Talk to Undergraduates — April 11, 2006
Relativity and Differential Geometry, Gauge Theories

1 Special Relativity

I am going to begin this discussion with special relativity, and in particular Einstein's motivation for investigating special relativity. Much argument has occurred over whether or not Einstein knew of the Michaelson-Morley experiment, which demonstrated that there was evidence that Newtonian kinematics could not be right, but Einstein always claimed that his principal motivation was not to explain why the Michaelson-Morley experiment gave a null result, but why two different reference frames could be used to give different explanations of the same physics. Let me quote from Einstein's famous first paper on relativity:

> "It is known that Maxwell's electrodynamics—as usually understood at the present time—when applied to moving bodies, leads to asymmetries which do not appear to be inherent in the phenomena. Take, for example, the reciprocal electrodynamic action of a magnet and a conductor. The observable phenomenon here depends only on the relative motion of the conductor and the magnet, whereas the customary view draws a sharp distinction between the two cases in which either the one or the other of these bodies is in motion. For if the magnet is in motion and the conductor at rest, there arises in the neighborhood of the magnet an electric field with a certain definite energy, producing a current at the places where parts of the conductor are situated. But if the magnet is stationary and the conductor in motion, no electric field arises in the neighborhood of the magnet. In the conductors, however, we find an electromotive force, to which in itself there is no corresponding energy, but which gives rise—assuming equality of relative motion in the two cases discussed—to electric currents of the same path and intensity as those produced by the electric forces in the former case.
>
> "Examples of this sort, together with the unsuccessful attempts to discover any motion of the earth relatively to the "light medium", suggest that the phenomena of electrodynamics as well as of mechanics possess no properties corresponding to the idea of absolute rest."

What is Einstein saying here? We have all seen that if a bar magnet is moved into a ring of wire, a current is induced in the wire. One way of saying this is that the magnetic flux inside the ring is changing, so there is an induced EMF, which generates the current. Or you could say that a changing magnetic field forces an electric field to develop due to the Maxwell equation

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}.$$

These are mathematically equivalent explanations.

But if we consider the situation from the reference frame of the magnet, $\partial \vec{B}/\partial t$ is zero everywhere, and there is no induced electric field. Instead, we have a wire surface which pushes the electrons within, giving them a velocity $\vec{v}$ so that the magnetic field produces a force $q\vec{v} \times \vec{B}$, causing the same current to flow in the ring of wire.

So uppermost in Einstein's mind was that if there are two different explanations given for the same effects by two observers using different coordinate systems, there must be a unifying way of thinking about the phenomena.

We know that the theory this led him to, special relativity, is a statement of how Physics in one coordinate system is related to the same Physics in another set of coordinates. In special relativity, the coordinate systems in question must both be inertial. His theory can be derived axiomatically from the two statements

- The laws of Physics are the same in all inertial coordinate systems.
- The speed of light in vacuum is a constant.

By carefully examining these laws, we find that the relationship between the coordinates of an event as described by an observer $\mathcal{O}'$, are related to the coordinates of the same event described by another observe $\mathcal{O}$ by a linear relation which mixes space and time in unexpected ways. That is, the coordinates x', y', z', t' are linear combinations of x, y, z, t, but with t' different from t, depending on the spacial coordinates. The relationship is given by the requirement that all observers agree on how much the hands of an inertial clock moves between two events. This is given by

$$\sqrt{(\Delta t)^2 - c^{-2}(\Delta x)^2 - c^{-2}(\Delta y)^2 - c^{-2}(\Delta z)^2},$$

which must therefore be the same in any inertial coordinate system. This looks like what is the same in the different coordinate systems is the length of a vector in 4 dimensional space, except that

- The units of time were chosen inconsistently with the units of space. This is just foolishness on our part, as if we measured North-South distances in miles and East-West distances in kilometers. This can be fixed simply by measuring time in meters/c, or space in light-seconds. We use the notation $x \to x^1$, $y \to x^2$, $z \to x^3$, where we use superscripts instead of subscripts, and define a measure of time $x^0 = ct$. That takes care of that.
- A more serious problem is the relative minus sign between the x^0 term and the others. This means that our invariant length is not a real length. If it were, the relationship between the different coordinate systems (ignoring translations) would have been a rotation in four dimensional space, given by

  $$x'^\mu = \sum_\nu \Lambda^\mu{}_\nu x^\nu.$$

where we have written the $\Lambda^\mu{}_\nu$ matrix in a funny way, but it would still be an orthogonal matrix. Because of the sign, that is not quite right, and $\Lambda^\mu{}_\nu$ is instead a pseudo-orthogonal matrix called a Lorentz transformation.

Thus the same physics can be described using different coordinate systems, and the explanations are equivalent, giving the same physical answers even though the coordinates used to express those answers may differ. Of course, not only the coordinates of points are required to describe Physics, but also other objects whose coordinates may depend on the observer, such as the electric field. In fact, we find that to describe the electric field in a "covariant" way, we need to combine it with the magnetic field into an object called the "field-strength tensor" $F^{\mu\nu}$. This is how an electric field can be seen in the moving magnet frame even though there is no electric field in the moving ring frame, where the magnet is stationary.

Notice that we have limited our attention to inertial reference frames, all of which have equivalent physical laws, and simple geometry. In fact, displacements in each frame form a vector space, in the sense that if the displacement from A to B is x^μ_{BA}, and the displacement from B to C is x^μ_{CB}, then the displacement from A to C is $x^\mu_{CA} = x^\mu_{BA} + x^\mu_{CB}$. We could think of this four dimensional space in terms of Euclidean geometry, in four dimensions, except that the natural length which is preserved in going from one frame of reference to another is

$$(\Delta s)^2 := (\Delta x^0)^2 - (\Delta x^1)^2 - (\Delta x^2)^2 - (\Delta x^3)^2 = \sum_{\mu\nu} \eta_{\mu\nu} \Delta x^\mu \Delta x^\nu,$$

where we have defined a "metric tensor"

$$\eta_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

which means $\eta_{00} = 1$, $\eta_{ij} = -\delta_{ij}$ for $i, j = 1, 2, 3$, and all other elements are 0. Except for the minus signs, Δs would just be the length of the vector Δx^μ.

2 Principle of Equivalence and Fictious Forces

2.1 Fictious Forces

Now let us review how physics is described in non-inertial frames of reference. In particular, let us consider a flying physicist $\mathcal{O}'$ in a closed room, out in outer space, which is accelerating at a constant acceleration a. Let $\mathcal{O}$ be an inertial observer, and let $v(t)$ be the velocity of $\mathcal{O}'$ with respect to $\mathcal{O}$. Let us forget special relativity for the moment, assuming $v(t) \ll c$, so in the inertial coordinates Newton's laws hold for the acceleration of each body α within the room,

$$F_\alpha = m\frac{dv_\alpha}{dt} = ma_\alpha,$$

where F_α is the force applied to particle α by all the other objects in the universe. The coordinates of the flying physicist are not the same as those of $\mathcal{O}$. In fact, the velocities are related by $v'_\alpha = v_\alpha - v(t)$, so if he tries to apply Newton's third law

$$F'_\alpha := m_\alpha a'_i = m_\alpha \frac{d}{dt}(v_\alpha - v(t)) = m_\alpha a_\alpha - m_\alpha a = F_\alpha - m_\alpha a,$$

he has, in addition to whatever forces are due to the various other objects in the universe, an extra force $-m_\alpha a$, with no discernable object to have the reaction force. So he might get the idea that this is evidence that he is not in an inertial system but is in fact accelerating.

But how does he know the force is not a real force? He could check, presumably, that no other object he could see in his box is responsible for this force, but how could he exclude the possibility that it was due to an action at a distance from some object outside his box? He might argue that the fact that the forces are consistent with being pseudoforces, in that each particle α experiences a force is equal to the mass of α times a universal vector $-\vec{a}$ is evidence that the force is a pseudoforce. He and his friend Eötvös might even set up elaborate experiments to check that the force really is proportional to the inertial mass, and find that it is true to at least one part in a trillion, so he might conclude that the force must be a pseudoforce and his coordinate system $\mathcal{O}'$ must be noninertial, and perhaps he should set up another one, $\mathcal{O}$, and use that instead, so Physics would simplify.

Yes, we say, you should look beyond the obvious, for a view in which the physics is simpler, as a better guide to what is really happening.

But wait! Perhaps we in this room K are that observer. In addition to the electromagnetic forces and their macroscopic effects, we see that there is a force on every object in the room given by its mass times a universal vector pointing downwards with a magnitude of 9.8 m/s^2. Shouldn't we conclude that this room is a spaceship accelerating upwards at 1 g, out in the middle of space? Oh, some of you may think that there is another explanation, that we are in an inertial frame, but there is a gravitational force due to the Earth, but the consequences seem to be the same.

2.2 Principle of Equivalence

So we have two different explanations of the same effect, depending on your coordinate system. But says Einstein (I divine) “last time I got good mileage out of saying these must be equivalent, so let’s try it again!”. Let us announce a “Principle of Equivalence”

> “we assume that the systems K and K' are physically exactly equivalent, that is, if we assume that we may just as well regard the system K as being in a space free from gravitational fields, if we then regard K as uniformly accelerated. This assumption of exact physical equivalence makes it impossible for us to speak of the absolute acceleration of the system of reference, just as the usual theory of relativity forbids us to talk of the absolute velocity of a system; and it makes the equal falling of all bodies in a gravitational field seem a matter of course.”

Now saying that the room in a spaceship is exactly like this room on the surface of the Earth is a slight overstatement. If I very carefully measure the accelerations of different particles with no obvious external forces on them, throughout this room, I will find that they are not exactly the same. The acceleration of the particles near the floor is slightly greater than those near the roof, and the directions of acceleration of particles on opposite sides of the room are not exactly parallel but intersect roughly 6×10^6 m away. The equivalence between a gravitational force and an accelerating reference frame is true only locally; at each point, there is an accelerating reference frame which, without gravity, is equivalent to our Earthbound frame with gravity, but only in an infinitesimal neighborhood of the particular point chosen. This can be done at each point, but it takes a different reference frame for each point. This means there is no global inertial reference frame. We need to develop some mathematical technology to handle such a complicated situation.

Just what mathematics do we need? We need to be able to refer to the points in our space, tell which are near which, and what the distances between nearby points are. Let’s forget for the moment about getting the distances right—all we want to begin with is a set of coordinates that describe the points of space (or space-time). We want continuity and smoothness in space to be represented by continuity and smoothness in the coordinates describing the space. In a small region around any point, the space is almost Euclidean (or Minkowskian), so there is no problem setting up such a coordinate system for a small part of the space. But as we try to extend the coordinate system there may be problems.

Think of making a Rand-McNalley map of Iowa. Although Iowa appears to be perfectly flat, it is really more like part of a perfect sphere. But we can easily project that sphere onto a flat plane, say tangent at Des Moines. In Des Moines, distances will be unchanged, but at the edges of the state the true distances towards Des Moines will be projected onto something slightly smaller. Still, it is all continuous. But if we try to continue this projection onto larger regions of the globe, we find that somewhere around Madrid cities start falling right on top of other cities. We can't have that—we want a set of coordinates to correspond in a 1 to 1 manner with points in physical space. There are better ways of projecting the sphere onto the plane, but no way to project it all.

What does the Rand-McNalley World Atlas do about this problem? They publish an atlas, which consists of a number of pages. Each page is a mapping of a portion of the space we want (the surface of a sphere) into a Euclidean plane. We call each one a chart, using seafaring terminology, rather than a map, a word which mathematicians use with overly great generality. So an atlas consists of a set of charts, with every point of the space mapped into at least one chart. The charts need to overlap some, so some points occur in several charts.

Now to describe physics we will need to consider fields, that is, functions on space such as temperature, which is a scalar, or perhaps vector and tensor valued fields. Each such function, say ψ, from space is also a function on each chart, so for each chart it is a function from Euclidean space into, say, the real numbers. We know how to ask whether such functions are continuous or differentiable, etc, so this gives us a way to ask whether the functions from space are continuous or differentiable. Of course, this requires the chart to be a suitable map between the space and the Euclidean paper. And as one point in space may occur in several charts, the question of whether ψ is continuous or differentiable had better have the same answer in all the charts. This is a requirement about overlaps of charts; if C_1 and C_2 both contain an image of a set Φ, then there is an induced map from part of C_1 to part of C_2 which is the composition of the map from C_1 into space with the map from space into C_2. If all such maps are differentiable as many times as we need to ask about the fields, the answers for the fields will be independant of which chart we use to get the answer.

This kind of framework is called differential geometry. An atlas with all overlap functions having r continuous derivatives is called a $\mathcal{C}^r$ manifold.

This gives us enough information so that, for example, if we have a path through space depending on a parameter (could be time), say $\mathcal{P}(t)$, and a scalar field $\psi(\mathcal{P})$, we can ask about $d\psi/dt$, for example. But we need to know more; we need to know about distances, how far apart are two points, at least for points close enough. This requires a concept of metric. If $\mathcal{P}$ and $\mathcal{P}'$ are close enough, we should be able to give a distance between them.

Now the equivalence principle told us that at any point $\mathcal{P}$ of spacetime, we can go to a reference frame, or coordinate system, or chart, in which there is no gravitational field in the vicinity of the point $\mathcal{P}$. Then if ξ^μ and ξ'^μ are the coordinates of $\mathcal{P}$ and $\mathcal{P}'$ in that chart, the distance Δs between them is given by

$$(\Delta s)^2 = (\Delta\xi^0)^2 - (\Delta\xi^1)^2 - (\Delta\xi^2)^2 - (\Delta\xi^3)^2.$$

If we were talking only about space, there would be no minus signs, and the manifold we are talking about would then be said to have a Riemannian metric. Because of the minus sign, spacetime is a pseudo-Riemannian manifold. Notice, however, that the coordinates ξ are tied to the particular chart which is Minkowskian at the particular point $\mathcal{P}$, and this form for infinitesimal distances does not hold anywhere else.

Let me give you an example. Suppose you were a two dimensional being, unable to conceive, except very formally in math class, the possibility of a third dimension. When you were young you experimented with rulers and protractors, and you found that if you laid out coordinates at 90°, the distance between two points was always, as near as you could measure, $\sqrt{(\Delta x)^2 + (\Delta y)^2}$, and, being a very smart bug, you wrote a series of books called "the Elements", in which you derived Euclidean geometry. But on your 17th birthday you got a car. Now you could drive long distances in straight lines, and you found that if you started from home and drove straight for 12,400 miles, you wound up at the same place regardless of which direction you started out in. In other words, the circumference of small circles was $2\pi r$, but for large circles, it was different, and in particular if $r = 12,400$ miles, the circumference is zero!

Is this possible? Sure! The bug is just on the surface of the Earth. It doesn't matter where his home is, but if it happens to be at the North Pole, his straight line trips are along circles of longitude, and he always winds up at the south pole.

In fact, the bugs find other interesting effects. They find, for example, that the sum of the angles of a triangle is greater than 180°, and the bigger the triangle the larger the extra angle. As an example, the triangle composed of two lines of longitude between the North Pole and the equator, together with a piece of the equator, form a triangle with two right angles and a third nonzero angle at the north pole.

Now perhaps you will say that the bugs are fooled because they don't realize that they are actually in a three dimensional Euclidean space, and that these straight lines of theirs aren't really straight but are great circles on the sphere. But from the bugs' point of view, they are the shortest distance between two points, they "lie evenly on themselves", as Euclid defined a straight line to be, so they are entitled to call them straight.

2.3 Need for Non-Euclidean Space

Once we accept that physics may be as appropriately discussed in noninertial frames as in inertial ones, we are forced to accept that we are in the same boat as the bugs. To see this, consider another set of bugs. They live on a flat turntable in the living room of our spacecraft, which is now not accelerating, so that our reference frame, $\mathcal{O}$, is inertial and Euclidean. Let us draw a circle and measure the circumference C and the radius r, and see if $C = 2\pi r$. Our radius is 2 rulers, so we expect $C = 12.6$ rulers. Is it?

Yup, it is.
The bugs, however, rotate with their turntable (according to us) and they insist on using a noninertial from $\mathcal{O}'$, which we regard as rotating. They also measure the circumference and radius. It might make us dizzy to watch, but if we rotate our head in sync with the turntable, everything looks the same as when we did the measurement, provided the turntable edge isn't moving too fast. But suppose it is rotating so fast that the perimeter is moving at $\frac{1}{2}c$. Then when the bugs carry their rulers from their storehouse at the center to the periphery, they appear to us to shrink, in this case to 87% of their "true" size, and therefore the original 13 rulers are not enough to span the circumference, which is now $14\frac{1}{2}$ rulers long. The rulers measuring the radius, however, have not shrunk, as they are moving perpendicular to their lenght. So $C \neq 2\pi r$, according to the bugs, and Euclid was wrong.

They decide to lay out a straight line between two points on the circumference. They do this by trying lots of curved paths between the points and measuring the lengths. Here is what they get, a curve which seems to us to head towards the center and curve away. We radio to the bugs what we think is a straight line and ask them to measure that. To do so, they need to move their rulers further from the center of the turntable, so they move faster and shrink, and it turns out that it takes more metersticks to traverse this curve than the first one they had. They radio back to us that we don't know what we are talking about and the first curve is straight, ours is longer and bent outwards.

Notice that now we cannot interpret the bugs disagreement with Euclid to be due to there 2-d-sightedness; there are no shortcuts they were missing by being confined to the two dimensional space. If they climbed up and down buildings on their turntable world, they would just find that things in a $z-r$ plane seemed Euclidean, but things in more general configurations did not.

So we see that, if we allow noninertial reference frames, we better learn how to deal with curved space.

2.4 Gravitational Red Shift

But of course if we want to discuss gravity, we must discuss noninertial frames, because they are equivalent. In fact, we will show that **space-time** is curved. Here is how.

Consider two clocks at the front and tail of a spaceship, a proper distance h apart. Suppose that the spaceship $\mathcal{O}'$ is accelerating with small acceleration a. These clocks are, of course, atomic, so they work by emitting a particular monochromatic wavelength of light and counting the wiggles. Now let the light emitted by the tail clock reach the head clock. To describe what is happening, we will sit in a companion spaceship $\mathcal{O}$ which is **not** accelerating, and just happens to be at rest with respect to $\mathcal{O}'$ at the instant the light is emitted. Now by the time the ray reaches the head clock, which is roughly at time h/c, the clock has reached a velocity ah/c, away from the source, so it sees a redshifted frequency

$$f_{\text{recv}}/f_{\text{emit}} = \sqrt{\frac{1 - ah/c^2}{1 + ah/c^2}} \approx 1 - ah/c^2 = 1 - ah,$$

because, of course, $c \equiv 1$ in God's units. So the front clock sees that the bottom clock has the period of its oscillations lengthened from what they ought to be, and is running slowly. You could do this gedanken experiment the other way, and the bottom clock would agree that the top clock is running more quickly.

Now suppose we walk into the Empire State Building, and ask the blond receptionist at the front desk if we can meet with the grayhaired CEO up on the 100th floor. Sure, she says, but you better get your mind in gear, because his thoughts run faster. After all, the Empire State Building is equivalent to our spaceship, with $a = 9.8$m/s and $h \approx 300$m, so he ages about one microsecond more each year than the receptionist does.

Note that spacetime is therefore not Minkowskian. If we stay downstairs for a year and then travel on a photon up to the top, and if our brother first travels on a photon to the top and then waits for a year, instead of meeting as you might expect for a parallelogram in spacetime, he will be annoyed at our being late (by a microsecond).

3 (Pseudo) Riemannian Manifolds

3.1 Metric Tensor

We saw that at any one point $\mathcal{P}$ we can find coordinates ξ so that the distance between two points very near $\mathcal{P}$ is

$$(ds)^2 = (d\xi^0)^2 - (d\xi^1)^2 - (d\xi^2)^2 - (d\xi^3)^2 = \sum_{\mu,\nu} \eta_{\mu\nu} d\xi^\mu d\xi^\nu$$

but away from $\mathcal{P}$ the distance is not given correctly by this expression. We can't very well change our chart every time we move any finite distance at all, so let us consider a chart with coordinates x^μ, and see if we can find what the expression for the distance between $\mathcal{P}$ and $\mathcal{P}'$ is in terms of the infinitesimal changes dx^μ. Now even though we want to answer in terms of the x^μ, we know that there is an inertial frame ξ for $\mathcal{P}$, with the distance given as above. Furthermore, the points $\mathcal{P}$ and $\mathcal{P}'$ are in both charts, so the overlap map gives ξ^μ as a function of the x^ν, and vice-versa. So the differentials $d\xi^\mu$ can be expanded by the chain rule

$$d\xi^\mu = \sum_\rho \frac{\partial \xi^\mu}{\partial x^\rho} dx^\rho.$$

Plugging this into the expression for the length squared, we have

$$\begin{aligned} (ds)^2 &= \sum_{\mu,\nu} \eta_{\mu\nu} d\xi^\mu d\xi^\nu = \sum_{\mu,\nu} \eta_{\mu\nu} \sum_\rho \frac{\partial \xi^\mu}{\partial x^\rho} dx^\rho \sum_\sigma \frac{\partial \xi^\nu}{\partial x^\sigma} dx^\sigma \\ &= \sum_{\rho,\sigma} g_{\rho\sigma} dx^\rho dx^\sigma \end{aligned}$$

with[1]

$$g_{\rho\sigma} := \sum_{\mu,\nu} \eta_{\mu\nu} \frac{\partial \xi^\mu}{\partial x^\rho} \frac{\partial \xi^\nu}{\partial x^\sigma}.$$

[1]

$$V^\alpha{}_\mu := \frac{\partial \xi^\alpha}{\partial x^\mu}$$

is called a *Vierbien* or *tetrad*, depending on how angry you still are with the German's for the two World Wars. The Vierbeins are in a sense more fundamental than the metric tensor, and are required to deal with fermionic spinors. They are also gauge fields, in the sense that physics is invariant under local lorentz transformations of the ξ^α.

So the distances on the chart x are not specified just by the coordinates, but in addition we need a tensor field $g_{\rho\sigma}$ which is known as the metric tensor, or Riemannian (really pseudo-Riemannian) metric.

Why is this metric so important? Well, physics deals with rates of change, and that means how quickly some quantity changes if we move a certain **distance** in a certain direction, **not** if we change the coordinate by a certain amount. Let me give you an example you have already seen. If Φ is a scalar field in flat space described by polar coordinates, the component of the gradient of Φ in the θ, or tangential direction, is

$$\left(\vec{\nabla}\Phi\right)_\theta = \frac{1}{r}\frac{\partial\Phi}{\partial\theta} \neq \frac{\partial\Phi}{\partial\theta}.$$

The reason is that changing θ by a given small amount $d\theta$ changes the position by $rd\theta$, and that is what we must divide by.

3.2 What is a Vector?

The idea we have just used involves the setting up of fixed unit vectors at all points. While this is useful in discussing fixed curvilinear coordinates for Euclidean space, such as polar, spherical, or prolate ellipsoidal coordinates, it is not very useful in discussing general spaces. In fact, we need to reexamine the very concept of what a vector is.

What do we mean by a vector? In flat space physics, we introduce vectors by talking about displacements. We recognize that positions themselves are not truly vectors, because there is no inherent physical meaning to adding two positions. But displacements, such as the displacement from $\vec{x}_A$ to $\vec{x}_B$, can be thought of as a translation of the whole space by a distance (and direction) $\vec{x}_B - \vec{x}_A$. These translations can be added by successive application to the space. But notice that this requires the idea that the displacement $\vec{x}_B - \vec{x}_A$ acts on the whole space, not just the position $\vec{x}_A$. This is why we say that a vector has length and direction, but not location. Also notice that while the translation changes each point of the space to another point in the space, it leaves the set of points in the space unchanged.

Now in curved space this is not generally possible. Consider the surface of the Earth again, as an example of a curved space. What do I mean by the displacement (SERC)−(Serin Physics Lab)? While you might be able to apply that displacement to Biology and turn it into alcohol studies, you would have more trouble applying it to the base of the Eiffel tower. As a displacement in the three dimensional Euclidean space in which this surface is embedded, that displacement is well defined and is called an elevator ride to the top, but this is not a displacement defined in the curved space we are trying to discuss, the **surface** of the Earth. Of course you might try to define the displacement as a displacement by 300 m to the East. This would be well defined, though expensive, to apply to the base of the Eiffel tower, but what would it do to the North pole?

We see that small local displacements are well defined, because locally we can always define them in a local Euclidean (or inertial) coordinate system, but globally they are not well defined. Do we need more than small displacements to describe physics? No! All we need is enough to define derivatives, and those displacements can be as small as we choose.

3.3 Covariant Derivative

Now suppose we have a vector field, say the electric field. To apply Maxwell's equations, we have to be able to differentiate $\vec{E}(\vec{r})$ with respect to position. What does that mean? In Euclidean space, I find $\frac{\partial E_x}{\partial y}$ by taking the limit of the difference of E_x's at two close points. In curved space, there is, of course a local Euclidean coordinate system, but at each of these points, the x axes can be chosen arbitrarily, so there is no guarantee that we are subtracting corresponding components. If you don't think this is a problem, look at an example even in flat space. In spherical coordinates, the electric field due to a point charge q at the origin is $(q/4\pi\epsilon_0)r^{-2}\hat{e}_r$. Everywhere except at the origin, this is supposed to have no divergence because there is no charge. But we see that the r component does vary with r, while the other components are zero everywhere. So how can the divergence be zero? The answer is that we have naïvely subtracted the wrong components, and the correct formula for the divergence of a vector $\vec{V}(r,\theta,\phi)$ is

$$\vec{\nabla}\cdot\vec{V} = \frac{\partial V_r}{\partial r} + \frac{2}{r}V_r + \frac{1}{r}\frac{\partial V_\theta}{\partial \theta} + \cot\theta V_\theta + \frac{1}{r\sin\theta}\frac{\partial V_\phi}{\partial \phi}.$$

We can understand the coefficients in front of the partial derivative terms by normalizing the variation with respect to θ or ϕ to a unit length change rather than a unit change in angle, but what is the origin of the terms that have no derivatives? They come from the fact that the unit vectors themselves change direction as we move from point to point.

How do we determine that the unit vectors are changing direction? What tells us that the $(\hat{e}_r, \hat{e}_\theta, \hat{e}_\phi)$ coordinate system is rotating from one point to another? We need to have a notion of **parallel transport** which defines what it means for the coordinate system **not** to rotate, so we can compare to the one we are using and see if it does. In flat space that parallel transport is provided by the requirement that **in cartesian coordinates**, the components of a vector do not change as you parallel transport it, but in other coordinate systems it can.

Now in curved space there is no *a priori* preferred coordinate system like cartesian coordinates for flat space. The question of parallel transport is something that needs to be specified, just like the metric needs to be specified. The operation which describes the real physical change in the 4-vector, say V^ρ, is given by the **covariant derivative**

$$(D_\mu V)^\rho = \frac{\partial V^\rho}{\partial x^\mu} + \sum_\nu \Gamma^\rho{}_{\nu\mu} V^\nu.$$

To parallel transport a vector means to make the components change with x^μ in such a way that the derivative terms cancel the terms involving the **connection** $\Gamma^\rho{}_{\nu\mu}$, so that the covariant derivative is zero.

Now locally things are supposed to behave like they do in flat space. One way to determine how one parallel transports the coordinate system is to say that as one moves along a straight line, the angles between the various coordinate axes and the straight line remain fixed. In curved space, we need to replace the words "straight line" by "geodesic", but keep the same notion. An involved argument then shows that the connection coefficients are determined in terms of the metric coefficients and their derivatives.

We don't need to work this out in order to point out one surprizing fact about parallel transport. If I wander about space, carefully holding a flag so as to parallel transport it, when I come back to where I started from, the flagpole may not be pointing in the same direction it was when I started. Let me give you an example.

I will start out at the North pole, and start hiking towards London with my flagpole pointing straight ahead (so towards London, along the 0'th meridian). I keep the flag along the direction I walk, a geodesic. So the flag is pointing South. When I get to the equator, I turn myself to the West, but am careful not to rotate the flagpole, which still points South, now to my left, as I hike along the surface of the atlantic ocean and the amazon forest until I get near Bogota. All this time the flag is 90 degrees to my left, pointing South. Now I turn North, again being careful not to rotate the flag, which, still pointing South, is now 180 degrees from my geodesic path. I hike due North, passing Piscataway, until I finally get back to the North pole. But my flagpole is pointing towards Bogota, not London, so it is at an angle of 75 degrees from what it was at when I started, even though I have been careful to parallel transport (not rotate) it at all times.

This rotation is a direct, coordinate independant measure of the curvature in the region around which we have parallel transported the vector. In four dimensions, there are a lot of ways the coordinate system can rotate, and it depends on the orientation of the path around which we parallel transport the vector, so the notion of curvature is more complicated. In two dimensions, the rotation of the flagpole is simply proportional to the integral of the curvature over the surface surrounded by the transport path.

In general relativity, it turns out that matter produces curvature of space-time. But I am not going to go any further. This is left for a course in general relativity.

4 Gauge Theories

The concept of corvariant derivative has other applications besides general relativity. It is always related to there being a local invariance. Consider what it is due to in general relativity: At each point we know there is a physical preference for locally inertial coordinates, in terms of which the physics looks like there is no gravity and no fictitious forces. But there are many choices for this locally inertial coordinate system, because if $\{\xi^\mu\}$ is any such system, so is any other set of coordinates ξ' related to ξ by a Lorentz transformation. So the physics could be equally well described by this coordinate system as well, and is invariant under the local change of $\xi \to \xi'$. As this is true at each point, this is called a local gauge invariance.

There are other situations in which it is necessary to have some sort of coordinate system to write down the components of a field, and sometimes this coordinate system has nothing to do with spacetime. In the theory of quantum chromodynamics (QCD), the fields live in a three dimensional space whose axes might be called redness, greenness, and blueness, having nothing to do with the three spatial dimensions. But the physics of quantum chromodynamics treats all directions in this space as equivalent, so that to say a particular quark is red is a statement about your coordinates as much as about the quark. In order to write down the physics of these quarks, we have to write a term which expresses how the quark field changes from point to point, but that requires comparing fields at two different points, with two arbitrarily chosen coordinate axes for the colors. Just as for gravity, this requires a connection to define how the coordinate system changes when it is "parallel transported". That is, for each direction in space-time that one can move (x^μ), there is a field $A_\mu(\vec{r}, t)$ which is an operator which rotates the color coordinate system appropriately. Then every derivative that appears in the physical equations appears in the combination

$$\frac{D}{Dx^\mu} := \frac{\partial}{\partial x^\mu} + \frac{e}{\hbar c} A_\mu$$

which is called the covariant derivative.

A simpler case is electrodynamics, where charged particles are represented as vectors in a two-dimensional real space (or by one complex number, which is equivalent). There is only one thing that can happen to the coordinate system, a rotation in the two-dimensional space, so A can be considered a number or to be a matrix proportional to $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Then A_μ is the electromagnetic 4-potential. In QCD, the situtation is more complicated, with A_μ representing a group operation in the group called $SU(3)$ of unitary 3×3 matrices with determinant 1. Then its physical interpretation is the quantum field of the gluons, the eight particles of glue which hold together the quarks so tightly that one can never free a quark.

In either case, the A_μ represents partly a choice of arbitrary coordinate systems, so it is not an object with direct physical meaning. But because of this gauge field, there is a strange affect: the covariant derivatives do not commute with each other the way ordinary derivates do. Instead

$$D_\mu D_\nu - D_\nu D_\mu = F_{\mu\nu},$$

where F is known as the field strength tensor. For electromagnetism this is just the object we mentioned before which consists of electric and magnetic fields, which are physical and independant of the coordinate system used to describe the wavefunctions of the charged particles. The nonzero value of the commutator of two covariant derivatives is directly connected to the net rotation of the flag after parallel transporting it around a path.