PART I

GENERAL RELATIVITY

1 Introduction

This is a book about gravity. Of the four fundamental interactions (strong, weak, electromagnetic, and gravitational), gravity is by far the weakest, characterized by a force that is intrinsically $\sim 10^{36}$ times more feeble than the electromagnetic force. Yet gravity determines *completely* the large-scale structure of the Universe. How can this be? In this introductory chapter we will look qualitatively at how gravity sets itself apart from all other fundamental interactions, how it can be best described in mathematical terms, and how Einstein's theory of general relativity revised its fundamental meaning and interpretation.

1.1 Gravity and the Universe on Large Scales

Gravity is intrinsically weak but it has some properties that distinguish it from all the other fundamental interactions.

1. *Gravity is long-ranged.* It is one of only two fundamental interactions that are long-ranged, the other being electromagnetism, with the gravitational force and the electromagnetic force each varying as the inverse square of the distance to the source of the corresponding field. In contrast, the strong and weak interactions act only over distances comparable to the size of a nucleus, a very short range indeed! It follows that the strong and weak forces are fundamental in determining the microscopic properties of matter but they have no direct bearing on the large-scale structure of the Universe. The race to determine that structure is now down to electromagnetism and gravity with, in the language of *The Tortoise and the Hare* from *Aesop's Fables,* the sleek, fast rabbit of electromagnetism pitted against the plodding, methodical tortoise of gravity (with the rabbit sporting a top speed 10^{36} times that of the tortoise). Surely only a fool would bet against the rabbit. But wait; I haven't told you everything yet!

2. *Gravity is unscreened.* Electrical charges can be positive or negative. Thus although in principle electromagnetic forces are long-ranged, in practice they tend to be short-ranged because positive and negative charges partially offset each other at shorter range and completely cancel each other at longer range; this is *screening,* and it implies that matter on larger scales (moons, planets, stars, galaxies, . . .) may under normal conditions be assumed *completely electrically neutral.* In contrast, a comparison of the equations for Newtonian gravity and for electrostatics indicates that *mass* is the gravitational "charge," but mass

has only one sign so the gravitational interaction is *unscreened* and *always attractive.*[1] An electron on the Moon feels no electrical force from a proton on the Earth because the force is completely screened by intervening matter. In contrast, the electron on the Moon feels the full gravitational force exerted by that same proton on the Earth because there is no screening of the gravitational force, even if there is intervening matter. Advantage tortoise!

3. *Gravity is universal,* acting between all masses (and energy, since $E = mc^2$) with the same attractive sign. This is fundamentally different from electromagnetism, where the Coulomb interaction between two objects depends on their charges, which can be positive, negative, or zero (even for unscreened matter). Advantage tortoise!

Because of points 1–3, the plodding tortoise carrying the banner of gravity easily wins the race to determine the large-scale structure of the Universe over the swift electromagnetic hare. The reason is the same reason that the tortoise wins in the original *Tortoise and Hare* fable: the relentless pursuit of a singular goal. Gravity can do only one thing, but it does it tirelessly and methodically.

On the other hand, the extreme weakness of gravity means that it can be neglected completely for the microscopic structure of matter: that of molecules, atoms, or nuclei. The lone caveat to this statement is that on incredibly short distance scales (many, many orders of magnitude below present measurement capabilities) gravity can become strong enough that it cannot be ignored in considering the quantum structure of matter. This *Planck scale* is the regime of *quantum gravity,* for which we do not yet have an adequate theory and are reduced to speculation and analogy.

1.2 Classical Newtonian Gravity

Having established the dominance of gravity in determining how the Universe operates on all but the shortest distance scales, it is of importance to ask how gravity can best be described in mathematical terms. A quite serviceable option has been available for three centuries. *Newtonian gravity* works remarkably well for just about everything. It describes the motion of rocks thrown at the Earth's surface and the orbits of the planets and moons and asteroids of the Solar System with almost arbitrary precision, and NASA engineers with confidence send astronauts to the Moon and back, and spacecraft to a precise rendezvous with bodies in the far reaches of the Solar System, based on its prowess. These are remarkable technical achievements, so why would anyone want anything better?

The basic answer is that the motivation and successful quest for a better theory of gravity grew from the remarkable physical intuition and work of one person, Albert Einstein, in the early years of the twentieth century. That better theory of gravity is called *general relativity.* The development of general relativity was different from the development of almost any other new scientific theory in two regards: (1) As just suggested, it was very much the work

[1] The discussion in this Introduction assumes the gravity of everyday experience. Later it will be shown that on cosmological scales it is possible for gravity to become effectively repulsive. But that is a story for later that has no bearing on daily life in our little corner of the Universe.

of a single person, unlike most scientific breakthroughs, which involve the direct work of more than one person "standing on the shoulders of giants" (in the words of Newton) who had paved the way before them. (2) Unlike for many paradigm shifts in science, there was no crying need for general relativity brought about by new experiments or observations (quite different from, say, quantum mechanics, which arose also in the early part of the twentieth century in response to a crucial need to understand what measurements in the new field of atomic physics implied about the structure of atoms).

For the theory of gravity it may fairly be argued that at the beginning of the twentieth century there was but a single fly in the ointment of Newtonian gravity, and it was an extremely tiny and arcane fly: the measured orbit of the planet Mercury showed a discrepancy with the predictions of Newtonian gravity in a certain measured angle called the *perihelion shift* that corresponded to a difference of 43 arcseconds per century.[2] You read correctly, *per century!* While this was a puzzling anomaly, one can imagine that very few scientists of the time lost sleep over this discrepancy in the perihelion advance of Mercury, and even fewer would have guessed that the resolution of this tiny anomaly would entail a seismic shift in our understanding of gravity and the nature of space and time.

The precession of the perihelion of Mercury was the first problem to which Einstein applied his new theory of general relativity, and Einstein himself said that he was so overcome with joy when he found that his new theory predicted exactly 43 arcseconds per century of precession over that of Newtonian theory that for several days he could hardly function and experienced heart palpitations [178]. However, the resolution of this problem in Mercury's orbit was *not* Einstein's motivation for developing general relativity. Instead, it seems that Einstein was motivated by more abstract reasoning to develop a new theory of gravity, only later applying the new theory to practical problems like Mercury's orbit. To understand this reasoning it is necessary to first consider the *special theory of relativity,* and to do that we must address the effect of transformations between coordinate systems on the laws of physics.

1.3 Transformations between Inertial Systems

In 1905 Einstein published the *special theory of relativity*, which revolutionized our understanding of space and time with its concepts of space contraction and time dilation, and that the simultaneity of two events was not an absolute thing but rather depended on the relative velocity of the observer. The motivation for the special theory was the central conviction of Einstein that the laws of physics cannot depend on the observer (the *principle of relativity*), meaning that the laws of physics must not depend on the coordinate system in which they are formulated, and that the transformations between inertial frames (coordinate systems not accelerated with respect to each other in which Newton's first law is valid) should be the same for particles and for light.

[2] An arcsecond is 1/3600 of an angular degree.

Newtonian mechanics already contained a principle of relativity for space (but not for time), in that the Newtonian laws of mechanics were unchanged by a transformation between inertial frames called a *Galilean transformation*. For motion along the x axis a Galilean transformation takes the form

$$x' = x - vt \qquad y' = y \qquad z' = z \qquad t' = t, \tag{1.1}$$

where primed coordinates and unprimed coordinates represent the two different inertial frames, the velocity in the x direction is v, and a single universal time $t = t'$ has been assumed for all observers. This is just the "common sense" notion that if you are moving in the x direction at constant speed on a railroad flatcar and throw a ball forward with some velocity as measured from the flatcar, the velocity of the ball in the x direction measured by an observer on the ground is the sum of velocities for the train and the ball relative to the train, the transverse y and z directions are unaffected, and the time t' measured on the train and the time t measured on the ground are the same. Obvious, right? And it *is obvious* for balls and trains moving at relatively low velocities, as has been confirmed by many experiments. But Einstein (and others) realized that there is a problem in that these common sense notions of relative motion between inertial frames were inconsistent with the theory of light, which was well understood in 1905 to be an electromagnetic wave described by the Maxwell equations (first published by James Clerk Maxwell in 1861, but put in their more modern form in the 1880s by Oliver Heaviside).

1.4 Maxwell, the Aether, and Galileo

One of the revolutionary features of the Maxwell theory was that wave disturbances could propagate in the electromagnetic field, and these waves traveled with a speed that was a constant of the theory and thus independent of inertial frames. When the constant was evaluated it was found to be equal numerically to the measured speed of light, which caused the Maxwell electromagnetic waves to be identified with light. The beauty of Maxwell's equations greatly impressed Einstein, but they presented a problem of interpretation for classical physics.

By the Galilean transformations, which worked well for mechanical problems, the speed of light should certainly depend upon the frame from which it was measured; but by Maxwell's equations it should not because it is a *constant* of the theory. Since the Maxwell equations and the Galilean transformations were valid in their respective domains, it was desirable to keep both. The standard interpretation that emerged to permit this was that there must be a medium through which the electromagnetic waves moved (Obvious, right? A wave can't just travel through nothing, can it?) This medium was called the *luminiferous aether* or *aether* for short, and it was assumed to be an invisible, rigid (because light waves are transverse and transverse waves don't propagate through fluids) substance permeating all of space but relevant only for light propagation. Then it was proposed that the constant speed of light was an artifact of the special aether rest frame in which light propagated.

The aether of course is fictitious and it is now well understood that light waves are propagating disturbances in electric and magnetic fields that do not require a physical medium, but in the latter part of the nineteenth century the aether was widely believed to be real and various attempts were made to detect the motion of the Earth relative to the aether.[3] The definitive experiments were those of Michelson and Morley, who showed in 1887 using light interferometry that there was no evidence for a different drift of the Earth with respect to the hypothetical aether when it was moving in different directions on its orbit around the Sun, thus casting serious doubt on the existence of the aether.[4]

1.5 The Special Theory of Relativity

With the aether discounted, the Maxwell equations and their constant speed of light were clearly incompatible with the Galilean invariance exhibited by material particles. Others had proposed hints of a solution (most notably Hendrik Lorentz and Henri Poincaré) but it was Einstein who bridged this impasse with the bold hypothesis that the Maxwell theory was correct and that it was the *Galilean transformations* that needed modification to bring them into accord with the Maxwell equations. Thus he proposed that the speed of light was constant for all observers, with no qualifications. This, along with the assumption of relativity (physical law does not depend on the coordinate system) yielded in 1905 the special theory of relativity. The special theory assumes the existence of global inertial frames, so it could not be applied to gravity, which is incompatible with the existence of global inertial frames because it is associated with curved spacetime.

In the special theory (of relativity), the requirement that the speed of light c be an invariant in all inertial frames necessitated replacement of the Galilean transformations with the *Lorentz transformations,*

$$x' = \gamma(x - vt) \qquad y' = y \qquad z' = z \qquad t' = \gamma\left(t - \frac{vx}{c^2}\right) \qquad \gamma \equiv \frac{1}{\sqrt{1 - v^2/c^2}}, \tag{1.2}$$

[3] This was called the *aether drift.* Einstein himself apparently was interested in constructing an experiment to measure the aether drift as a student. This never came to fruition because of lack of funds and equipment, and the opposition of his teachers, and in retrospect his methods likely would not have worked, even if the aether existed. It is not clear when Einstein abandoned the idea of the aether, but it was certainly before the publication of the special theory of relativity in 1905 [178].

[4] Nevertheless, efforts persisted for decades to salvage the aether hypothesis. For example, one idea was that a piece of the aether had been somehow trapped in the basement laboratory in which the Michelson–Morley experiment was carried out, thus explaining the null result. The reader will not be surprised to learn that experiments carried out at other locations also found no evidence for the aether [178]. It is not clear to what degree Einstein was influenced by the results of the Michelson–Morley experiment. Einstein claimed at various times that it had little effect on his reasoning. Scientific journals were not nearly as widely available then as they are today, and there is evidence that in his earlier years Einstein sometimes did not have access to important scientific papers. But it seems likely that Einstein knew of the Michelson–Morley results. Perhaps the most consistent interpretation is that Einstein knew of the null aether results but felt that this was not as important as his own reasoning in coming to the special theory of relativity. For example, Einstein placed special emphasis on the role of thought experiments that he began carrying out as a student concerning whether one could travel fast enough to catch up with a light wave, and what the observational consequences would be.

where γ is called the *Lorentz γ-factor*. Notice that in the limit $v/c \to 0$ the factor $\gamma \to 1$ and the Lorentz transformations (1.2) become equivalent to the Galilean transformations (1.1), so the Galilean transformations are quite correct in the low-velocity world of our normal experience. Notice also that, unlike in the Galilean transformation where there is a universal time shared by observers, time transforms non-trivially under Lorentz transformations with the consequence that the Lorentz transformations mix the space and time coordinates.

The mathematician Hermann Minkowski (who was once Einstein's teacher at what is now ETH Zurich) then noted that it is most natural to abandon separate notions of space and time and instead view special relativity in terms of a *4-dimensional spacetime* parameterized by *spacetime coordinates*

$$(x^0, x^1, x^2, x^3) \equiv (ct, x, y, z), \tag{1.3}$$

where the superscripts are indices (not exponents).[5] In a 1908 presentation entitled *Raum und Zeit* (*Space and Time*) that was delivered to the *80th Assembly of German Natural Scientists and Physicians* in Cologne, Minkowski introduced the idea of 4-dimensional spacetime using a phrasing that has now become legendary:[6]

> *The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.* (Hermann Minkowski (1908))

Now time (more precisely time scaled by the speed of light, so that it has the same units as the other three coordinates) becomes just another coordinate in the 4-dimensional spacetime. As Minkowski noted, special relativity is simple when viewed in 4-dimensional spacetime, but becomes more complicated when projected onto 3-dimensional space.

Minkowski also introduced the tensor formalism for special relativity (Einstein's 1905 paper did not use tensors), and introduced terminology such as *worldline* that is now standard. Einstein at first viewed Minkowski's formulation of special relativity using tensors as just a mathematical trick, but soon realized the power of these methods and adopted many of them in his later formulation of the general theory of relativity. Minkowski undoubtedly would have made further contributions to the development of relativity but he died unexpectedly of peritonitis only months after his famous Space and Time lecture.

1.6 Minkowski Space

The surface traced out by allowing the coordinates (x^0, x^1, x^2, x^3) to range over all their possible values defines the manifold of 4-dimensional spacetime. The resulting space is

[5] The reason that the indices are in an upper position will be made clear in due time; for now just view them as labels with a possibly eccentric placement, and don't confuse them with exponents.

[6] An English translation of the full presentation may be found at https://en.wikisource.org/wiki/Translation:Space_and_Time.

commonly called *Minkowski spacetime,* which is often shortened to *Minkowski space* or just *spacetime.* In Minkowski space the square of the infinitesimal distance ds^2 between two points (ct, x, y, z) and $(ct + cdt, x + dx, y + dy, z + dz)$ is given by

$$ds^2 = \sum_{\mu\nu} \eta_{\mu\nu} dx^\mu dx^\nu = -c^2dt^2 + dx^2 + dy^2 + dz^2,$$
$$= -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2, \tag{1.4}$$

which is called the *line element* of the Minkowski space.[7] The quantity $\eta_{\mu\nu}$ may be expressed as the diagonal matrix

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{1.5}$$

and is termed the *metric tensor* of the Minkowski space. The line element (1.4) or the metric tensor $\eta_{\nu\mu}$ determine the geometry of Minkowski space because they specify distances, distances can be used to define angles, and that is geometry. The pattern of signs on the right side of Eq. (1.4) defines the *signature of the metric.* For Minkowski space the signature is $(- + ++)$.[8]

The geometry of 4-dimensional Minkowski space differs from that of 4-dimensional euclidean space, so 4-dimensional Minkowski spacetime is *not* "just like ordinary space but with more dimensions." The difference is encoded in the signature of the metric, which for 4-dimensional euclidean space is $(+ + ++)$, compared with the signature $(- + ++)$ for the Minkowski metric. (That is, the metric tensor of 4-dimensional euclidean space is just the 4×4 unit matrix.) That change in sign for the first entry makes all the difference. Most of the unusual features of special relativity (space contraction, time dilation, relativity of simultaneity, the twin "paradox," ...) follow from this difference in geometry between 4-dimensional Minkowski spacetime and 4-dimensional euclidean space.

1.7 A New Theory of Gravity

Now that the transformation laws between inertial systems for both light and mechanical particles had been unified in special relativity, Einstein turned his attention to how this

[7] In this notation ds^2 means the square of ds [that is, $(ds)^2$], and dx^2 means $(dx)^2$, but in (x^0, x^1, x^2, x^3) the superscripts are indices and not powers. This is standard notation and you quickly will learn to distinguish whether a superscript is meant as an index or an exponent from the context. If there is potential ambiguity, use parentheses to make the intention clear, as in the second line of Eq. (1.4).

[8] The pattern of $+$ and -1 signs is referred to here as the signature. Some authors define the signature to be an integer that is the difference of the number of $+$ and $-$ signs. The two definitions convey similar information. A metric such as that of Minkowski space in which the signs in the sign pattern are not all the same is termed an *indefinite metric*. Some authors use instead the signature $(+ - --)$ for the sign pattern in Eq. (1.4). This leads to the same physical results as our choice as long as all signs are carried through consistently with either choice. The important point is that for Minkowski spacetime the last three terms have the same sign and the sign of the first term is different from that of the other three (provided that the usual modern convention of displaying the timelike coordinate in the first position and the spacelike coordinates in the last three positions is employed).

"special" relativity, which applied only to the restricted case of flat spacetime in which global inertial frames were valid, could be generalized to apply the same principles to the gravitational problem. This was a much more difficult task, so much so that it took Einstein almost a decade to solve it. (During this decade Einstein also made fundamental contributions to quantum mechanics and statistical mechanics, but that is not relevant for the present discussion.)

What Einstein sought was a new gravitational theory that would remove the inconsistencies of Newtonian gravity with respect to the principles of special relativity, but still recover the documented success of Newtonian gravity in a suitable limit. Newtonian gravity and special relativity are at odds in several respects. The most important are that Newtonian theory ascribes a physical reality to space and time coordinates, it treats space and time in an asymmetric way, and it implies that the gravitational force exerted by one mass on another is felt instantaneously by the second mass, no matter how large the distance between them. The first is inconsistent with the Einstein principle of relativity, the second is inconsistent with the Lorentz transformations, and the third is inconsistent with the constant (finite) speed of light in special relativity, which implies that the gravitational interaction cannot act instantaneously across space (no "action at a distance"). Einstein tried various approaches to the generalization of special relativity to incorporate gravity without much success until in late 1907 he hit upon the idea that would eventually lead to general relativity, though it was not until 1915 that he was able to elaborate the idea mathematically into a complete theory.

1.8 The Equivalence Principle

The starting point for this new gravitational theory is that the universality of gravity alluded to above is even more curious than just the generic statement that gravity differs from all other forces in that it acts attractively on all matter. It has been known since the days of Galileo – before Newton was even born – that objects of different mass and/or different composition *fall at the same rate in Earth's gravitational field* (neglecting the effect of friction with the air, of course). This may be stated somewhat more esoterically in terms of the *weak equivalence principle:* the *gravitational mass* of an object (the mass determined from Newton's law of gravity by observing its interaction with a gravitational field) and its *inertial mass* (the mass determined from Newton's second law of motion by pushing the object) are to high experimental precision *equivalent*.[9]

This is at first glance surprising, since there is no a priori reason to expect the two definitions of mass to coincide. In Newtonian theory the equivalence of gravitational and inertial mass is an interesting but unexplained coincidence that mostly gets ignored. For Einstein it became the key to understanding the true nature of gravity and thence to the formulation of general relativity. Einstein realized (what is in retrospect) the obvious

[9] More precisely they are proportional, but with a suitable choice of units the constant of proportionality can be chosen to be one.

implication of the gravitational acceleration being independent of any specific property of the mass being accelerated:

> If gravity acts universally on all mass, irrespective of its specific characteristics, then the gravitational force cannot be a property of the masses themselves and therefore must be a universal property of the spacetime in which gravity acts.

Specifically, Einstein realized that if he were in free fall in a gravitational field he would not be able to feel his own weight, so in a small freely falling reference frame the effects of gravity may be transformed away.[10] This led to the realization that (in a small region of spacetime) there was no operational way to distinguish an arbitrary acceleration from the effects of gravity, and this set of ideas came to be called the *(strong) equivalence principle.* By various thought experiments using the equivalence principle it became apparent to Einstein that gravity was associated with the *geometry of spacetime,* specifically through its *curvature:* In the absence of gravity spacetime is flat (Minkowski space); in the presence of gravity, spacetime becomes curved. Using these ideas Einstein was able to find some essential features of general relativity such as the gravitational deflection and redshift of light but the field equations describing the full effects of general relativity required substantial additional mathematical development and were revealed by Einstein for the first time only in late 1915, in a presentation to the Prussian Academy of Sciences.[11]

1.9 General Relativity

The theory of general relativity published by Einstein beginning in 1915 represents a radical new view of space, time, and gravity relative to our "common sense" intuition. It supersedes Newtonian mechanics and Newtonian gravity, but reduces to those theories in the limit of velocities that are small with respect to the speed of light and gravitational fields that are weak (with respect to criteria that will be specified later). It reduces to the theory of special relativity in the limit that gravity vanishes or, in a sense that will be specified more precisely later, for sufficiently local regions of spacetime even in the presence of strong gravitational fields.

General relativity revises fundamentally the very meaning of space, time, and gravity because the effects of gravity no longer appear as a force but as the motion of *free*

[10] This is presumably far more obvious to children of the space age accustomed to seeing weightless astronauts in orbital free fall on television than it would have been to Einstein in the early 1900s.

[11] It is sometimes claimed that the mathematician David Hilbert published the field equations for general relativity shortly before Einstein's presentation to the Prussian Academy, and thus should be given full or joint credit for the theory of general relativity. Historical research indicates that Hilbert and Einstein were in a race to complete the field equations, but that the version published by Hilbert before Einstein's presentation was later modified to be consistent with Einstein's correct version of the field equations. At any rate, Hilbert's contribution to the mathematics of general relativity was substantial but he had been motivated to work on the theory by lectures that Einstein gave in Göttingen, and Hilbert gave full credit to Einstein as the author of general relativity. Although Hilbert was a far better mathematician than Einstein, he understood that it was Einstein's deep physical intuition that was most essential to the basic formulation of general relativity.