins also exist, in nature as well as group theory, e.g., the electron and proton. (Unfortunately, almost all relativity textbooks have not been updated since 1916 or so, and thus ignore this topic even in special relativity.) This might be expected to complicate matters, but actually simplifies them, due to the well-known identity

$$\tfrac{1}{2} < 1$$

This means that a "spinor", describing spin 1/2, has only 2 components, compared to the 3 components of a vector, and its matrices (e.g., for rotations) are thus only 2×2 instead of 3×3.

We first consider spinors in matrix notation, then generalize to “spinor notation” (spinor indices). The simplest way to understand why rotations can be represented as 2×2 instead of 3×3 is to see why 3-vectors themselves can be understood as 2×2 matrices, which for some purposes is simpler. (This is equivalent to Hamilton’s “quaternions”, which predated Gibbs’ vector notation, and were used by Maxwell for his equations.)

Consider such matrices to be “hermitian”, which is natural from the quantum mechanical point of view. Then they have four real components, one too many for a 3-vector (but just right for a relativistic 4-vector), so we restrict them to also be traceless:

$$V = V^\dagger \equiv V^{*T}, \quad tr\ V = 0$$

The simplest way to get a single number out of a matrix, besides taking the trace, is to take the determinant. By expanding a general matrix identity to quadratic order we find an identity for 2×2 matrices

$$det(I+M) = e^{tr\,ln(I+M)} \quad \Rightarrow \quad -2\ det\ M = tr(M^2) - (tr\ M)^2$$

It is then clear that in our case $-det\ V$ is positive definite, as well as quadratic, so we can define the norm of this 3-vector as

$$|V|^2 = -2\ det\ V = tr(V^2)$$

This can be compared easily with conventional notation by picking a basis:

$$V = \tfrac{1}{\sqrt{2}} \begin{pmatrix} V^1 & V^2 - iV^3 \\ V^2 + iV^3 & -V^1 \end{pmatrix} = \vec{V}\cdot\vec{\sigma} \quad \Rightarrow \quad det\ V = -\tfrac{1}{2}(V^i)^2$$

where $\vec{\sigma}$ are the Pauli σ matrices, up to normalization. (The $\sqrt{2}$ is an arbitrary normalization factor, which we could eliminate here, but would just pop up someplace else.) As usual, the inner product follows from the norm:

$$|V+W|^2 = |V|^2 + |W|^2 + 2V\cdot W$$

$$\Rightarrow \quad V\cdot W = det\ V + det\ W - det(V+W) = tr(VW)$$

Some other useful identities for 2×2 determinants are

$$MCM^TC = I\ det\ M, \qquad M^{-1} = CM^TC(det\ M)^{-1}$$

where we now use the imaginary, hermitian matrix

$$C = \left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right)$$

(These can easily be derived from more general identities for determinants, but for 2×2 matrices the algebra is trivial to check directly.) If we make the replacement $M \to e^M$ and expand to linear order in M, we find

$$M + CM^T C = I\ tr\ M$$

This implies

$$tr\ V = 0 \quad \Leftrightarrow \quad (VC)^T = VC$$

i.e., the tracelessness of V is equivalent to the symmetry of VC. Furthermore, the combination of the trace and determinant identities tell us

$$M^2 = M\ tr\ M - I\ det\ M \quad \Rightarrow \quad V^2 = -I\ det\ V = I\tfrac{1}{2}|V|^2$$

Here by "V^2" we mean the square of the matrix, while "$|V|^2$"$= (V^i)^2$ is the square of the norm (neither of which should be confused with the component $V^2 = V^i \delta_i^2$.)

Again expressing the inner product in terms of the norm, we then find the "anticommutator"

$$\{V, W\} \equiv VW + WV = (V \cdot W)I$$

Also, since the commutator of two finite matrices is traceless, and picks up a minus sign under hermitian conjugation, we can define an outer product (vector×vector = vector) by

$$[V, W] \equiv VW - WV = \sqrt{2} i V \times W$$

Therefore, the cross product is a special case (2×2) of the commutator. Combining these two results,

$$VW = \tfrac{1}{2}(V \cdot W)I + \tfrac{1}{\sqrt{2}} i V \times W$$

In other words, the product of two traceless hermitian 2×2 matrices gives a real trace piece, symmetric in the two matrices, plus an antihermitian traceless piece, antisymmetric in the two. Thus, we have a simple relation between the matrix product, the inner ("dot") product and the outer ("cross") product.

Exercise 2.2.1

Check this result in two ways:

a Show the normalization agrees with the usual outer product. Using only the above definition of $V \times W$, along with $\{V, W\} = (V \cdot W)I$, show

$$-I|V \times W|^2 = ([V, W])^2 = -I[|V|^2|W|^2 - (V \cdot W)^2]$$

b Use components, with the above basis.

Exercise 2.2.2

Consider electromagnetism in 2×2 matrix notation: Define the field strength as a *complex* vector $F = \sqrt{2}(E + iB)$. Write partial derivatives as the sum of a (rotational) scalar plus a (3-)vector as $\partial = \frac{1}{\sqrt{2}} I\partial_t + \nabla$, where $\partial_t = \partial/\partial t$ is the time derivative and ∇ is the partial space derivatives written as a traceless matrix. Do the same for the charge density ρ and (3-)current j as $J = -\frac{1}{\sqrt{2}} I\rho + j$. Using the definition of dot and cross products in terms of matrix multiplication as discussed in this section, show that the simple matrix equation $\partial F = -J$, when separated into its trace and traceless pieces, and its hermitian and antihermitian pieces, gives the usual Maxwell equations

$$\nabla \cdot B = 0, \quad \nabla \cdot E = \rho, \quad \nabla \times E + \partial_t B = 0, \quad \nabla \times B - \partial_t E = j$$

(Note: Avoid the Pauli σ-matrices and explicit components.)

Since vectors are hermitian, we expect their transformations to be “unitary”:

$$V' = UVU^\dagger, \qquad U^\dagger = U^{-1}$$

It is easily checked that this preserves the properties of these matrices:

$$(V')^\dagger = (UVU^\dagger)^\dagger = V', \qquad tr(V') = tr(UVU^{-1}) = tr(U^{-1}UV) = tr(V) = 0$$

Furthermore, it also preserves the norm (and thus the inner product):

$$det(V') = det(UVU^{-1}) = det(U)det(V)(det\ U)^{-1} = det\ V$$

Unitary 2×2 matrices have 4 parameters; however, we can elimimate one by the condition

$$det\ U = 1$$

This eliminates only the phase factor in U, which cancels out in the transformation law anyway. This is the group SU(2) (“U” for unitary). Taking the product of two rotations now involves multiplying only 2×2 matrices, and not 3×3 matrices.

We can also write U in exponential notation, which is useful for going to the infinitesimal limit:

$$U = e^{iG} \quad \Rightarrow \quad G^\dagger = G, \quad tr\ G = 0$$

This means that G itself can be considered a vector. Rotations can be parametrized by a vector whose direction is the axis of rotation, and whose magnitude is $(1/\sqrt{2}\times)$ the angle of rotation:

$$V' = e^{iG} V e^{-iG} \quad \Rightarrow \quad \delta V = i[G, V] = -\sqrt{2} G \times V$$

The hermiticity condition on V can also be expressed as a reality condition:

$$V = V^\dagger \quad and \quad tr\ V = 0 \quad \Rightarrow \quad V^* = -CVC, \quad (VC)^* = C(VC)C$$

where " * " is the usual complex conjugate. A similar condition for U is

$$U^\dagger = U^{-1} \quad and \quad det\ U = 1 \quad \Rightarrow \quad U^* = CUC$$

which is also a consequence of the fact that we can write U in terms of a vector as $U = e^{iV}$. As a result, the transformation law for the vector can be written in terms of VC in a simple way, which manifestly preserves its symmetry:

$$(VC)' = UVU^{-1}C = U(VC)U^T$$

2.3. Nonrelativistic spinors

Note that the mapping of SU(2) to SO(3) is two-to-one: This follows from the fact $V' = V$ when U is a phase factor. We eliminated continuous phase factors from U by the condition $det\ U = 1$, which restricts U(2) to SU(2). However,

$$det(Ie^{i\theta}) = e^{2i\theta} = 1 \quad \Rightarrow \quad e^{i\theta} = \pm 1$$

for 2×2 matrices. More generally, for any SU(2) element U, $-U$ is also an element of SU(2), but acts the same way on a vector; i.e., these two SU(2) transformations give the same SO(3) transformation. Thus SU(2) is called a "double covering" of SO(3). However, this second transformation is not redundant, because it acts differently on half-integral spins.

The other convenience of using 2×2 matrices is that it makes obvious how to introduce spinors — Since a vector already transforms with two factors of U, we define a "square root" of a vector that transforms with just one U:

$$\psi' = U\psi \quad \Rightarrow \quad \psi^{\dagger\prime} = \psi^\dagger U^{-1}$$

where ψ is a two-component "vector", i.e., a 2×1 matrix (column vector). The complex conjugate of a spinor then transforms in essentially the same way:

$$(C\psi^*)' = CU^*\psi^* = U(C\psi^*)$$

Note that the antisymmetry of C implies that ψ must be complex: We might think that, since $C\psi^*$ transforms in the same way as ψ, we can identify the two consistently with the transformation law. But then we would have

$$\psi = C\psi^* = C(C\psi^*)^* = CC^*\psi = -\psi$$

Thus the representation is "pseudoreal". The fact that $C\psi^*$ transforms the same way under rotations as ψ leads us to consider the transformation

$$\psi' = C\psi^*$$

Since a vector transforms the same way under rotations as $\psi\psi^\dagger$, under this transformation we have

$$V' = CV^*C = -V$$

which identifies it as a reflection.

Another useful way to write rotations on ψ (like looking at VC instead of V) is

$$(\psi^T C)' = (\psi^T C)U^{-1}$$

This tells us how to take an invariant inner product of spinors:

$$\psi' = U\psi, \quad \chi' = U\chi \quad \Rightarrow \quad (\psi^T C\chi)' = (\psi^T C\chi)$$

In other words, C is the "metric" in the space of spinors. An important difference of this inner product from the familiar one for three-vectors is that it is antisymmetric. It can thus be used to raise and lower spinor indices, like the metric for vectors: In fact, we have already done that in defining the norm of a vector as a matrix with 2 spinor indices (using 2 C's). (We could also do that for a spinor, but it would have to be an "anticommuting c-number", describing a classical fermion, which we won't discuss here.) Of course, since rotations are unitary, we also have the usual $\psi^\dagger\psi$ as an invariant, positive definite, inner product.

Exercise 2.3.1

Consider a hermitian but *not* traceless 2×2 matrix M ($M = M^\dagger$, $tr\ M \neq 0$).

a Show

$$det\ M = 0 \quad \Rightarrow \quad M = \pm\psi \otimes \psi^\dagger$$

for some spinor (column vector) ψ (and some sign $\pm$).

b Define a vector by

$$V = \sqrt{2}(M - \tfrac{1}{2}I\ tr\ M)$$

Show $|V|$ (*not* $|V|^2$) is simply $\psi^\dagger\psi$.

Like vectors, we can use indices for spinors:

$$\psi^\alpha, \qquad \alpha = (\oplus, \ominus)$$

(We use $\oplus, \ominus$ to distinguish from 1,2,... for vectors, $+, -$ for lightcone vectors,...) Then we can also construct tensors from spinors by using multiple 2-spinor indices, but now they will include both integer and half-integer spin. There is a further advantage of spinor indices over vector indices: The Levi-Civita tensor for spinors is *the same* as the metric (up to normalization), since both are antisymmetric in their 2 2-spinor indices. In particular, antisymmetrizing in 2 2-valued indices is the same as contracting the indices with the 2-index Levi-Civita tensor, since anything antisymmetric in 2 2-valued indices must be equal to the Levi-Civita tensor times the corresponding object not carrying those 2 indices:

$$C_{\alpha\beta} = -C_{\beta\alpha} = -C^{\alpha\beta} = C^{\beta\alpha} = \left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right)$$

$$\Rightarrow \quad A_{[\alpha\beta]} = C_{\alpha\beta} C^{\gamma\delta} A_{\gamma\delta}$$

The end result is that all irreducible representations of rotations can be obtained as objects totally symmetric in their indices, where the number of indices is just $2s$:

$$T^{\alpha_1 \dots \alpha_{2s}} : \qquad T^{\dots\alpha\dots\beta\dots} = T^{\dots\beta\dots\alpha\dots}$$

The counting of components is trivial here: Since the order of indices doesn't matter, we just count the number of "$\oplus$" indices, which is any integer from 0 to $2s$, thus $2s+1$ components. (We have already studied the vector case, $s = 1$.)

2.4. Relativistic spinors

Tensor notation is more complicated in 4 dimensions than 3, since now the Levi-Civita tensor has 4 indices, so now 2 antisymmetric indices are converted by it back into 2 antisymmetric indices (although now 3 get converted to 1). Thus we need to consider sets of pairs of antisymmetric indices in addition to sets of single vector indices, and consider some kind of symmetrization, to get tensors irreducible under the Lorentz group. We won't consider this method for finding irreducible representations of the Lorentz group here, since again the spinor method is simpler.

Consider now a 2×2 matrix, whose elements we label as

$$(V)^{\alpha\dot{\beta}} = \begin{pmatrix} V^{\oplus\dot{\oplus}} & V^{\oplus\dot{\ominus}} \\ V^{\ominus\dot{\oplus}} & V^{\ominus\dot{\ominus}} \end{pmatrix} = \begin{pmatrix} V^{+} & V^{t*} \\ V^{t} & V^{-} \end{pmatrix}$$

$$= \tfrac{1}{\sqrt{2}} \begin{pmatrix} V^0 + V^1 & V^2 + iV^3 \\ V^2 - iV^3 & V^0 - V^1 \end{pmatrix} = V^a (\sigma_a)^{\alpha\dot{\beta}}$$

which we choose to be hermitian,

$$V = V^\dagger \quad \Rightarrow \quad V^{\alpha\dot{\beta}} = (V^\dagger)^{\alpha\dot{\beta}} \equiv (V^{\beta\dot{\alpha}})^*$$

where we distinguish the right spinor index by a dot because it will be chosen to transform differently from the left one. For comparison, lowering both spinor indices with the matrix C as for SU(2), and the vector indices with the Minkowski metric (in either the orthonormal or null basis, as appropriate), we find another hermitian matrix

$$(V)_{\alpha\dot\beta} = \begin{pmatrix} V_+ & V_t{}^* \\ V_t & V_- \end{pmatrix} = \tfrac{1}{\sqrt{2}} \begin{pmatrix} V_0 + V_1 & V_2 - iV_3 \\ V_2 + iV_3 & V_0 - V_1 \end{pmatrix} = V_a (\sigma^a)_{\alpha\dot\beta}$$

In the orthonormal basis, σ_a are the Pauli matrices and the identity, up to normalization. They are also the Clebsch-Gordan-Wigner coefficients for spinor⊗spinor = vector. In the null basis, they are completely trivial: 1 for one element, 0 for the rest, the usual basis for matrices. In other words, they are simply an arbitrary way (according to choice of basis) to translate a 2×2 (hermitian) matrix into a 4-component vector. We will sometimes treat a vector index "a" as an abbreviation for a spinor index pair "$\alpha\dot\alpha$":

$$V^a = V^{\alpha\dot\alpha}, \quad a = \alpha\dot\alpha = (\oplus\dot\oplus, \oplus\dot\ominus, \ominus\dot\oplus, \ominus\dot\ominus) = (+, \bar t, t, -)$$

where α and $\dot\alpha$ are understood to be independent indices ($\oplus \neq \dot\oplus$, etc.).

Examining the determinant of (either version of) V, we find the correct Minkowski norms:

$$-2\ det\ V = -2V^+V^- + 2V^tV^{t*} = -(V^0)^2 + (V^1)^2 + (V^2)^2 + (V^3)^2 = V^2$$

Thus Lorentz transformations will be those that preserve the hermiticity of this matrix and leave its determinant invariant:

$$V' = gVg^\dagger, \quad det\ g = 1$$

($det\ g$ could also have a phase, but that would cancel in the transformation.) Thus g is an element of "SL(2,C)". ("C" means complex, while "L" means linear, i.e., neither orthogonal nor unitary.) We can again use exponentials:

$$g = e^G, \quad tr\ G = 0$$

Thus the group space is 6-dimensional (G has three independent complex components), the same as SO(3,1).

Exercise 2.4.1

Consider some modifications of this definition of a vector:

a Instead of hermitian, consider a *real* 2×2 matrix. Again defining the norm in terms of the determinant, what changes? What is now the symmetry group?

b Instead of hermitian or real, consider the reality condition

$$V^* = CVC$$

Now what norm and symmetry group does the determinant give? Compare this to the conditions satisfied by an element of SU(2) as described previously.

In index notation, we write for this vector

$$V'_{\alpha\dot{\beta}} = g_\alpha{}^\gamma g^*{}_{\dot{\beta}}{}^{\dot{\delta}} V_{\gamma\dot{\delta}}$$

while for a ("Weyl") spinor we have

$$\psi'_\alpha = g_\alpha{}^\beta \psi_\beta$$

The metric of the group SL(2,C) is the two-index antisymmetric symbol, which is also the metric for Sp(2,C): In our conventions,

$$C_{\alpha\beta} = -C_{\beta\alpha} = -C^{\alpha\beta} = C_{\dot{\alpha}\dot{\beta}} = \left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right)$$

We also have the identities, as for SU(2),

$$A_{[\alpha\beta]} = C_{\alpha\beta} C^{\gamma\delta} A_{\gamma\delta}, \quad A_{[\alpha\beta\gamma]} = 0$$

where the last identity essentially says that 2-valued indices can't take 3 different values. We again use the metric to raise, lower, and contract indices:

$$\psi_\alpha = \psi^\beta C_{\beta\alpha}, \quad \psi_{\dot{\alpha}} = \psi^{\dot{\beta}} C_{\dot{\beta}\dot{\alpha}}$$
$$V \cdot W = V^{\alpha\dot{\beta}} W_{\alpha\dot{\beta}}$$

Antisymmetry in vector indices also implies some antisymmetry in spinor indices. For example, the antisymmetric Maxwell field strength $F_{ab} = -F_{ba}$, after translating vector indices into spinor, can be separated into its parts symmetric and antisymmetric in undotted indices; antisymmetry in vector indices (now spinor index pairs) then implies the opposite symmetry in dotted indices:

$$F_{\alpha\dot{\gamma},\beta\dot{\delta}} = -F_{\beta\dot{\delta},\alpha\dot{\gamma}} = \tfrac{1}{4}(F_{(\alpha\beta)[\dot{\gamma}\dot{\delta}]} + F_{[\alpha\beta](\dot{\gamma}\dot{\delta})}) = \bar{C}_{\dot{\gamma}\dot{\delta}} f_{\alpha\beta} + C_{\alpha\beta} \bar{f}_{\dot{\gamma}\dot{\delta}}, \qquad f_{\alpha\beta} = \tfrac{1}{2} F_{\alpha\dot{\gamma},\beta}{}^{\dot{\gamma}}$$

Thus, an antisymmetric tensor also can be written in terms of a (complex) 2×2 matrix.

Thus, the most general irreducible (finite-dimensional) representation of SL(2,C) (and thus SO(3,1)) has an arbitrary number of dotted and undotted indices, and is totally symmetric in each: $A_{(\alpha_1 \ldots \alpha_m)(\dot{\beta}_1 \ldots \dot{\beta}_n)}$. Treating a vector index directly as a dotted-undotted pair of indices (e.g., $a = \alpha\dot{\alpha}$, which is just a funny way of labeling a 4-valued index), we can translate into spinor notation the two constant tensors of SO(3,1): Since the only constant tensor of SL(2,C) is the antisymmetric symbol, they can be expressed in terms of it:

$$\eta_{\alpha\dot{\alpha},\beta\dot{\beta}} = C_{\alpha\beta}C_{\dot{\alpha}\dot{\beta}}, \qquad \epsilon_{\alpha\dot{\alpha},\beta\dot{\beta},\gamma\dot{\gamma},\delta\dot{\delta}} = i(C_{\alpha\beta}C_{\gamma\delta}C_{\dot{\alpha}\dot{\delta}}C_{\dot{\beta}\dot{\gamma}} - C_{\alpha\delta}C_{\beta\gamma}C_{\dot{\alpha}\dot{\beta}}C_{\dot{\gamma}\dot{\delta}})$$

When we work with just vectors, these can be expressed in matrix language: For this purpose, to avoid explicit C's, we use the matrices

$$V \rightarrow V_\alpha{}^{\dot{\beta}} \quad \Rightarrow \quad V^* \rightarrow -V^\beta{}_{\dot{\alpha}}$$

We can then write

$$V \cdot W = tr(VW^*) \quad \Rightarrow \quad VW^* + WV^* = (V \cdot W)I$$

$$\epsilon_{abcd}V^aW^bX^cY^d \equiv \epsilon(V,W,X,Y) = i\ tr(VW^*XY^* - Y^*XW^*V)$$

(We have assumed real vectors; for complex vectors we should really write $V \cdot W^* = ...$, etc.)

Since we have exhausted all possible linear transformations on spinors (except for scale, which relates to conformal transformations), the only way to represent discrete Lorentz transformations is as antilinear ones:

$$\psi'_\alpha = \sqrt{2} n_\alpha{}^{\dot{\beta}} \bar{\psi}_{\dot{\beta}} \qquad (\psi' = -\sqrt{2} n \psi^*)$$

From its index structure we see that n is a vector, representing the direction of the reflection. The product of two identical reflections is then, in matrix notation

$$\psi'' = 2n(n\psi^*)^* = n^2\psi \quad \Rightarrow \quad n^2 = \pm 1$$

where we have required closure on an SL(2,C) transformation (±1). Thus n is a unit vector, either spacelike or timelike. Applying the same transformation to a vector, where $V^{\alpha\dot{\beta}}$ transforms like $\psi^\alpha\chi^{\dot{\beta}}$, we write in matrix notation

$$V' = -2nV^*n = n^2V - 2(n \cdot V)n$$

(The overall sign is ambiguous.) This transformation thus describes parity. In particular, to describe purely parity without any additional rotation (i.e., exactly reflection of the 3 spatial axes), in our basis we must choose a unit vector in the time direction,

$$\sqrt{2} n^{\alpha\dot{\beta}} = \delta^{\alpha\dot{\beta}} \quad \Rightarrow \quad \psi'^\alpha = \bar{\psi}_{\dot{\alpha}} \qquad (\bar{\psi}'^{\dot{\alpha}} = -\psi_\alpha)$$

$$\Rightarrow \quad V'^{\alpha\dot{\beta}} = -V_{\beta\dot{\alpha}}$$

which corresponds to the usual in vector notation, since in our basis

$$\sigma_a^{\alpha\dot{\beta}} = \sigma^a_{\beta\dot{\alpha}}$$

To describe time reversal, we need a transformation that does not preserve the complex conjugation properties of spinors: For example, combined parity and time reversal is

$$\psi'^{\alpha} = \psi^{\alpha}, \quad \bar{\psi}'^{\dot{\beta}} = -\bar{\psi}^{\dot{\beta}} \quad \Rightarrow \quad V' = -V$$

(The overall sign on V is unambiguous.)

3. Actions

3.1. Equations of motion

A fundamental concept in physics, of as great importance as symmetry, is the action principle. In quantum physics the dynamics is necessarily formulated in terms of an action or an equivalent Hamiltonian. Action principles are also convenient and powerful for classical physics, allowing all field equations to be derived from a single function, and making symmetries simpler to check.

We begin with some general properties of actions. Generally, equations of motion are derived from actions by setting their variation with respect to their arguments to vanish:

$$\delta S[\phi] \equiv S[\phi+\delta\phi] - S[\phi] = \delta\phi_i \frac{\partial S[\phi_i]}{\partial \phi_i} = 0 \quad \Rightarrow \quad \frac{\partial S[\phi_i]}{\partial \phi_i} = 0$$

The solutions to this equation (find ϕ, given S) are "extrema" of the action; generally we want them to be minima, corresponding to minima of the energy, so that they will be stable under small perturbations.

Exercise 3.1.1

Often continuous coordinates are replaced with discrete ones, for calculational or conceptual purposes. Consider

$$S = -\sum_{n=-\infty}^{\infty} \tfrac{1}{2}(q_{n+1} - q_n)^2$$

The integer n is interpreted as a discrete time, in terms of some "small" unit.

a Show that

$$\delta S = 0 \quad \Rightarrow \quad q_{n+1} - 2q_n + q_{n-1} = 0$$

b Examine the continuum limit of the action and equations of motion: Introduce appropriate factors of ϵ, with $t = n\epsilon$, and take the limit $\epsilon \to 0$.

Now we take the variables ϕ to be functions of time; thus, S is a function of functions, a "functional". It just means that S is a function of an infinite set of variables. We can generalize properties of ordinary functions (derivatives, etc.) as usual by considering discrete time and taking a continuum limit:

$$i = 1, 2, ... \quad \to \quad t \in [-\infty, \infty]$$

$$\phi_i \quad \to \quad \phi(t)$$

$$\sum_i \quad \to \quad \int dt$$

$$\begin{aligned} \delta_{ij} \quad &\to \quad \delta(t-t') \\ \frac{\partial}{\partial\phi_i} \quad &\to \quad \frac{\delta}{\delta\phi(t)} \\ \int d\phi_i \quad &\to \quad \int D\phi(t) \end{aligned}$$

(the last, a "functional integral", will appear in quantum theory) where δ_{ij} is the usual Kronecker delta function, while $\delta(t-t')$ is the "Dirac delta function". It's not really a function, since it takes only the values 0 or ∞, but a "distribution", meaning it's defined only by integration:

$$\int dt'\ f(t')\delta(t-t') \equiv f(t)$$

Of course, the variable $\phi(t)$ can also carry an index (or indices). In field theory, it will also be a function of more coordinates, those of space.

For example, making these substitutions into the definition of a (partial) derivative to get a "functional derivative",

$$\frac{\partial f(\phi_i)}{\partial\phi_j} = \lim_{\epsilon\to 0}\frac{f(\phi_i+\epsilon\delta_{ij})-f(\phi_i)}{\epsilon} \quad \Rightarrow \quad \frac{\delta f[\phi(t)]}{\delta\phi(t')} = \lim_{\epsilon\to 0}\frac{f[\phi(t)+\epsilon\delta(t-t')]-f[\phi(t)]}{\epsilon}$$

Sometimes the functional derivative is defined in terms of that of the variable itself:

$$\frac{\delta\phi(t)}{\delta\phi(t')} = \delta(t-t')$$

If we apply this definition of the Dirac δ to $\delta\phi/\delta\phi$, we obtain the previous definition of the functional derivative. (Consider, e.g., varying $S = \int dt\ f(t)\phi(t)$ for a fixed function f.) However, in practice we never need to use these definitions of the functional derivative: The only thing for which we need a functional derivative is the action, whose functional derivative is defined by its variation,

$$\delta S[\phi] \equiv S[\phi+\delta\phi] - S[\phi] \equiv \int dt\ \delta\phi(t)\frac{\delta S}{\delta\phi(t)}$$

(The fact that the variation can always be written in this form is just the statement that it is linear in $\delta\phi$, since $\delta\phi$ is "infinitesimal".)

A general principle of mechanics is "locality", that events at one time directly affect only those events an infinitesimal time away. (In field theory these events can be also only an infinitesimal distance away in space.) This means that the action can be expressed in terms of a Lagrangian:

$$S[\phi] = \int dt\ L[\phi(t)]$$

where L at time t is a function of only $\phi(t)$ and a finite number of its derivatives. For more subtle reasons, this number of time derivatives is restricted to be no more than two for any term in L; after integration by parts, each derivative acts on a different factor of ϕ. The general form of the action is then

$$L(\phi) = -\tfrac{1}{2}\dot{\phi}^m \dot{\phi}^n g_{mn}(\phi) + \dot{\phi}^m A_m(\phi) + U(\phi)$$

where "$\dot{}$" means $\partial/\partial t$, and the "metric" g, "vector potential" A, and "scalar potential" U are not to be varied independently when deriving the equations of motion. (Specifically, $\delta U = (\delta\phi^m)(\partial U/\partial\phi^m)$, etc. Note that our definition of the Lagrangian differs in sign from the usual.) The equations of motion following from varying an action that can be written in terms of a Lagrangian are

$$0 = \delta S \equiv \int dt\ \delta\phi^m \frac{\delta S}{\delta\phi^m} \quad \Rightarrow \quad \frac{\delta S}{\delta\phi^m} = 0$$

where we have eliminated $\delta\dot{\phi}^m$ terms by integration by parts (assuming $\delta\phi = 0$ at the boundaries in t), and used the fact that $\delta\phi(t)$ is arbitrary at each value of t. For example,

$$S = -\int dt\ \tfrac{1}{2}\dot{q}^2 \quad \Rightarrow \quad 0 = \delta S = -\int dt\ \dot{q}\delta\dot{q} = \int dt\ (\delta q)\ddot{q} \quad \Rightarrow \quad \frac{\delta S}{\delta q} = \ddot{q} = 0$$

Exercise 3.1.2

Find the equations of motion for ϕ^m from the above general action in terms of the external fields g, A, and U (and their partial derivatives with respect to ϕ).

The Hamiltonian form of an action is only first-order in derivatives; Lagrangians are usually no more than second-order. A second-order Lagrangian can be converted into first-order by doubling the number of variables: Assuming the metric above can be inverted, we can write

$$L_H(\phi, \pi) = -\dot{\phi}^m \pi_m + H(\phi, \pi)$$

$$H = \tfrac{1}{2} g^{mn}(\phi)[\pi_m + A_m(\phi)][\pi_n + A_n(\phi)] + U(\phi)$$

where g^{mn} is the inverse of g_{mn}. (There is no "η_{mn}" here.) Since the "canonical momentum" π_m appears without derivatives, its equations of motion are algebraic, and can be solved without ambiguity:

$$\pi_m = \dot{\phi}^n g_{nm}(\phi) - A_m(\phi)$$

Substitution back into L_H returns the original L.

Actions for field theories are just a special case (*not* a generalization) of the actions we have just considered: We just treat spatial coordinates x^i as part of the indices carried by the variables appearing in the action. Writing M for this all-inclusive index and I for the index on the field,

$$M \to (I, x^i)$$
$$\varPhi^M(t) \to \varPhi^I(t, x^i)$$

Then spatial derivatives are just certain matrices with respect to the M index, $\int d^3x$ comes from summation over M, etc.

3.2. Conservation

Action principles directly relate symmetries to conservation laws. For example, we have considered coordinate symmetries, such as the Poincaré group. (We also considered Galilean boosts, but because of certain complications these are actually easier to understand as a limit of Lorentz boosts.) Let's look at an arbitrary coordinate transformation of an action, but not involving time:

$$\delta\phi(t) = f(\phi(t))$$

where time is not varied, and f is the same function (not a functional) for any time. Now we include variations at the boundaries, because the symmetry holds even at the boundary, even "off shell" (without applying the equations of motion), so we keep boundary terms from integration by parts:

$$\delta S = \int dt\ \delta L(\phi, \dot{\phi}) = \int dt\ \left[\delta\phi\, \frac{\partial L(\phi, \dot{\phi})}{\partial \phi} + \delta\dot{\phi}\, \frac{\partial L(\phi, \dot{\phi})}{\partial \dot{\phi}} \right]$$

$$= \int dt\ \partial_t \left[\delta\phi\, \frac{\partial L(\phi, \dot{\phi})}{\partial \dot{\phi}} \right]$$

where we have dropped the usual field equation terms

$$\int dt\ \delta\phi \left[\frac{\partial L(\phi, \dot{\phi})}{\partial \phi} - \partial_t\, \frac{\partial L(\phi, \dot{\phi})}{\partial \dot{\phi}} \right] = \int dt\ \delta\phi\, \frac{\delta S}{\delta \phi} = 0$$

Requiring invariance of the action under this symmetry, where $\int dt$ is evaluated between some initial and final times, implies the conservation law

$$Q \equiv \delta\phi\, \frac{\partial L(\phi, \dot{\phi})}{\partial \dot{\phi}} \quad \Rightarrow \quad Q(t_f) - Q_(t_i) = 0 \quad \Rightarrow \quad \dot{Q} = 0$$

In mechanics, $\partial L(x,\dot{x})/\partial\dot{x} = p$ is the momentum (with an irrelevant minus sign in our conventions), so we have the conserved quantities

$$Q = (\delta x)\cdot p$$

In the nonrelativistic case, we have from rotations and spatial translations

$$\delta x^i = x^j a_j{}^i + b^i \quad \Rightarrow \quad Q = a_j{}^i x^j p_i + b^i p_i$$

Since b is arbitrary while a must be antisymmetric, this tells us momentum and angular momentum are conserved:

$$P_i = p_i, \qquad J_{ij} = x_{[i} p_{j]}$$

These must generalize in the obvious way for the Poincaré group, although we have yet to write an explicit action for the relativistic particle. (This only means we have not yet discussed what "p" is in that case.)

In field theory, this generalizes in an obvious way to internal symmetries, where ϕ is now a field, and

$$\delta\phi(x^m) = f(\phi(x^m))$$

involves neither time nor space (but again ϕ can carry an index).

Poincaré (or conformal) and internal symmetries of the above form are called "global", since they act the same way at all points: They have parameters (like a and b) that do not depend on spacetime. There are also "local", or "gauge" symmetries, whose parameters do depend on x^m, and they are common in field theory.

In field theory we have local conservation laws, since the action for a field is written as an integral $\int d^D x$ of a Lagrangian density that depends only on fields at x, and a finite number of their spacetime derivatives. A simple way to derive the local conservation laws is by coupling gauge fields: The method is to first find the field equations for the gauge field, then set the gauge field to vanish.

For example, in electromagnetism (to which we will return later) we have the "gauge transformation" for the (4-)vector potential

$$\delta A_m(x) = -\partial_m \lambda(x)$$

in terms of the "gauge parameter" λ. We couple the electromagnetic field A_m to arbitrary charged matter fields ϕ and demand gauge invariance of the matter part of the action, the matter-free part of the action being separately invariant. We then have

$$0 = \delta S_M = \int dx \left[(\delta A_m)\frac{\delta S_M}{\delta A_m} + (\delta\phi)\frac{\delta S_M}{\delta\phi} \right]$$

using just the definition of the functional derivative δ/δ. Applying the matter field equations $\delta S_M/\delta\phi = 0$, integration by parts, and the gauge transformation, we find the locally conserved "current":

$$0 = \int dx\ \lambda \left(\partial_m \frac{\delta S_M}{\delta A_m} \right) \quad \Rightarrow \quad J^m = \frac{\delta S_M}{\delta A_m}, \quad \partial_m J^m = 0$$

The local conservation law implies a global one, since

$$\partial_m J^m = 0 \quad \Rightarrow \quad 0 = \int d^D x\ \partial_m J^m = \frac{d}{dt} \int d^{D-1}x\ J^0 = \dot{Q} = 0$$

where we have integrated over a volume whose boundaries in space are at infinity (where J vanishes), and whose boundaries in time are infinitesimally separated. Equivalently, the global symmetry is a special case of the local one. Often global conservation laws are easiest to derive from local ones; later we'll see how to do this even in the case of particles (mechanics).

Similar remarks apply to gravity, but only if we evaluate the "current", in this case the energy-momentum tensor, in flat space $g_{mn} = \eta_{mn}$, since gravity is self-interacting. Thus, without loss of generality, we can start with a "weak-field" approximation, where

$$\delta g_{mn} \approx \partial_{(m}\lambda_{n)}$$

We then find

$$T^{mn} = -2\, \frac{\delta S_M}{\delta g_{mn}} \bigg|_{g_{mn}=\eta_{mn}}, \quad \partial_m T^{mn} = 0$$

where the normalization factor of -2 will be found later for consistency with the particle. In this case the corresponding "charge" is the D-momentum:

$$P^m = \int d^{D-1}x\ T^{0m}$$

Exercise 3.2.1

Show that the local conservation of the energy-momentum tensor allows definition of a conserved angular momentum

$$J^{mn} = \int d^{D-1}x\ x^{[m}T^{n]0}$$

4. Particles

4.1. Momentum

In quantum mechanics or field theory, “momentum” p^m is defined in terms of spacetime derivatives; we’ll come back to that later. In classical mechanics, momentum can be defined generally from the action; in the Hamiltonian approach, it is considered as a variable independent from the coordinates, being related only by the equations of motion.

A free, spinless, relativistic particle is completely described by the constraint

$$p^2 + m^2 = 0$$

The main qualitative distinction from the nonrelativistic case in the constraint

$$\begin{aligned} nonrelativistic: &\quad -2mE + (p^i)^2 = 0 \\ relativistic: &\quad m^2 - E^2 + (p^i)^2 = 0 \end{aligned}$$

is that the equation for the energy $E \equiv p^0$ is now quadratic, and thus has two solutions:

$$p^0 = \pm\omega, \qquad \omega = \sqrt{(p^i)^2 + m^2}$$

Later we’ll see how the second solution is interpreted as an “antiparticle”.

Exercise 4.1.1

Show that for $p^2 + m^2 = 0$ ($m^2 \geq 0$, $p^a \neq 0$), the signs of p^+ and p^- are always the same as the sign of the energy p^0.

Note that the nonrelativistic constraint in 4D spacetime looks just like the massless, relativistic constraint in 5D spacetime, with the replacements

$$p^+ \to m, \qquad p^- \to E; \qquad x^+ \to t$$

This is a way to derive the Galilean group without taking a limit. The only restrictions are that the mass m is positive, and to consider only transformations that leave it invariant. This also gives an easier way to derive conserved quantities: The Galilean boosts come from part of $J^{ab} = x^{[a}p^{b]}$,

$$J^{+i} = x^+p^i - x^ip^+ \to tp^i - x^im$$

while conservation of $p^\pm$ leads to independent conservation of (rest) mass and energy.

For the massive case, we also have

$$p^a = m\frac{dx^a}{ds}$$

For the massless case $ds = 0$: Massless particles travel along lightlike lines. However, we can define a new parameter τ such that

$$p^a = \frac{dx^a}{d\tau}$$

is well-defined in the massless case. In general, we then have

$$s = m\tau$$

While this fixes $\tau = s/m$ in the massive case, in the massless case it instead restricts $s = 0$. Thus, proper time does not provide a useful parametrization of the world line of a classical massless particle, while τ does: For any piece of such a line, $d\tau$ is given in terms of (any component of) p^a and dx^a. Later we'll see how this parameter appears in relativistic classical mechanics.

Exercise 4.1.2

The relation between x and p is closely related to the Poincaré conservation laws:

a Show that

$$dP_a = dJ_{ab} = 0 \quad \Rightarrow \quad p_{[a}dx_{b]} = 0$$

and use this to prove that conservation of P and J imply the existence of a parameter τ such that $p^a = dx^a/d\tau$.

b Consider a multiparticle system (but still without spin) where some of the particles can interact only when at the same point (i.e., by collision; they act as free particles otherwise). Define $P_a = \sum_I p^I_a$ and $J_{ab} = \sum_I x^I_{[a}p^I_{b]}$ as the sum of the individual momenta and angular momenta (where we label the particle with "I"). Show that momentum conservation implies angular momentum conservation,

$$\Delta P_a = 0 \quad \Rightarrow \quad \Delta J_{ab} = 0$$

where "Δ" refers to the change from before to after the collision(s).

Special relativity can also be stated as the fact that the only physically observable quantities are those that are Poincaré invariant. (Other objects, such as vectors, depend on the choice of reference frame.) For example, consider two spinless particles that interact by collision, producing two spinless particles (which may differ from the originals). Consider just the momenta. (Quantum mechanically, this is a complete description.) All invariants can be expressed in terms of the masses and the "Mandelstam variables" (not to be confused with time and proper time)

$$s = -(p_1 + p_2)^2, \qquad t = -(p_1 - p_3)^2, \qquad u = -(p_1 - p_4)^2$$

where we have used momentum conservation, which shows that even these three quantities are not independent:

$$p_I^2 = -m_I^2, \quad p_1 + p_2 = p_3 + p_4 \quad \Rightarrow \quad s + t + u = \sum_{I=1}^{4} m_I^2$$

(The explicit index now labels the particle, for the process 1+2→3+4.)

4.2. Antiparticles

While the metric η_{mn} is invariant under all Lorentz transformations (by definition), the 4D Levi-Civita tensor

$$\epsilon_{mnpq} \; totally\ antisymmetric, \qquad \epsilon_{0123} = -\epsilon^{0123} = 1$$

(with the $-$ from raising indices, since $det(\eta) = -1$) is invariant under only proper Lorentz transformations: It has an odd number of space indices and of time indices, so it changes sign under parity "P" or time reversal. (More precisely, under parity or time reversal the Levi-Civita tensor does not suffer the expected sign change, since it's constant, so there is an "extra" sign compared to the one expected for a tensor.) Consequently, we can use it to define "pseudotensors": Given polar vectors, whose signs change as position or momentum under improper Lorentz transformations, and scalars, which are invariant, we can define axial vectors and "pseudoscalars" as

$$V_a = \epsilon_{abcd} B^b C^c D^d, \qquad \phi = \epsilon_{abcd} A^a B^b C^c D^d$$

which get an extra sign change under such transformations (parity or time reversal, but not their product).

There is another such discrete transformation that is defined on phase space, but which does not affect spacetime. It changes the sign of all components of the momentum, while leaving the spacetime coordinates unchanged. This transformation is called "charge conjugation (C)", and is also only an approximate symmetry in nature. (Quantum mechanically, complex conjugation of the position-space wave function changes the sign of the momentum.) The misnomer "CT" for time reversal follows historically from the fact that the combination of reversing the time axis and charge conjugation preserves the sign of the energy. The physical meaning of charge conjugation is clear from the spacetime-momentum relation of relativistic classical mechanics $p = m\ dx/ds$: It is proper-time reversal, changing the sign of s. The relation to charge follows from "minimal coupling": The "covariant momentum"

$m\ dx/ds = p + qA$ (for charge q) appears in the constraint $(p+qA)^2 + m^2 = 0$ in an electromagnetic background; $p \to -p$ then has the same effect as $q \to -q$.

Previously we mentioned how negative energies were associated with "antiparticles". Now we can better see the relation in terms of charge conjugation. Note that charge conjugation, since it only changes the sign of τ but does not effect the coordinates, does not change the path of the particle, but only how it is parametrized. This is also true in terms of momentum, since the velocity is given by p^i/p^0. Thus, the only observable property that is changed is charge; spacetime properties (path, velocity, mass; also spin) remain the same. Another way to say this is that charge conjugation commutes with the Poincaré group. One way to identify an antiparticle is that it has all the same kinematical properties (mass, spin) as the corresponding particle, but opposite sign for internal quantum numbers (like charge). (Another way is pair creation and annihilation: See below.)

All these transformations are summarized in the table:

	C	CT	P	T	CP	PT	CPT
s	$-$	$+$	$+$	$-$	$-$	$-$	$+$
t	$+$	$-$	$+$	$-$	$+$	$-$	$-$
x^i	$+$	$+$	$-$	$+$	$-$	$-$	$-$
E	$-$	$-$	$+$	$+$	$-$	$+$	$-$
p^i	$-$	$+$	$-$	$-$	$+$	$+$	$-$

(The upper-left 3×3 matrix contains the definitions, the rest is implied.) In terms of complex wave functions, we see that C is just complex conjugation (no effect on coordinates, but momentum and energy change sign because of the "i" in the Fourier transform). On the other hand, for CT and P there is no complex conjugation, but changes in sign of the coordinates that are arguments of the wave functions, and also on the corresponding indices — the "orbital" and "spin" parts of these discrete transformations. (For example, derivatives ∂_a have sign changes because x^a does, so a vector wave function ψ^a must have the same sign changes on its indices for $\partial_a \psi^a$ to transform as a scalar.) The other transformations follow as products of these.

Exercise IA5.1

Find the effect of each of these 7 transformations on wave functions that are: **a** scalars, **b** pseudoscalars, **c** vectors, **d** axial vectors.

However, from the point of view of the "particle" there *is* some kind of kinematic change, since the proper time has changed sign: If we think of the mechanics of a particle as a one-dimensional theory in τ space (the worldline), where $x(\tau)$ (as well as any such variables describing spin or internal symmetry) is a wave function or field on that space, then $\tau \to -\tau$ is T on that one-dimensional space. (The fact we don't get CT can be seen when we add additional variables for internal symmetry: C mixes them on both the worldline and spacetime. So, on the worldline we have the "pure" worldline geometric symmetry CT × C = T.) Thus,

$$\textit{worldline } T \leftrightarrow \textit{spacetime } C$$

On the other hand, *spacetime* P and CT are simply internal symmetries with respect to the worldline (as are proper Poincaré transformations).

4.3. Action

The simplest relativistic actions are those for the *mechanics* (as opposed to field theory) of particles. These also give the simplest examples of gauge invariance in relativistic theories.

For nonrelativistic mechanics, the fact that the energy is expressed as a function of the three-momentum is conjugate to the fact that the spatial coordinates are expressed as functions of the time coordinate. In the relativistic generalization, all the spacetime coordinates are expressed as functions of a parameter τ: All the points that a particle occupies in spacetime form a curve, or "worldline", and we can parametrize this curve in an arbitrary way. Such parameters generally can be useful to describe curves: A circle is better described by $x(\theta), y(\theta)$ than $y(x)$ (avoiding ambiguities in square roots), and a cycloid can be described explicitly only this way.

The action for a free, spinless particle then can be written in relativistic Hamiltonian form as

$$S_H = \int d\tau [-\dot{x}^m p_m + v\tfrac{1}{2}(p^2 + m^2)]$$

where v is a "Lagrange multiplier", which upon variation enforces the constraint $p^2 + m^2 = 0$. This action is very similar to nonrelativistic ones, but instead of $x^i(t), p_i(t)$ we now have $x^m(\tau), p_m(\tau), v(\tau)$ (where "$\dot{}$ " now means $d/d\tau$). v also acts as a 1D gauge field (with respect to the 1 dimension of the worldline):

$$\delta x = \zeta p, \quad \delta p = 0, \quad \delta v = \dot{\zeta}$$

A more recognizable form of this invariance can be obtained by noting that any action $S(\phi^A)$ has invariances of the form

$$\delta \phi^A = \epsilon^{AB} \frac{\delta S}{\delta \phi^B}, \quad \epsilon^{AB} = -\epsilon^{BA}$$

which have no physical significance, since they vanish by the equations of motion. In this case we can add

$$\delta x = \epsilon(\dot{x} - vp), \quad \delta p = \epsilon\dot{p}, \quad \delta v = 0$$

and set $\zeta = v\epsilon$ to get

$$\delta x = \epsilon\dot{x}, \quad \delta p = \epsilon\dot{p}, \quad \delta v = \dot{(\epsilon v)}$$

We then can recognize this as a (infinitesimal) coordinate transformation for τ:

$$x'(\tau') = x(\tau), \quad p'(\tau') = p(\tau), \quad d\tau' v'(\tau') = d\tau\ v(\tau); \qquad \tau' = \tau - \epsilon(\tau)$$

The transformation laws for x and p identify them as "scalars" with respect to these "one-dimensional" (worldline) coordinate transformations (but they are vectors with respect to D-dimensional spacetime). On the other hand, v transforms as a "density": The "volume element" $d\tau\ v$ of the worldline transforms as a scalar. This gives us a way to measure length on the worldline in a way independent of the choice of τ parametrization. Because of this geometric interpretation, we are led to constrain

$$v > 0$$

so that any segment of the worldline will have positive length. Equivalently, we can interpret v as the square root of 1D metric, since $(d\tau\ v)^2 = d\tau^2 v^2$ is also a scalar.

The Lagrangian form of the free particle action follows from eliminating p by its equation of motion

$$vp = \dot{x} \quad \Rightarrow \quad S_L = \int d\tau\ \tfrac{1}{2}(vm^2 - v^{-1}\dot{x}^2)$$

For $m \neq 0$, we can also eliminate v by its equation of motion

$$v^{-2}\dot{x}^2 + m^2 = 0 \quad \Rightarrow \quad S = m\int d\tau\sqrt{-\dot{x}^2} = m\int\sqrt{-dx^2} = m\int ds = ms$$

The action then has the purely geometrical interpretation as the proper time; however, this last form of the action is awkward to use because of the square root (e.g., in quantization), and doesn't apply to the massless case. Note that the v equation also implies

$$ds = m(d\tau\ v)$$

relating the "intrinsic" length of the worldline (as measured with the worldline volume element) to its "extrinsic" length (as measured by the spacetime metric). As a consequence of this and the p equation, in the massive case we also have the usual relation between momentum and "velocity"

$$p^m = m\frac{dx^m}{ds}$$

Exercise 4.3.1

Take the nonrelativistic limit of the spinless particle action, in the form ms. (Note that, while the relativistic action is positive, the nonrelativistic one is negative.)

The (D+2)-dimensional (conformal) representation of the massless particle can be derived from the action

$$S = \int d\tau\ \tfrac{1}{2}(-\dot{y}^2 + \lambda y^2)$$

where λ is a Lagrange multiplier. This action is gauge invariant under

$$\delta y = \epsilon\dot{y} - \tfrac{1}{2}\dot{\epsilon}y, \quad \delta\lambda = \epsilon\dot{\lambda} + 2\dot{\epsilon}\lambda + \tfrac{1}{2}\dddot{\epsilon}$$

If we vary λ to eliminate it and y^- as previously, the action becomes

$$S = -\int d\tau\ \tfrac{1}{2}e^2\dot{x}^2$$

which agrees with the previous result, identifying $v = e^{-2}$, which also guarantees $v > 0$.

Rather than use the equation of motion to eliminate v it's more convenient to use a "gauge choice", i.e., choosing the gauge parameter to determine part of the gauge field. In this case the gauge parameter and gauge field both have just 1 component, so we can fix the gauge field completely. The gauge $v = 1$ is called "affine parametrization" of the worldline. Since $T = \int d\tau\ v$, the intrinsic length, is gauge invariant, that part of v still remains when the length is finite, but it can be incorporated into the limits of integration: The gauge $v = 1$ is maintained by $\dot{\zeta} = 0$, and this constant ζ can be used to gauge one limit of integration to zero, completely fixing the gauge (i.e., the choice of τ). We then integrate $\int_0^T$, where $T \geq 0$ (since originally $v > 0$), and T is a variable to vary in the action. The gauge-fixed action is then

$$S_{H,AP} = \int_0^T d\tau[-\dot{x}^m p_m + \tfrac{1}{2}(p^2 + m^2)]$$

In the massive case, we can instead choose the gauge $v = 1/m$; then the equations of motion imply that τ is the proper time. The Hamiltonian $p^2/2m$ + constant then resembles the nonrelativistic one.

Another useful gauge is the "lightcone gauge"

$$\tau = \frac{x^+}{p^+}$$

which, unlike the Poincaré covariant gauge $v = 1$, fixes τ completely; since the gauge variation $\delta(x^+/p^+) = \zeta$, we must set $\zeta = 0$ to maintain the gauge. In lightcone gauges we always assume $p^+ \neq 0$, since we often divide by it. This is usually not too dangerous an assumption, since we can treat $p^+ = 0$ as a limiting case (in D>2).

For every degree of freedom we can gauge away, the conjugate variable can be fixed by the constraint imposed by its gauge field. In lightcone gauges the constraints are almost linear: The gauge condition is $x^+ = p^+\tau$ and the constraint is $p^- = ...$, so the Lagrange multiplier v is varied to determine p^-. On the other hand, varying p^- gives

$$\delta p^- \quad \Rightarrow \quad v = 1$$

so this gauge is a special case of the gauge $v = 1$. An important point is that we used only "auxiliary" equations of motion: those not involving time derivatives. (A slight trick involves the factor of p^+: This is a constant by the equations of motion, so we can ignore $\dot{p}^+$ terms. However, technically we should not use that equation of motion; instead, we can redefine $x^- \to x^- + ...$, which will generate terms to cancel any $\dot{p}^+$ terms.) The net result of gauge fixing and the auxiliary equation on the action is

$$S_{H,LC} = \int_{-\infty}^{\infty} d\tau [\dot{x}^- p^+ - \dot{x}^i p^i + \tfrac{1}{2}(p^{i2} + m^2)]$$

where $x^a = (x^+, x^-, x^i)$, etc. In particular, since we have fixed one more gauge degree of freedom in the Lagrangian (corresponding to constant ζ), we have also eliminated one more constraint variable (T, the constant part of v). This is one of the main advantages of lightcone gauges: They are "unitary", eliminating all unphysical degrees of freedom.

4.4. Interactions

One way to introduce external fields into the mechanics action is by considering the most general Lagrangian quadratic in τ derivatives:

$$S_L = \int d\tau [-\tfrac{1}{2} v^{-1} g_{mn}(x) \dot{x}^m \dot{x}^n + A_m(x)\dot{x}^m + v\phi(x)]$$

In the free case we have constant fields $g_{mn} = \eta_{mn}$, $A_m = 0$, and $\phi = \frac{1}{2}m^2$. The v dependence has been assigned consistent with worldline coordinate invariance. The curved-space metric tensor g_{mn} describes gravity, the D-vector potential A_m describes electromagnetism, and ϕ is a scalar field that can be used to introduce mass by interaction.

Exercise 4.4.1

Derive the relativistic Lorentz force law

$$\partial_\tau(v^{-1}\dot{x}_m) + F_{mn}\dot{x}^n = 0$$

by varying the Lagrangian form of the action for the relativistic particle, in an external electromagnetic field (but flat metric and constant ϕ), with respect to x.

This action also has very simple transformation properties under D-dimensional gauge transformations on the external fields:

$$\delta g_{mn} = \epsilon^p\partial_p g_{mn} + g_{p(m}\partial_{n)}\epsilon^p, \quad \delta A_m = \epsilon^p\partial_p A_m + A_p\partial_m\epsilon^p - \partial_m\lambda, \quad \delta\phi = \epsilon^p\partial_p\phi$$

$$\Rightarrow \quad S_L[x] + \delta S_L[x] = S_L[x+\epsilon] - \lambda(x_f) + \lambda(x_i)$$

where we have integrated the action $\int_{\tau_i}^{\tau_f} d\tau$ and set $x(\tau_i) = x_i$, $x(\tau_f) = x_f$.

We now look for conservation laws by varying this action with respect to the external fields, as described previously. We first must distinguish the particle coordinates $X(\tau)$ from coordinates x for all of spacetime: The particle exists only at $x = X(\tau)$ for some τ, but the fields exist at all x. In this notation we can write the mechanics action as

$$S_L = \int d^Dx \left[-g_{mn}(x)\int d\tau\ \delta^D(x-X)\tfrac{1}{2}v^{-1}\dot{X}^m\dot{X}^n \right.$$
$$\left. +A_m(x)\int d\tau\ \delta^D(x-X)\dot{X}^m + \phi(x)\int d\tau\ \delta^D(x-X)v \right]$$

using $\int d^Dx\ \delta^D(x - X(\tau)) = 1$. We then have

$$J^m(x) = \int d\tau\ \delta^D(x-X)\dot{X}^m$$

$$T^{mn} = \int d\tau\ \delta^D(x-X)v^{-1}\dot{X}^m\dot{X}^n$$

Note that $T^{00} \geq 0$ (since $v > 0$). Integrating to find the charge and momentum:

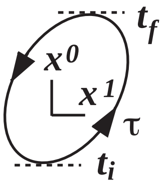
$$Q = \int d\tau\ \delta(x^0 - X^0)\dot{X}^0 = \int dX^0\ \epsilon(\dot{X}^0)\delta(x^0-X^0) = \epsilon(p^0)$$

$$P^m = \int d\tau\ \delta(x^0-X^0)v^{-1}\dot{X}^0\dot{X}^m = \int dX^0\ \epsilon(\dot{X}^0)\delta(x^0-X^0)v^{-1}\dot{X}^m = \epsilon(p^0)p^m$$

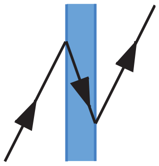
where we have used $p = v^{-1}\dot{X}$ (for the free particle), where p is the momentum conjugate to X, not to be confused with P. The factor of $\epsilon(p^0)$ ($\epsilon(u) = u/|u|$ is the sign of u) comes from the Jacobian from changing integration variables from τ to X^0.

The result is that our naive expectations for the momentum and charge of the particle can differ from the correct result by a sign. In particular p^0, which semiclassically is identified with the angular frequency of the corresponding wave, can be either positive or negative, while the true energy $P^0 = |p^0|$ is always positive, as physically required. (Otherwise all states could decay into lower-energy ones: There would be no lowest-energy state, the "vacuum".) When p^0 is negative, the charge Q and $dX^0/d\tau$ are also negative. In the massive case, we also have dX^0/ds negative. This means that as the proper time s increases, X^0 decreases. Since the proper time is the time as measured in the rest frame of the particle, this means that the particle is traveling backward in time: Its clock changes in the direction opposite to that of the coordinate system x^m. Particles traveling backward in time are called "antiparticles", and have charges opposite to their corresponding particles. They have positive true energy, but the "energy" p^0 conjugate to the time is negative.

4.5. Pair creation



Free particles travel in straight lines. Nonrelativistically, external fields can alter the motion of a particle to the extent of changing the signs of spatial components of the momentum. Relativistically, we might then expect that interactions could also change the sign of the energy, or at least the canonical energy p^0. As an extreme case, consider a worldline that is a closed loop: We can pick τ as an angular coordinate around the loop. As τ increases, X^0 will either increase or decrease. For example, a circle in the x^0-x^1 plane will be viewed by the particle as repeating its history after some finite τ, moving forward with respect to time x^0 until reaching a latest time t_f, and then backward until some earliest time t_i. On the other hand, from the point of view of an observer at rest with respect to the x^m coordinate system, there are no particles until $x^0 = t_i$, at which time both a particle and an antiparticle appear at the same position in space, move away from each other, and then come back together and disappear. This process is known as "pair creation and annihilation".



Whether such a process can actually occur is determined by solving the equations of motion. A simple example is a particle in the presence of only a static electric field, produced by the time component A^0 of the potential. We consider the case of a piecewise constant potential, vanishing outside a certain region and constant inside. Then the electric field vanishes except at the boundaries, so the particle travels in straight lines except at the boundaries. For simplicity we reduce the problem to two dimensions:

$$A^0 = -V \ for \ 0 \leq x^1 \leq L, \quad 0 \ otherwise$$

for some constant V. The action is, in Hamiltonian form,

$$S_H = \int d\tau \ \{-\dot{x}^m p_m + v\tfrac{1}{2}[(p+A)^2 + m^2]\}$$

and the equations of motion are

$$\begin{gathered}
\dot{p}_m = -v(p+A)^n \partial_m A_n \quad \Rightarrow \quad p^0 = E \\
(p+A)^2 = -m^2 \quad \Rightarrow \quad p^1 = \pm\sqrt{(E+A^0)^2 - m^2} \\
v^{-1}\dot{x} = p + A \quad \Rightarrow \quad v^{-1}\dot{x}^1 = p^1, \quad v^{-1}\dot{x}^0 = E + A^0
\end{gathered}$$

where E is a constant (the canonical energy at $x^1 = \infty$) and the equation $\dot{p}^1 = ...$ is redundant because of gauge invariance. We assume $E > 0$, so initially we have a particle and not an antiparticle.

We look only at the cases where the worldline begins at $x^0 = x^1 = -\infty$ (lower left) and continues toward the right till it reaches $x^0 = x^1 = +\infty$ (upper right), so that $p^1 = v^{-1}\dot{x}^1 > 0$ everywhere (no reflection). However, the worldline might bend backward in time ($\dot{x}^0 < 0$) inside the potential: To the outside viewer, this looks like pair creation at the right edge before the first particle reaches the left edge; the antiparticle then annihilates the original particle when it reaches the left edge, while the new particle continues on to the right. From the particle's point of view, it has simply traveled backward in time so that it exits the right of the potential before it enters the left, but it is the same particle that travels out the right as came in the left. The velocity of the particle outside and inside the potential is

$$\frac{dx^1}{dx^0} = \begin{cases} \dfrac{\sqrt{E^2 - m^2}}{E} & \text{outside} \\[2ex] \dfrac{\sqrt{(E-V)^2 - m^2}}{E - V} & \text{inside} \end{cases}$$

From the sign of the velocity we then see that we have normal transmission (no antiparticles) for $E > m + V$ and $E > m$, and pair creation/annihilation when

$$V - m > E > m \quad \Rightarrow \quad V > 2m$$

The true “kinetic” energy of the antiparticle (which appears only inside the potential) is then $-(E - V) > m$.

Exercise 4.4.1

This solution might seem to violate causality. However, in mechanics as well as field theory, causality is related to boundary conditions at infinite times. Describe another solution to the equations of motion that would be interpreted by an outside observer as pair creation *without any initial particles*: What happens ultimately to the particle and antiparticle? What are the allowed values of their *kinetic* energies (maximum and minimum)? Since many such pairs can be created by the potential alone, it can be accidental (and not acausal) that an external particle meets up with such an antiparticle. Note that the generator of the potential, to maintain its value, continuously loses energy (and charge) by emitting these particles.

5. Waves

5.1. Correspondence

The correspondence principle relates quantum particles/waves to both classical particles and classical waves. A classical field satisfies the same type of wave equation as a quantum mechanical wave function. In particular, we have plane wave solutions of the form

$$\phi(x) \sim e^{i(p^i x^i - Et)} = e^{i(p^m x^n \eta_{mn})}$$

in both relativistic and nonrelativistic theories, where the minus sign is obvious relativistically but mysterious nonrelativistically. Here p^m satisfies either the relativistic or nonrelativistic relation, as appropriate. We have used “natural units” $\hbar = 1$. More generally, we have the Fourier transform

$$\phi(x) = \int d^D p \ \tilde{\phi}(p) e^{ip\cdot x}$$

from which we see that the operator $-i\partial_m$ on $\phi(x)$ creates a factor of $p^n \eta_{nm}$ on $\tilde{\phi}(p)$.

In classical field theory, we can identify a “particle” with its “antiparticle” by requiring the field to be invariant under charge conjugation: For example, for a scalar field (spinless particle), we can have the reality condition

$$\phi(x) = \phi^*(x)$$

or in momentum space, by Fourier transformation,

$$\tilde{\phi}(p) = [\tilde{\phi}(-p)]^*$$

which implies the particle has charge zero (neutral).

Field equations are derived by the straightforward generalization of the variation of actions defined previously: As follows from treating the spatial coordinates in the same way as discrete indices,

$$\delta S \equiv \int d^D x \ \delta\phi^I(t, x^i) \frac{\delta S}{\delta\phi^I(t, x^i)}$$

$$\frac{\delta}{\delta\phi^I(x)} \phi^J(x') = \delta^I_J \delta^D(x - x')$$

For example,

$$S = -\int dt \ d^3x \ \tfrac{1}{2}\dot{\phi}^2 \quad \Rightarrow \quad \frac{\delta S}{\delta\phi} = \ddot{\phi}$$

Exercise 5.1.1

Consider the action

$$S[\phi] = \int dt\ d^{D-1}x\ [-\tfrac{1}{2}\dot{\phi}^2 + V(\phi)]$$

for potential $V(\phi)$ (a function, not a functional).

a Find the field equations.

b Assume $V(\phi) = \lambda\phi^n$ for some positive integer n and constant, *dimensionless* λ, in units $\hbar = c = 1$. Use dimensional analysis to relate n and D (of course, also a positive integer), and list all paired possibilities of (n, D).

The action for a real scalar is then (in some normalization conventions)

$$S = \int d^D x\ L, \qquad L = \tfrac{1}{4}(\partial\phi)^2 + V(\phi)$$

where the "potential" $V(\phi) \geq 0$, and L is the "Lagrangian density". In particular, $V = \frac{1}{4}m^2\phi^2$ for the free theory. The free field equation is then the "Klein-Gordon equation"

$$(-\Box + m^2)\phi(x) = 0, \qquad (p^2 + m^2)\tilde{\phi}(p) = 0$$

(where the "d'Alembertian operator" $\Box \equiv \partial^2$). The energy is given by the Hamiltonian; from the Lagrangian (not the Lagrangian density), it can be found by changing the sign of the term ("kinetic energy") with 2 time derivatives (and dropping any terms with a single time derivative). In this case, we thus have

$$E = \int d^{D-1}x\ [\tfrac{1}{4}\dot{\phi}^2 + \tfrac{1}{4}(\partial_i\phi)^2 + V(\phi)]$$

which is positive for positive potential. This condition fixes the signs for the 2 terms in the action; notice that it is the same condition as fixing the action itself to be positive in Euclidean space (changing the sign for the time direction in the metric).

For a complex scalar, we replace $\frac{1}{2}\phi\phi \to \chi^*\chi$ in both terms (when $V(\phi) = W(\frac{1}{2}\phi^2)$ for some polynomial W):

$$L \to \tfrac{1}{2}|\partial\chi|^2 + W(|\chi|^2)$$

This introduces a global internal symmetry

$$\chi' = e^{i\theta}\chi, \qquad \chi^{*\prime} = e^{-i\theta}\chi^*$$

corresponding to conservation of electric charge.

The 4D action for a massless Weyl spinor is

$$S = \int d^4x\ (-i\bar{\psi}^{\dot{\beta}}\partial_{\alpha\dot{\beta}}\psi^{\alpha})$$

For various reasons it is convenient to double the number of spinors; a mass term can then easily be added:

$$L = (\bar\psi_L^{\dot\alpha} i\partial^\alpha{}_{\dot\alpha}\psi_{L\alpha} + \bar\psi_R^{\dot\alpha} i\partial^\alpha{}_{\dot\alpha}\psi_{R\alpha}) + \tfrac{m}{\sqrt{2}}(\psi_L^\alpha\psi_{R\alpha} + \bar\psi_L^{\dot\alpha}\bar\psi_{R\dot\alpha})$$

where L and R denote the 2 spinors. (Mass can also be introduced for a single spinor, but that would require a discussion of statistics for classical fields, which we will avoid here.)

5.2. External fields

A scalar field must be complex to be charged (i.e., a representation of U(1)): From the gauge transformation

$$\chi' = e^{i\lambda}\chi$$

(where now λ is a function of x) we find the minimal coupling (for $q = 1$)

$$S_\chi = \int dx\ [\tfrac{1}{2}|(\partial + iA)\chi|^2 + \tfrac{1}{2}m^2|\chi|^2]$$

This action is also invariant under charge conjugation

$$C : \chi \to \chi^*, \qquad A \to -A$$

which changes the sign of the charge, since $\chi^{*\prime} = e^{-i\lambda}\chi^*$.

Exercise 5.2.1

Let's consider the semiclassical interpretation of a charged particle as described by a complex scalar field ψ, with Lagrangian

$$L = \tfrac{1}{2}(|\nabla\psi|^2 + m^2|\psi|^2)$$

a Use the semiclassical expansion in $\hbar$ defined by

$$\nabla \to \hbar\partial + iqA, \qquad \psi \to \sqrt{\rho}e^{-i\mathcal{S}/\hbar}$$

Find the Lagrangian in terms of ρ and $\mathcal{S}$ (and the background field A), order-by-order in $\hbar$ (in this case, just $\hbar^0$ and $\hbar^2$).

b Take the semiclassical limit by dropping the $\hbar^2$ term in L, to find

$$L \to \rho\tfrac{1}{2}[(-\partial\mathcal{S} + qA)^2 + m^2]$$

Vary with respect to $\mathcal{S}$ and ρ to find the equations of motion. Defining

$$p \equiv -\partial\mathcal{S}$$

show that these field equations can be interpreted as the "mass-shell condition" (energy-momentum relation) and current conservation. Show that A couples to this current by varying L with respect to A.

The spinor field also needs doubling for charge. The gauge transformations are similar to the scalar case, and the action again follows from minimal coupling, to an action that has the global invariance ($\lambda =$ constant in the absence of A):

$$\psi_L'^{\alpha} = e^{i\lambda}\psi_L^{\alpha}, \quad \psi_R'^{\alpha} = e^{-i\lambda}\psi_R^{\alpha}$$

$$S_e = \int dx\ [\bar{\psi}_L^{\dot{\beta}}(-i\partial_{\alpha\dot{\beta}} + A_{\alpha\dot{\beta}})\psi_L^{\alpha} + \bar{\psi}_R^{\dot{\beta}}(-i\partial_{\alpha\dot{\beta}} - A_{\alpha\dot{\beta}})\psi_R^{\alpha} + \tfrac{m}{\sqrt{2}}(\psi_L^{\alpha}\psi_{R\alpha} + \bar{\psi}_L^{\dot{\alpha}}\bar{\psi}_{R\dot{\alpha}})]$$

The current is found from varying with respect to A:

$$J^{\alpha\dot{\beta}} = \bar{\psi}_L^{\dot{\beta}}\psi_L^{\alpha} - \bar{\psi}_R^{\dot{\beta}}\psi_R^{\alpha}$$

Charge conjugation

$$C : \psi_L^{\alpha} \leftrightarrow \psi_R^{\alpha}, \qquad A \to -A$$

(which commutes with Poincaré transformations) changes the sign of the charge and current.

An interesting distinction between gravity and electromagnetism is that static bodies always attract gravitationally, whereas electrically they repel if they are like and attract if they are opposite. This is a direct consequence of the fact that the graviton has spin 2 while the photon has spin 1: The Lagrangian for a field of integer spin s coupled to a current, in an appropriate gauge and the weak-field approximation, is

$$-\tfrac{1}{4s!}\phi^{a_1\ldots a_s}\Box\phi_{a_1\ldots a_s} + \tfrac{1}{s!}\phi_{a_1\ldots a_s}J^{a_1\ldots a_s}$$

From a scalar field in the semiclassical approximation (see above), starting with

$$J^{a_1\ldots a_s} \sim \psi^* \overleftrightarrow{\partial}^{a_1}\cdots\overleftrightarrow{\partial}^{a_s}\psi$$

where "$A\overleftrightarrow{\partial}B$" means "$A\partial B - (\partial A)B$", we see that the current will be of the form

$$J^{a_1\ldots a_s} \sim \rho p^{a_1}\cdots p^{a_s}$$

for a scalar particle, for some ρ. (The same follows from comparing the expressions for currents and energy-momentum tensors for particles as above. The only way to get vector indices out of a scalar particle, to couple to the vector indices for the spin of the force field, is from momentum.) In the static approximation, only time components contribute: We then can write this Lagrangian as, taking into account $\eta_{00} = -1$,

$$-(-1)^s \tfrac{1}{4s!} \phi_{0\ldots 0} \Box \phi_{0\ldots 0} + \tfrac{1}{s!} \phi_{0\ldots 0} \rho (p^0)^s$$

where $E = p^0 > 0$ for a particle and < 0 for an antiparticle. Thus the spin-dependence of the potential/force between two particles goes as $(-E_1 E_2)^s$. It then follows that all particles attract by forces mediated by even-spin particles, and a particle and its antiparticle attract under all forces, while repulsion will occur for odd-spin forces between two identical particles. (We can substitute "particles of the same sign charge" for "identical particles", and "particles of opposite sign charge" for "particle and its antiparticle", where the charge is the coupling constant appropriate for that force.)

Exercise 5.2.2

Show that the above current is conserved,

$$\partial_{a_1} J^{a_1 \cdots a_s} = 0$$

(and the same for the other indices, by symmetry) if ψ satisfies the free Klein-Gordon equation (massless or massive).

5.3. Electromagnetism

As a last example, we consider the action for electromagnetism itself. We first translate the theory into spinor notation: The Maxwell field strength F_{ab} is expressed in terms of the vector potential ("gauge field") A_a, with a "gauge invariance" in terms of a "gauge parameter" λ with spacetime dependence. The gauge transformation $\delta A_a = -\partial_a \lambda$ becomes

$$A'_{\alpha\dot\beta} = A_{\alpha\dot\beta} - \partial_{\alpha\dot\beta}\lambda$$

where $\partial_{\alpha\dot\beta} = \partial/\partial x^{\alpha\dot\beta}$. It leaves invariant the field strength $F_{ab} = \partial_{[a} A_{b]}$:

$$F_{\alpha\dot\gamma,\beta\dot\delta} = \partial_{\alpha\dot\gamma} A_{\beta\dot\delta} - \partial_{\beta\dot\delta} A_{\alpha\dot\gamma} = \tfrac{1}{4}(F_{(\alpha\beta)[\dot\gamma\dot\delta]} + F_{[\alpha\beta](\dot\gamma\dot\delta)})$$

$$= \bar{C}_{\dot\gamma\dot\delta} f_{\alpha\beta} + C_{\alpha\beta} \bar{f}_{\dot\gamma\dot\delta}, \quad f_{\alpha\beta} = \tfrac{1}{2} \partial_{(\alpha\dot\gamma} A_{\beta)}{}^{\dot\gamma}$$

Maxwell's equations are

$$\partial^\beta{}_{\dot\gamma} f_{\beta\alpha} \sim J_{\alpha\dot\gamma}$$

They include both the field equations (the hermitian part) and the "Bianchi identities" (the antihermitian part). $f_{\alpha\beta}$ and $\bar{f}_{\dot\alpha\dot\beta}$ can be identified with the left and right circular polarizations of the electromagnetic field.

Exercise 5.3.1

Write Maxwell's equations, and the expression for the field strength in terms of the gauge vector, in 2×2 matrix notation, without using C's. Combine them to derive the wave equation for A.

We can write the action for pure electromagnetism as

$$S_A = \int d^4x \ \tfrac{1}{2e^2} f^{\alpha\beta} f_{\alpha\beta} = \int d^4x \ \tfrac{1}{2e^2} \bar{f}^{\dot\alpha\dot\beta} \bar{f}_{\dot\alpha\dot\beta} = \int d^4x \ \tfrac{1}{8e^2} F^{ab} F_{ab}$$

dropping boundary terms, where e is the electromagnetic coupling constant, i.e., the charge of the proton. (Other normalizations can be used by rescaling $A_{\alpha\dot\beta}$.) Maxwell's equations follow from varying the action with a source term added:

$$S = S_A + \int d^4x \ A^{\alpha\dot\beta} J_{\alpha\dot\beta} \qquad \Rightarrow \qquad \tfrac{1}{e^2} \partial^\beta{}_{\dot\gamma} f_{\beta\alpha} = J_{\alpha\dot\gamma}$$

Exercise 5.3.2

By plugging in the appropriate expressions in terms of A_a (and repeatedly integrating by parts), show that all of the above expressions for the electromagnetism action can be written as

$$S_A = -\int d^4x \ \tfrac{1}{4e^2} [A \cdot \Box A + (\partial \cdot A)^2]$$

Exercise 5.3.3

Find all the field equations for all the fields, found from adding to S_A the minimally coupled complex scalar and spinor above.

Having seen many of the standard examples of relativistic field theory actions, we now introduce one of the most important principles in field theory; unfortunately, it can be justified only at the quantum level:

Good ultraviolet behavior: All quantum field theories should have only couplings with nonnegative mass (engineering) dimension.

Exercise 5.3.4

Show in D=4 using dimensional analysis that this restriction (in addition to Poincaré invariance and locality) on scalars or vectors, written collectively as ϕ (with kinetic term $\phi\Box\phi$ free of couplings), restricts terms in the action to be of the form

$$\phi, \phi^2, \phi^3, \phi^4, \phi\partial\phi, \phi^2\partial\phi, \phi\partial\partial\phi$$

and find the dimensions of all the corresponding coupling constants (coefficients of each term).

The energy-momentum tensor for electromagnetism is much simpler in this spinor notation, and follows (up to normalization) from gauge invariance, dimensional analysis, Lorentz invariance, and the vanishing of its trace:

$$T_{\alpha\beta\dot\gamma\dot\delta} = -\tfrac{1}{e^2} f_{\alpha\beta}\bar f_{\dot\gamma\dot\delta}$$

We have used conventions where e appears multiplying only the action S_A, and not in the "covariant derivative"

$$\nabla = \partial + iqA$$

where q is the charge in units of e: e.g., $q = 1$ for the proton, $q = -1$ for the electron. Alternatively, we can scale A, as a field redefinition, to produce the opposite situation:

$$A \to eA: \qquad S_A \to \int d^D x\ \tfrac{1}{8}F^2, \qquad \nabla \to \partial + iqeA$$

The former form has the advantage that the coupling appears only in the one term S_A, while the latter has the advantage that the kinetic (free) term for A is normalized the same way as for scalars. The former form has the further advantage that e appears in the gauge transformations of none of the fields, making it clear that the group theory does not depend on the value of e.

Exercise 5.3.5

Using vector notation, minimal coupling, and dimensional analysis, find the mass dimensions of the electric charge e in arbitrary spacetime dimensions, and show it is dimensionless only in $D = 4$.

Maxwell's equations now can be easily generalized to include magnetic charge by allowing the current J to be complex. (However, the expression for F in terms of A is no longer valid.) This is because the "duality transformation" that switches electric and magnetic fields is much simpler in spinor notation: Using the expression given above for the 4D Levi-Civita tensor using spinor indices,

$$F'_{ab} = \tfrac{1}{2}\epsilon_{abcd}F^{cd} \quad \Rightarrow \quad f'_{\alpha\beta} = -if_{\alpha\beta}$$

More generally, Maxwell's equations in free space (but not the expression for F in terms of A) are invariant under the continuous duality transformation

$$f'_{\alpha\beta} = e^{i\theta}f_{\alpha\beta}$$

(and $J'_{\alpha\dot\beta} = e^{i\theta}J_{\alpha\dot\beta}$ in the presence of both electric and magnetic charges). Note that the energy-momentum tensor is invariant under this transformation.

Exercise 5.3.6

Prove the relation between duality in vector and spinor notation. Show that $F_{ab} + i\frac{1}{2}\epsilon_{abcd}F^{cd}$ contains only $f_{\alpha\beta}$ and not $f_{\dot{\alpha}\dot{\beta}}$.

Exercise 5.3.7

How does complexifying $J_{\alpha\dot{\beta}}$ modify Maxwell's equations in *vector* notation?

6. Cosmology

6.1. Dilaton

Some of the ideas in general relativity can be introduced by a simple model that involves introducing only a scalar field. Although this model does not correctly describe gravitational forces within our solar system, it does give an accurate description of cosmology. The basic idea is to introduce a dynamical length scale in terms of a real scalar field $\phi(x)$ called the "dilaton" by redefining lengths as

$$-ds^2 = dx^m dx^n \phi^2(x)\eta_{mn}$$

(Squaring ϕ preserves the sign of ds^2; we assume ϕ vanishes nowhere.) As explained in our discussion of conformal symmetry, this field changes only how we measure lengths, not angles (which is why it is insufficient to describe gravity): At any point in spacetime, it changes the length scale by the same amount in all directions. In fact, it allows us to introduce conformal invariance as a symmetry: We have already seen that under a conformal transformation the usual proper time of special relativity changes as

$$dx'^m dx'^n \eta_{mn} = \xi(x) dx^m dx^n \eta_{mn}$$

Thus, by transforming ϕ as

$$\phi'(x') = [\xi(x)]^{-1/2}\phi(x)$$

we have

$$ds'^2 = ds^2$$

for our new definition above of proper time. This transformation law for for the dilaton allows any Poincaré invariant action to be made conformally invariant. This definition of length is a special case of the general relativistic definition,

$$-ds^2 = dx^m dx^n g_{mn}(x) \quad \Rightarrow \quad g_{mn} = \phi^2 \eta_{mn}$$

The action for a particle is easily modified: For example,

$$S_L = \int d\tau \ \tfrac{1}{2}(vm^2 - v^{-1}\dot{x}^2) \quad \rightarrow \quad \int d\tau \ \tfrac{1}{2}[vm^2 - v^{-1}\phi^2(x)\dot{x}^2]$$

since $\dot{x}^2 = dx^2/d\tau^2$ (or by using our previous coupling to the metric tensor g_{mn}). It is convenient to rewrite this action by redefining

$$v(\tau) \rightarrow v(\tau)\phi^2(x(\tau))$$

The resulting form of the action

$$S_L \quad \to \quad \int d\tau \; \tfrac{1}{2}(vm^2\phi^2(x) - v^{-1}\dot{x}^2)$$

makes it clear that there is no change in the case $m = 0$: A massless (spinless) particle is automatically conformally invariant. We have seen this action before: It is the coupling of a massless particle to an external scalar field $\frac{1}{2}m^2\phi^2$. (What we call the scalar field is irrelevant until we write the terms in the action for that field itself.)

Exercise 6.1.1

Let's examine these actions in more detail:

a Find the equations of motion following form both forms of the particle action with background dilaton $\phi(x)$.

b Find the action that results from eliminating v by its equation of motion from both actions for $m \neq 0$, and show they are the same.

c By a different redefinition of v, find a form of the action that is completely linear in ϕ.

The corresponding change in field theory is obvious if we look at the Hamiltonian form of the particle action

$$S_H \quad \to \quad \int d\tau [-\dot{x}^m p_m + v\tfrac{1}{2}(p^2 + m^2\phi^2)]$$

Using the correspondence principle, we see that the Klein-Gordon equation for a scalar field ψ has changed to

$$(\Box - m^2\phi^2)\psi = 0$$

Since conformal invariance includes scale invariance, it is now natural to associate dimensions of mass with ϕ (or inverse length, if we do classical field theory) instead of m, since in scale invariant theories all constants in the field equations (or action) must be dimensionless (otherwise they would set the scale). The corresponding modification to the field theory action is

$$S \quad \to \quad \int d^D x \; \tfrac{1}{4}[(\partial\psi)^2 + m^2\phi^2\psi^2]$$

Since this is supposed to describe gravity, at least in some crude approximation that applies to cosmology, where is the (Newton's) gravitational constant? Since ϕ must be nonvanishing, "empty space" must be described by ϕ taking some constant value: We therefore write

$$\langle\phi\rangle = \frac{1}{\kappa}, \qquad \kappa^2 = G$$

where "$\langle\ \ \rangle$" means vacuum value, or asymptotic value, or weak-field limit (the value ϕ takes far away from matter). (There is actually an extra factor here of $4\pi/3$ in the definition of Newton's constant G, so we effectively use units $4\pi G/3 = 1$. We will ignore this factor here, since it can't be determined without introducing true gravity.) Thus, the usual mass in the Klein-Gordon equation arises in this way as m/κ. This phenomenon, where the vacuum solution breaks an invariance of the field equations or action, is known as "spontaneous symmetry breaking", and is also important in the Standard Model of unified electromagnetic and weak interactions. The dilaton ϕ is defined as the field that spontaneously breaks scale invariance.

In natural ("Planck") units $\kappa = 1$: Fixing $c = \hbar = \kappa = 1$ completely determines the units of length, time, and mass. These units are the convenient ones for quantum gravity; they are also the most obvious universal ones, since special relativity, quantum theory, and gravity apply to everything. In relation to standard units, these are approximately

$$\sqrt{\frac{G\hbar}{c^3}} = 1.61605(10) \cdot 10^{-35} m$$

$$\sqrt{\frac{G\hbar}{c^5}} = 5.39056(34) \cdot 10^{-44} s$$

$$\sqrt{\frac{\hbar c}{G}} = 2.17671(14) \cdot 10^{-8} kg$$

(where the numbers in parentheses refer to errors in the last digits).

Exercise 6.1.2

There is another Planck unit, for temperature. Evaluate it in standard units (Kelvins) by setting to 1 the Boltzmann constant k.

We have yet to determine the action for ϕ itself: We write the usual action for a massless scalar in D=4 (for other D we need to replace ϕ with a power by dimensional analysis),

$$S_\phi = -\int d^4x \ \tfrac{1}{2}(\partial\phi)^2$$

but we have written it with the "wrong" sign, for reasons we cannot justify without recourse to the complete theory of gravity. However, without this sign change we would not be able to get reasonable cosmological solutions, ones for which ϕ never vanishes.

6.2. Expansion

To a good approximation the universe can be described by a spacetime which is (spatially) rotationally invariant ("isotropic") with respect to a preferred time direction. Furthermore, it should be (spatially) translationally invariant ("homogeneous"), so the dilaton should depend only on that time coordinate. We therefore look for solutions of the equations of motion which depend only on time. Thus the proper time is given by

$$-ds^2 = \phi^2(t)[-dt^2 + (dx^i)^2]$$

By a simple redefinition of the time coordinate, this can be put in a form

$$-ds^2 = -dT^2 + \phi^2(T)(dx^i)^2$$

where by "$\phi(T)$" we really mean "$\phi(t(T))$", and the two time coordinates are related by

$$dT = dt\ \phi \quad \Rightarrow \quad T = \int dt\ \phi(t) \quad or \quad t = \int dT \frac{1}{\phi(t(T))}$$

In this latter form of ds we can recognize T as the usual time, as measured by a clock at rest with respect to this preferred time frame. It will prove convenient to calculate with time t, so we will work with that coordinate from now on, unless otherwise stated; in the end we will transform to T for comparison to quantities measured by experiment.

To a good approximation the matter in the universe can be approximated as a "dust", a collection of noninteracting particles. It should also be rotationally invariant with respect to the preferred time direction, so the momenta of the particles should be aligned in that time direction. (Really it is this matter that defines the time direction, since it generates the solution for ϕ.) Furthermore, the dust should be translationally invariant, so all the momenta should be the same (assuming they all have the same mass), and the distribution should be independent of time. Varying the Hamiltonian form of the action for a single particle with respect to ϕ, we find

$$\frac{\delta S_M}{\delta \phi(x)} = m^2 \int d\tau\ v\phi\delta^4(x - X)$$

Using the equations of motion following from that action, we also have

$$vm\phi = \sqrt{-\dot{x}^2} = \left| \frac{dt}{d\tau} \right|$$

where we have used $dx^i = 0$ for this dust, and the fact that v, m, ϕ are all positive by definition. We thus have

$$\frac{\delta S_M}{\delta \phi(x)} = m\delta^3(x - X)$$

We can compare this to the energy density, derived as before (since the $\dot{x}^2$ term in the action, which would contain the metric, is unmodified):

$$T_M^{00} = \int d\tau\ \delta^4(x-X)v^{-1}\dot{t}^2 = \int d\tau\ \delta^4(x-X)vm^2\phi^2 = m\phi\delta^3(x-X)$$

The relation between these 2 quantities is no accident: Our original introduction of ϕ was as $g_{mn} = \phi^2\eta_{mn}$. If we introduce both ϕ and metric independently, so as to calculate both of the above quantities, in the combination $\phi^2 g_{mn}$, then we automatically have

$$\phi\frac{\delta S_M}{\delta\phi} = 2g_{mn}\frac{\delta S_M}{\delta g_{mn}} = -T_M{}^m{}_m$$

which is T^{00} in this case (since the other components vanish).

Of course, the dust consists of more than one particle: It is a collection of particles, each at fixed x^i. That means we should replace $\delta^3(x-X)$ with some constant, independent of both x^i (because of homogeneity) and t (because of isotropy; the particles don't move). Actually, we need to average over particles of different masses: The result is then

$$\frac{\delta S_M}{\delta\phi(x)} = a, \qquad T_M^{00} = a\phi$$

for some constant a. The equations of motion for ϕ are now very simple; since $\partial_i\phi = 0$, we now have simply

$$\ddot{\phi} = a$$

where the dots now refer to t derivatives. If we take this equation and multiply both sides by $\dot{\phi}$, we get an obvious total derivative. Integrating this equation, we get

$$\tfrac{1}{2}\dot{\phi}^2 = a\phi + \tfrac{1}{2}b$$

for some constant $\frac{1}{2}b$. This equation has a simple interpretation: Recognizing $a\phi$ as the energy density T_M^{00} of the dust, and $-\frac{1}{2}\dot{\phi}^2$ as the energy density of ϕ (from our earlier discussion of Hamiltonian densities), we see it implies that the total energy density of the Universe is a constant.

We can also identify the source of this constant energy: We evaluated the energy density of dust and its coupling to ϕ. However, there can also be radiation: massless particles. As we saw, massless particles do not couple to ϕ. Also, we have neglected any interaction of particles with each other. Thus massless particles in this approximation are totally free; their energy consists totally of kinetic energy, and thus is constant. (They also move at the speed of light, so components of T^{ab} other than T^{00} are nonvanishing. However, we average over massless particles moving in all directions to preserve isotropy.) Therefore we can identify

$$T_R^{00} = \tfrac{1}{2}b$$

Exercise 6.2.1

Consider general forms of the energy-momentum tensor that have the right symmetry:

a Show that the most general form that has spatial isotropy and homogeneity is

$$T^{mn} = \rho(t)u^m u^n + P(t)(\eta^{mn} + u^m u^n), \qquad u^m \equiv \delta_0^m$$

(or the equivalent). ρ is the energy density, while P is the pressure. This general form is called a "perfect fluid" (e.g., an ideal gas).

b Relate pressure to energy density for radiation by using the fact that it doesn't couple to ϕ.

These equations are easily solved. There is an unavoidable "singularity" (the "Big Bang") $\phi = 0$ (all lengths vanish) at some time: Imposing the initial condition $\phi(0) = 0$ (i.e., we set it to be $t = 0$) and $\dot{\phi}(0) > 0$ (so $\phi \geq 0$),

$$\phi = a\tfrac{1}{2}t^2 + \sqrt{b}t$$

The "physical" time coordinate is then

$$T = \int_0 dt\ \phi = a\tfrac{1}{6}t^3 + \sqrt{b}\tfrac{1}{2}t^2$$

Since ϕ can't be expressed simply in terms of T, we use the expressions for both in terms of t. Simple expressions can be found for $a = 0$ ($\phi \sim \sqrt{T}$) and $b = 0$ ($\phi \sim T^{2/3}$).

For the case of pure matter ($b = 0$), the energy conservation equation written in terms of the T coordinate becomes, using $dT = \phi\ dt$,

$$\tfrac{1}{2}\left(\frac{d\phi}{dT}\right)^2 - \frac{a}{\phi} = 0$$

This is the same as the Newtonian equation for the radial motion of a particle under the influence of a fixed point mass (or the relative motion of 2 point particles).

Since ϕ increases with time, distances (as measured by ds) between slowly moving objects (such as the dust particles, but also the stars and galaxies to which they are an approximation) also increase. This is true in spite of the fact that such objects are not moving with respect to the natural rest frame.

6.3. Red shift

The most obvious effect of the cosmological expansion is the cosmological “red shift”. The expansion of the universe causes photons to lose energy, including those of the black-body radiation of the universe as well as those emitted long ago from distant sources.

Since the Universe is approximately translation invariant in the spatial directions, spatial momentum p^i is conserved. (For example, vary the particle action with respect to x^i.) This tells us nothing for the dust, but for the radiation we still have

$$0 = p^2 = -E^2 + (p^i)^2$$

and thus E also is conserved. But this is E as defined with respect to t, not T. (For example, it appeared in the action as $\dot{t}E$. Also, the above equation is for $p^m = v^{-1}dx^m/d\tau$ with $dx^2 = 0$.) However, the time measured by clocks at rest is T, and thus the energy $\hat{E}$ that is measured is with respect to T. In terms of canonical conjugates as defined in a Lagrangian or Hamiltonian, we see this as

$$\dot{t}E = \dot{T}\hat{E} \quad \Rightarrow \quad \hat{E} = \phi^{-1}E$$

using $dT = \phi\ dt$. In particular, for the dust particles we have $\hat{E} = m$.

Actually, this is true for all components of the (4-)momentum: At any fixed point $\mathring{x}^m$, we always choose coordinates near that point such that the proper time looks like the usual one, i.e., $\phi(\mathring{x}) = 1$. This can always be accomplished by a scale transformation: Since we have conformal invariance, we are allowed to choose a reference frame by not only choosing an origin (translation) and orientation of the axes (Lorentz transformation), but also the scale (and even acceleration, via conformal boost). Rather than make this scale transformation explicitly, we simply note that the measured momentum is actually

$$\hat{p}^m = \phi^{-1}p^m$$

For example, for massive particles we then have $\hat{p}^2 + m^2 = 0$.

Since E is conserved but $\hat{E}$ is measured, we thus have $\hat{E} \sim \phi^{-1}$. Therefore, observers measure the photon’s energy, frequency, and corresponding black-body radiation (whose distribution depends only on energy/temperature) as having time dependence $\sim \phi^{-1}$ (and wavelength as ϕ). The spectrum of radiation emitted by a distant object is then shifted by this energy loss, so the amount of shift determines how long ago it was emitted, and thus the distance of the emitter.

Similar remarks apply to observed energy densities: When using variations with respect to external fields, we used $\delta^4(x-X)$'s: For the observer's coordinates, this will be multiplied by ϕ^{-4} (since dx is multiplied by ϕ). Thus the observed energy density is

$$\hat{T}^{00} = T^{00}\phi^{-4} = a\phi^{-3} + \tfrac{1}{2}b\phi^{-4}$$

Astronomers use 3 parameters which are more directly observable. The "size" of the Universe ϕ is difficult to measure, but we can measure the change in time of this scale through red shifts: Comparing lengths at different times, we measure $\phi(T_2)/\phi(T_1)$, more conveniently represented in terms of the difference of the ln: In terms of the derivative, we have

$$ln\left(\frac{\phi(T_2)}{\phi(T_1)}\right) \equiv \int_{T_1}^{T_2} dT\ H(T)$$

or

$$H \equiv \frac{d\phi/dT}{\phi}$$

The "Hubble constant" H (constant in space, not time) measures the expansion rate, and gives an inverse length (time) scale. Thus it is not predicted, but determined from observations. As for all cosmological quantities, it is difficult to measure, its value is based on various astrophysical assumptions, and its quoted value has changed often and by large amounts over the years. A recent estimate for its present value is

$$H^{-1} = 14(2)\cdot 10^9\ yrs.$$

In "natural (Planck) units," $c = G = \hbar = 1$, $H^{-1} = 8(1)\cdot 10^{60}$.

We can also define a dimensionless "density parameter" σ by using H^{-1} as a length scale: However, in the simplified model we have used, it is already fixed

$$\sigma \equiv \frac{\hat{T}^{00}}{H^2} = \tfrac{1}{2}$$

In the more general (relativity) case, this parameter measures energy density with respect to the amount needed to "close" the universe; in this case, it takes the "critical" value, bordering between open and closed. However, this value agrees with observations to within experimental error. This alone shows that the dilaton is sufficient to give an accurate cosmological model (although ingredients other than those discussed so far may be needed).

The rate of change of the Hubble constant can be defined in terms of a dimensionless quantity by comparing its inverse with the true time:

$$q \equiv \frac{d(H^{-1})}{dT} - 1$$

or

$$q \equiv -\frac{\phi\ d^2\phi/dT^2}{(d\phi/dT)^2}$$

The "deceleration parameter" q tells how fast the expansion rate is slowing down. In the case of pure dust $q = \frac{1}{2}$, while for pure radiation $q = 1$; otherwise, it's somewhere in between. Various methods of observation indicate a value

$$0 \lesssim q < 2$$

Recent supernova observations suggest q might even be negative (based on the assumption of the supernova as a "standard candle").

Exercise 6.3.1

Calculate H, q and σ in terms of a, b, and t.

Although this "experimental" value for q is highly unreliable, and its estimate varies widely from year to year based on methods of measurement and choice of assumptions (as well as author), the existence of measurements indicating $q < \frac{1}{2}$ suggests the above model of energy coming from just dust and radiation may be too simple. In fact, other observations indicate the vast majority of energy in the Universe (about 95%!) is not in any known form. While some forms of proposed missing matter ("dark matter") seem to fit into the above types (but are simply not observed by non-gravitational methods), others ("dark energy") do not, and seem to form the majority of the missing energy. One simple remedy is to introduce a "cosmological constant" term (or its equivalent) into the action: In the language of the dilaton, it takes the form

$$S_\Lambda = \Lambda \int d^4x\ \phi^4$$

where Λ is the cosmological constant. This term preserves conformal invariance. (Its scale invariance is obvious by dimensional analysis.) Unfortunately, it makes the dilaton field equation nonlinear, so we no longer have a simple closed solution as before. (Numerical methods are required.) Furthermore, the observed value of this constant corresponds to a length scale of the order of the size of the observed Universe. While this can be explained for the Hubble constant, since it varies with time, there is no "natural" way to explain why a true constant should just happen to set a scale comparable to the present value of the Hubble constant (i.e., there is an unexplained 10^{60} floating around). One possibility is that it is dynamically generated as a vacuum value of another scalar field, and thus might vary with time.

Exercise 6.3.2

Show explicitly that the cosmological term is invariant under a conformal boost.

7. Schwarzschild solution

7.1. Metric

We have seen a metric tensor $g_{mn}(x)$ appearing in the action for a particle, and considered its effect on the particle's equation of motion. It appears by modifying the definition of proper time as

$$-ds^2 = dx^m dx^n g_{mn}(x)$$

Unlike the electromagnetic potential $A_m(x)$, whose coupling to the particle we have also considered, this "field" is nonvanishing even in empty ("flat") space, where it takes as its ("vacuum") value the Minkowski metric, $g_{mn}(x) = \eta_{mn}$. (We also studied the metric in more detail in the one-dimensional case, where the "worldline metric" v defined an intrinsic length for the worldline, and was necessary for writing even the free action for the particle.)

Einstein's field equations for the metric tensor in general relativity are similar to Maxwell's equations for the vector potential, except that energy-momentum is the source of the field instead of charge. (We have seen this previously in defining the energy-momentum tensor by varying the matter action with respect to the metric.) Although the electromagnetic field itself has no charge, the gravitational field has energy-momentum, so it interacts with itself. Rather than derive and solve these nonlinear equations, we first state a solution and consider its consequences.

All gravitational experiments outside of cosmology are based on the "Schwarzschild solution" to Einstein's equations. It describes spherical symmetry in "empty space", outside any region with matter (planet, star, ...). Time independence is then a consequence of spherical symmetry (Birkhoff's theorem). The explicit form of the metric is given by

$$-ds^2 = -\left(1 - \frac{2GM}{r}\right) dt^2 + \left(1 - \frac{2GM}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + sin^2\theta\ d\phi^2)$$

where the normalization $2GM$ (in units $c = 1$) was fixed by comparing to the nonrelativistic result at large distances. The coordinates used, named "t, r, θ, ϕ", reduce to the usual spherical coordinates in the absence of matter ($M = 0$). However, we have seen that the "gauge invariance" for the metric tensor is coordinate transformations, and there is considerable leeway in redefining these coordinates (as we'll see later) while still taking a simple form that relates to spherical coordinates.

This solution can be generalized with electric charge e and magnetic charge g. The result is modified because the electromagnetic field (which also carries energy-momentum) of the charge distribution extends outside of the matter, to infinity. The only change is the replacement

$$1 - \frac{2GM}{r} \to 1 - \frac{2GM}{r} + \frac{G(e^2 + g^2)}{r^2}$$

When comparing to the real world, it is useful to know some astrophysical radii:

(1) Earth's orbit (1 AU): 1.5 $\times 10^8$ km

(2) Solar radius: 7 $\times 10^5$ km

(3) Earth radius: 6000 km

(4) Solar gravitational (Schwarzschild) radius ($2GM$): 3 km

(5) Earth gravitational radius: 0.9 cm (1 shoe size).

Exercise 7.1.1

Consider the following very crude approximations to various types of stars:

a Assume a star has the density of a neutron, i.e., of a sphere with the mass M of the neutron and radius equal to the Compton radius $\hbar/Mc$. Assume also that the radius of this (spherical) star is equal to its gravitational radius. (This is roughly a "neutron star".) Find the mass and radius, in terms of both physical constants and conventional units. Note the appearance of the large dimensionless number, the ratio of the Planck mass to the neutron mass.

b Assume a star has the density of a "compressed" hydrogen atom, a sphere with the mass of the hydrogen atom (which we can take as roughly equal to the neutron mass) and radius equal to the Compton radius of the *electron*, $\hbar/mc$ for electron mass m. Assume the mass of this star is equal to that of the neutron star found in the previous example. (This is roughly a "white dwarf".) Find the radius, again in terms of both physical constants and conventional units.

c Assume the same mass again, but now assume the density of an ordinary hydrogen atom, which has the Bohr radius $\hbar^2/me^2$. (This is roughly an "ordinary" star.) Compare to the mass and radius of the Sun.

7.2. Gravitational redshift

All experiments (excluding cosmology) are based on the Schwarzschild metric. The first type of experiment involves gravitational redshift, but unlike the cosmological case, the relevant reference frames of observation are not frames of "free fall" but the static reference frame in which the Schwarzschild metric is defined, which requires a non-gravitational force (like standing on the Earth) to maintain. (There are also measurements of redshift from airplanes, whose reference frame is defined with respect to the Schwarzschild one.) In this reference frame the relevant conserved quantity is the one which expresses the fact that the space is static, the energy. As in the cosmological case, the observer uses coordinates that are flat at his position.

We thus need to consider 3 properties of a particle, all of which we have loosely considered as "momentum:"

$$p^m = v^{-1}\frac{dx^m}{d\tau} \quad \Rightarrow \quad p^m p^n g_{mn} + m^2 = 0$$

$$p_m = -\frac{\partial L_H}{\partial \dot{x}^m} \quad \Rightarrow \quad p_m p_n g^{mn} + m^2 = 0$$

$$\widehat{p}^a : \quad \widehat{p}^a \widehat{p}^b \eta_{ab} + m^2 = 0$$

The first is defined directly in terms of the path; it is "kinetic". The second is "canonically conjugate" to position: Since it appears in the Hamiltonian form of the action as $-\dot{x}^m p_m$, it is the quantity that has simple commutation relations with x quantum mechanically. (Indices are raised and lowered with the metric g_{mn} and its inverse g^{mn}.) More importantly, if the Hamiltonian is independent of some component of x^m (a symmetry), it is the corresponding component of p_m that is conserved (as is obvious from varying the action with respect to that component of x^m). The third is simply some linear combination of either p^m or p_m. As in the cosmological case, it corresponds to a different coordinate choice, but since we need these coordinates only at the point of observation, we just find the desired linear combination, the one that makes it satisfy the flat-space energy-momentum relation at that point only: In this case,

$$-m^2 = g^{mn} p_m p_n = -\left(1 - \frac{2GM}{r}\right)^{-1} p_t^2 + \left(1 - \frac{2GM}{r}\right) p_r^2 + \frac{1}{r^2} p_\theta^2 + \frac{1}{r^2 sin^2\theta} p_\phi^2$$

$$\Rightarrow \quad \hat{p}_t = \left(1 - \frac{2GM}{r}\right)^{-1/2} p_t, \quad \hat{p}_r = \left(1 - \frac{2GM}{r}\right)^{1/2} p_r,$$

$$\hat{p}_\theta = \frac{1}{r} p_\theta, \quad \hat{p}_\phi = \frac{1}{r\ sin\ \theta} p_\phi$$

We can therefore write

$$E = \sqrt{1 - \frac{2GM}{r}}\widehat{E}$$

where $\widehat{E} = -\hat{p}_t$ is the energy of a particle as measured by the observer, while $E = -p_t$ is the conserved quantity. Thus, conservation of E for a photon gives the r-dependence of the observed energy $\widehat{E}$ (and thus the frequency, which in turn determines the wavelength, since $\hat{p}^2 = 0$).

To compare with nonrelativistic mechanics, we instead evaluate E for a massive particle in the Newtonian limit:

$$E \approx \left(1 - \frac{GM}{r}\right)(m + K) \approx m + K - \frac{GMm}{r}$$

giving the "conserved energy" E in terms of the "particle energy" $\widehat{E}$ (rest mass m + kinetic K), including the potential energy.

7.3. Geodesics

The other type of experiment involves properties of "geodesics", paths of extremal length, where length s is now defined by the metric g_{mn}. So we need to solve the geodesic equations of motion, as found from the particle action. Without loss of generality, we can choose the angular coordinates such that the initial position and direction of the particle is in the equatorial plane $\theta = \pi/2$, where it remains because of the symmetry $\theta \leftrightarrow \pi - \theta$, as in the nonrelativistic case. Also as in the nonrelativistic case, we can find constants of the motion corresponding to the energy $E = -p_t$ and (z-component of) angular momentum $L = p_\phi$. Using the affine parametrization $v = 1$ without loss of generality (the v's would cancel anyway when we canceled $d\tau$'s),

$$E \equiv -g_{tm}\dot{x}^m = \left(1 - \frac{2GM}{r}\right)\dot{t}, \quad L \equiv g_{\phi m}\dot{x}^m = r^2\dot{\phi}$$

In the case where the particles come from infinity, these are the initial kinetic energy and angular momentum. We also need the energy-momentum relation

$$-m^2 = g_{mn}\dot{x}^m\dot{x}^n = -\left(1 - \frac{2GM}{r}\right)\dot{t}^2 + \left(1 - \frac{2GM}{r}\right)^{-1}\dot{r}^2 + r^2\dot{\phi}^2$$

Solving the previous equations for $\dot{t}$ and $\dot{\phi}$, this reduces to the radial equation

$$0 = -E^2 + \dot{r}^2 + \left(1 - \frac{2GM}{r}\right)\left(\frac{L^2}{r^2} + m^2\right)$$

$$\Rightarrow \quad \tfrac{1}{2}\dot{r}^2 + \left(-\frac{GMm^2}{r} + \frac{L^2}{2r^2} - \frac{GML^2}{r^3}\right) = \tfrac{1}{2}(E^2 - m^2)$$

This looks like a typical nonrelativistic Hamiltonian for "energy" $\frac{1}{2}(E^2 - m^2)$ with the same terms as in the Newtonian case but with an extra r^{-3} term. (To take the nonrelativistic limit for the massive case, first scale the affine parameter $\tau \rightarrow s/m$.) Since there are good coordinate systems for a "black hole" using r as a coordinate (e.g., see the following subsection: r and $r'' + t''$, as seen from the figure for Kruskal-Szkeres), this equation can even be used to descibe a fall into a black hole. (For example, for $L = 0$ we get the same cycloid solution as in cosmology and in Newtonian gravity, reaching the singularity at $r = 0$ in finite proper time.)

Because of the r^{-3} term in the potential, noncircular orbits are no longer closed. In particular, let's consider orbits which are close to circular. Circular orbits are found by minimizing the potential for the r-equation:

$$0 = \frac{dV}{dr} = \frac{GMm^2}{r^2} - \frac{L^2}{r^3} + \frac{3GML^2}{r^4}$$

$$0 < \frac{d^2V}{dr^2} = -\frac{2GMm^2}{r^3} + \frac{3L^2}{r^4} - \frac{12GML^2}{r^5}$$

The near-circular orbits are described by small (harmonic) oscillations about this minimum, with angular frequency given by

$$\omega_r^2 = \frac{d^2V}{dr^2} = \frac{GMm^2(r - 6GM)}{r^3(r - 3GM)}$$

from solving for $L^2 = GMm^2r^2/(r - 3GM)$. On the other hand, the frequency of the circular orbit itself in terms of its angular dependence is just $\dot{\phi} = L/r^2$, giving

$$\omega_\phi^2 = \frac{GMm^2}{r^2(r - 3GM)}$$

This means that the perihelion (closest approach to the Sun) of an orbit, which occurs every period $2\pi/\omega_r$ of the radial motion, results in the change of angle

$$2\pi + \delta\phi = \int_0^{2\pi/\omega_r} d\tau \frac{d\phi}{d\tau} = \frac{2\pi}{\omega_r}\omega_\phi = 2\pi\left(1 - 6\frac{GM}{r}\right)^{-1/2}$$

$$\Rightarrow \quad \delta\phi \approx 6\pi\frac{GM}{r}$$

in the weak-field approximation. This effect contributes to the measurement of the precession of the perihelion of the (elliptical) orbit of Mercury, but so do the precession of Earth's axis, the oblateness of the Sun, and gravitational interaction with other planets. As a result, this relativistic effect contributes less than 1% to the observed precession. In particular, the solar oblateness is difficult to measure.

The effects on geodesics of photons are much easier to measure, since there are no Newtonian effects. As a result, the weak field approximation is sufficient. We first consider bending of light by the Sun: A photon comes in from infinity and goes back out to infinity (actually to the Earth, which we assume is much farther from the Sun than the photon's closest approach to it), and we measure what angle its trajectory was bent by. (For example, we look at the apparent change of position in stars when the Sun passes in their direction during an eclipse.) Starting with the exact solution for a photon's geodesic (case $m^2 = 0$ above), we use the equations for $\dot{r}$ and $\dot{\phi}$ to find

$$\frac{dr}{d\phi} = \sqrt{\frac{E^2}{L^2}r^4 - r^2 + 2GMr}$$

Changing variables,

$$u \equiv \frac{b}{r}, \quad b \equiv \frac{L}{E}, \quad a \equiv \frac{GM}{b} \qquad \Rightarrow \qquad d\phi = \frac{du}{\sqrt{1 - u^2 + 2au^3}}$$

The *impact parameter* $b \equiv L/E$ would be the closest approach to the Sun neglecting gravitational effects ($L = rp = bE$). We now make the weak field approximation: For a small,

$$\begin{aligned} d\phi &\approx \frac{du}{\sqrt{1-u^2}}\left(1 - a\frac{u^3}{1-u^2}\right) \\ &= d\chi\left(1 - a\frac{sin^3\chi}{cos^2\chi}\right) \qquad (u \equiv sin\ \chi) \\ &= d\left[\chi - a\left(cos\ \chi + \frac{1}{cos\ \chi}\right)\right] \end{aligned}$$

Defining $\phi = 0$ at $r = \infty$, the integral is

$$\phi \approx \chi - a\frac{(1 - cos\ \chi)^2}{cos\ \chi} \qquad \Rightarrow \qquad \chi \approx \phi + a\frac{(1 - cos\ \phi)^2}{cos\ \phi}$$

$$\Rightarrow \qquad u = sin\ \chi \approx sin\ \phi + a(1 - cos\ \phi)^2$$

The change in ϕ from incoming photon to outgoing photon follows from the two solutions for $r = \infty$:

$$u = 0 \qquad \Rightarrow \qquad \phi = 0,\ \pi + 4a$$

Therefore the deviation of ϕ from a straight line is $4GME/L$. (Mathematical note: All variable changes were those suggested by the flat space case $a = 0$: E.g., $b/r = sin\ \chi$, where χ is what ϕ would be in flat space.)

A similar experiment involves measuring the round-trip travel time for radio waves from Earth to some reflector (on another planet or an artificial solar satellite), with and without the Sun near the path of the waves. Now, instead of $dr/d\phi$ we want, in units $b = 1$

$$\frac{dr}{dt} = \left(1 - \frac{2a}{r}\right)\sqrt{1 - \frac{1}{r^2} + \frac{2a}{r^3}}$$

$$\begin{aligned} \Rightarrow \quad dt &\approx \frac{rdr}{\sqrt{r^2-1}} + 2a\frac{dr}{\sqrt{r^2-1}} \\ &= d\left[\sqrt{r^2-1} + 2a\ cosh^{-1}r\right] \\ &= d\left[\sqrt{r^2-b^2} + 2GM\ cosh^{-1}\frac{r}{b}\right] \end{aligned}$$

putting the b's back. The first term in dt is the nongravitational piece; it is the length of the side of a triangle whose other side has length b and whose hypotenuse has length r. For simplicity we assume both orbits are circular, so their r's are fixed; the change in b then comes from those radii differing, so they revolve around the sun at different rates.

We have neglected the gravitational effect on $\dot{r}$ (the a term inside the original $\sqrt{\ }$), which is negligible compared to that on $\dot{t}$, since for most of the path $r \gg b$: It can be estimated by (1) noting the argument of the square root is exactly 0 at r_{min}, (2) looking at $d(r\sqrt{...})$, and noting its deviation from the exact result goes as a/r^3 times the usual, which is less than a/b^3, giving a contribution of order $2GMb/r_{max}$, and thus negligible.

We then integrate from $r = r_{min} \approx b$ to $r = r_{Earth}$, add the integral from $r = r_{min}$ to $r = r_{reflector}$, multiply by 2 for the round trip, and throw in a factor to convert to the proper time s of the observer (which turns out to have a negligible effect to this order in a). This result is then compared to the same measurement when both observer and reflector have revolved further about the Sun, so b changes significantly (but not r_{Earth} nor $r_{reflector}$). For $x \gg 1$, $cosh^{-1}x \approx ln(2x)$, so the contribution to $\Delta t/\Delta b$, for $b \ll r_{Earth}$ and $r_{reflector}$, coming from the *cosh* term is given by

$$\Delta s \approx -8GM\Delta(ln\ b)$$

7.4. Black holes

For physical massive bodies the Schwarzschild solution applies only outside the body, where $T_{ab} = 0$. The form of the solution inside the body depends on the distribution of matter, which is determined by its dynamics. Generally the surface of the body is at $r \gg GM$, but we can try to find a solution corresponding to a point mass by extending the coordinates as far as possible. The Schwarzschild metric is singular at $r = 2GM$. In fact, r and t switch their roles as space and time coordinates there. This unphysical singularity can be eliminated by first making the coordinate transformation, for $r > 2GM$,

$$r' = \int dr \left(1 - \frac{2GM}{r}\right)^{-1} = r + 2GM\ ln\left(\frac{r}{2GM} - 1\right)$$

and then making a second coordinate transformation by rescaling the "lightcone" coordinates as

$$r'' \pm t'' = 4GMe^{(r' \pm t)/4GM} = 4GM\sqrt{\frac{r}{2GM} - 1}\ e^{(r \pm t)/4GM}$$

The result is the "Kruskal-Szekeres coordinates"

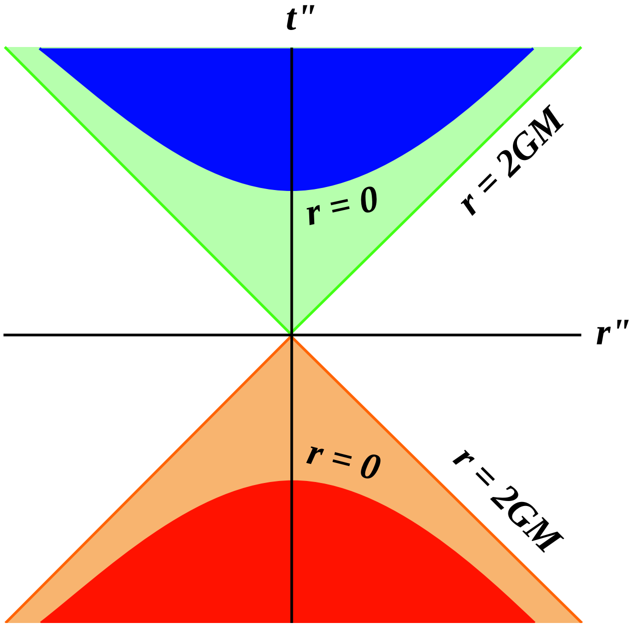
$$-ds^2 = \frac{2GM}{r} e^{-r/2GM}(-dt''^2 + dr''^2) + r^2(d\theta^2 + sin^2\theta\ d\phi^2)$$

where $r(r'', t'')$ is defined by

$$r''^2 - t''^2 = (4GM)^2 \left(\frac{r}{2GM} - 1\right) e^{r/2GM}$$

This can now be extended past $r = 2GM$ down to the physical singularity at $r = 0$.

The complete space now looks like (plotting just r'' and t''):



In this diagram lines at 45° to the axes represent radial lightlike geodesics. Since nothing travels faster than light, this indicates the allowed paths of physical objects. Curves of fixed r are hyperbolas: In particular, the physical singularity is the curve

$t''^2 - r''^2 = (4GM)^2$ $(r = 0)$, while $t''^2 - r''^2 = 0$ $(r = 2GM)$ is the "event horizon" which allows things to go only one way (out from the bottom half or into the top half), and $r = \infty$ is both $r'' = \pm\infty$. Nothing can communicate between the 2 "outside worlds" of the left and right 90° wedges. In particular, a star which collapses ("gravitational collapse") inside its "gravitational radius" $2GM$ is crushed to a singularity, and the spherically symmetric approximation to this collapse must be represented by part of the Kruskal-Szekeres solution (outside the star) by Birkhoff's theorem, patched to another solution inside the star representing the contribution of the matter (energy) there to the field equations. This means using just the top and right 90° wedges, with parts near the left edge of this modified appropriately. The top wedge is called a "black hole". (If a situation should exist described by just the bottom and right wedges, the bottom wedge would be called a "white hole".) Similarly, stable stars are described by just the right wedge, patched to some interior solution. This right wedge represents the original Schwarzschild solution in the region $r > 2GM$ where its coordinates are nonsingular. In that region lines of constant t are just "straight" radial lines in the Kruskal-Szekeres coordinate system $(r'' \sim t'')$.

Besides the fact that nothing can get out, another interesting feature of the black hole is that an outside observer never sees something falling in actually reach the event horizon: Consider an observer at fixed $r > 2GM$ using Schwarzschild coordinates, so his proper time $s \sim t$. Then light radiating radially from an in-falling object is received later and later, up till $t = \infty$, by the observer as the object approaches the event horizon, although it takes the object a finite amount of proper time to reach the event horizon and the physical singularity.

Exercise 7.4.1

> Apply the equations of motion in a Schwarzschild metric of the previous subsection for a massive object falling straight into a black hole (angular momentum $L = 0$): Solve for r, t, τ in an appropriate parametrization to show that it takes a finite proper time to reach the event horizon from any finite r outside it.

There are also more complicated black-hole solutions with spin and electric charge.

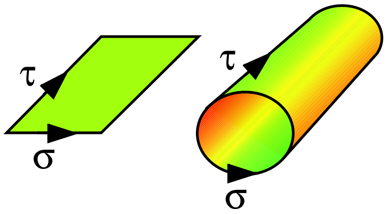
Another interesting effect of the event horizon is the eventual decay of the black hole ("Hawking radiation"): Pair creation can result in a similar way to that in an electrostatic potential of sufficient strength (see exercise 4.4.1). Particles are emitted near the event horizon (the edge of the gravitational barrier), carrying energy off to infinity, while their antiparticles fall into the singularity.

There are two features of the black hole that are less than desirable: the existence of singularities indicates a breakdown in the field equations, and the existence of event horizons results in an “information loss”. Both these properties might be avoidable quantum mechanically: Quantum effects can modify the short-distance behavior of the theory. One might think that such effects would be only at short distances away from the singularity, and thus remove the singularities but not the event horizons. However, it is possible (and examples of such solutions have been given) that the prevention of the creation of the singularity in stellar collapse would eventually result in a reversal of the collapse (“gravitational bounce”): The would-be black hole solution is patched to a would-have-been white hole by short-distance modifications, resulting in an exploding star that initially resembled a black hole but has no true event horizon.

8. Strings

8.1. Geometry

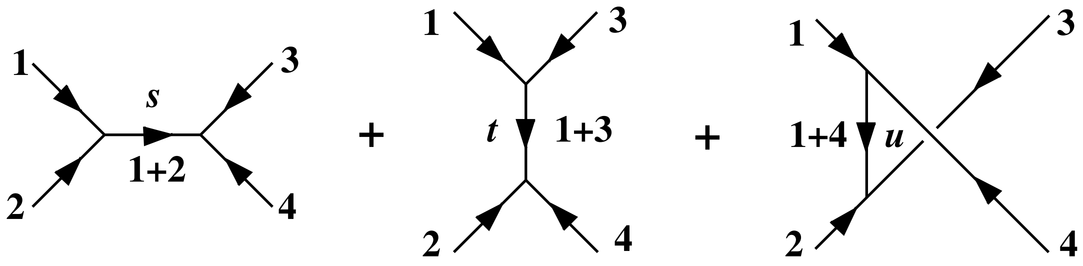
The defining concept of the string is that it is a two-dimensional object: Just as the particle is defined as a point object whose trajectory through spacetime is one-dimensional (a worldline), the string has as its trajectory a two-dimensional surface, the "worldsheet". There are two types of free strings: open (two ends) and closed (no boundary). Their worldsheets are a rectangle and a tube (cylinder).



To get some motivation for the string, we consider how strings might be formed as bound states of ordinary particles. There is some geometry associated with the interactions of particles/waves in quantum field theory: We draw a "Feynman diagram/graph", with (1) vertices representing the collisions of the particles, where they interact, (2) "internal" lines connecting the vertices representing the paths of the particles, where they act free, and (3) "external" lines with one end attached to a vertex and one end unattached, representing particles in the initial or final state. If there are any closed paths in this diagram, it is called a "loop" graph; otherwise, it is a "tree" graph. (Graphs can also be disconnected, but for simplicity we can consider just connected ones.)

Each diagram is associated with a quantum-mechanical probability amplitude. These diagrams are generally evaluated in momentum space: We then can associate a particular momentum with each line, and momentum is conserved at each vertex. An arrow is drawn on each line to indicate the direction of "flow" of the momentum. (Otherwise there is a sign ambiguity, since complex conjugation in position space changes the sign of the momentum.) Then the sum of all momenta flowing into (or all out of) a vertex vanishes. The momentum associated with a line is then interpreted as the momentum of that particle, with the arrow indicating the direction of flow of the proper time τ. (p changes sign with τ.)

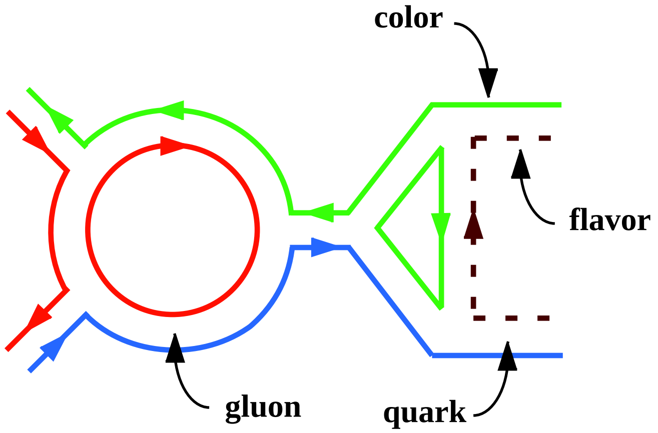
When evaluating an amplitude ("S-matrix element"), we assign a "propagtor" (Green function, for solution of the free wave equation) to each internal line, a single-particle wave function to each external line, and to each vertex a factor (perhaps with momentum dependence) corresponding to an interaction term in the action. The fact that the external-line wave functions satisfy the free wave equation means the external momenta are on-(mass-)shell ($p^2 + m^2 = 0$); on the other hand, this is not true of the momenta on the internal lines, even though those particles are treated as free.



The simplest nontrivial tree graphs are 4-point amplitudes (4 external lines). We now label all momenta as incoming, which is convenient for symmetry, and corresponds naturally to using incoming (initial) states with positive energy and outgoing (final) states with negative energy (as from the complex-conjugate final wave functions). These momenta are conveniently expressed in terms of the Mandelstam variables we considered earlier: with these signs,

$$s = -(p_1 + p_2)^2, \qquad t = -(p_1 + p_3)^2, \qquad u = -(p_1 + p_4)^2$$

We also use the convention that s is defined in terms of the momenta of the two initial particles (and we also use this same definition when there are more than two final particles); t and u are then more or less interchangable, but if initial and final particles are pairwise related we choose t in terms of the momenta of such a pair.

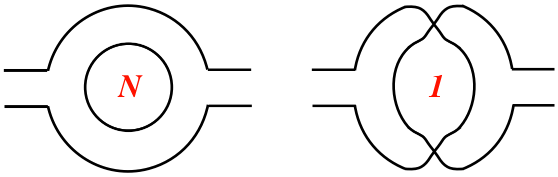


In some field theories the fields appearing in the action can carry indices not associated with spin. These are related to “internal symmetry”, such as isospin. In many cases of interest, the field carries 2 of these indices, not necessarily of the same type, making the field a matrix (not necessarily square). Terms in the action then involve a trace of products of these matrices. When we draw a Feynman diagram for such field theories, instead of a single line for each propagator, we draw a double (parallel) line, each line corresponding to one of the two indices on the matrix field. Because of the trace in the vertex, the propagator lines connect up there in such a way that effectively we have continuous lines that travel on through the vertices, although the two lines paired in a propagator go their separate ways at the vertex. These lines never split or join, and begin or end only on external fields. We can also draw arrows on the lines, pointing in the same direction everywhere along a single line, but pointing in opposite directions on the two lines in any propagator pair: This keeps track of the fact that ϕ appears in the trace always multipled as $\phi\phi$ and never as $\phi\phi^T$.

A physical picture we can associate with this is to think of the original particle as represented by the field as a bound-state of a particle-antiparticle pair, with one line associated with the particle and another with the antiparticle; the arrows are then oriented in the direction of time of this particle (which is the opposite of the direction of time for the antiparticle). These internal symmetry lines are usually associated with either “color” (associated with the strong interactions) or “flavor” (associated with the electroweak interactions). If the original particle was a meson, then the symmetry is flavor, and the constituent particles are a quark and an antiquark, with only the flavor of the quark explicitly showing. (Their color is “confined”.) On the other hand, if the original particle was a quark (one flavor line, one color) or gluon (2 color lines), then the constituent particles are a “preon” and antipreon, each of which carries just color or just flavor (with perhaps some new hidden symmetry causing their binding).

Group theory factors associated with this internal symmetry are trivial to follow in these diagrams: For example, in a diagram with just mesons, the same flavor quark continues along the extent of a “flavor line”; thus, there is a Kronecker δ for the two indices appearing at the ends of the flavor line (at external fields); each flavor of quark is conserved (by interactions involving only mesons).

To study both color and flavor (and for definiteness), we consider quantum chromodynamics (QCD), whose fundamental particles are quarks and gluons. Gluons have self-interactions, but quarks interact only with gluons. We now notice that in some loop graphs, depending on how the lines are connected, some of the color lines (e.g.) form closed loops. The group theory is trivial: There is a factor of N for each such loop, from the sum over the N colors.



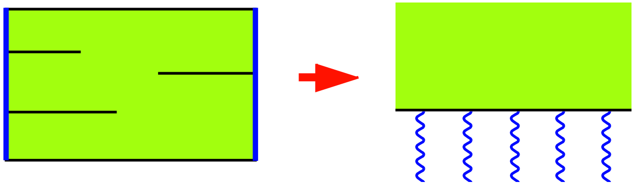
We can also give a physical picture to these numerical factors: Since we draw the propagator as preon and antipreon lines with finite separation, think of the particle as a (very short) string, with a preon at one end and antipreon at the other. This gives a two-dimensional structure to the diagram, by associating a surface with the area between the preons and antipreons (including the area at the vertices). We can extend this picture by associating a surface also with the area inside (i.e., on the other side of) each closed preon loop. In particular, for any "planar" diagram, i.e., any diagram that can be drawn on a sheet of paper without crossing any lines, and with all external lines on the outside of the diagram, the entire diagram forms an open sheet without holes, and with the topology of a disc (simply connected). It is also clear that, for a fixed number of loops and a fixed number of external lines, a planar diagram has the greatest number of factors of N, since crossing lines combines color loops and reduces the power of N.

A more careful analysis shows a direct relationship between powers of 1/N and topology: For any given number of vertices, for a diagram with just gluons, the leading order in 1/N comes from a diagram that can be drawn on a sphere. The next order is smaller by $(1/N)^2$, and can be drawn on a torus (doughnut), which is topologically the same as a sphere with a "handle". Similar remarks apply to higher orders, so the power of $(1/N)^2$ counts the number of handles added to the sphere.

When a quark field makes a closed loop, it looks like a planar loop of a gluon, except that a closed color line is missing, along with a corresponding factor of N. Thus, there is effectively a "hole" in the surface. Since only one factor of N is missing, a hole counts as half a handle. However, there is a closed flavor line in place of the missing closed color line, so we get instead a factor of M (for M flavors) for each quark loop. (Any quark loop gives a closed flavor line, since in QCD there are no flavor interactions: The interactions are all mediated by gluons, which carry only color.) For example, one hole in a sphere gives a disk: A single quark loop on a planar diagram gives a disk, with its one closed flavor line forming its boundary.

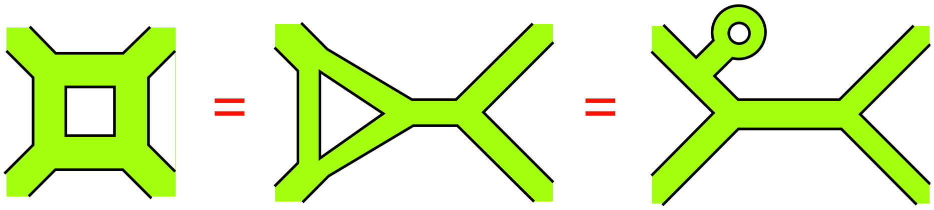
We can then classify diagrams by a perturbation expansion in 1/N, with the planar diagrams representing the leading order. The fact that the 1/N expansion is topological (the power of 1/N is the number of holes plus twice the number of handles) closely ties in with the experimental observation that hadrons (in this case, mesons) act like strings. Thus, we can expand in 1/N as well as in loops. While the leading order in the loop expansion is classical (particle) field theory, the leading order in the 1/N expansion is classical open-string theory (planar graphs). However, seeing the dynamical string properties requires summing to all orders in loops for leading order in 1/N.

Thus, 1/N acts as the string coupling constant. The experimental fact that the hadronic spectrum and scattering amplitudes follow so closely that of a string indicates that the perturbative expansion in 1/N is accurate, i.e., that quantum corrections are "small" in that sense. One application of the smallness of 1/N (largeness of N) is the "Okubo-Zweig-Iizuka rule": A planar graph describes classical scattering of open strings (mesons). It corresponds topologically to a disc, which is a sphere with one hole, and is therefore order 1/N. Compare this to two planar graphs connected by a handle. It describes classical scattering of open strings with one intermediate closed string (”glueball”), where the handle is a closed-string propagator connecting two otherwise-disconnected classical open-string graphs. It corresponds to a cylinder, which is a sphere with two holes, and is therefore order $1/N^2$. In terms of *flavor* lines, the latter graph differs from the former in that it has an intermediate state (the glueball) with no flavor lines. The OZI rule is that amplitudes containing an intermediate glueball are always smaller than those with an intermediate meson. This rule also has been verified experimentally, giving a further justification of the 1/N expansion (though not necessarily of string behavior).

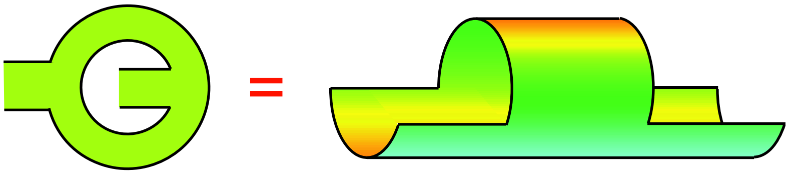


This leads to a much simpler picture of interactions for strings than for particles: For interacting particles, the geometric picture of a worldline becomes a graph, whose geometry is singular at the interaction points. For interacting strings, we have instead a worldsheet with nontrivial topology: sphere, disk, torus, etc. For example, a string tree graph now looks more like a real tree, in that the branches now have thickness, and they join smoothly to the rest of the tree.

Because of the local scale invariance of string theory, different string graphs are distinguished by their topology rather than their geometry: Any two graphs that can be "stretched" into one another are equivalent. In particular, in any loop graph any hole or window can be pulled out so that it appears as a "tadpole". The result is that any graph is equivalent to a tree graph with insertions of some one-loop open- or closed-string tadpoles.



However, this does not mean that any graph constructed with only open-string propagators and interactions can be expressed as an open-string tree graph with tadpole insertions: The one-loop open-string graph with two "half-twists" on the open-string propagators in the loop is equivalent to a tree graph with a closed-string intermediate state, as can be seen by stretching the surface, or by tracing the routes of the boundaries. (For example, drawing this graph in a psuedo-planar way, as a flat ring with external states connected to both the inner and outer edges, pulling the inner edge out of the plane reveals a closed string connecting the two edges.)



8.2. Classical mechanics

We now consider string theory as derived by first quantization. As for particles, the first step is to study the classical mechanics, which determines the appropriate set of variables, the kinetic term of the field theoretic action, some properties of the interactions, and some techniques useful for perturbation. Just as the simplest such action for the particle produces only the relatively uninteresting case of the scalar, the most obvious action for the string yields a model that is not only too simple, but quantum mechanically consistent only in 26 dimensions. However, this toy model exhibits many relevant qualitative features of string theory.

The simplest classical mechanics action for the string is a direct generalization of that for the massless scalar particle: For the Lagrangian form of this action we write

$$S_L = \frac{1}{\alpha'}\int \frac{d^2\sigma}{2\pi}\sqrt{-g}g^{mn}\tfrac{1}{2}(\partial_m X^a)\cdot(\partial_n X^b)\eta_{ab}$$

where $X^a(\sigma^m)$ is the position in spacetime of a point at worldsheet coordinates $\sigma^m = (\sigma^0,\sigma^1) = (\tau,\sigma)$, $g^{mn}(\sigma^m)$ is the (inverse) worldsheet metric, g is the determinant of g_{mn}, and α' is a normalization constant, the string tension. It can also be associated with the flat-space spacetime metric η_{ab}; if we couple a spacetime metric, then its vacuum value can be taken as η_{ab}/α', where α' is the gravitational coupling.

Exercise 8.2.1

Analyze the classical mechanics of the string by approximating σ by a set of discrete points, so $\partial_\sigma X(\sigma) \to X_{n+1} - X_n$, etc. Choose the gauge $X^0 = \tau$. Show that the string then acts as a bunch of particles connected by springs, and find all the usual spring properties: tension, speed of wave propagation, etc.

A new feature of this action (compared to the particle's) is that it is (2D) Weyl (local) scale invariant:

$$g'_{mn}(\sigma^p) = \lambda(\sigma^p)g_{mn}(\sigma^p) \quad\Rightarrow\quad (\sqrt{-g}g^{mn})' = \sqrt{-g}g^{mn}$$

This gauge invariance can be used to gauge away one component of the metric, in addition to the two that can be gauged away using 2D general coordinate invariance. The net result is that the worldsheet metric can be completely gauged away (except for some bits at boundaries), just as for the particle. However, this same invariance prevents the addition of a worldsheet cosmological term: In the particle case, such a term was needed to introduce mass. Here, mass is introduced through the coefficient $1/\alpha'$ of the $(\partial X)^2$ term: The same scale invariance that prevents use of a cosmological term also prevents this coefficient from being absorbed into the definition of the worldsheet metric.

Just as for the particle, the metric can be eliminated by its equation of motion, resulting in a more geometrical, but less useful, form of the action: In this case the equation of motion ("Virasoro constraints")

$$(\partial_m X)\cdot(\partial_n X) = \tfrac{1}{2}g_{mn}g^{pq}(\partial_p X)\cdot(\partial_q X)$$

after taking the determinant of both sides, gives

$$S = \frac{1}{\alpha'}\int \frac{d^2\sigma}{2\pi}\sqrt{-\tilde{g}}, \qquad \tilde{g}_{mn} = (\partial_m X)\cdot(\partial_n X)$$

This is the area of the string in terms of the "induced" metric $\tilde{g}_{mn}$, analogously to the particle case. The induced metric measures length as usually measured in spacetime:

$$d\sigma^m d\sigma^n \tilde{g}_{mn} = (d\sigma^m \partial_m X) \cdot (d\sigma^n \partial_n X) = (dX)^2$$

Equivalently, this action can be written in terms of the area element $dX^a \wedge dX^b$:

$$S = \frac{1}{2\pi\alpha'} \int \sqrt{-\tfrac{1}{2}(dX^a \wedge dX^b)^2}, \qquad dX^a \wedge dX^b = (d\sigma^0 \partial_0 X^{[a})(d\sigma^1 \partial_1 X^{b]})$$

For purposes of quantization, it's also useful to have the Hamiltonian form of the action. This also allows us to see how the Virasoro constraints generalize the Klein-Gordon equation. By the usual methods of converting from Lagrangian to Hamiltonian, we find

$$S_H = \int \frac{d^2\sigma}{2\pi}(-\dot{X} \cdot P + \mathcal{H}), \qquad \mathcal{H} = \frac{\sqrt{-g}}{g_{11}} \tfrac{1}{2}(\alpha' P^2 + \alpha'^{-1} X'^2) + \frac{g_{01}}{g_{11}} X' \cdot P$$

where $\dot{} = \partial_0$ and $' = \partial_1$. Various combinations of components of the worldsheet metric now appear explicitly as Lagrange multipliers. If we define

$$\hat{P}_{(\pm)} = \tfrac{1}{\sqrt{2}}(\alpha'^{1/2} P \pm \alpha'^{-1/2} X')$$

the constraints can be written as two independent sets $\hat{P}^2_{(\pm)}$.

Since 2D general coordinate (and even just Lorentz) invariance is no longer manifest, for some purposes we need to generalize this to a form that is first-order with respect to both τ and σ derivatives:

$$S_1 = -\frac{1}{\alpha'} \int \frac{d^2\sigma}{2\pi}[(\partial_m X) \cdot P^m + (-g)^{-1/2} g_{mn} \tfrac{1}{2} P^m \cdot P^n]$$

obviously reproduces S_L after eliminating P^m. Eliminating just P^1 gives a simpler way of deriving S_H (with $P^0 = \alpha' P$).

Since open strings have boundaries, the action implies boundary conditions, originating from integration by parts when deriving the field equations. In the last form of the action variation of the first term gives, in addition to the $\int d^2\sigma$ terms $(\delta P) \cdot \partial X$ and $-(\delta X) \cdot \partial P$ for the field equations, a boundary term $\int d\sigma^m \epsilon_{mn} (\delta X) \cdot P^n$, where $d\sigma^m$ is a line integral along the boundary, and the ϵ_{mn} picks the component of P^m normal to the boundary. We thus have

$$n_m P^m = 0 \textit{ at boundaries}$$

where n_m is a vector normal to the boundary. From the constraint imposed by varying g_{mn}, it then follows that

$$(t_m P^m)^2 = 0 \textit{ at boundaries}$$

where t_m is a vector tangent to the boundary (or any vector, for that matter). Since by the field equations $P^m \sim g^{mn} \partial_n X$, this means that the boundary is lightlike in spacetime: The ends of the string travel at the speed of light.

8.3. Gauges

In direct analogy to the particle, the two most useful gauges are the "conformal gauge", defined by completely fixing the worldsheet metric, and the lightcone gauge, which is not manifestly globally covariant but is a complete fixing of the residual gauge invariance of the conformal gauge. In the conformal gauge we set

$$g_{mn} = \eta_{mn}$$

by using the 2 coordinate invariances and the 1 scale invariance to fix the 3 components of the symmetric tensor g_{mn}. The coordinate part of this gauge is essentially the temporal gauge $g_{0m} = \eta_{0m}$, just as for the particle ($-g_{00} = v^2 = 1$). Also as for the particle, this gauge can't be fixed everywhere, but the equation of motion from the metric is implied everywhere by imposing it at the just the boundaries in τ. In this gauge the equations of motion for X are just the 2D Klein-Gordon equation, which is easy to solve in 2D lightcone coordinates:

$$\partial_+\partial_- X = 0 \quad \Rightarrow \quad X = X_{(+)}(\tau+\sigma) + X_{(-)}(\tau-\sigma)$$

(We have used $\tau \pm \sigma$ in place of $\sigma^\pm$ for later convenience.) The constraints are then $\hat{P}^2_{(\pm)} \sim (X'_{(\pm)})^2 = 0$. This directly relates to the form of 2D conformal transformations, which are infinite-dimensional in D=2:

$$ds^2 = 2d\sigma^+ d\sigma^- \quad \Rightarrow \quad \sigma'^+ = f_{(+)}(\sigma^+), \quad \sigma'^- = f_{(-)}(\sigma^-)$$

For the lightcone gauge, we again fix the (spacetime) +-components of the variables, and solve the +-components of the equations of motion (found by varying the −-components). Looking at the equations of motion first, using the first-order form of the action,

$$0 = \frac{\delta S}{P^{-m}} \sim \partial_m X^+ + (-g)^{-1/2} g_{mn} P^{+n}$$

$$\Rightarrow \quad (-g)^{-1/2} g_{mn} = (A \cdot B)^{-1}(\epsilon_{mp}A^p \epsilon_{nq}A^q - B_m B_n); \qquad A^m = P^{+m}, \quad B_m = \partial_m X^+$$

(as seen, e.g., by using $\epsilon_{mn}A^n, B_m$ as a basis), and

$$0 = \frac{\delta S}{X^-} \sim \partial_m P^{+m} \quad \Rightarrow \quad \frac{d}{d\tau}\int d\sigma \ P^{+0} = 0$$

which identifies $\int d\sigma \ P^{+0}$ as the conserved momentum p^+, up to a factor of $2\pi\alpha'$ (since p is really the coefficient of $\dot{x}$ in the action, where $X(\sigma) = x + ...$). Similarly, $\delta S/\delta g^{mn}$ determines P^{-m}, and thus X^-.

We then choose as our main set of gauge conditions

$$X^+ = k\tau, \qquad P^{+0} = k$$

for some constant k, which explicitly determines τ, and determines σ up to a function of τ: An equivalent way to define the lightcone σ in terms of an arbitrary spacelike coordinate σ' is

$$\sigma = k^{-1} \int_0^\sigma d\sigma' \ P^{+0}(\sigma')$$

which identifies σ as the amount of momentum p^+ between that value of σ and $\sigma = 0$ (at fixed τ). We thus have that the length of the string (the range of σ, not the physical length) is

$$l = k^{-1} \int d\sigma \ P^{+0} = 2\pi\alpha' p^+ k^{-1}$$

We then need to fix the location of $\sigma = 0$ as some function $\sigma'(\tau)$: Since in this gauge

$$\partial_1 P^{+1} = 0$$

so P^{+1} is also a function of just τ, we further fix the gauge for σ by choosing

$$P^{+1} = 0 \quad \Rightarrow \quad (-g)^{-1/2} g_{mn} = \eta_{mn}$$

Thus the lightcone gauge is a special case of the conformal gauge, after also fixing scale gauge $g = -1$. For the open string, this almost fixes $\sigma'(\tau)$ at $\sigma = 0$, which we can take as one boundary: The boundary condition for X^+ is now

$$0 = n \cdot \partial X^+ \sim n_0$$

since in this (and any conformal) gauge $\partial_m X \sim \eta_{mn} P^n$. Thus the normal to the boundary must be in the σ direction, so the boundary is at constant σ. This means we have one constant left to fix:

$$\sigma = 0 \ \textit{at one boundary (open string)}$$

This invariance was left because all our previous gauge conditions preserved global σ translation. Unfortunately, there is no corresponding convenient gauge choice for the closed string, so there we leave just this one invariance. In summary, our complete set of ligtcone gauge conditions is now:

$$\textit{gauge}: \qquad X^+ = k\tau, \qquad P^{+m} = k\delta_0^m, \qquad \sigma = 0 \ \textit{at one boundary (open string)}$$

The lightcone action is now, in Hamiltonian form,

$$S_{lc} = -\int d\tau \left\{ \dot{x}^- p^+ + \int \frac{d\sigma}{2\pi} \left[\dot{X}_i P_i - \tfrac{1}{2}(\alpha' P_i^2 + \alpha'^{-1} X_i'^2) \right] \right\}$$

The only distinction between open and closed strings is the boundary condition (since closed strings by definition have no boundary). For closed strings we have only periodicity in σ (by definition of “closed”), while for open strings we have

$$X'(\tau,0) = X'(\tau,l) = 0$$

One consequence, as we just saw, is that closed strings have one residual gauge invariance in the lightcone gauge. These two strings can be made to resemble each other more closely by extending the open string to twice its length, defining X for negative σ by

$$X(\tau,-\sigma) = X(\tau,\sigma)$$

This is the known as the “method of images”: $X(\tau,-\sigma)$ is identified with its mirror image in the τ axis, $X(\tau,-\sigma)$. Then the two strings both satisfy periodic boundary conditions, while the open string has this one additional condition. We also choose

$$k = 2\kappa\alpha' p^+, \qquad \kappa = \begin{cases} 1 & (open) \\ \frac{1}{2} & (closed) \end{cases} \qquad \Rightarrow \qquad l = \frac{\pi}{\kappa}$$

so the length of the closed string is 2π, while the open string has original length π that has now been doubled to match the closed string. Our choice of “phase” in relating X for positive and negative σ for the open string automatically enforces the boundary condition $X'(\tau,0) = 0$ at one end of the string, while the condition $X'(\tau,\pi) = 0$ at the other end is implied in the same way by the closed string “boundary condition” of periodicity, which can be written as $X(\tau,\pi) = X(\tau,-\pi)$. The picture is then that the open string is a closed string that has collapsed on itself, so that for half of the range of σ X doubles back over the path it covered for the other half.

Because σ has a finite range, X can always be expanded in Fourier modes in that variable; the boundary conditions slightly restrict the form of this expansion. We saw that the equations of motion, being second-order in τ-derivatives, gave two modes for each initial state: a left-handed one and a right-handed one. We need to be a bit more precise about the zero-modes (independent of σ): We can separate them out as

$$X(\tau,\sigma) = x + \frac{2\pi\alpha'}{l}p\tau + \sqrt{\tfrac{\alpha'}{2}}[Y_{(+)}(\tau+\sigma) + Y_{(-)}(\tau-\sigma)], \qquad \int d\sigma\ Y_{(\pm)} = 0$$

where Y contains only nonzero-modes. (The normalization of p, conjugate to x, comes from the $-\dot{x}\cdot p$ term in the Lagrangian.) Then x represents the “center of mass” of the string, and p its total momentum. Note that this implies $X_{(\pm)}$ aren’t quite periodic:

$$X_{(\pm)}(\sigma + 2\pi) = X_{(\pm)}(\sigma) + 2\pi\kappa\alpha' p$$

Now the periodicity boundary conditions shared by open and closed strings imply

$$Y_{(\pm)}(\sigma+2\pi)=Y_{(\pm)}(\sigma)$$

while the extra boundary condition for the open string implies

$$Y_{(+)}(\sigma)=Y_{(-)}(\sigma)=Y(\sigma)$$

allowing us to drop the subscript in that case. Thus, the closed string has twice as many modes as the open, except for the nonperiodic part, corresponding to the total momentum and average position. This is related to the interpretation that the open string is a closed string with its two halves occupying the same path. This doubling also shows up in the constraints: For the closed string we have $\hat{P}^2_{(\pm)}$, while for the open string we can consider just $\hat{P}^2_{(+)}$, since $\hat{P}^2_{(-)}(\sigma)=\hat{P}^2_{(+)}(-\sigma)$. In the lightcone gauge we solve these constraints for X^-, by integrating

$$0=\hat{P}^2_{(\pm)}\sim\dot{X}^2_{(\pm)}=\left(\dot{X}^i_{(\pm)}\right)^2-k\dot{X}^-_{(\pm)}\sim(\dot{Y}_{(\pm)}+\kappa\sqrt{2\alpha'}p)^2$$

This separation of zero-modes from nonzero-modes also allows us to find the spin and mass of the string: In any conformal gauge,

$$0=p^2+M^2=\frac{1}{\kappa\alpha'^2}\int_0^{\pi/\kappa}\frac{d\sigma}{2\pi}\ \tfrac{1}{2}(\dot{X}^2+X'^2)\qquad\Rightarrow\qquad M^2=\frac{1}{2\kappa\alpha'}\sum_{\pm}\int\frac{d\sigma}{2\pi}\ \dot{Y}^2_{(\pm)}$$

$$J^{ab}=x^{[a}p^{b]}+S^{ab}=\frac{1}{\alpha'}\int_0^{\pi/\kappa}\frac{d\sigma}{2\pi}\ X^{[a}\dot{X}^{b]}\qquad\Rightarrow\qquad S^{ab}=\sum_{\pm}\int\frac{d\sigma}{2\pi}\ Y^a_{(\pm)}\dot{Y}^b_{(\pm)}$$

where for the open string we can replace

$$\sum_{\pm}\int_0^{\pi}\frac{d\sigma}{2\pi}\to\int_0^{2\pi}\frac{d\sigma}{2\pi},\qquad Y_{(\pm)}\to Y$$

For the lightcone gauge we then have the gauge condition to determine X^+ and Virasoro constraints to determine X^-:

$$Y^+_{(\pm)}=0,\qquad\kappa\alpha'p^-+\sqrt{\tfrac{1}{2}\alpha'}\,\dot{Y}^-_{(\pm)}=\frac{1}{2\kappa\alpha'p^+}\left(\kappa\alpha'p^i+\sqrt{\tfrac{1}{2}\alpha'}\,\dot{Y}^i_{(\pm)}\right)^2$$

Exercise 8.3.1

Consider gauge fixing in the temporal gauge $X^0=\tau$, replacing X^+ with X^0. The classical interpretation is now simpler, since τ and X^0 can now be identified with the usual time. Everything is similar except that the Virasoro constraints can't be solved (e.g., for X^1) in general without square roots.

a Show that some 3D solutions (2 space, 1 time) for the open string are given by

$$\tfrac{1}{\sqrt{2}}(\dot{Y}^1 - i\dot{Y}^2)(\tau) = ce^{-in\tau},$$

for nonzero integer n. (Without loss of generality, we can choose c real and positive.) Find the mass (energy) and spin as

$$M = \frac{c}{\sqrt{\alpha'}}, \qquad S^{12} = \frac{c^2}{n} = \frac{\alpha'}{n}M^2$$

Find X explicitly, and show it describes an "n-fold spinning rod".

b Show that the above solution can be generalized to closed strings by using two such Y's, and fixing the relative magnitude of the two c's. Consider the special cases where $n_- = \pm n_+$. Find the explicit masses, spins, and X's, and show that one describes another n-fold spinning rod, while the other is an "n-fold oscillating ring".

8.4. Quantum mechanics

The more interesting features of the string don't appear until quantization. In particular, we can already see at the free level the discrete mass spectrum characteristic of bound states.

Canonical quantization is simplest in the lightcone gauge. As for particles, canonical quantization is convenient only in mechanics (first quantization), not field theory (second quantization). As can be seen from the lightcone action, the Hamiltonian is part of the constraints: For the spinless particle, we had only the constraint $p^2+m^2=0$, which became $E=H$ in the lightcone gauge $X^+ = p^+\tau$ after identifying the lightcone "energy" $E = p^+p^-$ and its Hamiltonian $H = \frac{1}{2}(p_i^2 + m^2)$. The string Hamiltonian can be rewritten conveniently in terms of $\hat{P}$. Since the closed string is effectively just a doubling of the open string, we treat the open string first. The Hamiltonian is simply

$$H = \int_{-\pi}^{\pi} \frac{d\sigma}{2\pi} \tfrac{1}{2}\hat{P}_i^2$$

where $\hat{P} = \hat{P}_{(+)}$. Since we have chosen $X^+ = 2\alpha' p^+\tau$, we have $E = 2\alpha' p^+ p^-$.

To identify the individual particle states, we Fourier expand the worldsheet variables in σ. As for the particle, we can work at $\tau = 0$, since all the dynamics is contained in the constraints. Equivalently, from the nonrelativistic view of the lightcone formalism, we can work in the Schrödinger picture where the τ dependence is in the wave function instead of the operators. We expand as

$$\hat{P}(\sigma) = \sum_{n=-\infty}^{\infty} \tilde{a}_n e^{-in\sigma}, \qquad \tilde{a}_0 = \sqrt{2\alpha'}p, \qquad \tilde{a}_{-n} = \tilde{a}_n{}^\dagger$$

The canonical commutation relations for P and X are

$$[P_i(\sigma_1), X_j(\sigma_2)] = -2\pi i\delta(\sigma_2 - \sigma_1)\delta_{ij}$$

as the direct generalization of the usual $[p, q] = -i$. (The 2π is from our normalization $d\sigma/2\pi$.) From the definition of $\hat{P}$, we then have

$$[\hat{P}_i(\sigma_1), \hat{P}_j(\sigma_2)] = -2\pi i\delta'(\sigma_2 - \sigma_1)\delta_{ij}$$

We can then decompose this into modes by multiplying by $e^{i(m\sigma_1+n\sigma_2)}$ and integrating, where

$$\int \frac{d\sigma}{2\pi} e^{in\sigma} = \delta_{n0}$$

We then find

$$[\tilde{a}_{im}, \tilde{a}_{jn}] = m\delta_{m+n,0}\delta_{ij}$$

as well as the usual $[p_i, x_j] = -i\delta_{ij}$, and thus can relate the modes to the usual harmonic oscillator creation and annihilation operators:

$$\tilde{a}_n = \sqrt{n}a_n, \qquad \tilde{a}_{-n} = \sqrt{n}a_n{}^\dagger \quad \Rightarrow \quad [a_m, a_n{}^\dagger] = \delta_{mn}$$

for positive n. After normal ordering, we find for the Hamiltonian

$$H = \alpha' p_i^2 + N - \alpha_0, \qquad N = \sum_{n=1}^{\infty} n a_{in}{}^\dagger a_{in}$$

$$\Rightarrow \quad H - E = \alpha'(p_a^2 + M^2), \qquad M^2 = \alpha'^{-1}(N - \alpha_0)$$

for some constant α_0.

From the expression for the mass in terms of the number operator N, we see that the nth oscillator $a_{in}{}^\dagger$ raises the mass-squared of the ground state $|0\rangle$ by n (and similarly for multiple applications of these oscillators). For any given mass, the highest-spin state is the symmetric, traceless tensor part of multiple $a_{i1}{}^\dagger$'s acting on $|0\rangle$: This describes spins

$$j = \alpha' M^2 + \alpha_0$$

the "leading Regge trajectory". Let's look first at the first excited level, obtained by acting on the scalar ground state $|0\rangle$ with the lowest-mass oscillators $a_{i1}{}^\dagger$. Clearly this describes a (lightcone) transverse vector, not a massive vector: It has only D−2 components, not the D−1 necessary for a massive vector. Thus this state describes a massless vector, so

$$\alpha_0 = 1$$

The ground state is then a scalar tachyon with $M^2 = -\alpha'^{-1}$. For any given level past the first excited level, one can check explicitly that the states coming from the various oscillators include the necessary fields for making massive states. For example, at the second excited level, ${a_{i1}}^\dagger {a_{j1}}^\dagger$ contains a traceless, symmetric tensor and a scalar (coming from the trace), while ${a_{i2}}^\dagger$ is a vector; they combine to describe a massive tensor.

The closed string works similarly to the open, but with two sets of harmonic oscillators, but with

$$p_{(+)} = p_{(-)} = \tfrac{1}{2}p$$

In that case we find

$$M^2 = 2\alpha'^{-1}(N_{(+)} + N_{(-)} - 2)$$

where $N_{(+)}$ and $N_{(-)}$ are the number operators for the two independent sets of oscillators. In the lightcone gauge the closed string has the residual gauge invariance generated by $\int d\sigma\ X' \cdot \delta/\delta X$; this gives the residual constraint

$$N_{(+)} = N_{(-)}$$

The closed-string states are thus the direct product of two open-string states of the same mass: For example, the ground state is a scalar tachyon with $M^2 = -2\alpha'^{-1}$, while the first excited states are massless ones from the product of two vectors — a scalar, an antisymmetric tensor, and a symmetric, traceless tensor. The leading Regge trajectory consists of states created with equal numbers of ${a_{i1(+)}}^\dagger$'s and ${a_{i1(-)}}^\dagger$'s, with

$$j = \tfrac{1}{2}\alpha' M^2 + 2$$

In summary, the leading trajectory for open or closed string is given by

$$j = \kappa\alpha' M^2 + \frac{1}{\kappa}$$

9. Yang-Mills

9.1. Lie algebra

The concept of a “covariant derivative” allows the straightforward generalization of electromagnetism to a self-interacting theory, once U(1) has been generalized to a nonabelian group. The simplest case is Yang-Mills theory, which is an essential part of the Standard Model. Later we’ll further generalize this concept to describe gravity.

Commutators are defined for finite matrices and quantum mechanics as

$$[A, B] = AB - BA$$

They have certain properties, following from the usual properties of multiplication (distributivity and associativity), that can be abstracted and used without referring back to the original product:

$$[\alpha A + \beta B, C] = \alpha[A, C] + \beta[B, C] \quad \text{for numbers } \alpha, \beta \qquad \text{(distributivity)}$$

$$[A, B] = -[B, A] \qquad \text{(antisymmetry)}$$

$$[A, [B, C]] + [B, [C, A]] + [C, [A, B]] = 0 \qquad \text{(Jacobi identity)}$$

These properties also give an abstract definition of a form of multiplication, the “Lie bracket”, which defines a “Lie algebra”. (The first property is true of algebras in general.) Another familiar example in physics is the “cross” product for three-vectors. (However, we saw this can also be expressed in terms of matrix multiplication using spinor notation.) The most important use of Lie algebras for physics is for describing (continuous) infinitesimal transformations, especially those describing symmetries.

Infinitesimal symmetry transformations are then written as

$$\delta A = i[G, A], \qquad A' = A + \delta A$$

where G is the “generator” of the transformation. More explicitly, infinitesimal generators will contain infinitesimal parameters: For example, for translations we have

$$G = \epsilon^i p_i \quad \Rightarrow \quad \delta x^i = i[G, x^i] = \epsilon^i, \quad \delta p_i = 0$$

where ϵ^i are infinitesimal numbers, and we have used $[p_i, x^j] = -i\delta_i^j$.

The most evident physical symmetries are those involving spacetime. For nonrelativistic particles, these symmetries form the Galilean group. For the free particle, those infinitesimal transformations are linear combinations of

$$M = m, \quad P_i = p_i, \quad J_{ij} = x_{[i}p_{j]} \equiv x_i p_j - x_j p_i, \quad E = H = \frac{p_i^2}{2m}, \quad V_i = mx_i - p_i t$$

in terms of the position x^i, momenta p_i, and (nonvanishing) mass m. These transformations are the space translations (momentum) P, rotations (angular momentum — just orbital for the spinless case) J, time translations (energy) E, and velocity transformations (Galilean boosts) V. (The mass M is not normally associated with a symmetry, and is not conserved relativistically.)

Exercise 9.1.1

Let's examine the Galilean group more closely. Using just the relations for $[x, p]$ and

$$[A, BC] = [A, B]C + B[A, C]$$

(and the antisymmetry of the bracket):

a Find the action on x_i of each kind of infinitesimal Galilean transformation.

b Show that the nonvanishing commutation relations for the generators are

$$[J_{ij}, P_k] = i\delta_{k[i}P_{j]}, \qquad [J_{ij}, V_k] = i\delta_{k[i}V_{j]}, \qquad [J_{ij}, J^{kl}] = i\delta_{[i}^{[k}J_{j]}{}^{l]}$$

$$[P_i, V_j] = -i\delta_{ij}M, \qquad [H, V_i] = -iP_i$$

For more than one free particle, we introduce an m, x^i, and p_i for each particle (but the same t), and the generators are the sums over all particles of the above expressions. If the particles interact with each other the expression for H is modified, in such a way as to preserve the commutation relations. If the particles also interact with dynamical fields, field-dependent terms must be added to the generators. (External, nondynamical fields break the invariance. For example, a particle in a Coulomb potential is not translation invariant since the potential is centered about some point.) Note that for N particles there are 3N coordinates describing the particles, but still only 3 translations: The particles interact in the same 3-dimensional space. We can use translational invariance to fix the position of any one particle at a given time, but not the rest: The differences in position are translationally invariant. On the other hand, it is often useful not to fix the position of any particle, since keeping this invariance (and the corresponding redundant variables) allows all particles to be treated equally. We might also consider using the differences of positions themselves as the variables, allowing a symmetric treatment of the particles in terms of translationally invariant variables: However, this would require applying constraints on the variables, since there are 3N(N−1)/2 differences, of which only 3(N−1) are independent. We find similar features for local invariances: In general, the most convenient description of a theory is with the invariance; the invariance can then be fixed, or invariant combinations of variables used, appropriately for the particular application.

The rotations (or at least their “orbital” parts) and space translations are examples of coordinate transformations. In general, generators of coordinate transformations are of the form

$$G = \lambda^i(x)p_i \quad \Rightarrow \quad \delta\phi(x) = i[G,\phi] = \lambda^i\partial_i\phi$$

where $\phi(x)$ is a scalar field. Thus, for coordinate transformations the Lie derivative is really a derivative with respect to the coordinates. For more general transformations, we can still think of the bracket as a derivative: The “Lie derivative” of B with respect to A is defined as

$$\mathcal{L}_A B = [A,B]$$

As a consequence of the properties of the Lie bracket, this derivative satisfies the usual properties of a derivative, including the Leibniz rule.

We can now define finite transformations by exponentiating infinitesimal ones:

$$A' \approx (1 + i\epsilon\mathcal{L}_G)A \quad \Rightarrow \quad A' = \lim_{\epsilon\to 0}(1 + i\epsilon\mathcal{L}_G)^{1/\epsilon}A = e^{i\mathcal{L}_G}A$$

In cases where we have $[A,B] = AB - BA$, we can also write

$$e^{i\mathcal{L}_G}A = e^{iG}Ae^{-iG}$$

This follows from replacing G on both sides with αG and taking the derivative with respect to α, to see that both satisfy the same differential equation with the same initial condition. We then can recognize this as the way transformations are performed in quantum mechanics: A linear transformation that preserves the Hilbert-space inner product must be unitary, which means it can be written as the exponential of an antihermitian operator.

Just as infinitesimal transformations define a Lie algebra with elements G, finite ones define a “Lie group” with elements

$$g = e^{iG}$$

The multiplication law of two group elements follows from the fact the product of two exponentials can be expressed in terms of multiple commutators:

$$e^A e^B = e^{A+B+\frac{1}{2}[A,B]+...}$$

We now have the mathematical properties that define a group, namely:

(1) a product, so that for two group elements g_1 and g_2, we can define g_1g_2, which is another element of the group (closure),

(2) an identity element, so $gI = Ig = g$,

(3) an inverse, where $gg^{-1} = g^{-1}g = I$, and

(4) associativity, $g_1(g_2g_3) = (g_1g_2)g_3$.

In this case the identity is $1 = e^0$, while the inverse is $(e^A)^{-1} = e^{-A}$.

Since the elements of a Lie algebra form a vector space (we can add them and multiply by numbers), it's useful to define a basis:

$$G = \alpha^i G_i \quad \Rightarrow \quad g = e^{i\alpha^i G_i}$$

The parameters α^i then also give a set of (redundant) coordinates for the Lie group. (Previously they were required to be infinitesimal, for infinitesimal transformations; now they are finite, but may be periodic, as determined by topological considerations that we will mostly ignore.) Now the multiplication rules for both the algebra and the group are given by those of the basis:

$$[G_i, G_j] = -if_{ij}{}^k G_k$$

for the ("structure") constants $f_{ij}{}^k = -f_{ji}{}^k$, which define the algebra/group (but are ambiguous up to a change of basis). They satisfy the Jacobi identity

$$[[G_{[i}, G_j], G_{k]}] = 0 \quad \Rightarrow \quad f_{[ij}{}^l f_{k]l}{}^m = 0$$

A familiar example is SO(3) (SU(2)), 3D rotations, where $f_{ij}{}^k = \epsilon_{ijk}$ if we use $G_i = \frac{1}{2}\epsilon_{ijk}J_{jk}$. An "Abelian" group is one for which all the generators commute, and thus the structure constants vanish; otherwise, it is "nonabelian".

Another useful concept is a "subgroup": If some subset of the elements of a group also form a group, that is called a "subgroup" of the original group. In particular, for a Lie group the basis of that subgroup will be a subset of some basis for the original group. For example, for the Galilean group J_{ij} generate the rotation subgroup.

For some purposes it is more convenient to absorb the "i" in the infinitesimal transformation into the definition of the generator:

$$G \to -iG \quad \Rightarrow \quad \delta A = [G, A] = \mathcal{L}_G A, \quad g = e^G, \quad [G_i, G_j] = f_{ij}{}^k G_k$$

This affects the reality properties of G: In particular, if g is unitary ($gg^\dagger = I$), as usually required in quantum mechanics, $g = e^{iG}$ makes G hermitian ($G = G^\dagger$), while $g = e^G$ makes G antihermitian ($G = -G^\dagger$). In some cases anithermiticity can be an advantage: For example, for translations we would then have $P_i = \partial_i$ and for rotations $J_{ij} = x_{[i}\partial_{j]}$, which is more convenient since we know the i's in these (and any) coordinate transformations must cancel anyway. On the other hand, the U(1) transformations of electrodynamics (on the wave function for a charged particle) are just phase transformations $g = e^{i\theta}$ (where θ is a real number), so clearly we want the explicit i; then the only generator has the representation $G_i = 1$. In general we'll find that for our purposes absorbing the i's into the generators is more convenient for just spacetime symmetries, while explicit i's are more convenient for internal symmetries.

Matrices are defined by the way they act on some vector space; an n×n matrix takes one n-component vector to another. Given some group, and its multiplication table (which defines the group completely), there is more than one way to represent it by matrices. Any set of matrices we find that has the same multiplication table as the group elements is called a "representation" of that group, and the vector space on which those matrices act is called the "representation space." The representation of the algebra or group in terms of explicit matrices is given by choosing a basis for the vector space. If we include infinite-dimensional representations, then a representation of a group is simply a way to write its transformations that is linear: $\psi' = M\psi$ is linear in ψ. More generally, we can also have a "realization" of a group, where the transformations can be nonlinear. These tend to be more cumbersome, so we usually try to make redefinitions of the variables that make the realization linear. A precise definition of "manifest symmetry" is that all the realizations used are linear. (One possible exception is "affine" or "inhomogeneous" transformations $\psi' = M_1\psi + M_2$, such as the usual coordinate representation of Poincaré transformations, since these transformations are still very simple, because they are really still linear, though not homogeneous.)

Exercise 9.1.2

Consider a general real affine transformation $\psi' = M\psi + V$ on an n-component vector ψ for arbitrary real $n \times n$ matrices M and real n-vectors V. A general group element is thus (M, V).

a Perform 2 such transformations consecutively, and give the resulting "group multiplication" rule for (M_1, V_1) "×" $(M_2, V_2) = (M_3, V_3)$.

b Find the infinitesimal form of this transformation. Define the n^2+n generators as operators on ψ, in terms of ψ^a and $\partial/\partial\psi^a$.

c Find the commutation relations of these generators.

d Compare all the above with (nonrelativistic) rotations and translations.

For example, we always have the "adjoint" representation of a Lie group/algebra, which is how the algebra acts on its own generators:

$$\text{(1) adjoint as operator:}\quad G = \alpha^i G_i, \quad A = \beta^i G_i \quad \Rightarrow \quad \delta A = i[G,A] = \beta^j \alpha^i f_{ij}{}^k G_k$$

$$\Rightarrow \quad \delta\beta^i = -i\beta^k \alpha^j (G_j)_k{}^i, \quad (G_i)_j{}^k = i f_{ij}{}^k$$

This gives us two ways to represent the adjoint representation space: as either the usual vector space, or in terms of the generators. Thus, we either use the matrix $A = \beta^i G_i$ (for arbitrary representation of the matrices G_i, or treating G_i as just abstract generators), or we can write A as a row vector:

$$\text{(2) adjoint as vector:}\quad \langle A| = \beta^i \langle_i| \quad \Rightarrow \quad \delta\langle A| = -i\langle A|G$$

$$\Rightarrow \quad \delta\beta^i \langle_i| = -i\beta^k \alpha^j (G_j)_k{}^i \langle_i|$$

We now give some simple examples of finite matrix groups. The simplest example is the "General Linear" group GL(n), where the generators are arbitrary real n×n matrices. The most convenient notation is to label the generators by a pair of indices: We choose as a basis matrices with a 1 as one entry and 0's everywhere else, and label that generator by the row and column where the 1 appears. Explicitly,

$$GL(n): \quad (G_I{}^J)_K{}^L = \delta_I^L \delta_K^J$$

where the "${}_I{}^J$" labels which generator, while the "${}_K{}^L$" labels which matrix element. This basis applies as well for GL(n,C), arbitrary complex matrices, the only difference being that the coefficients α in $G = \alpha_I{}^J G_J{}^I$ are complex instead of real. The next simplest case is U(n), unitary matrices: We can again use this basis, although the matrices $G_I{}^J$ are not all hermitian, by requiring that $\alpha_I{}^J$ be a hermitian matrix. This turns out to be more convenient in practice than using a hermitian basis for the generators. A well known example is SU(2), where the two generators with the 1 as an off-diagonal element (and 0's elsewhere) are known as the "raising and lowering operators" $J_\pm$, and are more convenient than their hermitian parts for purposes of contructing representations. (This generalizes to other unitary groups, where all the generators on one side of the diagonal are raising, all those on the other side are lowering, and those along the diagonal give the maximal Abelian subalgebra, or "Cartan subalgebra".)

Representations for the other "classical groups" follow from applying their definitions to the GL(n) basis. We thus find

$$SL(n): \quad (G_I{}^J)_K{}^L = \delta_I^L \delta_K^J - \tfrac{1}{n}\delta_I^J \delta_K^L$$

$$SO(n): \quad (G_{IJ})^{KL} = \delta^K_{[I}\delta^L_{J]}$$
$$Sp(n): \quad (G_{IJ})^{KL} = \delta^K_{(I}\delta^L_{J)}$$

for the "Special Linear" groups SL(n) (determinant 1, so traceless generators), Special Orthogonal groups, and "Symplectic groups" Sp(n), which are similar to the orthogonal groups but have an antisymmetric metric. As before, SL(n,C) and SU(n) use the same basis as SL(n), etc. For SO(n) and Sp(n) we have raised and lowered indices with the appropriate metric (so SO(n) includes SO(n_+,n_-)). For some purposes (especially for SL(n)), it's more convenient to impose tracelessness or (anti)symmetry on the matrix α, and use the simpler GL(n) basis.

9.2. Nonabelian gauge theory

The group U(1) of electromagnetism is Abelian: Group elements commute, which makes group multiplication equivalent to multiplication of real numbers, or addition if we write $U = e^{iG}$. The linearity of this addition is directly related to the linearity of the field equations for electromagnetism without matter. On the other hand, the nonlinearity of nonabelian groups causes the corresponding particles to interact with themselves: Photons are neutral, but "gluons" have charge and "gravitons" have weight.

In coupling electromagnetism to the particle, the relation of the canonical momentum to the velocity is modified: Classically, the covariant momentum is $dx/d\tau = p + qA$ for a particle of charge q (e.g., $q = 1$ for the proton). Quantum mechanically, the net effect is that the wave equation is modified by the replacement

$$\partial \to \nabla = \partial + iqA$$

which accounts for all dependence on A ("minimal coupling"). This "covariant derivative" has a fundamental role in the formulation of gauge theories, including gravity. Its main purpose is to preserve gauge invariance of the action that gives the wave equation, which would otherwise be spoiled by derivatives acting on the coordinate-dependent gauge parameters: In electromagnetism,

$$\psi' = e^{iq\lambda}\psi, \quad A' = A - \partial\lambda \quad \Rightarrow \quad (\nabla\psi)' = e^{iq\lambda}(\nabla\psi)$$

or more simply

$$\nabla' = e^{iq\lambda}\nabla e^{-iq\lambda}$$

(More generally, q is some Hermitian matrix when ψ is a reducible representation of U(1).)

Yang-Mills theory then can be obtained as a straightforward generalization of electromagnetism, the only difference being that the gauge transformation, and therefore the covariant derivative, now depends on the generators of some nonabelian group. We begin with the hermitian generators

$$[G_i, G_j] = -if_{ij}{}^k G_k, \qquad G_i{}^\dagger = G_i$$

and exponentiate linear combinations of them to obtain the unitary group elements

$$\mathbf{g} = e^{i\lambda}, \quad \lambda = \lambda^i G_i; \qquad \lambda^{i*} = \lambda^i \quad \Rightarrow \quad \mathbf{g}^\dagger = \mathbf{g}^{-1}$$

We then can define representations of the group:

$$\psi' = e^{i\lambda}\psi, \quad \psi^{\dagger\prime} = \psi^\dagger e^{-i\lambda}; \qquad (G_i\psi)_A = (G_i)_A{}^B \psi_B$$

“Compact” groups are those for which the “invariant volume” of the group (the size of the space defined by its parameters) is finite. For example, for SU(2) we can rotate over angles that have a finite range. Abelian groups are not compact, since for each generator the parameter is an arbitrary real number, which has infinite range. For compact groups charge is quantized: For example, for SU(2) the spin (or, for internal symmetry, “isospin”) is integral or half-integral. On the other hand, with Abelian groups the charge can take continuous values: For example, in principle the proton might decay into a particle of charge π and another of charge $1-\pi$. The experimental fact that charge is quantized suggests already semiclassically that all interactions should be descibed by compact groups.

If λ is coordinate dependent (a local, or “gauge” transformation), the ordinary partial derivative spoils gauge covariance, so we introduce the covariant derivative

$$\nabla_a = \partial_a + iA_a, \qquad A_a = A_a{}^i G_i$$

Thus, the covariant derivative acts on matter in a way similar to the infinitesimal gauge transformation,

$$\delta\psi_A = i\lambda^i G_{iA}{}^B \psi_B, \qquad \nabla_a \psi_A = \partial_a \psi_A + iA_a{}^i G_{iA}{}^B \psi_B$$

Gauge covariance is preserved by demanding it have a covariant transformation law

$$\nabla' = e^{i\lambda} \nabla e^{-i\lambda} \quad \Rightarrow \quad \delta A = -[\nabla, \lambda] = -\partial\lambda - i[A, \lambda]$$

The gauge covariance of the field strength follows from defining it in a manifestly covariant way:

$$[\nabla_a, \nabla_b] = iF_{ab} \quad \Rightarrow \quad F' = e^{i\lambda} F e^{-i\lambda}, \quad F_{ab} = F_{ab}{}^i G_i = \partial_{[a} A_{b]} + i[A_a, A_b]$$

$$\Rightarrow \quad F_{ab}{}^{i} = \partial_{[a} A_{b]}{}^{i} + A_{a}{}^{j} A_{b}{}^{k} f_{jk}{}^{i}$$

The Jacobi identity for the covariant derivative is the “Bianchi identity” for the field strength:

$$0 = [\nabla_{[a}, [\nabla_{b}, \nabla_{c]}]] = i[\nabla_{[a}, F_{bc]}]$$

(If we choose instead to use antihermitian generators, all the explicit i’s go away; however, with hermitian generators the i’s will cancel with those from the derivatives when we Fourier transform for purposes of quantization.) Since the adjoint representation can be treated as either matrices or vectors, the covariant derivative on it can be written as either a commutator or multiplication: For example, we may write either $[\nabla, F]$ or ∇F, depending on the context.

Actions then can be constructed in a manifestly covariant way: For matter, we take a Lagrangian $L_{M,0}(\partial, \psi)$ that is invariant under global (constant) group transformations, and couple to Yang-Mills as

$$L_{M,0}(\partial, \psi) \to L_{M,A} = L_{M,0}(\nabla, \psi)$$

(This is the analog of minimal coupling in electrodynamics.) The representation we use for G_i in $\nabla_a = \partial_a + iA_a^i G_i$ is determined by how ψ represents the group. (For an Abelian group factor U(1), G is just the charge q, in multiples of the coupling for that factor.) For example, the Lagrangian for a massless scalar is simply

$$L_0 = \tfrac{1}{2}(\nabla^a \phi)^\dagger (\nabla_a \phi)$$

(normalized for a complex representation).

For the part of the action describing Yang-Mills itself we take (in analogy to the U(1) case)

$$L_A(A_a^i) = \tfrac{1}{8g^2} tr(F^{ab} F_{ab})$$

for some particular matrix representation of the generators. It is therefore usually convenient to normalize these generators so that

$$tr(G_i G_j) = \delta_{ij}$$

We have chosen a normalization where the Yang-Mills coupling constant g appears only as an overall factor multiplying the F^2 term (and similarly for the electromagnetic coupling, as discussed in previous chapters). An alternative is to rescale $A \to gA$ and $F \to gF$ everywhere; then $\nabla = \partial + igA$ and $F = \partial A + ig[A, A]$, and the F^2 term has no extra factor. This allows the Yang-Mills coupling to be treated similarly to other couplings, which are usually not written multiplying kinetic terms (unless analogies to

Yang-Mills are being drawn), since (almost) only for Yang-Mills is there a nonlinear symmetry relating kinetic and interaction terms.

Current conservation works a bit differently in the nonabelian case: Applying the same argument as for electromagnetism, but taking into account the modified (infinitesimal) gauge transformation law, we find

$$J^m = \frac{\delta S_M}{\delta A_m}, \qquad \nabla_m J^m = 0$$

Since $\partial_m J^m \neq 0$, there is no corresponding covariant conserved charge.

Exercise 9.2.1

Let's look at the field equations:

a Using properties of the trace, show the entire covariant derivative can be integrated by parts as

$$\int dx\ tr(\mathcal{A}[\nabla, \mathcal{B}]) = -\int dx\ tr([\nabla, \mathcal{A}]\mathcal{B}), \qquad \int dx\ \psi^\dagger \nabla \chi = -\int dx\ (\nabla\psi)^\dagger \chi$$

for matrices $\mathcal{A}, \mathcal{B}$ and column vectors ψ, χ.

b Show

$$\delta F_{ab} = \nabla_{[a} \delta A_{b]}$$

c Using the definition of the current as for electromagnetism, derive the field equations with arbitrary matter,

$$\tfrac{1}{g^2}\tfrac{1}{2}\nabla^b F_{ba} = J_a$$

d Show that gauge invariance of the action S_A implies

$$\nabla^a(\nabla^b F_{ba}) = 0$$

Also show this is true directly, using the Jacobi identity, but not the field equations. (Hint: Write the covariant derivatives as commutators.)

Exercise 9.2.2

Expand the left-hand side of the field equation (given in the previous excercise) in the field, as

$$\tfrac{1}{g^2}\tfrac{1}{2}\nabla^b F_{ba} = \tfrac{1}{g^2}\tfrac{1}{2}\partial^b \partial_{[b} A_{a]} - j_a$$

where j contains the quadratic and higher-order terms. Show the *noncovariant* current

$$\mathcal{J}_a = J_a + j_a$$

is conserved. The j term can be considered the gluon contribution to the current: Unlike photons, gluons are charged. Although the current is gauge dependent, and thus physically meaningless, the corresponding charge can be gauge independent under situations where the boundary conditions are suitable.

9.3. Lightcone gauge

Since gauge parameters are always of the same form as the gauge field, but with one less vector index, an obvious type of gauge choice (at least from the point of view of counting components) is to require the gauge field to vanish when one vector index is fixed to a certain value. Explicitly, in terms of the covariant derivative we set

$$n \cdot \nabla = n \cdot \partial \quad \Rightarrow \quad n \cdot A = 0$$

for some constant vector n^a. We then can distinguish three types of "axial gauges":

(1) "Arnowitt-Fickler", or spacelike ($n^2 > 0$),

(2) "lightcone", or lightlike ($n^2 = 0$), and

(3) "temporal", or timelike ($n^2 < 0$).

By appropriate choice of reference frame, and with the usual notation, we can write these gauge conditions as $\nabla^1 = \partial^1$, $\nabla^+ = \partial^+$, and $\nabla^0 = \partial^0$.

One way to apply this gauge in the action is to keep the same set of fields, but have explicit n dependence. A much simpler choice is to use a gauge choice such as $A_0 = 0$ simply to eliminate A_0 explicitly from the action. For example, for Yang-Mills we find

$$A_0 = 0 \quad \Rightarrow \quad F_{0i} = \dot{A}_i \quad \Rightarrow \quad \tfrac{1}{8}(F_{ab})^2 = -\tfrac{1}{4}(\dot{A}_i)^2 + \tfrac{1}{8}(F_{ij})^2$$

where " $\dot{}$ " here refers to the time derivative. Canonical quantization is simple in this gauge, because we have the canonical time-derivative term. However, the gauge condition can't be imposed everywhere, as seen for the corresponding gauge for the one-dimensional metric in the case of the particle.

In the case of the lightcone gauge we can carry this analysis one step further. Gauge fixing alone gives us, again for the example of pure Yang-Mills,

$$A^+ = 0 \quad \Rightarrow \quad F^{+i} = \partial^+ A^i, \quad F^{+-} = \partial^+ A^-, \quad F^{-i} = \partial^- A^i - [\nabla^i, A^-]$$

$$\Rightarrow \quad \tfrac{1}{8}(F^{ab})^2 = -\tfrac{1}{4}(\partial^+ A^-)^2 - \tfrac{1}{2}(\partial^+ A^i)(\partial^- A^i - [\nabla^i, A^-]) + \tfrac{1}{8}(F^{ij})^2$$

In the lightcone formalism ∂^- $(-\partial_+)$ is to be treated as a time derivative, while ∂^+ can be freely inverted (i.e., modes propagate to infinity in the x^+ direction, but boundary conditions set them to vanish in the x^- direction). Thus, we can treat A^- as an auxiliary field. The solution to its field equation is

$$A^- = \frac{1}{\partial^{+2}}[\nabla^i, \partial^+ A^i]$$

which can be substituted directly into the action:

$$\begin{aligned} \tfrac{1}{8}(F^{ab})^2 &= \tfrac{1}{2}A^i\partial^+\partial^- A^i + \tfrac{1}{8}(F^{ij})^2 - \tfrac{1}{4}[\nabla^i, \partial^+ A^i]\frac{1}{\partial^{+2}}[\nabla^j, \partial^+ A^j] \\ &= -\tfrac{1}{4}A^i\Box A^i + i\tfrac{1}{2}[A^i, A^j]\partial^i A^j + i\tfrac{1}{2}(\partial^i A^i)\frac{1}{\partial^+}[A^j, \partial^+ A^j] \\ &\quad - \tfrac{1}{8}[A^i, A^j]^2 + \tfrac{1}{4}[A^i, \partial^+ A^i]\frac{1}{\partial^{+2}}[A^j, \partial^+ A^j] \end{aligned}$$

We can save a couple of steps in this derivation by noting that elimination of any auxiliary field, appearing quadratically (as in going from Hamiltonian to Lagrangian formalisms), has the effect

$$L = \tfrac{1}{2}ax^2 + bx + c \to -\tfrac{1}{2}ax^2|_{\partial L/\partial x = 0} + L|_{x=0}$$

In this case, the quadratic term is $(F^{-+})^2$, and we have

$$\tfrac{1}{8}(F^{ab})^2 = \tfrac{1}{8}(F^{ij})^2 - \tfrac{1}{2}F^{+i}F^{-i} - \tfrac{1}{4}(F^{+-})^2 \to \tfrac{1}{8}(F^{ij})^2 - \tfrac{1}{2}(\partial^+ A^i)(\partial^- A^i) + \tfrac{1}{4}(F^{+-})^2$$

where the last term is evaluated at

$$0 = [\nabla_a, F^{+a}] = -\partial^+ F^{+-} + [\nabla^i, F^{+i}] \quad \Rightarrow \quad F^{+-} = \frac{1}{\partial^+}[\nabla^i, F^{+i}]$$

$$\Rightarrow \quad L = \tfrac{1}{8}(F^{ij})^2 + \tfrac{1}{2}A^i\partial^+\partial^- A^i - \tfrac{1}{4}[\nabla^i, \partial^+ A^i]\frac{1}{\partial^{+2}}[\nabla^j, \partial^+ A^j]$$

as above.

In this case, canonical quantization is even simpler, since interpreting ∂^- as the time derivative makes the action look like that for a nonrelativistic field theory, with a kinetic term linear in time derivatives (as well as interactions without them). The free part of the field equation is also simpler, since the kinetic operator is now just $\Box$. In general, lightcone gauges are the simplest for analyzing physical degrees of freedom (within perturbation theory), since the maximum number of degrees of freedom is eliminated, and thus kinetic operators look like those of scalars. On the other hand, interaction terms are more complicated because of the nonlocal Coulomb-like terms involving $1/\partial^+$: The inverse of a derivative is an integral. (However, in practice we often work in momentum space, where $1/p^+$ is local, but Fourier transformation itself introduces multiple integrals.) This makes lightcone gauges useful for discussing unitarity (they are "unitary gauges"), but inconvenient for explicit calculations. (In the literature, "lightcone gauge" is sometimes used to refer to an axial gauge where A^+ is set to vanish but A^- is not eliminated, and D-vector notation is still used, so unitarity is not manifest. Here we always eliminate both components and explicitly use $(D-2)$-vectors, which has distinct technical advantages.)

Although spin 1/2 has no gauge invariance, the second step of the lightcone formalism, eliminating auxiliary fields, can also be applied there: For example, for a massless spinor in D=4, identifying $\partial^{\ominus\dot\ominus} = \partial^-$ as the lightcone "time" derivative, we vary $\bar\psi^{\dot\ominus}$ (or $\psi^\ominus$) as the auxiliary field:

$$-iL = \bar\psi^{\dot\ominus}\partial^{\oplus\dot\oplus}\psi^\ominus + \bar\psi^{\dot\oplus}\partial^{\ominus\dot\ominus}\psi^\oplus - \bar\psi^{\dot\ominus}\partial^{\ominus\dot\oplus}\psi^\oplus - \bar\psi^{\dot\oplus}\partial^{\oplus\dot\ominus}\psi^\ominus$$

$$\Rightarrow \quad \psi^\ominus = \frac{1}{\partial^{\oplus\dot\oplus}}\partial^{\ominus\dot\oplus}\psi^\oplus$$

$$\Rightarrow \quad L = \bar\psi^{\dot\oplus}\frac{\frac{1}{2}\Box}{i\partial^{\oplus\dot\oplus}}\psi^\oplus$$

This tells us that a 4D massless spinor, like a 4D massless vector (or a *complex* scalar) has only 1 complex (2 real) degree of freedom, describing a particle of helicity +1/2 and its antiparticle of helicity −1/2 (±1 for the vector, 0 for the scalar). ("Helicity" refers to circular polarization, or how the wave function transforms under a rotation about the velocity axis. It thus takes values $\pm s$ for what would have been spin s if it were massive.) On the other hand, in the massive case we can always go to a rest frame, so the analysis is in terms of spin (SU(2) for D=4) rather than helicity. For a massive Weyl spinor we can perform the same analysis as above, with the modifications

$$L \to L + \frac{im}{\sqrt{2}}(\psi^\oplus\psi^\ominus + \bar\psi^{\dot\oplus}\bar\psi^{\dot\ominus}) \quad \Rightarrow \quad L = \bar\psi^{\dot\oplus}\frac{\frac{1}{2}(\Box - m^2)}{i\partial^{\oplus\dot\oplus}}\psi^\oplus$$

where we have dropped some terms that vanish upon using integration by parts. So now we have the two states of an SU(2) spinor, but these are identified with their antiparticles. This differs from the vector: While for the spinor we have 2 states of a given energy for both the massless and massive cases, for a vector we have 2 for the massless but 3 for the massive, since for SU(2) spin s has 2s+1 states:

4D states of given E:

spin	0	$\frac{1}{2}$	1	$\frac{3}{2}$	...
$m = 0$	1	2	2	2	...
$m > 0$	1	2	3	4	...

10. General relativity

10.1. Coordinate tensors

General relativity can be described by a simple extension of the methods used to describe Yang-Mills theory. The first thing to understand is the gauge group. We start with coordinate transformations, which are the local generalization of translations, since gravity is defined to be the force that couples to energy-momentum in the same way that electromagnetism couples to charge. Depending on the choice of coordinates, the coordinates may transform nonlinearly (i.e., as a realization, not a representation), as for the D-dimensional conformal group in terms of D (not D+2) coordinates. However, given the nonlinear transformation of the coordinates, there are always representations other than the defining one (scalar field) that we can immediately write down (such as the adjoint). We now consider such representations: These are useful not only for the spacetime symmetries we have already considered, but also for general relativity, where the symmetry group consists of arbitrary coordinate transformations. Furthermore, these considerations are useful for describing coordinate transformations that are not symmetries, such as the change from Cartesian to polar coordinates in nonrelativistic theories.

When applied to quantum mechanics, we write the action of a symmetry on a state as $\delta\psi = iG\psi$ (or $\psi' = e^{iG}\psi$), but on an operator as $\delta A = i[G, A]$ (or $A' = e^{iG}Ae^{-iG}$). However, if $G = \lambda^m\partial_m$ is a coordinate transformation (e.g., a rotation) and ϕ is a scalar field, then in quantum notation we can write

$$\delta\phi(x) = [G, \phi] = G\phi = \lambda^m\partial_m\phi \qquad (\phi' = e^G\phi e^{-G} = e^G\phi)$$

since the derivatives in G just differentiate ϕ. (For this discussion of coordinate transformations we switch to absorbing the i's into the generators.) The coordinate transformation G has the usual properties of a derivative:

$$[G, f(x)] = Gf \quad \Rightarrow \quad Gf_1f_2 = [G, f_1f_2] = (Gf_1)f_2 + f_1Gf_2$$

$$e^Gf_1f_2 = e^Gf_1f_2e^{-G} = (e^Gf_1e^{-G})(e^Gf_2e^{-G}) = (e^Gf_1)(e^Gf_2)$$

and similarly for products of more functions.

The adjoint representation of coordinate transformations is a "vector field" (in the sense of a spatial vector), a function that has general dependence on the coordinates (like a scalar field) but is also linear in the momenta (as are the Poincaré generators):

$$G = \lambda^m(x)\partial_m, \quad V = V^m(x)\partial_m \quad \Rightarrow \quad \delta V = [G, V] = (\lambda^m\partial_mV^n - V^m\partial_m\lambda^n)\partial_n$$

$$\Rightarrow \quad \delta V^m = \lambda^n \partial_n V^m - V^n \partial_n \lambda^m$$

Finite transformations can also be expressed in terms of transformed coordinates themselves, instead of the transformation parameter:

$$\phi(x) = e^{-\lambda^m \partial_m} \phi'(x) = \phi'(e^{-\lambda^m \partial_m} x)$$

as seen, for example, from a Taylor expansion of ϕ', using $e^{-G}\phi' = e^{-G}\phi' e^{G}$. We then define

$$\phi'(x') = \phi(x) \quad \Rightarrow \quad x' = e^{-\lambda^m \partial_m} x$$

This is essentially the statement that the active and passive transformations cancel. However, in general this method of defining coordinate transformations is not convenient for applications: When we make a coordinate transformation, we want to know $\phi'(x)$. Working with the "inverse" transformation on the coordinates, i.e., our original e^{+G},

$$\tilde{x} \equiv e^{+\lambda^m \partial_m} x \quad \Rightarrow \quad \phi'(x) = e^G \phi(x) = \phi(\tilde{x}(x))$$

So, for finite transformations, we work directly in terms of $\tilde{x}(x)$, and simply plug this into ϕ in place of x ($x \to \tilde{x}(x)$) to find ϕ' as a function of x.

Similar remarks apply for the vector, and for derivatives in general. We then use

$$x' = e^{-G} x \quad \Rightarrow \quad \partial' = e^{-G} \partial e^{G}$$

where $\partial' = \partial/\partial x'$, since $\partial' x' = \partial x = \delta$. This tells us

$$V^m(x)\partial_m = e^{-G} V'^m(x) \partial_m e^G = V'^m(x')\partial'_m$$

or $V'(x') = V(x)$. Acting with both sides on x'^m,

$$V'^m(x') = V^n(x) \frac{\partial x'^m}{\partial x^n}$$

On the other hand, working in terms of $\tilde{x}$ is again more convenient: Changing the transformation for the vector operator in the same way as the scalar

$$V'(x) = V(\tilde{x})$$

$$\Rightarrow \quad V'^m(x)\partial_m = V^m(\tilde{x})\tilde{\partial}_m$$

$$\Rightarrow \quad V'^m(x) = V^n(\tilde{x}) \frac{\partial x^m}{\partial \tilde{x}^n}$$

where $\tilde{\partial} = \partial/\partial\tilde{x}$ and as usual

$$(\tilde{\partial}_m x^n)(\partial_n \tilde{x}^p) = \delta_m^p \quad \Rightarrow \quad \frac{\partial x^n}{\partial \tilde{x}^m} = \left[\left(\frac{\partial \tilde{x}(x)}{\partial x} \right)^{-1} \right]_m{}^n$$

We can also use

$$V'(x) = e^G V(x) e^{-G}$$

$$\Rightarrow \quad V'^m(x)\partial_m = (e^G V^m(x) e^{-G})(e^G \partial_m e^{-G}) = V^m(\tilde{x})\tilde{\partial}_m$$

A “differential form” is defined as an infinitesimal $W = dx^m W_m(x)$. Its transformation law under coordinate transformations, like that of scalar and vector fields, is defined by $W'(x') = W(x)$. For any vector field $V = V^m(x)\partial_m$, $V^m W_m$ transforms as a scalar, as follows from the “chain rule” $d = dx'^m \partial'_m = dx^m \partial_m$. Explicitly,

$$W'_m(x') = W_n(x)\frac{\partial x^n}{\partial x'^m}$$

or in infinitesimal form

$$\delta W_m = \lambda^n \partial_n W_m + W_n \partial_m \lambda^n$$

Thus a differential form is dual to a vector, at least as far as the matrix part of coordinate transformations is concerned. They transform the same way under rotations, because rotations are orthogonal; however, more generally they transform differently, and in the absence of a metric there is not even a way to relate the two by raising or lowering indices.

10.2. Gauge invariance

Coordinate transformations are not enough to define spinors. This is easy to see already from the linear part of coordinate transformations: Whereas SO(3,1) is the same Lie group as SL(2,C), GL(4) (a Wick rotation of U(4)) does not have a corresponding covering group; there is no way to take the square root of a vector under coordinate transformations. So we include Lorentz transformations as an additional local group. We therefore have a coordinate transformation group, which includes translations and the orbital part of Lorentz transformations, and a local Lorentz group, which includes the spin part of Lorentz transformations.

Clearly the coordinates x^m themselves, and therefore their partial derivatives ∂_m, are not affected by the (spin) Lorentz generators. We indicate this by use of “curved” vector indices $m, n, ...$. On the other hand, all spinors should be acted on by the Lorentz generators, so we give them “flat” indices $\alpha, \beta, ..$, and we also have flat vector indices $a, b, ...$ for vectors that appear by squaring spinors. Flat indices can be treated the same way as in flat space, with metrics $C_{\alpha\beta}$ and η_{ab} to raise, lower, and contract them.

Some gravity texts, particularly the more mathematical ones, emphasize the use of “index-free notation”. An example of such notation is matrix notation: Matrix notation is useful only for objects with two indices or fewer. Such mathematical texts consider the use of indices as tantamount to specifying a choice of basis; on the contrary, as we have seen in previous chapters, indices in covariant equations usually act only (1) as place holders, indicating where contractions are made and how to associate tensors on either side of equations, and (2) as mnemonics, reminding us of representations and transformation properties. Thus, the full content of the equation can be seen at a glance. In contrast, many mathematical-style equations (when indeed equal signs are actually used) say little more than "$A = B$", with the real content of the equation buried in the text of preceding paragraphs.

We therefore define the elements of the group as

$$g = e^{\lambda}, \qquad \lambda = \lambda^m \partial_m + \tfrac{1}{2}\lambda^{ab} M_{ba}$$

where ∂_m acts on all coordinates, including the arguments of the real gauge parameters λ^m and λ^{ab} and any fields. $M_{ab} = -M_{ba}$ are the Lorentz generators: They act on all flat indices, including those on λ^{ab} and any fields that carry flat indices. The action of the Lorentz generators on vector indices is given by

$$[M_{ab}, V_c] = V_{[a}\eta_{b]c} \quad \Rightarrow \quad \tfrac{1}{2}\lambda^{bc}[M_{cb}, V_a] = \lambda_a{}^b V_b$$

This implies the commutation relations

$$[M_{ab}, M^{cd}] = -\delta_{[a}^{[c} M_{b]}{}^{d]}$$

As for derivatives, when acting on functions instead of operators we can write the action of the Lorentz generator as simply $M_{ab}V_c$ without the commutator.

Matter representations of the group work similarly to Yang-Mills. We define such fields to have only flat indices. Then their transformation law is

$$\psi' = e^{\lambda}\psi$$

where the transformation of a general Lorentz representation follows from that for a vector (or spinor, if we include them), as defined above. Alternatively, the transformation of a vector could be defined with curved indices, being the adjoint representation of the coordinate group:

$$V = V^m \partial_m \quad \Rightarrow \quad V' = V'^m \partial_m = e^{\lambda^m \partial_m} V e^{-\lambda^m \partial_m}$$

However, as in Yang-Mills theory, it is more convenient to identify only the gauge field as an operator in the group. In any case, only the adjoint representation (and direct products of it) has such a nice operator interpretation.

As an example of this algebra, we now work out the commutator of two transformations in gory detail: We first recall that the coordinate transformation commutator was already worked out above, using the usual quantum mechanical relations

$$[f, f] = [\partial, \partial] = 0, \qquad [\partial, f] = (\partial f)$$

for any function f. For the Lorentz algebra we will use the additional identities

$$[M_{ab}, \partial_m] = [M_{ab}, \lambda^m] = [M_{ab}, \lambda^{cd} M_{dc}] = 0$$

all expressing the fact the Lorentz generators commute with anything lacking free flat indices (i.e., Lorentz scalars). The commutator algebra is then

$$\begin{aligned}
&[\lambda_1^m \partial_m + \tfrac{1}{2}\lambda_1^{ab} M_{ba}, \lambda_2^n \partial_n + \tfrac{1}{2}\lambda_2^{cd} M_{dc}] \\
&\qquad = \lambda_{[1}^m [\partial_m, \lambda_{2]}^n] \partial_n + \lambda_{[1}^m [\partial_m, \tfrac{1}{2}\lambda_{2]}^{ab}] M_{ba} + \tfrac{1}{2}\lambda_1^{ab} [M_{ba}, \tfrac{1}{2}\lambda_2^{cd}] M_{dc} \\
&\qquad = (\lambda_{[1}^n \partial_n \lambda_{2]}^m) \partial_m + \tfrac{1}{2}(\lambda_{[1}^m \partial_m \lambda_{2]}^{ab} + \lambda_{[1}^{ac} \lambda_{2]c}{}^b) M_{ba}
\end{aligned}$$

One fine point to worry about: We may consider spaces with nontrivial topologies, where it is not possible to choose a single coordinate system for the entire space. For example, on a sphere spherical coordinates have singularities at the two poles, where varying the longitude gives the same point and not a line. (However, the sphere can be described by coordinates with only one singular point.) We then either treat such points by a limiting procedure, or choose different sets of nonsingular coordinates on different regions ("patches") and join them to cover the space.

10.3. Covariant derivatives

We can also define covariant derivatives in a manner similar to Yang-Mills theory; however, since ∂_m is now one of the generators, the "∂" term can be absorbed into the "A" term of $\nabla = \partial + A$:

$$\nabla_a = e_a{}^m \partial_m + \tfrac{1}{2}\omega_a{}^{bc} M_{cb}$$

in terms of the"vierbein (tetrad)" $e_a{}^m$ and "Lorentz connection" $\omega_a{}^{bc}$. Now the action of the covariant derivative on matter fields looks even more similar to the gauge transformations: e.g.,

$$\delta\phi = \lambda^m \partial_m \phi, \qquad \nabla_a \phi = e_a{}^m \partial_m \phi$$

$$\delta V_a = \lambda^m \partial_m V_a + \lambda_a{}^b V_b, \qquad \nabla_a V_b = e_a{}^m \partial_m V_b + \omega_{ab}{}^c V_c$$

I.e., the covariant derivative ∇_a is essentially D elements (labeled by "a") of the gauge algebra.

Note that the free index on the covariant derivative is flat so that it transforms nontrivially under

$$\nabla'_a = e^{\lambda} \nabla_a e^{-\lambda}$$

Explicitly, for an infinitesimal transformation $\delta\nabla = [\lambda, \nabla]$ we have

$$\delta e_a{}^m = (\lambda^n \partial_n e_a{}^m - e_a{}^n \partial_n \lambda^m) + \lambda_a{}^b e_b{}^m$$

$$\delta\omega_a{}^{bc} = \lambda^m \partial_m \omega_a{}^{bc} + (-e_a{}^m \partial_m \lambda^{bc} + \omega_a{}^{d[b} \lambda_d{}^{c]} + \lambda_a{}^d \omega_d{}^{bc})$$

This commutator is the same as for $[\lambda_1, \lambda_2]$ in the previous subsection, except for the two additional terms coming from the Lorentz generators acting on the free index on ∇_a. In particular, the vierbein $e_a{}^m$ transforms on its flat index as the vector (defining) representation of the local Lorentz group, and on its curved index (and argument) as the vector (adjoint) representation of the coordinate group. Also, it should be invertible, since originally we had $\nabla = \partial + A$: We want to be able to separate out the flat space part as $e_a{}^m = \delta_a^m + h_a{}^m$ for perturbation theory or weak gravitational fields. That means we can use it to convert between curved and flat indices:

$$V^m = V^a e_a{}^m \quad \Leftrightarrow \quad V^a = V^m e_m{}^a$$

where $e_m{}^a$ is the inverse of $e_a{}^m$. Furthermore, if we want to define the covariant derivative of an object with curved indices, we can simply flatten its indices, take the covariant derivative with ∇, and then unflatten its indices.

Flat indices are the natural way to describe tensors: (1) They are the *only* way to describe half-integral spin. (2) Even for integer spin, they correspond to the way components are actually measured. In fact, the above conversion of vectors from curved to flat indices is exactly the one you learned in your freshman physics course! The special cases you saw there were curvilinear coordinates (polar or spherical) for flat space. Then $e_a{}^m$ was the usual orthonormal basis. Only the notation was different: Using Gibbs' notation for the curved but not the flat indices, $\vec{V} = V^a \vec{e}_a$, where, e.g., $a = (r, \theta, \phi)$ for spherical coordinates and $\vec{e}_a = (\hat{r}, \hat{\theta}, \hat{\phi})$ are the usual orthonormal basis. Thus, you probably learned about the vierbein years before you ever saw a "metric tensor". Similarly, when you learned how to integrate over the volume element of spherical coordinates, you found it from this basis, and only learned much later (if yet) to express it in terms of the square root of the determinant of the metric. (With the orthonormal basis, there was no square root to take; the determinant came from the cross product.) You also learned how to do this for curved space: Considering again the sphere, vectors in the sphere itself can be expressed in terms of just $\hat{\theta}$ and $\hat{\phi}$. And the area element of the sphere (the volume element of this smaller space) you again found from this basis.

For example, consider nonrelativistic momentum in two flat spatial dimensions, but in polar coordinates: Now using x^m to represent just the nonrelativistic spatial coordinates,

$$x^m = (r, \theta), \qquad p^m = m\frac{dx^m}{dt} = (m\dot{r}, m\dot{\theta}) = p^a e_a{}^m$$

$$e_a{}^m = \begin{pmatrix} 1 & 0 \\ 0 & r^{-1} \end{pmatrix}, \qquad p^a = (m\dot{r}, mr\dot{\theta})$$

Then the two components of p^a (with the simplest choice of $e_a{}^m$) are the usual components of momentum in the radial and angular directions. On the other hand, one component of p^m is still the radial component of the momentum, while the other component of p^m is the *angular* momentum — a useful quantity, but not normally considered as a component along with the radial momentum, which doesn't even have the same engineering dimensions. In writing the Hamiltonian, one simply squares p^a in the naive way, whereas squaring p^m would require use of the metric.

Exercise 10.3.1

Show that the above choice of $e_a{}^m$ actually describes flat space: Use the fact that p^a transforms as a scalar under the coordinate transformations that express r and θ in terms of Cartesian coordinates x and y, and as a vector under local "Lorentz" transformations, which are in this case just 2D rotations, to transform it to the usual Cartesian $p'^a = (m\dot{x}, m\dot{y})$.

This direct conversion between curved and flat indices also leads directly to the covariant generalization of length: In terms of momentum (as would appear in the action for the classical mechanics of the particle),

$$p^m = m\frac{dx^m}{ds}, \qquad -m^2 = p^2 = p^a p^b \eta_{ab} \quad \Rightarrow \quad -ds^2 = dx^m dx^n e_m{}^a e_n{}^b \eta_{ab} \equiv dx^m dx^n g_{mn}$$

Equivalently, the metric tensor g_{mn} is just the conversion of the flat-space metric η_{ab} to curved indices. Also, in terms of differential forms,

$$\Omega^a = dx^m e_m{}^a \quad \Rightarrow \quad -ds^2 = \Omega^a \Omega^b \eta_{ab}$$

The field strengths are also defined as in Yang-Mills:

$$[\nabla_a, \nabla_b] = T_{ab}{}^c \nabla_c + \tfrac{1}{2} R_{ab}{}^{cd} M_{dc}$$

where we have expanded the field strengths over ∇ and M rather than ∂ and M so that the "torsion" T and "curvature" R are manifestly covariant:

$$M' = e^{\lambda} M e^{-\lambda} = M \quad \Rightarrow \quad T' = e^{\lambda} T e^{-\lambda}, \quad R' = e^{\lambda} R e^{-\lambda}$$

The commutator can be evaluated as before, with the same change as for going from $[\lambda_1, \lambda_2]$ to $[\lambda, \nabla_a]$ (i.e., now there are two free indices on which the Lorentz generators can act), except that now we rearrange terms to convert $\partial_m \to e_a{}^m \partial_m \to \nabla_a$. Making the further definitions

$$e_a = e_a{}^m \partial_m, \quad [e_a, e_b] = c_{ab}{}^c e_c \quad \Rightarrow \quad c_{ab}{}^c = (e_{[a} e_{b]}{}^m) e_m{}^c = -e_a{}^m e_b{}^n \partial_{[m} e_{n]}{}^c$$

for the "structure functions" $c_{ab}{}^c$, we find the explicit expressions

$$T_{ab}{}^c = c_{ab}{}^c + \omega_{[ab]}{}^c = -e_a{}^m e_b{}^n (\partial_{[m} e_{n]}{}^c + e_{[m}{}^d \omega_{n]d}{}^c)$$

$$R_{ab}{}^{cd} = e_{[a} \omega_{b]}{}^{cd} - c_{ab}{}^e \omega_e{}^{cd} + \omega_{[a}{}^{ce} \omega_{b]e}{}^d = e_a{}^m e_b{}^n (\partial_{[m} \omega_{n]}{}^{cd} + \omega_{[m}{}^{ce} \omega_{n]e}{}^d)$$

The covariant derivative satisfies the Bianchi (Jacobi) identities

$$0 = [\nabla_{[a}, [\nabla_b, \nabla_{c]}]] = [\nabla_{[a}, T_{bc]}{}^d \nabla_d + \tfrac{1}{2} R_{bc]}{}^{de} M_{ed}]$$

$$= (\nabla_{[a} T_{bc]}{}^d) \nabla_d + \tfrac{1}{2} (\nabla_{[a} R_{bc]}{}^{de}) M_{ed} - T_{[ab|}{}^e (T_{e|c]}{}^f \nabla_f + \tfrac{1}{2} R_{e|c]}{}^{fg} M_{gf}) - R_{[abc]}{}^d \nabla_d$$

$$\Rightarrow \quad R_{[abc]}{}^d = \nabla_{[a} T_{bc]}{}^d - T_{[ab|}{}^e T_{e|c]}{}^d, \qquad \nabla_{[a} R_{bc]}{}^{de} - T_{[ab|}{}^f R_{f|c]}{}^{de} = 0$$

To make the transformation laws manifestly covariant we can define instead

$$\lambda = \lambda^a \nabla_a + \tfrac{1}{2} \lambda^{ab} M_{ba}$$

which is just a redefinition of the gauge parameters. The infinitesimal transformation law of the covariant derivative is then

$$\delta \nabla_a = [(\delta e_a{}^m) e_m{}^b] \nabla_b + \tfrac{1}{2} (e_a{}^m \delta \omega_m{}^{bc}) M_{cb} = [\lambda^b \nabla_b + \tfrac{1}{2} \lambda^{bc} M_{cb}, \nabla_a]$$
$$= (-\nabla_a \lambda^b + \lambda^c T_{ca}{}^b + \lambda_a{}^b) \nabla_b + \tfrac{1}{2} (-\nabla_a \lambda^{bc} + \lambda^d R_{da}{}^{bc}) M_{cb}$$

$$\Rightarrow \quad (\delta e_a{}^m) e_m{}^b = -\nabla_a \lambda^b + \lambda^c T_{ca}{}^b + \lambda_a{}^b, \quad e_a{}^m \delta \omega_m{}^{bc} = -\nabla_a \lambda^{bc} + \lambda^d R_{da}{}^{bc}$$