General Relativity

Nirmalya Kajuri

School of Physical Sciences,
IIT Mandi,
Himachal Pradesh,
India

nirmalya@iitmandi.ac.in

Contents

1 Introduction

General relativity(GR) is our best-tested theory of gravitation. Since its discovery by Einstein in 1915, GR has been confirmed in every observational test that we have conducted. Predictions once viewed with suspicion by physicists(including even Einstein himself), such as black holes and gravitational waves, have proven to be accurate. The tremendous progress we have made in understanding the history of the universe would have been impossible without general relativity. The theory has even found its way into our daily lives via its application in GPS navigation.

The theory itself is striking in its beauty and simplicty. The seemingly disparate concepts of space-time and gravitation are united in it. GR's central equation is succinct enough to be printed on a T-shirt, yet it contains literal universes.

There was a time when general relativity had a reputation of being a difficult subject to learn. As you will see, this is far from the truth! Yes, you will encounter some novel physics and maths concepts in this course, and yes, some of them take a little time to get used to. But you will find that the rules are quite simple, straightforward and easy to follow.

The main idea of general relativity is easy to state: "gravity is geometry". Or, in the words of John Wheeler: "matter tells spacetime how to curve, spacetime tells matter how to move." In this course we will spend a lot of time decoding these slogans, particlularly the meaning of the words 'geometry' and 'curve.'

In this chapter, we will cover some preliminaries. We will start by asking why general relativity was needed in the first place? We will then recollect Newton's laws of force and find hints that gravity may not even be a force. We will then introduce the idea of different geometries. Finally we will revisit the idea of a vector.

1.1 Why did Newton's Law of Gravitation Need an Upgrade?

For over 200 years, Newton's law was our best-tested theory of gravitation. It stated that the gravitational force follows an inverse square law. In details: the force on a particle of mass m_1 located at $\vec{r_1}$, due to a particle of mass m_2 located at $\vec{r_2}$ is given by:

$$\vec{F}_{12} = \frac{GMm(\vec{r_2} - \vec{r_1})}{|\vec{r_2} - \vec{r_1}|^3} \tag{1.1}$$

where G is a constant, called Newton's constant.

Using this simple rule, one could make accurate predictions for a whole host of phenomena. From the speed of an apple falling to the earth to the orbits of planets around the Sun– Newton's theory could match all the observed data.

In fact, even when Einstein started working on general relativity there was no data that contradicted Newton's law of gravity[1] So how did Einstein and others realise that it needed to be replaced?

The tension was not between theory and observation, but between theory and theory. To everyone who understood special relativity, it was immediately clear that Newton's laws were in contradiction with special relativity. We will see this in more detail after we review special relativity, but we can spot the tension immediately from (1.1).

Let one of the masses be the Sun and the other, Earth. It takes eight minutes for light to travel from Sun to the Earth. Nothing travels faster than light as per special relativity–so no information from the Sun can reach the Earth in less than 8 minutes. If the Sun exploded away or was jolted from its orbit, we should not have any idea for 8 minutes.

But this time gap does not appear in Newton's law of gravitation. According to (1.1) is correct, the force on the earth due to the Sun depends on the *current position* of the Sun. This means that if I measure the direction and the magnitude of this force, I can use (1.1) to deduce the current position of the Sun. So the information about the position of the Sun is transmitted faster than the speed of light. It was transmitted *instantaneously* i.e with zero time delay.

The property that that the gravitational force propagates instantaneously in Newton's law of gravity is called 'action at a distance', i.e one body can influence another from a distance without any physical signal (like light) going from one to thee other. Action at a distance is incompatible with special relativity.

It is useful to contrast Newton's law of gravitation with Maxwell's electrodynamics. One might think that the same problem will crop up in electrodynamics. After all Coulomb's law looks a lot like Newton's law:

$$\vec{F} = \frac{K q_1 q_2 (\vec{r_1} - \vec{r_2})}{|\vec{r_1} - \vec{r_2}|^3} \tag{1.2}$$

But Coulomb's law holds only for electrostatics, when all charges are at rest. When we consider charges in motion, the formula is different!

The formula for the field is complicated, but the electric potential has a relatively simple expression. The electric potential due to a charge moving with velocity $\vec{v}(t)$ at a point that is

[1]There was one observation at the time that did not match with the predictions of Newtonian gravity–the recession of perihelion of Mercury. However, the popular explanation was that the discrepancy was caused by a small, hidden planet.

currently r distance away, calculated in Lorentz gauge, is[2]:

$$\phi_{\text{electric}}(r,t) = -\frac{q}{\left(1 - \frac{\vec{v}(t)\cdot\hat{R}}{c^2}\right) R(r,t)} \tag{1.3}$$

Here $R(r,t)$ is the *retarded distance*, i.e the distance from the charge at a time $t - r/c$ where c is the speed of light. This is the past location of the charge from where the light has just reached the point in question. In our example, this would be the position of the Sun 8 minutes earlier.

In contrast, the gravitational potential is

$$\phi_{\text{grav}}(r,t) = -\frac{Gm}{r}.$$

The key difference between the two cases is that the force in the electrodynamic case always depends on the retarded position of the source and not the current/instantaneous position(The factor $(1 - \frac{\vec{v}(t)\cdot\hat{R}}{c^2})$ is just a relativistic length contraction). Hence, there is no faster than light information transfer in electrodynamics.[3]

Newton's laws needed an upgrade to become compatible with relativity. This is the starting point of the journey towards general relativity.

(Aside: This does not mean Newton's law of gravity is wrong or useless. Every new theory gives us a better and better approximation of nature. Newton's gravity works perfectly well when the gravitational potential is sufficiently weak and the objects under consideration move at non-relativistic speeds. We will see later that we can derive Newtonian gravity as an approximation of Einstein's theory in this regime.)

1.2 Inertial Frames and Newton's Laws of Motion

There is one concept from Newtonian physics that will continue to be useful even in general relativity: the concept of reference frames and specifically inertial frames of reference. Let us revisit this concept.

To locate a point in 3 dimensional space, we typically use a Cartesian coordinate system and specify its (x, y, z) coordinates. There can be different coordinate systems depending on the choice of origin and choice of axes. A reference frame plays the same role as a coordinate

[2]Caveat: I showed the electric potential here because it has a simple expression compared to the field. But actually it is the field that has this property of propagating at the speed of light, the potential is an unphysical quantity that need not respect causality. Indeed it would not if we chose a different gauge.

[3]Maxwell's equations of electrodynamics have special relativity built into it, even though special relativity was not known during Maxwell's time. Electrodynamics came with the 'relativity upgrade', but the the gravitational law did not!

Figure 1. Principle of relativity: there is no way to distinguish between different inertial frames; laws of physics are identical in all of them.

system, but to locate points in both space and time. So it consists of a Cartesian coordinate system (x, y, z) and a clock measuring time t. A reference frame is then labelled by 4 numbers (x, y, z, t).

Inertial Frames are a special type of reference frame. Their special property is this: free objects do not accelerate in an inertial frame. A free object is one which is not subject to any external force.

How would one verify if they are in an inertial frame? If you see some object is accelerating in an inertial frame, you are guaranteed to find some external agent exerting force on it. It could be a charge that is causing it, or a mass, or a magnet–if you look around you are guaranteed (in principle) to find some source. This is not the case in non-inertial frames, where you will find stuff accelerating without any external agent in the picture.

The key point here is that the definition of an inertial frame only talks about acceleration, not velocity. If you find one inertial frame, then any other frame moving with constant velocity with respect to it will also be inertial. This is because an object that is not accelerating in the first frame will not be accelerating in any of these frames.

In other words, the definition of an inertial frame makes no distinction between rest and motion with constant speed. This fact is called the principle of relativity. It says that there is no experiment to tell you if something is at rest, or moving with constant speed. Put differently, the laws of motion are the same in all inertial frames. A non-inertal reference frame is one that is itself accelerating with respect to an inertial frame. In such an accelerating frame, free objects will accelerate.

Inertial frames are defined via Newton's first law. Indeed, Newton's first law simply says that

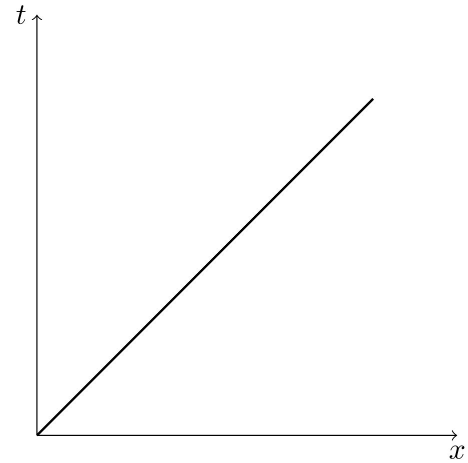

Figure 2. Plot of the motion of a free object: a straight line spacetime

free objects don't accelerate, which is the same as saying inertial frames exist.[4]

It will be useful for the future to phrase Newton's first law in the language of 'spacetime'. The time vs distance plot of a constant velocity object is a straight line. If we refer to this graph as 'spacetime', Newton's first law says, a free object moves in a straight line in spacetime (as shown in Fig 1.2. This formulation will become meaningful after we study relativity and understand why spacetime is a fundamental concept.

We will see that Newton's first law will continue to hold in general relativity.

While we are here, let us quickly recap Newton's second law (this will be useful in a moment). This tells us that in an inertial frame, we have

$$m\vec{a} = \vec{F} \tag{1.4}$$

where $\vec{F}$ follows a *force law*. The force law could be the gravitational force law

$$\vec{F} = \frac{GMm(\vec{r_2} - \vec{r_1})}{|\vec{r_2} - \vec{r_1}|^3} \tag{1.5}$$

or Coulomb's law of electrostatics:

$$\vec{F} = \frac{Kq_1q_2(\vec{r_1} - \vec{r_2})}{|\vec{r_1} - \vec{r_2}|^3} \tag{1.6}$$

or something else.

(Aside: It is important that there is a force law for every force, otherwise (1.4) would have been kind of circular and useless. Without the force law, we would never be able to predict the acceleration.)

[4]Because it is often taught with more stress on the exact words "objects in state of rest or uniform motion bla bla" than the meaning, this is not always clear to students.

1.2.1 Pseudo Forces

If I am accelerating with respect to an intertial frame, I will see free objects accelerate. So Newton's second law as stated above would not work in an accelerating frame.

The trick to keep using Newton's second law in an accelerated frame by introducing fictitious forces or 'pseudo forces.'

If a frame is accelerating at $\vec{g}$, then a free object will be have an acceleration

$$\vec{a} = -\vec{g} \tag{1.7}$$

in this frame.

This acceleration is purely kinematical, it does not involve dynamics.[5]

But one can still put this acceleration in the language of Newton's second law by inventing a 'fake' or fictitious force law:

$$\vec{F} = -m\vec{g} \tag{1.8}$$

While (1.8) looks like a force law, all we did was to multiply both sides of (1.7) with the mass m. Such forces are called pseudo-forces or fictitious forces.

The thing to note here is that we are not gaining any extra information from (1.8) that was not present in (1.7). Introducing fictitious forces is dressing up kinematics as dynamics.

Another example of a pseudo force is the centrifugal force. In a frame rotating uniformly with angular velocity ω, a free object will be seen to accelerate with

$$\vec{a} = \omega^2\vec{r} \tag{1.9}$$

Again we can introduce the fictitious centrifugal force with the law:

$$\vec{F} = m\omega^2\vec{r} \tag{1.10}$$

Again, we don't lose any info if we use the acceleration equation (1.9) instead of the force law (1.10).

The way to identify a pseudo force is that you can always cancel the mass out on both sides and get an expression for acceleration that only depends on the position or velocity of the objects. This is simply the fact that the acceleration will be exactly the same for any object, no matter how heavy.

[5] Recall that kinematics is the study of the properties of motion without trying to understand what causes the motion. Dynamics is the study of what causes the motion i.e forces. Kinematics studies formulae like $v^2 = u^2 + 2as$, dynamics studies force laws.

1.3 Gravity is not a Real Force?!

Now I am going to try to convince you that there is something different about gravity compared to pseudo forces. In a way, it is similar to pseudo-forces. This will open the way to general relativity.

This clue to gravity's secret kinematic origin comes from the observation we can cancel the mass on both sides of (1.1) to get:

$$\vec{a} = \frac{GM(\vec{r_1} - \vec{r_2})}{|\vec{r_1} - \vec{r_2}|^3} \tag{1.11}$$

This observation has a name – it is called the principle of equivalence.

Note that there is no force in this equation. It is an equation telling us the acceleration experienced by a body entirely in terms of its location. It will be exactly the same for every object at the same position, no matter how heavy.

This is just the same as in the case of fictitous forces. Like in those cases, we would not loss any information if we worked with the acceleration equation (1.11) instead of the force law (1.1).

What this is telling us that the action of gravity depends not on what the object is(i.e how heavy), but on *where* it is.

All of this drives home the point that gravity is in some ways more of a kinematic effect than a dynamic one. As we will see, gravity is really a *geometric* effect.

1.3.1 Gravity as geometry

Gravity originates from the shape of spacetime. We will unravel this statement over the course of this course. Now we consider a baby example of how shapes can affect motion.

Consider two situations. In the first situation two ants are moving on a plane. They start moving north and follow their nose. These ants will keep moving parallely and never meet.

But what if the ants were moving on a spherical surface, like a globe? Suppose two ants again start out from different points on the equator and start moving north, once again following their nose. These two ants would end up meeting at the North Pole.

If you zoom closely enough to any surface, it looks flat. To the ant, it would therefore have looked like it was always moving in a straight line on a flat surface. A flat-earther ant would be very surprised at meeting the other ant at the North Pole. It might try to explain this by positing some force attracting them to the North Pole.

But in reality, it was the shape of the surface. Even though the ants were trying to move in a straight line at any point, their path became more complicated because of the shape of the surface they were moving on.

We will see that gravity works similarly to this. It is the shape of space-time that becomes different. As we will see, a fully empty spacetime has a shape similar to a plane. But the presence of massive objects causes the shape to change from plane to curved.

Recall our spacetime formulation of Newton's first law: every object moves in a straight line in spacetime. It will turn out to still hold. But when spacetime is itself curved, the object will end up following a different overall trajectory, like an orbit. Gravity is not a force, just the effect of objects trying to move in a straight line in a space-time that is curved.

I gave you a vague, hand-waving explanation of gravity here. The aim of this course is to make this idea mathematically precise.

1.4 Life in Two Dimensions

We will not look at the two geometries from the last section in a little more detail and introduce a key concept in the study of geometry.

How do we mathematically distinguish between a 2 dimensional plane('flat' space) and a sphere ('curved' space)?[6] In 3-dimensions it is easy: a sphere would be described by an equation like $x^2 + y^2 + z^2 = R^2$, whereas a plane would have an equation like $x + y + z = a$.

But suppose we were 2D creatures ourselves, inhabiting a 2D world. What mathematical tool would we use then?

It turns out that we can use the distance between two very close(infinitesimally close) points to make a distinction.

In a 2D plane, the distance between two points (x, y) and $(x + \Delta x, y + \Delta y)$ is given by:

$$\Delta S^2 = \Delta x^2 + \Delta y^2. \tag{1.12}$$

When the two points are infinitesimally close–(x, y) and $(x + dx, y + dy)$–the distance between them is:

$$dS^2 = dx^2 + dy^2. \tag{1.13}$$

[6] When we say sphere we mean a spherical surface. Not the inside of the surface. This is usual language in math, which can be confusing at first.

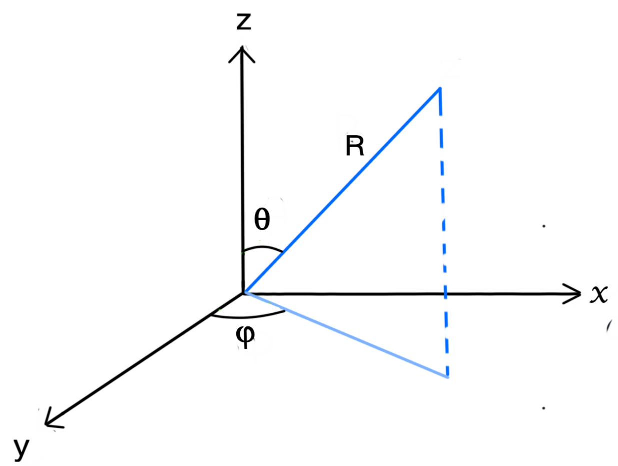

Figure 3. Spherical Polar Coordinates

This is called the 'infinitesimal interval' or just the interval. This is an important concept in the study of geometry of spaces and will be central to this course.

Let us rewrite the interval in polar coordinates:

$$dS^2 = dr^2 + r^2 d\theta^2. \tag{1.14}$$

Now let us try to deduce the infinitesimal interval on a sphere. One way to do this is to start from 3-dimensional space and use spherical polar coordinates (R, θ, ϕ) as shown in Figure 3.

In these coordinates, the infinitesimal distance in 3D is:

$$ds^2 = dR^2 + R^2(d\theta^2 + \sin^2 d\phi^2). \tag{1.15}$$

We can define a sphere around the origin by fixing some radius $R = R_0$ around the origin. So the distance between any two points on the sphere will be given by (1.15), but with R fixed to be R_0.

Then $dR = 0$ on the surface.

So we have

$$ds^2 = R_0^2(d\theta^2 + \sin^2\theta d\phi^2). \tag{1.16}$$

Introducing $r = R_0\theta$, we rewrite this as:

$$ds^2 = dr^2 + R_0^2 \sin^2(r/R_0) d\phi^2. \tag{1.17}$$

Let's take $R_0 = 1$ for simplicity. We have then:

$$dS^2 = dr^2 + \sin^2 r d\phi^2. \tag{1.18}$$

This is different from the infinitesimal interval on a plane (1.14).

If we zoom into a tiny region of a sphere, it looks flat like a plane. Just like the earth appears flat at small distances. This is reflected in the formula above–when r is small, then $\sin r \approx r$ and the the two intervals match.

So far this is purely mathematical. If we were 2D creatures living in these spaces, we would want to know how this math relates to what we observe.

Let us compare how distant objects appear to the inhabitants of these two worlds. For the plane, our intuition says that far-away objects appears smaller and smaller. Let us see this mathematically.

Let us put our observer at the origin $r = 0$ and consider a small object at a radial distance r. Let the object is small enough compared to its distance from the origin to be considered of an infinitesimal size dS.

Then we have from (1.14):

$$dS^2 = r^2 d\theta^2 => d\theta = dS/r \tag{1.19}$$

where we put $dr = 0$ because we are interested in the angular size only(which is what we would observe). So the angular size of an object decreases inversely as its distance in the 2D planar world (In 3 dimensions the solid angle decreases as $1/r^2$).

Now for the sphere. Using a similar logic as before, we see that the angular distance goes as:

$$d\phi = \frac{dS}{\sin r} \tag{1.20}$$

As $sinr$ increases with r from 0 (North Pole) to $\pi/2$(equator), the angular distance decreases till this point. After that it starts increasing, eventually hitting infinity at South Pole.[7]

We can also see that close to $r = 0$, $sinr \approx r$, so the angular distance falls off in the same way as a plane, as we expected.

An even simpler thing to compre are the circumference-to-radius ratios of a circle.

[7]The reason is that no matter which direction the observer looks, their line of sight will end at the South Pole.

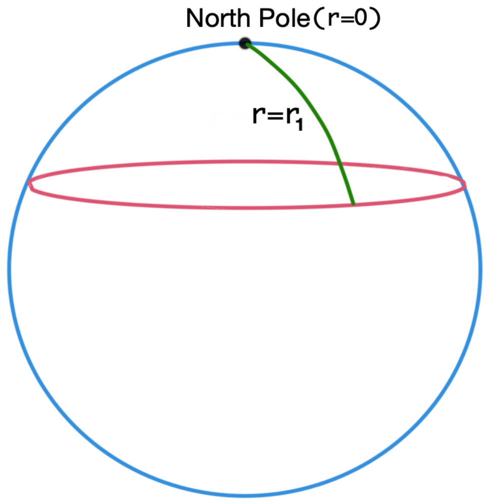

Figure 4. Circle on a sphere. Its origin is the North Pole and radius is r_1

Circumference of a circle in a plane is

$$C = \int_0^{2\pi} r d\theta = 2\pi r. \tag{1.21}$$

Circumference-to-radius ratio on a plane is

$$C/r = 2\pi.$$

On a sphere, let us draw a circle of radius r_1 around the North Pole ($r = 0$). Its radius is the distance of the points from North pole which is $r = r_1$. This is shown in Figure 4.

Circumference $= \int ds$ where ds is given by (1.18) with $r = r_1$. So we get:

$$\int dS = \int \sin r_1 d\phi = 2\pi \sin r_1 \tag{1.22}$$

Then circumference/radius ratio is $\frac{2\pi \sin r}{r}$, which is different from 2π.

Notice that when r is very small, then we have:

$$\frac{2\pi \sin r}{r} \approx 2\pi. \tag{1.23}$$

So for a very small circle we get the same circumference/radius ratio as that of a plane, as expected.

In both the examples we used the infinitesimal interval. In fact, the infinitesimal inerval contains all the information needed about the geometrical properties of a space. We can deduce everything from it.

We are used to describing spheres and other 2-dimensional curved spaces in terms of their embedding in three dimensions. But we saw here that the properties of the 2D sphere can be described by the 2D interval (1.18) without referring to its embedding in higher dimensions.

This is important because our 4-dimensional spacetime is curved, but there is no evidence that it is embedded in a higher dimensional spacetime. So it is important to be able to describe curved spaces using intrinsic properties like the interval.

1.4.1 Invariance of the Interval

When dealing with any complex physics problem, we make our life easy by choosing a coordinate system. However, any physical quantity is independent of our choice of coordinate system. In the last section, we described the interval dS^2 as an intrinsic geometrical property of the space. This too must be independent of the coordinate system.

This is quite obvious, because dS^2 is the distance between 2 points–it does not know about what coordinate system we have employed.

To see this explicitly, let us consider coordinate systems related by rotations in two dimensions. After a rotation by angle θ, the new coordinates (x', y') are related to the old ones (x, y) by:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{1.24}$$

It is easy to check from (1.24) that

$$dx^2 + dy^2 = {x'}^2 + {y'}^2 \tag{1.25}$$

Quantities that are unchanged by a coordinate system are called **invariants**.

Our description of physics must be in terms of such geometric objects, which are independent of coordinates.

1.4.2 Physical vs Coordinate distance

This is a good time to drive home the point that physical distance and coordinate distance are different.

This is obvious from polar coordinates: the coordinate distance between two points on a plane maybe $(dr, d\theta)$ but the physical distance is not $dr^2 + d\theta^2$, it is $dr^2 + r^2 d\theta^2$. For two points on a sphere, the same coordinate distance corresponds to a different physical distance $dr^2 + \sin^2 r d\theta^2$.

We will see many examples of the difference between coordinate distance and physical distance in this course. The physical distance is the one which is invariant under coordinate changes. What we actually measure is physical distance, never coordinate distance.

An analogy to keep in mind is maps vs reality. Coordinate distances are like distances in maps. They have to be multiplied by some scale factor to get the real physical distances, which can be measured.

1.5 Vectors and Scalars

Vectors: Another example of geometric objects are vectors. Take a moving car. Its velocity does not depend on how we choose our x and $y$axes. Nor does the electric field of a charge, or the magnetic field of a magnet.

But vectors are different from an invariant like distance, because they have a direction. So when we use a coordinate system, we describe vectorsin terms of their components. When we make a change of coordinates, the components of a vector also change.

If we are given the components of a vector in two different coordinate systems, how would we figure if they are the components of the same vector.

To answer this, consider the displacement vector in 2D. In a given coordinate system (x, y) the dispacement vector is defined by its components $(\Delta x,\ \Delta y)$. Then if I consider a different coordinate system rotated by an angle θ, the transformation of the components of the dispacement vector follows from (1.24):

$$\begin{pmatrix} \Delta x' \\ \Delta y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \tag{1.26}$$

Explicitly, they are given by:

$$\Delta x' = \cos\theta \Delta x + \sin\theta \Delta y \tag{1.27}$$

$$\Delta y' = -\sin\theta \Delta x + \cos\theta \Delta y \tag{1.28}$$

(Aside: Note that the square of the magnitude of an infinitesimal displacement vector is the infinitesimal interval $dS^2 : |\vec{dr}|^2 = \vec{dr} \cdot \vec{dr} = dx^2 + dy^2 = dS^2$)

A general vector A in two dimensions (which could be an electric field, a direction of wind etc) has the same transformation law:

$$\begin{pmatrix} A_{x'} \\ A_{y'} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} A_x \\ A_y \end{pmatrix} \tag{1.29}$$

This is how we can answer our question above–components of a vector in two different coordinate systems would be related to one another by a transformation law like (1.29).

(1.29) is a key property of vectors that will be extremely useful to us. We will see that this property generalises to tensors, which are also coordinate independent, geometric objects.

One can even define vectors using this property as follows: a n-component vector(under rotations) is an n-component object whose compotents transform under rotations like the components of a displacement vector in n-dimensions.

This definition is more useful for us than the ones we see in school (like magnitude + direction, or triangle law) . It gives us a way to define tensors later. However, we will provide a more useful definition later.

Note that what (1.29) defines are vectors under rotation. Later we will encounter quantities that behave like vectors under more general coordinate transformations.

Scalars: Another kind of coordinate independent object are scalars. You would know scalars as quantities that have a magnitude and no direction, like temperature. Obviously the value of temperature at a point is independent of what coordinates we use. Therefore, the new definition of a scalar is as an object whose value is unchanged by a coordinate transformation.

1.5.1 Vector indices and Einstein summation convention

In (x, y, z) coordinates, it is usual to write the components of a vector as $\vec{A} \equiv (A^x, A^y, A^z)$. In relativistic physics, a different convention is typically used where the indices are denoted by numbers: $(A^x, A^y, A^z) = (A^0, A^1, A^2)$. Coordinates x, y, z are likewise written as x^1, x^2, x^3. When it is not specified which particular component (i.e x or y or z) we are referring to, we will denote the vector component by A^i, where i can take any value between 1-3.

The unit vectors (which you encountered as i, j, k) will be denoted by e_i, where $i$again takes values between 1-3 . Note the index is downstairs. The full vector is then

$$\vec{A} = \sum_i A^i e_i = A^1 e_1 + A^2 e_2 + A^3 e_3. \tag{1.30}$$

We use the convention of writing the indices of vector components 'upstairs' and for unit vectors, 'downstairs.' The usefulness for this convention will become clear when we introduce the Einstein summation convention, which we now introduce.

Consider this equation where a square matrix C multiplies a row matrix D:

$$\mathbf{A} = \mathbf{CD}$$

In index notation, we write:

$$A^i = \sum_j C^{ij} D^j$$

It is pretty obvious which index is being summed over, it is the repeated index. So we could just get rid of the summation sign:

$$A^ß = C^i_j D^j$$

and simply remember the convention: the repeated index is being summed over.

To make this even easier to remember, we can lower the column index:

$$A^ß = \sum_j C^i_j D^j$$

A easy way to remember this is to imagine that the same index upstairs and downstairs should 'cancel.' The index that is not cancelled (in this case i) should match on both sides.

This is the **Einstein summation convention**. It will be extremely useful when we deal with multiple indices. We will follow it throughout the rest of the book, unless otherwise specified.

Consider the transformation of a 2-dimensional vector under rotation:

$$\begin{pmatrix} A^{1'} \\ A^{2'} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} A^1 \\ A^2 \end{pmatrix} \tag{1.31}$$

If we denote the rotation matrix components by R^i_j, the above can be written using Einstein summation convention as:

$$A^{i'} = R^{i'}_i A^i \tag{1.32}$$

The expression for a vector (1.30) can be written in this convention as:

$$\vec{A} = A^i e_i \tag{1.33}$$

Now let us write the dot product between two vectors in index notation:

$$\vec{A} \cdot \vec{B} = (A^i e_i) \cdot (B^j e_j) = (A^i B^j)(e_i \cdot e_j) \tag{1.34}$$

We chose orthonormal basis vectors, so we have:

$$e_i \cdot e_j = \delta_{ij} \tag{1.35}$$

This is the Kronecker delta function. This means that the LHS is 1 when we take dot product between the same two basis vectors and zero otherwise.

So we get:

$$\vec{A} \cdot \vec{B} = (A^i B^j)\delta_{ij} \tag{1.36}$$

Particularly, the invariant distance $dS^2 = |d\vec{r}|^2$ becomes:

$$dS^2 = \delta_{ij} dx^i dx^j \tag{1.37}$$

This may seem like a fancy way of writing something simple, but we will see the importance