General Relativity

Philosophy of general relativity.—As with any major theory in physics, GR has been framed and derived in many different ways, each giving their own insight. **Ask class:** can they think of other examples in physics? $F = ma$ versus Lagrangian or Hamiltonian mechanics; wave versus matrix versus path integral quantum mechanics; quaternions(!) versus vector electromagnetism. In the case of GR, there is the geometric approach, good for insight and reasoning, and the action approach, probably better for trying to unify gravity with the other forces. In all these examples, a common theme is that the predictions had better be the same. Similarly, although, say Newtonian mechanics is based on a completely different set of philosophical principles than quantum mechanics or relativity, in the big, slow-moving, weak-gravity limit the predictions of all those theories are the same. This, the contact with observables, is the most fundamental point of theories, in my opinion. Therefore, I will present things in a way designed for calculation.

Another point about general relativity is that it is the least confirmed of our current fundamental theories. A major reason for this is that its most dramatic effects only show up in extremely strong gravity, such as near black holes and neutron stars. This gives it special status, and means that astronomical observations may have the most to contribute to fundamental physical understanding in the realm of strong gravity.

Finally, let me say that I plan to go into a little more detail about the formalism and equations of GR than I did into particle interactions. The reason is that you aren't necessarily going to see GR anywhere else, so I'd like this part of the course to be more self-contained.

Fundamental GR concepts

(1) As in special relativity, space and time are both considered as aspects of spacetime. However, whereas in special relativity spacetime is "flat" (in a sense to be defined later), in general relativity the presence of gravity warps spacetime.

(2) The natural motion of objects is to follow the warps in spacetime. "Matter tells space how to curve and space tells matter how to move." An object that is freely falling (i.e., following spacetime's warps) does not "feel" force, meaning that an accelerometer would measure zero. The path of a freely falling particle is called a geodesic.

(3) The only "force" in this sense that can be exerted by gravity is tidal force. That is, if an object has finite size, different parts of it want to follow different geodesics, and these deviate. Geodesic deviation is the GR equivalent of tidal forces.

(4) Because of this deviation, global spacetime is not flat and there is no coordinate transformation that will make it look flat everywhere.

(5) However, THE most important principle of GR is that in a sufficiently small region

of spacetime (small spatial scale, small time interval), the spacetime looks flat. This means that there is a local inertial frame that can be defined in that small patch of spacetime. In that local inertial frame, all the laws of physics are the same as they are in special relativity (electrodynamic, hydrodynamic, strong+weak nuclear, ...)!! This is called the *equivalence principle*. There is a classic elevator analogy for this principle, which says that if you are in an elevator and you feel like you are being pushed towards its floor, you can't tell whether you are at rest in a gravitational field or are being accelerated in flat spacetime. The equivalence principle means that in practice one of the best ways to do calculations in GR is to do them in the local inertial frame and then use well-defined transformations between the local and the global frame.

(6) All forms of energy gravitate. In the Newtonian limit, rest mass is overwhelmingly the dominant component, but in ultradense matter other forms can be important as well.

The Mathematics of Curved Spacetime

Consider a two-dimensional space. We know that there are differences between, e.g., a flat plane and the surface of a sphere. One example of this is that on a plane, the interior angles of a triangle always add to 180°, whereas on the surface of a sphere the angles always add to something larger than 180°, but the actual value depends on the size of the triangle. Another example is that if you take a vector on a flat plane and transport it parallel to itself, you can move it around the plane to your heart's content and when you bring it back to the starting point it will have the same orientation it did before. This is not the case on a sphere!

Note, however, that (in good analogy to GR!), on a small enough region of a sphere you can treat it as flat. We need to develop a formalism that can handle curvature like this, except in four dimensions (three spatial, one time). This is the formalism of geometry in curved spacetime.

Geometric objects.—Let's start by defining some geometric objects. Bear with me for the first couple, which seem obvious but lay the groundwork for the less obvious sequels.

Event.—First we have an *event*. An event is effectively a "point" in spacetime. More generally, if you have an N-dimensional space, you need N numbers to label it uniquely. For example, in two dimensions you need two numbers; e.g., x and y for a plane, or θ and ϕ for the surface of a sphere. For spacetime, you need four numbers: e.g., t, x, y, and z. Naturally, the essence of the event isn't changed if you relabel the coordinates. I want to stress this, because something that is obvious for events but may not be obvious for some other geometric objects is that although when you finally calculate something you may choose a coordinate system and break things into components, there is also an independent reality (well, within the math at least!) of the objects. Going with the coordinate-free representation has proved very helpful in proving theorems about GR, but when doing astrophysics it is usually best to investigate components in some given system.

Vector.—Next, consider a *vector*. In flat space, this is easy. Using our previous definition, we can simply think of a vector as an arrow connecting two events. As long as we define the arrow to be a straight line, there is no ambiguity, regardless of how far separated the two events are. Again for concreteness, let's think of two dimensions and Cartesian coordinates, so the events are labeled by their x and y coordinates and a vector between them also has an x and y component. Now, think of two points (aka events) on the surface of a sphere. How do we now define a vector? Not easy. First, we need to decide what a straight line is. A good choice is a great circle. Then, however, we have a problem. Every pair of points on a sphere is connected by at least two great circles, and antipodal points are connected by an infinite number of great circles! That shows that in curved spacetime, vectors can't be defined for two points that are distant from each other. Therefore, we define vectors only as infinitesimal quantities. Here again, we've defined a vector as a line between two points (now infinitesimally close), but a vector has its own existence independent of points or events or coordinate systems. In higher dimensions, in any case, vectors are only locally defined. They can have a magnitude, like a gradient, but they don't extend over more than an infinitesimal region.

Also, of course, the direction and magnitude of a vector field can change with position. Think of an electric or magnetic field. At each point you can define the vector (direction plus magnitude) of the field, but both direction and magnitude change with position.

One-form.—Events and vectors are pretty familiar. Not so with the next object. This is a *one-form*. Think again about a Euclidean plane, and two points very close to each other. You can define a vector between them. But you could equally well define something perpendicular to that vector (draw). If you imagined two points in three dimensions, then after drawing a vector you could draw a plane perpendicular to it. In four dimensions (spacetime), you would have a three-dimensional thing perpendicular to the vector. This is called a one-form. One-forms are written differently than vectors. For example, consider the radial component of a velocity $\mathbf{v}$. It is written v^r, with a superscript. The r-component of the corresponding one-form would be written v_r, with a subscript. These are also called, respectively, *contravariant* (up) and *covariant* (down) components. We will later get into how to transform between the two, by raising and lowering indicies.

A common and reasonable question at this stage is: what the heck is the *use* of one-forms? At the level we'll be using them they just allow us to do bookkeeping, but in differential geometry their importance is greater. To get a sense for that, let's use a concrete example.

Suppose that you are in a room with a temperature gradient. You move from one part of the room to another and the question is: how large a temperature change did you experience? This depends on the distance and direction you traveled *and* on the magnitude and direction of the temperature gradient. The first part (distance and direction) is best represented using a vector. But the second part (magnitude and direction of the gradient) is best described by a one-form. Why? Think about the temperature gradient as a series of surfaces of constant temperature; to make things simpler, suppose that those surfaces are planes that are parallel to each other. For example, maybe you have one plane at 22° C, one at 21° C, one at 20° C, and so on. The closer the planes, the steeper the gradient. For example, in one room the planes might be 1 m apart, so that the gradient is 1° C/m. In another room the planes might be 0.5 m apart, so that the gradient is 2° C/m. This gradient is not conveniently represented by an arrow (vector), so instead we use a one-form. The total temperature change you experience is the dot product between the vector and the one-form. In the same way that an arrow in space (our common picture for a vector) has a coordinate-free identity, the planes (or more generally surfaces) of constant temperature in our example have a coordinate-free identity.

But there are differences as well. One of them goes back to the strange terminology of "covariant" and "contravariant". These refer to how quantities, in a given system of units, change when the units are changed. To be concrete again, suppose that we have a temperature gradient of 1° C/m in the x direction, and an arrow of length 1 m in the x direction. Say that the unit of length in our system is 1 m. This means that our temperature gradient is 1° C/unit and our arrow has a length of 1 unit. Now suppose that we have the *same* temperature gradient and arrow, but we switch to a system in which the unit of length is 2 m. Then our arrow has a length of 0.5 units. Since increasing the length of the unit decreases the length in those units (same actual length, but measured in units the value has decreased), the length changes in a way opposite the unit change. Thus vectors are *contra*variant. In contrast, the temperature gradient increases in the new units, to 2° C/unit. Since the gradient in our new unit system increases when the unit increases, the gradient changes in the same way as the unit changes and thus gradient are *co*variant. One-forms, which represent gradients (among many other things), are covariant, whereas vectors, which represent distances and directions (among many other things) are contravariant.

I hope this explains things at least a bit. If you want to pursue differential geometry more thoroughly you'll get a better sense of why there is a distinction between one-forms and vectors.

Tensor.—One can generalize with the further concept of *tensors*. Think, for example, of the gradient of an electric field: $K \equiv \nabla \mathbf{E}$. At a given point, if you want to know the components of this you can't just specify its x or y component, for example. Instead, you need to specify things like the gradient of the x component of E, in the y direction. Then K might have components like K^{xy} or K_{zz}, depending on whether one wanted to go with a contravariant or covariant description. We'll get to how to manipulate tensors and their indices a little later. For now, it is also useful to think of a tensor as a machine that can take vectors or one-forms as inputs in its "slots", one slot per index, and return a number. It is a linear machine. As with vectors and one-forms, a tensor has a mathematical existence of its own, independent of coordinate systems, but when calculating it is usually convenient to select a particular coordinate system and compute with components.

The "rank" of the tensor is the number of separate indices it has. For example, $T^{\mu\nu}$ is a second-rank tensor and $R^{\alpha}_{\beta\gamma\delta}$ is a fourth-rank tensor. One especially important second-rank tensor is the metric tensor, which we'll talk about now.

Metrics

Now let's move a little from those basic definitions to how they are used in curved spacetime. A great way to characterize the curvature is through the use of a *metric*. This effectively tells you the "distance" between two events. For example, in two dimensions, what is the distance between two events separated by dx in the x direction and dy in the y direction? The distance ds is just given by $ds^2 = dx^2 + dy^2$. In polar coordinates we would write $ds^2 = dr^2 + r^2 d\theta^2$, but it's the same thing. One point (obvious here, not so obvious later) is that this distance is the same for two given events even if the coordinates are redefined (e.g., by rotation of the coordinate system). Therefore, we can call this the invariant *interval* between the events.

In four dimensions and flat space, special relativity tells us that the invariant interval is defined as

$$ds^2 = -c^2 dt^2 + dx^2 + dy^2 + dz^2 \ . \tag{1}$$

For two arbitrary events, ds^2 can be positive, negative, or zero. **Ask class:** consider just dt and dx, so that $ds^2 = -c^2 dt^2 + dx^2$. What is the condition that $ds^2 = 0$? This is the condition that the two events could be connected by a photon going from one to the other. Therefore, $ds^2 = 0$ is the path of a light ray.

We can represent this more compactly, using the *metric tensor* $g_{\alpha\beta}$, as

$$ds^2 = g_{\alpha\beta} dx^\alpha dx^\beta \ . \tag{2}$$

Here we have introduced two conventions. The first is the use of greek indices to represent indices that might be any of the four in spacetime (for example, t, x, y, and z). The second is the Einstein summation convention. Whenever you see a symbol used as an up and a down index in the same expression, you are supposed to sum over the four possibilities. For example, $v^\alpha u_\alpha = v^t u_t + v^x u_x + v^y u_y + v^z u_z$. This also means that if you sum over indices, you don't count them in the rank of the tensor. For example, $v^\alpha u_\alpha$ is a scalar, $T^{\alpha\beta} u_\alpha$ is a rank one tensor, and $R^{\alpha}_{\beta\alpha\gamma}$ is a rank two tensor. This means that ds^2 is a scalar, so for example it transforms as a scalar does under Lorentz transformations, i.e., it's unchanged (that's why it is an "invariant" interval!).

Let's take a moment to emphasize that point a bit more. *Scalars are unchanged under Lorentz transformations!* At a given "point" (i.e., event) in spacetime, *everyone* will agree on the value of a scalar. If this isn't clear, just think about a trivial example of a scalar: a pure number. For example, suppose a box contains 3 particles. The number 3 is a scalar. Obviously everyone will agree that the number of the count is 3: countest thou not to 2 unless thou countest also to 3; 5 is right out; and so on. **Ask class:** what are other examples of scalars? Since scalars are invariant, it often gives insight to construct scalars out of the geometrical quantities of interest.

Of course, the particular indices used are dummy indices; we could as well have written $ds^2 = g_{\mu\nu} dx^\mu dx^\nu$. Therefore, dx^μ (or dx^ν or dx^α or whatever, it's the same thing) is the μth component of the separation between the events, in this particular coordinate system.

The metric tensor is symmetric: $g_{\alpha\beta} = g_{\beta\alpha}$.

Going back to the particular metric we wrote earlier: $g_{\alpha\beta} = (-1, 1, 1, 1)$ down the diagonal. This is called the Minkowski metric, and is usually given a special symbol: $\eta_{\alpha\beta}$. The Minkowski metric is of special importance. It describes flat spacetime, which is spacetime without gravity. Any metric that can be put in the Minkowski form by a transformation also describes flat spacetime. Here, by the way, is a place to point out the difference between *spacetime* and a particular metric used to describe it. Spacetime has some particular geometric characteristics (for example, it's flat). A metric is what you get when you pick a coordinate system to use with that spacetime. Seems trivial, but as we'll see in a later lecture, sometimes a change in the coordinates (which therefore does not change the spacetime) has made a big difference in how things are perceived.

So, in flat spacetime one could easily write $ds^2 = -c^2 dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)$, and recover the Minkowski metric by the usual spherical to Cartesian transformation. However, there are some metrics for which a global transformation to Minkowski is not possible, e.g.,

$$ds^2 = -(1 - 2GM/rc^2)dt^2 + dr^2/(1 - 2GM/rc^2) + r^2(d\theta^2 + \sin^2\theta d\phi^2) \ . \tag{3}$$

Such a metric describes a fundamentally different spacetime (in this case, the spacetime around a spherically symmetric gravitating object, such as a black hole or the Earth for that matter). The nontransformability to Minkowski tells us that this spacetime is curved.

However, although you can't make such a transformation globally, you *can* make it in a small enough region of spacetime. That is, if you go into this spacetime and select an itty bitty region just Δt by Δx by Δy by Δz across, where all those are small, then in just that region you can devise a coordinate system that looks like Minkowski, except for terms of order $(\Delta t)^2$ and so on. This is an exact analogy with the fact that on a sphere, if you look at a small enough patch, you can invent a coordinate system that looks just like Euclidean plane geometry except for little terms of second order. This is one of the aspects of the equivalence principle: in a small enough region of spacetime, the geometry looks flat, which means that there is a reference frame in which the effects of gravity don't exist! The only deviations from the experiments you'd do in perfectly flat space are second-order. These are "tidal forces", known in this context as geodesic deviation. A key to doing GR calculations is to (1) do all the physics in the local Minkowski frame, where things are simple, and (2) know how to transform from that local frame to the global frame.