Contents

5 The Schwarzschild solution

5.1 The Schwarzschild metric

5.1.1 Deriving the metric

A large number of exact solutions of the Einstein equation have been found over the years. The first to be discovered (apart from the trivial Minkowski space) is the **Schwarzschild solution**, derived by Karl Schwarzschild in December 1915 and published in January 1916, only a month and two months after the November 1915 publication of the Einstein equation by Einstein and the Einstein–Hilbert action by Hilbert. The Schwarzschild solution is the general spherically symmetric vacuum solution of GR. It is central in the study of the gravity of astrophysical objects such as planets, stars, and black holes.

Let us derive the solution. If a (four-dimensional spacetime) metric is spherically symmetric, there exists a two-dimensional hypersurface that is conformal to the unit two-sphere and orthogonal to the other two directions. So, adopting coordinates (t, r, θ, φ), we can write the metric as (recall that $g_{\alpha\beta} = \underset{\sim}{e}_\alpha \cdot \underset{\sim}{e}_\beta$)

$$\mathrm{d}s^2 = -A(t,r)\mathrm{d}t^2 + B(t,r)\mathrm{d}r^2 + 2C(t,r)\mathrm{d}t\mathrm{d}r + e^{2\delta(t,r)}(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\varphi^2) \ , \quad (5.1)$$

where $0 < \theta < \pi$, $0 \leq \varphi < 2\pi$ as usual for the two-sphere. The area of the two-sphere is $S = 4\pi e^{2\delta(t,r)}$. We now choose a new radial coordinate $\tilde{r}$ defined so that the area of the two-sphere is $4\pi\tilde{r}^2$, i.e. $\tilde{r} \equiv \sqrt{S/(4\pi)} = e^{\delta(t,r)}$. In three-dimensional Euclidean space, it is a result that the area of a two-sphere is $4\pi r^2$, where r is the proper distance from the centre to the sphere. Here we have only made a coordinate choice, and we have not yet determined the relation of $\tilde{r}$ to the proper distance. We now drop the ˜ and simply denote the new radial coordinate by r.

Having fixed the radial coordinate, we can still change the time coordinate to simplify the metric. Under the transformation $t \to \tilde{t}(t,r)$, we have

$$\mathrm{d}t \quad \to \quad \mathrm{d}\tilde{t} = \dot{\tilde{t}}\mathrm{d}t + \tilde{t}'\mathrm{d}r \ , \tag{5.2}$$

where dot denotes ∂_t and prime denotes ∂_r. The metric (5.1) becomes

$$\begin{aligned} \mathrm{d}s^2 &= -A\dot{\tilde{t}}^2\mathrm{d}t^2 + (B - A\tilde{t}'^2 + 2C\tilde{t}')\mathrm{d}r^2 + 2\dot{\tilde{t}}(C - A\tilde{t}')\mathrm{d}t\mathrm{d}r \\ &\quad + r^2(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\varphi^2) \ . \end{aligned} \tag{5.3}$$

The function $\tilde{t}(t,r)$ is so far arbitrary. We now demand that $g_{tr} = 0$, i.e. choose $\tilde{t}$ in such a way that $\tilde{t}' = C/A$. The only coordinate freedom left in the function $\tilde{t}$ is then $\tilde{t} \to \tilde{t} + D(\tilde{t})$, where $D(\tilde{t})$ is an arbitrary function. This corresponds to the freedom to redefine the time coordinate as function of itself. Dropping the $\tilde{}$, we can now write the metric as (because the metric is diagonal, we have $g_{tt} < 0$ and $g_{rr} > 0$)

$$\mathrm{d}s^2 = -e^{2\alpha(t,r)}\mathrm{d}t^2 + e^{2\beta(t,r)}\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\varphi^2) \ . \tag{5.4}$$

We have exhausted the freedom to fix components of the metric using coordinate transformations. The functions $\alpha(t,r)$ and $\beta(t,r)$ are determined by the equation of motion.

The non-zero connection coefficients corresponding to the metric (5.4) are

$$\Gamma^0_{00} = \dot{\alpha} \qquad \Gamma^0_{01} = \alpha' \qquad \Gamma^0_{11} = e^{-2(\alpha-\beta)}\dot{\beta} \tag{5.5}$$

$$\Gamma^1_{00} = e^{2(\alpha-\beta)}\alpha' \qquad \Gamma^1_{01} = \dot{\beta} \qquad \Gamma^1_{11} = \beta' \tag{5.6}$$

$$\Gamma^1_{22} = -re^{-2\beta} \qquad \Gamma^1_{33} = -re^{-2\beta}\sin^2\theta \tag{5.7}$$

$$\Gamma^2_{12} = \frac{1}{r} \qquad \Gamma^2_{33} = -\sin\theta\cos\theta \tag{5.8}$$

$$\Gamma^3_{13} = \frac{1}{r} \qquad \Gamma^3_{23} = \frac{\cos\theta}{\sin\theta} \ . \tag{5.9}$$

The corresponding non-zero components of the Riemann tensor are

$$R^0{}_{101} = (\ddot{\beta} + \dot{\beta}^2 - \dot{\alpha}\dot{\beta})e^{-2(\alpha-\beta)} - \alpha'' - \alpha'^2 + \alpha'\beta' \tag{5.10}$$

$$R^0{}_{202} = -r\alpha' e^{-2\beta} \tag{5.11}$$

$$R^0{}_{303} = \sin^2\theta R^0{}_{202} \tag{5.12}$$

$$R^0{}_{212} = -r\dot{\beta}e^{-2\alpha} \tag{5.13}$$

$$R^0{}_{313} = \sin^2\theta R^0{}_{212} \tag{5.14}$$

$$R^1{}_{212} = r\beta' e^{-2\beta} \tag{5.15}$$

$$R^1{}_{313} = \sin^2\theta R^1{}_{212} \tag{5.16}$$

$$R^3{}_{323} = \sin^2\theta(1 - e^{-2\beta}) \ , \tag{5.17}$$

so the non-zero components of the Ricci tensor are

$$R^0{}_0 = (\ddot\beta + \dot\beta^2 - \dot\alpha\dot\beta)e^{-2\alpha} - \left(\alpha'' + \alpha'^2 - \alpha'\beta' + \frac{2}{r}\alpha'\right)e^{-2\beta} \quad (5.18)$$

$$R^0{}_1 = -\frac{2\dot\beta}{r}e^{-2\alpha} \quad (5.19)$$

$$R^1{}_1 = (\ddot\beta + \dot\beta^2 - \dot\alpha\dot\beta)e^{-2\alpha} - \left(\alpha'' + \alpha'^2 - \alpha'\beta' - \frac{2}{r}\beta'\right)e^{-2\beta} \quad (5.20)$$

$$R^2{}_2 = R^3{}_3 = \left(-\frac{1}{r}\alpha' + \frac{1}{r}\beta' - \frac{1}{r^2}\right)e^{-2\beta} + \frac{1}{r^2} \ . \quad (5.21)$$

The Einstein equation in vacuum is $G_{\alpha\beta} = 0$, which is equivalent to $R_{\alpha\beta} = 0$. (We take the cosmological constant to be zero. We will later come back to the case when it is nonzero.) As we noted in the chapter 4, not all components of the Einstein equation are independent as far as the functional degrees of freedom are concerned, as they are related by the contracted Bianchi identity $\nabla^\alpha G_{\alpha\beta} = 0$. (They are independent if we count the number of degrees of freedom at a point.)

To start with, equating the component (5.19) to zero gives $\beta = \beta(r)$. Equating the components (5.18) and (5.20) to zero and subtracting them from each other gives $\alpha' = -\beta'$, i.e. $\alpha(t,r) = -\beta(r) + E(t)$, where $E(t)$ is an arbitrary function. We now redefine the time coordinate as $\mathrm{d}\tilde t = e^E \mathrm{d}t$ and drop the tilde. (In other words, we put $E = 0$ without loss of generality.) The only thing left to determine is the function $\beta(r)$. It's easiest to solve $\beta(r)$ from the component (5.21), because it involves only first derivatives. Equating it to zero gives

$$2\beta' r = 1 - e^{2\beta} \ , \quad (5.22)$$

so we get

$$\begin{aligned} \int \frac{\mathrm{d}r}{r} &= 2\int \frac{\mathrm{d}\beta}{1 - e^{2\beta}} \\ &= \int \frac{\mathrm{d}x}{x(1-x)} \\ &= \int \mathrm{d}x \left(\frac{1}{x} - \frac{1}{x-1}\right) \ , \end{aligned} \quad (5.23)$$

where we have used the change of variables $x \equiv e^{2\beta} \Rightarrow \mathrm{d}\beta = \mathrm{d}x/(2x)$. Integrating and solving for x, we get $x = e^{2\beta} = (1 - r_\mathrm{s}/r)^{-1}$, where r_s is an integration constant with the dimension of length.

We thus obtain the **Schwarzschild metric**:

$$\mathrm{d}s^2 = -\left(1 - \frac{r_\mathrm{s}}{r}\right)\mathrm{d}t^2 + \frac{1}{1 - \frac{r_\mathrm{s}}{r}}\mathrm{d}r^2 + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\varphi^2\right) \ . \quad (5.24)$$

The metric is uniquely determined up to the single constant r_s, called the **Schwarzschild radius**. For $r \gg r_\mathrm{s}$, the difference between the metric (5.24) and the Minkowski metric in spherical coordinates is $\delta g_{\alpha\beta} \simeq r_\mathrm{s}/r$ for $\alpha = \beta = t, r$, and 0 otherwise. The difference approaches zero as $r \to \infty$, so the spacetime is **asymptotically flat**.

The value of r_s is fixed by the Newtonian limit. For the perturbed Minkowski metric discussed in chapter 4, a spherically symmetric Newtonian mass M creates a metric perturbation equal to -2 times the Newtonian potential, i.e. $2G_N M/r$. We thus identify $r_s = 2G_N M$.[1] This identification is the definition of mass in GR for a spherically symmetric source. We can also meaningfully say that a sphere around the centre with $r > r_s$ has total energy M. This can be generalised to a general asymptotically flat spacetime: we can look at how fast the metric falls off at infinity and identify the coefficient of the $1/r$ term as $2G_N M$. (The actual procedure can be quite complicated, but that's the idea.) If there is an energy flux coming to asymptotic infinity, the mass defined this way will decrease accordingly, so that the mass plus escaped energy stays constant. (We will discuss such an energy flux when we come to gravitational waves in chapter 8.) Note that if we defined energy by projecting the energy-momentum tensor onto an observer's four-velocity u^α and integrating over the hypersurface orthogonal to u^α (assuming it exists), the result would instead be zero, because the spacetime is empty everywhere where the metric (5.24) applies.

It is remarkable that the metric (5.24) has an extra symmetry that was not assumed: it is independent of time (in addition to having no g_{0i} components), so the spacetime is **static**. (We will define the property of being static in more generality when discussing symmetries in chapter 9.) This result has been given its own name: the fact that a spherically symmetric vacuum solution is static is called **Birkhoff's theorem**. So a spherically pulsating mass distribution has no effect on the spacetime outside it. The only information that can be measured from the outside (using gravity) is the mass. This result is stable to small perturbations: if a spacetime is nearly spherically symmetric and nearly empty, it is close to the Schwarzschild solution.

The tt component of the metric (5.24) vanishes at $r = r_s$, while the rr component diverges. This means that the coordinates do not apply at $r \leq r_s$. So the range for the radial coordinate is $r_s < r < \infty$, while the solution is eternal both to the past and the future, $-\infty < t < \infty$. To get an idea of the relevant radial scale, consider the Schwarzschild radius for the Earth and the Sun:

$$\begin{aligned} &\text{Earth} \quad M_\oplus = 5.98 \times 10^{24}\ \text{kg} && r_{s\oplus} = 0.886\ \text{cm} \\ &\text{Sun} \quad M_\odot = 1.99 \times 10^{30}\ \text{kg} && r_{s\odot} = 2.95\ \text{km} \end{aligned} \tag{5.25}$$

If the radius of the object is larger than r_s, the vacuum solution ceases to apply before we get to r_s. If this is not the case it is not obvious from the metric (5.24) whether the spacetime ends at $r = r_s$. After all, for Euclidean space written in spherical coordinates, $g_{\theta\theta}$ and $g_{\varphi\varphi}$ both vanish at $r = 0$, but this does not mean that the point $r = 0$ would be physically special, although it is true that we cannot continue the manifold beyond it. We will see that for the Schwarzschild metric the situation is the opposite: the radius $r = r_s$ is special, but we can continue the spacetime beyond it.

[1] In chapter 4 we used Cartesian coordinates, where all directions were perturbed, here we use spherical coordinates such that the angular directions are not perturbed. So we should transform to the same coordinate system to compare the metric perturbations. The result of our sloppy argumentation is, however, correct. (**Exercise:** Show this.)

5.1.2 Including the cosmological constant

Including the cosmological constant, the vacuum Einstein equation reads

$$G^{\alpha}{}_{\beta} + \Lambda \delta^{\alpha}{}_{\beta} = 0 \quad \Leftrightarrow \quad R^{\alpha}{}_{\beta} = \Lambda \delta^{\alpha}{}_{\beta} \ . \tag{5.26}$$

Therefore the results $\beta = \beta(r)$ and $\alpha = -\beta$, which were derived from the 01 component and from the difference between the 00 and 11 components are unchanged, only the form of $\beta(r)$ changes. The result is (**Exercise:** Show this.)

$$\mathrm{d}s^2 = -\left(1 - \frac{r_\mathrm{s}}{r} - \frac{\Lambda}{3} r^2\right)\mathrm{d}t^2 + \frac{1}{1 - \frac{r_\mathrm{s}}{r} - \frac{\Lambda}{3} r^2}\mathrm{d}r^2 + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\varphi^2\right) \ . \tag{5.27}$$

When $\Lambda > 0$, this is called the **de Sitter–Schwarzschild metric**, because it reduces to the Schwarzschild metric when $\Lambda = 0$ and to the **de Sitter metric** when $r_\mathrm{s} = 0$. In the Newtonian limit, the cosmological constant introduces a potential that rises like r^2, corresponding to a linearly growing force. Such a term could be incorporated into Newtonian gravity, and was in fact already considered by Isaac Newton.

In the case $\Lambda > 0$, the radial coordinate is bounded not only from below, but also from above (by approximately $\sqrt{3/\Lambda}$ when $\sqrt{3/\Lambda} \gg r_\mathrm{s}$). We will discuss the physical meaning of this in chapter 9 when we consider maximally symmetric spacetimes, including **de Sitter space**. We take $\Lambda = 0$ for the rest of this chapter, and look at the physical structure of the Schwarzschild metric.

5.1.3 Spatial structure

Let us consider the hypersurface of constant time. Its induced metric is

$$\mathrm{d}s^2 = g_{ij}\mathrm{d}x^i\mathrm{d}x^j = \frac{\mathrm{d}r^2}{1 - \frac{r_\mathrm{s}}{r}} + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\mathrm{d}\varphi^2 \ . \tag{5.28}$$

The hypersurfaces where both t and r are constant are two-spheres, by construction. The proper distance along a radial line (radial lines are geodesics, as it is easy to show) from radial coordinate r_1 to radial coordinate r_2 is

$$\begin{aligned} L &= \int_1^2 \mathrm{d}s = \int_{r_1}^{r_2} \frac{\mathrm{d}r}{\sqrt{1 - \frac{r_\mathrm{s}}{r}}} \\ &= \sqrt{r_2(r_2 - r_\mathrm{s})} - \sqrt{r_1(r_1 - r_\mathrm{s})} + r_\mathrm{s}\ln\frac{\sqrt{r_2} + \sqrt{r_2 - r_\mathrm{s}}}{\sqrt{r_1} + \sqrt{r_1 - r_\mathrm{s}}} \\ &> r_2 - r_1 \ , \end{aligned} \tag{5.29}$$

Thus, while the circumference of a sphere is $2\pi r$ and its area is $4\pi r^2$, its volume is larger than $\frac{4}{3}\pi r^3$, because the proper distance from the centre to the sphere is larger than r. This is an expression of the fact that the space is non-Euclidean. We could have chosen the proper distance from the centre to the sphere as the radial coordinate (i.e. $g_{rr} = 1$), or defined the radial distance so that we get the same volume as in the

Euclidean case. The coordinate-invariant statement is that the volume of a sphere in the Schwarzschild spacetime is larger than in Euclidean (or Minkowski) space, if the surface area is equal. The lesson is that there are different ways to generalise the spherical coordinates used in the Euclidean case to a non-Euclidean space.

In the limit $r_1, r_2 \gg r_\mathrm{s}$ the distance (5.29) reduces to $L \approx r_2 - r_1 + \frac{1}{2} r_\mathrm{s} \ln \frac{r_2}{r_1}$. The distance does not reduce to the Euclidean case for radial coordinates far away from the Schwarzschild radius. The extra term can be sizeable if the difference between r_2 and r_1 is large (in units of r_s). Even though the space becomes flatter with larger r, the residual curvature has an effect if we consider points far enough apart.

Let us now consider the cut of the three-dimensional space along the equator, $\theta = \frac{\pi}{2}$. (Because of rotational symmetry, all hypersurfaces where both t and θ are constant are identical.) The metric of the equatorial plane is

$$\mathrm{d}s^2 = g_{ij} \mathrm{d}x^i \mathrm{d}x^j = \frac{\mathrm{d}r^2}{1 - \frac{r_\mathrm{s}}{r}} + r^2 \mathrm{d}\varphi^2 \ . \tag{5.30}$$

Exercise: Show that the equatorial plane has the geometry of a two-dimensional surface in three-dimensional Euclidean space that you get by starting with a parabola lying sideways off the z-axis and rotating it around the z-axis. Find the equation for this surface in the cylindrical coordinates (r, z, ϕ).

5.1.4 Time dilation

Consider a line with $(r, \theta, \varphi) =$ constant. The infinitesimal proper time interval along the line is

$$\mathrm{d}\tau = \sqrt{1 - \frac{r_\mathrm{s}}{r}} \mathrm{d}t < \mathrm{d}t \ . \tag{5.31}$$

So the coordinate time t is the proper time measured by observers at constant (r, θ, φ) in the limit $r \to \infty$, when the spacetime is flat. The proper time measured by an observer at constant (r, θ, φ) at finite r is smaller: clocks closer to $r = r_\mathrm{s}$ run slower. This is an example of **gravitational time dilation**. The deeper you are in the gravitational well, the slower time passes. This effect grows without limit when approaching the Schwarzschild radius. From (5.31) it looks as if time stood still at $r = r_\mathrm{s}$, but recall that the metric we have used does not apply there. While time dilation does not exist in Newtonian theory, a closely related effect, the **gravitational redshift**, can be understood in Newtonian terms.

5.1.5 Gravitational redshift

Let us look how the frequency of light changes as it travels in the Schwarzschild metric. Without extra effort, we can do the calculation for a general static metric ($\partial_t g_{\alpha\beta} = 0$, $g_{0i} = 0$),

$$\mathrm{d}s^2 \quad = \quad -|g_{00}(x^k)| \mathrm{d}t^2 + g_{ij}(x^k) \mathrm{d}x^i \mathrm{d}x^j \ , \tag{5.32}$$

of which the Schwarzschild metric is a special case. We want to find the relation between the proper time intervals measured by an emitter sitting at constant spatial

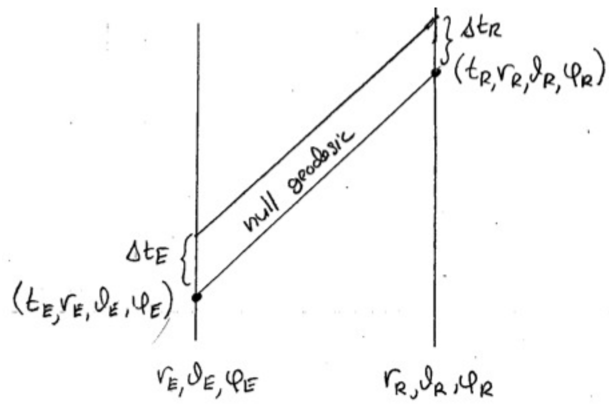

Figure 1: Light emission and reception at fixed spatial points E and R, respectively.

coordinates x^i_E and a receiver sitting at constant spatial coordinates x^i_R, as illustrated in figure 1. Observers at constant spatial coordinates are called **stationary**. In the geometrical optics approximation, light travels on null geodesics, $\mathrm{d}s^2 = 0$. Therefore the coordinate time that elapses between emission and reception is

$$\begin{aligned} t_R - t_E &= \int_E^R \sqrt{\frac{g_{ij}(x^k)}{|g_{00}(x^k)|}\mathrm{d}x^i\mathrm{d}x^j} \\ &= \int_{\lambda_E}^{\lambda_R} \mathrm{d}\lambda \sqrt{\frac{g_{ij}(x^k)}{|g_{00}(x^k)|}\frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}} \, , \end{aligned} \tag{5.33}$$

where the spatial coordinates along the path are $x^i(\lambda)$. Consider now a signal sent at $t_R + \delta t_R$ and received at $t_E + \delta t_E$. Because the integrand does not depend on time, and the spatial positions of the observer and receiver do not change, the result is the same, $t_R + \delta t_R - (t_E + \delta t_E) = t_R - t_E$, so $\delta t_R = \delta t_E$. So the coordinate time interval between two wavecrests is the same as the receiver and the emitter. For a stationary observer, the relation between proper time and coordinate time is $\mathrm{d}\tau = \sqrt{|g_{00}|}\mathrm{d}t$, so

$$\begin{aligned} \frac{\delta\tau_R}{\delta\tau_E} = \frac{f_E}{f_R} = \frac{E_E}{E_R} &= \sqrt{\frac{|g_{00}(x_R)|}{|g_{00}(x_E)|}} \\ &= \sqrt{\frac{1 - r_\mathrm{s}/r_R}{1 - r_\mathrm{s}/r_E}} \, , \end{aligned} \tag{5.34}$$

where on the last line we have applied the Schwarzschild metric (5.24). Here f is the frequency and E is the energy of the lightwave.

So just as clocks deeper down run slower, the wave frequency is larger. If a signal is sent from r_R to $r_E > r_R$, it will arrive with less energy than it had when it was sent. Correspondingly, a signal sent down into the gravitational well arrives with more energy than it had initially. This phenomenon is familiar from the Newtonian physics of massive particles: a particle loses kinetic energy when climbing up from a gravitational well, because its kinetic plus potential energy is conserved. In the limit $r \gg r_\mathrm{s}$, we in fact recover the Newtonian result quantitatively if we identify the initial energy with mass (but note that we consider light, not massive particles):

$$\frac{E_E - E_R}{E_E} \simeq G_\mathrm{N} M \left(\frac{1}{r_E} - \frac{1}{r_R} \right) . \tag{5.35}$$

5.2 Geodesics

5.2.1 The geodesic equation

Let us look at geodesic motion of both massive and massless particles (i.e. time-like and null geodesics) in the Schwarzschild geometry. As discussed in chapter 3, geodesic equations are easily obtained from the variational principle with the Lagrangian $L = \frac{1}{2} g_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta$. For the Schwarzschild metric (5.24) we have

$$L = -\frac{1}{2}\left(1 - \frac{r_\mathrm{s}}{r}\right)\dot{t}^2 + \frac{1}{2}\left(1 - \frac{r_\mathrm{s}}{r}\right)^{-1}\dot{r}^2 + \frac{1}{2}r^2\dot{\theta}^2 + \frac{1}{2}r^2 \sin^2\theta\dot{\varphi}^2 , \tag{5.36}$$

where $\dot{} \equiv \frac{\mathrm{d}}{\mathrm{d}\lambda}$, and λ is an affine parameter along the geodesic.

The Euler–Lagrange equation is

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{\partial L}{\partial \dot{x}^\alpha} - \frac{\partial L}{\partial x^\alpha} = 0 . \tag{5.37}$$

This shows that if the Lagrangian does not depend on the coordinate x^α, the quantity $\frac{\partial L}{\partial \dot{x}^\alpha}$ is conserved along the path, an application of Noether's theorem. In particular, the Schwarzschild metric is independent of the time t, so we have

$$\frac{\partial L}{\partial \dot{t}} = -\left(1 - \frac{r_\mathrm{s}}{r}\right)\dot{t} = \text{constant} \equiv -k . \tag{5.38}$$

The constant k is related (we will later see precisely how) to the energy of the particle, and also determines (together with r_s/r) the time dilation $\dot{t}$. In Newtonian mechanics, energy is conserved because the Hamiltonian (and therefore the Lagrangian) is independent of time. In GR, we get a similar conservation law along a geodesic for any static metric.

The Schwarzschild metric is also independent of the angle φ, so we get another conserved quantity:

$$\frac{\partial L}{\partial \dot{\varphi}} = r^2 \sin^2\theta\dot{\varphi} = \text{constant} \equiv h . \tag{5.39}$$

This quantity corresponds to the conserved angular momentum per unit mass in Newtonian mechanics.

Let us now look at the θ component of the Euler–Lagrange equation. We have

$$\frac{\partial L}{\partial \theta} = r^2 \sin\theta\cos\theta\dot{\varphi}^2 \tag{5.40}$$

$$\frac{\partial L}{\partial \dot{\theta}} = r^2\dot{\theta} . \tag{5.41}$$

Inserting these into (5.37) gives

$$\ddot{\theta} + \frac{2}{r}\dot{r}\dot{\theta} - \sin\theta\cos\theta\dot{\varphi}^2 = 0 \ . \tag{5.42}$$

If we choose coordinates such that that the initial particle location and velocity lie on the equatorial plane, $\theta = \frac{\pi}{2}, \dot{\theta} = 0$, the above equation shows that $\theta = \frac{\pi}{2}$ at all times. (That geodesic motion is confined to a plane is obvious from the spherical symmetry, but it's nice to see how it comes out from the equations of motion.) We now choose such coordinates.

The radial component of the equation of motion remains. We have

$$\frac{\partial L}{\partial r} = -\frac{1}{2}\frac{r_s}{r^2}\dot{t}^2 - \frac{1}{2}\frac{r_s}{\left(1-\frac{r_s}{r}\right)^2}\frac{\dot{r}^2}{r^2} + r\dot{\varphi}^2 \tag{5.43}$$

$$\frac{\partial L}{\partial \dot{r}} = \frac{\dot{r}}{1-\frac{r_s}{r}} \ . \tag{5.44}$$

Inserting these into the Euler–Lagrange equation (5.37), we get the radial equation of motion:

$$\ddot{r} + \frac{1}{2}\left(1-\frac{r_s}{r}\right)\frac{r_s}{r^2}\dot{t}^2 - \frac{1}{2}\frac{r_s}{1-\frac{r_s}{r}}\frac{\dot{r}^2}{r^2} - \left(1-\frac{r_s}{r}\right)r\dot{\varphi}^2 = 0 \ . \tag{5.45}$$

We can now solve $\dot{t}$ from (5.38) and $\dot{\varphi}$ from (5.39) and insert them into (5.45) to get an ordinary differential equation for one unknown $r(\lambda)$. However, this equation is second order and non-linear, making it difficult to solve. It is often easier to instead use the extra condition on $\dot{x}^\alpha$ that arises from the normalisation of the four-velocity. We have (with $u^\alpha = \frac{\mathrm{d}x^\alpha}{\mathrm{d}\lambda}$ as usual)

$$g_{\alpha\beta}u^\alpha u^\beta = -\left(1-\frac{r_s}{r}\right)\dot{t}^2 + \frac{1}{1-\frac{r_s}{r}}\dot{r}^2 + r^2\dot{\varphi}^2 = \begin{cases} -1 & \text{for a timelike curve} \\ 0 & \text{for a null curve} \end{cases} , \tag{5.46}$$

where in the timelike case λ is taken to be the proper time τ.

Let us first look at some simple cases, before analysing the leading GR corrections to the motion of Mercury and motion of light in the Solar system.

5.2.2 Circular motion

Consider circular motion, i.e. take $\dot{r} = 0$. The radial equation of motion (5.45) reduces to

$$\frac{1}{2}\frac{r_s}{r^2}\dot{t}^2 - r\dot{\varphi}^2 = 0 \ , \tag{5.47}$$

which gives, writing $r_s = 2G_N M$,

$$\left(\frac{\mathrm{d}\varphi}{\mathrm{d}t}\right)^2 = \frac{G_N M}{r^3} \ . \tag{5.48}$$

Over one period $\Delta\varphi = 2\pi$, so

$$(\Delta t)^2 \quad = \quad \frac{4\pi^2}{G_{\rm N}M} r^3 \ . \tag{5.49}$$

This is similar to Kepler's third law: the square of the orbital period is proportional to the third power of the radius. Kepler's law of course holds also for elliptical orbits, and we here consider only circular orbits. Another difference is that t here is the coordinate time, which equals the proper time measured by a stationary observer at asymptotic infinity. The proper time interval measured by an observer at finite r is $\Delta\tau = \sqrt{1 - 2G_{\rm N}M/r}\Delta t$.

We haven't specified whether the orbit is timelike or null. Consider now the proper time measured by an an observer on a timelike orbit. Combining (5.47) with the normalisation condition (5.46) gives

$$\Delta\tau = \sqrt{1 - \frac{3G_{\rm N}M}{r}}\Delta t = 2\pi\sqrt{\frac{r^3}{G_{\rm N}M}\left(1 - \frac{3G_{\rm N}M}{r}\right)} \ . \tag{5.50}$$

The orbital period goes to zero as r approaches $3G_{\rm N}M = \frac{3}{2}r_{\rm s}$. There are no circular timelike geodesic orbits closer than this to the Schwarzschild radius. The spacetime is so curved that geodesics will turn and spiral towards the Schwarzschild radius. It is still possible to move on a circular non-geodesic curve by applying sufficient force. If we consider a stationary observer hovering above the Schwarzschild radius (such an observer is non-geodesic), the proper period they measure is

$$\Delta\tau_{\rm hov} = 2\pi\sqrt{\frac{r^3}{G_{\rm N}M}\left(1 - \frac{2G_{\rm N}M}{r}\right)} \ . \tag{5.51}$$

Note that this is longer than the proper period measured by a geodesic observer. Timelike geodesics give a local maximum of the proper time between two points, but not necessarily the global maximum. Here the worldline of a stationary (i.e. accelerated) observer is not a small perturbation of the worldline of a geodesic observer.

What about null geodesics? Combining the equation of motion (5.47) with the normalisation condition (5.46) in the null case, a solution exists only for $r = 3G_{\rm N}M$. Above this radius photon trajectories are not sufficiently curved to remain in orbit, below it photons spiral down towards the Schwarzschild radius just like massive particles.

5.2.3 Radial motion

Let us now look at radial motion, i.e. $\dot\varphi = 0$. Considering first timelike motion, combining energy conservation (5.39) and the normalisation (5.46) gives

$$\frac{1}{2}(k^2 - 1) = \frac{1}{2}\dot r^2 - \frac{G_{\rm N}M}{r} \ . \tag{5.52}$$

Comparison to the corresponding Newtonian equations shows that $(k^2 - 1)/2$ corresponds to the Newtonian total energy per unit mass. (In this analogy, there is no energy associated with the mass of the particle, i.e. no rest energy.) If $k \geq 1$, the particle can escape to infinity, whereas solutions with $k < 1$ represent gravitationally bound particles. As the equation (5.52) has the same form as in Newtonian theory, the solutions are identical in terms of the proper time. The only difference comes from the distinction between proper time and coordinate time.

As we know from the Newtonian theory, a bound particle falls down to any radius smaller than the initial radius (in this case to the Schwarzschild radius, where our coordinate system ends) in a finite time. To find how this looks to an observer at fixed spatial coordinates, let's find coordinate time as a function of coordinate radius for a particle dropped from $r = r_0 > r_\mathrm{s}$ with initial zero velocity. (The particle can also be initially heading up, in which case r_0 is the maximum radius it reaches before turning around.) Using (5.52), we can determine the constant k in terms of the initial condition r_0:

$$\frac{1}{2}\dot{r}^2 = G_\mathrm{N} M \left(\frac{1}{r} - \frac{1}{r_0} \right) \ , \tag{5.53}$$

i.e. $k^2 = 1 - \frac{2G_\mathrm{N}M}{r_0}$. The coordinate time in terms of the coordinate radius is

$$\frac{\mathrm{d}t}{\mathrm{d}r} = \frac{\mathrm{d}t}{\mathrm{d}\tau}\frac{\mathrm{d}\tau}{\mathrm{d}r} = \frac{\dot{t}}{\dot{r}} \ , \tag{5.54}$$

so inputting $\dot{t}$ from the energy constraint (5.38) and $\dot{r}$ from (5.53), the coordinate time to fall from coordinate radius r_0 to the Schwarzschild radius r_s is

$$\begin{aligned} \Delta t &= \sqrt{\frac{r_0 - r_\mathrm{s}}{r_\mathrm{s}}} \int_{r_\mathrm{s}}^{r_0} \mathrm{d}r \frac{r^{3/2}}{\sqrt{r_0 - r}(r - r_\mathrm{s})} \\ &> \sqrt{\frac{r_0 - r_\mathrm{s}}{r_\mathrm{s}}} \frac{r_\mathrm{s}^{3/2}}{\sqrt{r_0 - r_\mathrm{s}}} \int_{r_\mathrm{s}}^{r_0} \mathrm{d}r \frac{1}{r - r_\mathrm{s}} \\ &= \infty \ . \end{aligned} \tag{5.55}$$

The gravitational time dilation between the time t measured by stationary observers at asymptotic infinity and stationary observers at a finite radius larger than r_s is finite. However, an infalling observer reaches the Schwarzschild radius in a finite time according to their own clock, while from the point of view of an observer sitting still, the infalling observer never reaches the Schwarzschild radius, approaching it ever more slowly. Because the redshift grows without limit when approaching the Schwarzschild radius, the infalling observer is lost to sight in a finite time: beyond some point, they no longer have enough energy to send signals up (assuming observers outside have a limit on how small energy signals they can detect).

Let us now consider null geodesics. The energy constraint (5.38) and the normalisation (5.46) give

$$\frac{\mathrm{d}r}{\mathrm{d}t} = 1 - \frac{r_\mathrm{s}}{r} \ , \tag{5.56}$$

so the time to fall from radius r_1 to r_s is

$$\Delta t = \int_{r_\mathrm{s}}^{r_1} \mathrm{d}r \frac{r}{r - r_\mathrm{s}} = \infty \ . \tag{5.57}$$

So like massive objects, light never reaches the Schwarzschild radius from the point of view of a stationary observer outside. From their point of view, nothing can reach r_s (and nothing can come up from r_s; we will look at this in the next chapter when we consider coordinates that extend beyond r_s).

Let us finally comment on the relation of $(k^2-1)/2$ as a measure of energy to the energy measured by an observer moving at four-velocity v^α,[2] $E = -v_\alpha p^\alpha$, where $p^\alpha = m u^\alpha$, and m is particle mass. The quantity E depends on the observer. Let us consider stationary observers, $v^i = 0$. The normalisation condition $-1 = g^{\alpha\beta} v_\alpha v_\beta$ then gives $v_0 = -1/\sqrt{-g^{00}} = -\sqrt{-g_{00}} = -\sqrt{1 - r_\mathrm{s}/r}$, so we have, applying (5.38),

$$\frac{E}{m} = -v_0 u^0 = -v_0 \dot{t} = \frac{k}{\sqrt{1 - r_\mathrm{s}/r}} \ . \tag{5.58}$$

This quantity depends on position, as it measures only the kinetic and rest energy of the particle, unlike $(k^2-1)/2$, which also contains gravitational energy (but not rest energy), and which is defined as a constant of motion.

5.2.4 General non-radial orbits

Let us now consider general orbits where $\dot{\varphi} \neq 0$. Except in the case $\dot{r} = 0$, which we have already considered, they are curves parametrised by λ in effectively three dimensions $(t(\lambda), r(\lambda), \varphi(\lambda))$, as the motion is on the equatorial plane. The curve $x^\alpha(\lambda)$ never intersects itself, as $t(\lambda)$ is monotonic. If we project the trajectory on the (r, φ) plane, we get a curve that can intersect itself. For a general curve, neither $r(\lambda)$ nor $\varphi(\lambda)$ is monotonic, and so they cannot be inverted. However, for trajectories that do not change the rotation direction on the (r, φ)-plane, there is a function $r(\varphi)$ if we extend the range of φ from $0 \leq \varphi < 2\pi$ to $-\infty < \varphi < \infty$. Let us consider such trajectories and find the equation for the function $r(\varphi)$.

We use the conservations equations (5.38) and (5.39) and the normalisation condition (5.46),

$$\left(1 - \frac{r_\mathrm{s}}{r}\right)\dot{t} = k \tag{5.59}$$

$$r^2 \dot{\varphi} = h \tag{5.60}$$

$$\left(1 - \frac{r_\mathrm{s}}{r}\right)\dot{t}^2 - \frac{\dot{r}^2}{1 - \frac{r_\mathrm{s}}{r}} - r^2\dot{\varphi}^2 = \begin{cases} 1 & \text{for massive particles} \\ 0 & \text{for photons} \end{cases} \ . \tag{5.61}$$

Dividing (5.61) by $\dot{\varphi}^2$, we get

$$\left(1 - \frac{r_\mathrm{s}}{r}\right)\left(\frac{\dot{t}}{\dot{\varphi}}\right)^2 - \frac{1}{1 - \frac{r_\mathrm{s}}{r}}\left(\frac{\mathrm{d}r}{\mathrm{d}\varphi}\right)^2 - r^2 = \begin{cases} \frac{1}{\dot{\varphi}^2} \\ 0 \end{cases} \ . \tag{5.62}$$

Inputting now $\dot{t}$ from (5.59) and $\dot{\varphi}$ from (5.60), we get

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\varphi}\right)^2 + r^2\left(1 - \frac{r_\mathrm{s}}{r}\right)\left(1 + \begin{Bmatrix} r^2/h^2 \\ 0 \end{Bmatrix}\right) - \frac{k^2}{h^2} r^4 = 0 \ . \tag{5.63}$$

[2] Not to be confused with u^α, which is the four-velocity of the particle considered.

We can make the equation nicer by switching to the variable $u \equiv 1/r$ ($\Rightarrow$ $\mathrm{d}u = -\mathrm{d}r/r^2$) and multiplying by $1/r^4$ to get

$$\left(\frac{\mathrm{d}u}{\mathrm{d}\varphi}\right)^2 + u^2 = \frac{k^2}{h^2} - \begin{Bmatrix} 1/h^2 \\ 0 \end{Bmatrix} + \begin{Bmatrix} r_\mathrm{s}/h^2 \\ 0 \end{Bmatrix} u + r_\mathrm{s} u^3 \ . \tag{5.64}$$

We can write this separately for timelike curves as

$$\left(\frac{\mathrm{d}u}{\mathrm{d}\varphi}\right)^2 + u^2 = \frac{k^2 - 1}{h^2} + \frac{r_\mathrm{s}}{h^2} u + r_\mathrm{s} u^3 \ , \tag{5.65}$$

and for null curves as

$$\left(\frac{\mathrm{d}u}{\mathrm{d}\varphi}\right)^2 + u^2 = \frac{k^2}{h^2} + r_\mathrm{s} u^3 \ . \tag{5.66}$$

5.2.5 Precession of the perihelion of Mercury

The equation (5.65) provides a convenient starting point for evaluating GR corrections to trajectories that are close to Newtonian. Let us first consider the timelike case and bound orbits. In particular, we will look at the GR correction to the orbit of Mercury. Mercury is the closest planet to the the Sun and therefore its orbit is most affected by post-Newtonian effects. An observation of particular importance is the change in the **perihelion distance**, i.e. shortest distance from the Sun. If a planet were to move under Newtonian gravity in the gravitational field of the Sun alone (and the Sun were perfectly spherical), the orbit would be an ellipse. GR effects change this in two ways: they deform the shape of the closed orbit (such corrections are periodic in φ), and also turn the orbit on the orbital plane so that it no longer closes (such corrections are non-periodic in φ). The non-periodic corrections lead to the near-ellipse rotating on the orbital plane, a phenomenon called **precession**. This precession of the perihelion of Mercury is the first observational GR effect that was calculated; Einstein did the calculation already before the publication of GR. Let us do the calculation and compare to observation.

We consider a situation where the GR correction is a small perturbation to the Newtonian solution. To that end, it is convenient to use the dimensionless variable

$$x \equiv \frac{2h^2 u}{r_\mathrm{s}} \ . \tag{5.67}$$

In terms of x, the equation of motion (5.65) reads

$$x'^2 + x^2 = \frac{4h^2(k^2-1)}{r_\mathrm{s}^2} + 2x + \frac{r_\mathrm{s}^2}{2h^2} x^3 \ , \tag{5.68}$$

where prime denotes derivative with respect to φ. We can simplify this equation by taking a derivative with respect to φ:

$$x'' + x - 1 = \varepsilon x^2 \ , \tag{5.69}$$

where

$$\varepsilon \equiv \frac{3r_\mathrm{s}^2}{4h^2} \ . \tag{5.70}$$

Notice that the dependence on the constant k has disappeared; it is encoded in the extra initial condition that we have to provide (now that the equation is second order rather than first order).

The parameter ε is tiny for all Solar system orbits. To see this in a simple manner, consider a circular orbit, $x' = 0$. The equation (5.69) is then a quadratic algebraic equation for x, with the solution

$$x = \frac{1}{2\varepsilon}\left(1 - \sqrt{1-4\varepsilon}\right) . \tag{5.71}$$

Inputting the definition of x in (5.67) into the definition of ε in (5.70), we get

$$\varepsilon = \frac{3r_s^2}{4h^2} = \frac{3r_s}{2xr} , \tag{5.72}$$

so using (5.71) we get

$$\frac{3r_s}{2r} = \varepsilon x = \frac{1}{2}\left(1 - \sqrt{1-4\varepsilon}\right) . \tag{5.73}$$

Taking into account $r_s/r \ll 1$, we get

$$\varepsilon = \frac{3r_s}{2r} . \tag{5.74}$$

Given that the Schwarzschild radius of the Sun is 2.95 km and the perihelion distance of Mercury is 46.0×10^6 km, we have $\varepsilon \sim 10^{-7}$, so the approximation $\varepsilon \ll 1$ is pretty good. The orbit of Mercury is not quite circular, but taking the ellipticity into account does not change the order of magnitude of ε.

The Newtonian orbit corresponds to $\varepsilon = 0$, and the post-Newtonian corrections can be organised in a power series in ε. We consider the first order correction, and write

$$x(\varphi) = x_0(\varphi) + \varepsilon x_1(\varphi) . \tag{5.75}$$

Inputting this into the equation of motion (5.69), we get (dropping terms higher than first order in ε),

$$x_0'' + x_0 - 1 + \varepsilon(x_1'' + x_1) = \varepsilon x_0^2 . \tag{5.76}$$

We first solve x_0 from the above equation neglecting ε, and then use this solution as the source term for ε. So (5.76) splits into two equations:

$$\begin{aligned} x_0'' + x_0 &= 1 \\ x_1'' + x_1 &= x_0^2 . \end{aligned} \tag{5.77}$$

Let us first consider the background equation. It is linear, so the general solution is the sum of the general solution of the homogeneous equation plus any one solution of the inhomogeneous equation. One solution of the inhomogeneous equation is simply $x_0 = 1$, so we have

$$\begin{aligned} x_0 &= A\cos\varphi + \tilde{A}\sin\varphi + 1 \\ &= 1 + e\sin(\varphi - \varphi_0) \\ &= 1 + e\sin\varphi , \end{aligned} \tag{5.78}$$

where A and $\tilde{A}$ are constants (as are e and φ_0). Without loss of generality, we have chosen the phase so that $\varphi_0 = 0$ (we can rotate the plane), i.e. $x_0(0) = 1$. We will use this choice also for the full solution, $x(0) = 1$, so $x_1(0) = 0$.

The parameter e is called the **eccentricity**, and its value for Mercury is $e = 0.2056$. The perihelion distance is the smallest value of r, i.e. the largest value of x, which is at $\varphi = \frac{\pi}{2}$: $x_0 = 1 + e$. Let us denote the perihelion distance by r_-. From the definition of x (5.67) we then get the value of h^2:

$$h^2 = \frac{1}{2}(1+e) r_{\rm s}\, r_- \ . \tag{5.79}$$

Inputting the solution (5.78) into the first order equation (5.68) and using the result (5.79) for h gives the constant k:

$$1 - k^2 = \frac{1}{2}(1-e)\frac{r_{\rm s}}{r_-} \ . \tag{5.80}$$

We can check that we have $k < 1$ for the bound orbit, as $0 \leq e < 1$ (because $x_0 > 0$ in (5.78)).

This completes the solution of the Newtonian trajectory. Now let's find the GR correction. With the solution (5.78) for x_0, the equation (5.77) for x_1 reads

$$x_1'' + x_1 \ = \ (1 + e\sin\varphi)^2 \ . \tag{5.81}$$

First, as (5.81) is linear in x_1, the general solution is again a sum of the general solution of the homogeneous equation plus one solution of the inhomogeneous equation. As the perturbed homogeneous equation (5.81) is the same as the background homogeneous equation (5.78), the general solution solution of (5.81) just adds a perturbation to the constants of the solution of (5.78), which can be absorbed into a redefinition of e. So we only need a single solution of the inhomogeneous equation. For $e = 0$, $x_1 = 1$ is a solution. So let's write

$$x_1(\varphi) = 1 + f(\varphi)\cos\varphi \ , \tag{5.82}$$

where the factor $\cos\varphi$ has been chosen so that f will drop out of (5.81), and only f' and f'' will appear. We have

$$\begin{aligned} x_1' &= f'\cos\varphi - f\sin\varphi \\ x_1'' &= f''\cos\varphi - 2f'\sin\varphi - f\cos\varphi \ . \end{aligned} \tag{5.83}$$

Inserting this into (5.81), we get

$$f''\cos\varphi - 2f'\sin\varphi = 2e\sin\varphi + e^2\sin^2\varphi \ . \tag{5.84}$$

This is a linear first-order equation for f', which can be readily integrated. The general solution for f is

$$f = -e\varphi + C + D\tan\varphi + \frac{1}{3}e^2\left(\cos\varphi + \frac{1}{\cos\varphi}\right) \ , \tag{5.85}$$

where C and D are integration constants. The solution for the perturbation is thus

$$x_1 = 1 - e\varphi\cos\varphi + C\cos\varphi + D\sin\varphi + \frac{1}{3}e^2\left(1+\cos^2\varphi\right) \ . \tag{5.86}$$

The term $D\sin\varphi$ can be arbsorbed into a redefinition of the ellipticity e in the background solution, so we drop it. The condition $x_1(0) = 0$ gives $C = -1 - \frac{2}{3}e^2$, so we get

$$x_1 = -e\varphi\cos\varphi + 1 - \cos\varphi + \frac{1}{3}e^2(1-\cos\varphi)^2 \ . \tag{5.87}$$

The full solution is thus

$$\begin{aligned} x(\varphi) &= x_0(\varphi) + \varepsilon x_1(\varphi) \\ &= 1 + e\sin\varphi + \varepsilon\left[-e\varphi\cos\varphi + 1 - \cos\varphi + \frac{1}{3}e^2(1-\cos\varphi)^2\right] \\ &= 1 + e\sin\varphi - e\varepsilon\varphi\cos\varphi + \varepsilon(\text{terms with a period of } 2\pi) \ . \end{aligned} \tag{5.88}$$

The periodic correction terms correspond to deformations of the orbit that keep it closed (orbit here referring to the projection of the curve (t, r, φ) onto the (r, φ) plane). The first correction term, in contrast, prevents closure of the orbit, so the planet comes to a slightly different position every period. This term is secular, the deviation from the Newtonian prediction grows with every orbit, so it is easier to detect than the periodic corrections – recall that $\varepsilon \sim 10^{-7}$.

The relative change of radius over one period is

$$\begin{aligned} \frac{r(2\pi) - r(0)}{r(0)} &= \frac{x^{-1}(2\pi) - x^{-1}(0)}{x^{-1}(0)} \\ &= (1 - 2\pi e\varepsilon)^{-1} - 1 \\ &\simeq 2\pi e\varepsilon \ . \end{aligned} \tag{5.89}$$

As the perihelion distance of Mercury is 46.0×10^6 km, the change in distance is 4.7 km (we put in the numbers below). However, traditionally the change is expressed in terms of the change of the viewing angle, which is readily observable. Because ε is small, we can write

$$\begin{aligned} x(\varphi) &= 1 + e\sin\varphi - e\varepsilon\varphi\cos\varphi + (\text{terms with a period of } 2\pi) \\ &\simeq 1 + e\sin\varphi\cos(\varepsilon\varphi) - e\cos\varphi\sin(\varepsilon\varphi) + (\text{terms with a period of } 2\pi) \\ &= 1 + e\sin(\varphi - \varepsilon\varphi) + (\text{terms with a period of } 2\pi) \ , \end{aligned} \tag{5.90}$$

where we have used the relation $\sin(x+y) = \sin x\cos y + \sin y\cos x$. The first order perturbative solution (5.88) is valid provided $\varepsilon\varphi \ll 1$, and since each period gives 2π, we get a limit on the number of periods n this approximation can handle as $10^{-7}2\pi n \ll 1$, which gives $n \ll 10^6$. The orbital period of Mercury is 88.0 Earth days, so the approximation becomes invalid after about 100 000 Earth years. The change in the angular position of the planet over one period is

$$\begin{aligned}\Delta\varphi &= 2\pi\varepsilon \\ &= \frac{3}{2}\pi\frac{r_s^2}{h^2} \\ &= \frac{3\pi}{1+e}\frac{r_s}{r_-} \\ &\approx \frac{3\pi}{1+0.2056}\frac{2.95\text{ km}}{46.0\times10^6\text{ km}} \\ &\approx 5.01\times10^{-7}\text{ rad} \\ &\approx 2.87\times10^{-5}\,{}^\circ \\ &\approx 0.103'' \ . \end{aligned} \tag{5.91}$$

where we have inserted the definition of ε from (5.70) and the value of h^2 from (5.79), and on the last line expressed the change in terms of arcseconds. (One arcminute is 1/60th of a degree, and one arcsecond is 1/60th of an arcminute.) The change per century is $0.103'' \times 100 \times 365/88 \approx 43''$. The observed change is 575 arcsec/century, of which 532 arcsec/century is due to other planets, leaving 43 arcsec/century to be explained by GR.

So the GR correction matches the observations, but it is more accurate to call this a **postdiction** rather than a prediction, because the extra precession was known since 1859, over 50 years before it was explained by GR. Explanations of longstanding observational anomalies can offer strong support for a new theory. First, if the anomaly has been known for long, there has been time to minimise the role of systematic errors in the observations. Second, if no one has found a convincing explanation for decades, it is less likely that there is a simple alternative explanation.

In the case of the precession of the perihelion of Mercury, it was proposed that the missing $43''$ would be explained by a new planet between the Sun and Mercury. This was a well justified first hypothesis: if $532''$ is explained by other planets, why not the whole $575''$? (These are the modern values, which are more accurate than those in the 19th century, but the difference is not large.) Furthermore, it was these kind of disturbances in the orbit of Uranus that had led to the discovery of Neptune in 1846. So the planet Vulcan was proposed. This explanation was quickly confirmed by the observation of this new planet. However, as others failed to replicate the observation (though further sightings of Vulcan were reported), doubts grew, and the situation remained anomalous until GR explained the missing $43''$. Vulcan was thus consigned to the dustbin of science (although it has since enjoyed a successful career in popular culture).

It is remarkable how straightforwardly GR explains the anomalous precession of Mercury: we just take the spherically symmetric vacuum solution, consider timelike geodesics that are close to Newtonian, and the correct result comes out. Unlike in particle physics, there is little room for small adjustments to GR: we could for example allow for non-zero torsion and complicate the action by including torsion terms, but this is a large change in the theory, and most such changes lead to unstable and unviable theories.

Nevertheless, since the perihelion precession was known when GR was developed, we could argue that it's not a test of the theory, since if it had predicted something else, a different theory might have been proposed. Predictions are important, and the first new effect predicted by GR and observationally confirmed was bending of light.

5.2.6 Bending of light by the Sun

After the timelike case, let us now look at the GR correction to the trajectory of null geodesics. The calculation is easier than in the timelike case, because the Newtonian solution is just a straight line. We want to find out how much the light ray bends if its shortest distance from the centre of the Sun (called the **impact parameter**) is r_0.

We start from the equation of motion (5.66). As in the timelike case, we define a new dimensionless variable, $z \equiv hu/k$. The equation (5.66) then reads

$$z'^2 + z^2 = 1 + \varepsilon z^3 \ , \tag{5.92}$$

where we have defined

$$\varepsilon \equiv \frac{k}{h} r_{\rm s} \ . \tag{5.93}$$

As in the timelike case, $\varepsilon = 0$ gives the classical trajectory. Let us find the value of h/k on the classical trajectory in terms of the impact parameter r_0. As the point of nearest approach is a minimum of the distance, the derivative of the distance is zero there. Putting $z' = 0$ in (5.92) and approximating $\varepsilon = 0$ gives $z = 1$, i.e. $r_0 = h/k$.

In the Solar system, the ε term is a small perturbation, so we write

$$z(\varphi) = z_0(\varphi) + \varepsilon z_1(\varphi) \ , \tag{5.94}$$

where z_0 is the classical solution. Inputting this ansatz into (5.92) and linearising with regard to ε, we have

$$z_0'^2 + z_0^2 + 2\varepsilon(z_0' z_1' + z_0 z_1) = 1 + \varepsilon z_0^3 \ , \tag{5.95}$$

so we get the background and perturbed equation of motion:

$$z_0'^2 + z_0^2 = 1 \tag{5.96}$$

$$2(z_0' z_1' + z_0 z_1) = z_0^3 \ . \tag{5.97}$$

We see immediately that the solution of the background equation is

$$z_0(\varphi) = \sin(\varphi - \varphi_0) = \sin\varphi \ , \tag{5.98}$$

where we have, without loss of generality, chosen the angular coordinate so that $\varphi = 0$ corresponds to $x = 0$, i.e. $r = \infty$, so $\varphi_0 = 0$. This solution is just a straight line: $y = r\sin\varphi \propto z^{-1}\sin\varphi = 1$.

Let us now turn to the perturbation. The same initial condition applies: $z_1(0) = 0$, i.e. we choose $\varphi_0 = 0$ to correspond to the initial direction of the full perturbed light ray. Given the background solution (5.98), the perturbed equation of motion is

$$\begin{aligned}\sin^3\varphi &= 2\cos\varphi z_1' + 2\sin\varphi z_1 \\ &= 2\cos^2\varphi\frac{\mathrm{d}}{\mathrm{d}\varphi}\left(\frac{z_1}{\cos\varphi}\right) \ ,\end{aligned} \tag{5.99}$$

which can be rewritten as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\varphi}\left(\frac{z_1}{\cos\varphi}\right) &= \frac{1}{2}\frac{\sin^3\varphi}{\cos^2\varphi} \\ &= \frac{1}{2}\sin\varphi\frac{1-\cos^2\varphi}{\cos^2\varphi} \\ &= \frac{1}{2}\left(\frac{\sin\varphi}{\cos^2\varphi} - \sin\varphi\right),\end{aligned} \tag{5.100}$$

which we can straightforwardly integrate to get

$$z_1(\varphi) = \frac{1}{2}\left(1+\cos^2\varphi\right) + A\cos\varphi \ , \tag{5.101}$$

where A is an integration constant. The initial condition $z_1(0) = 0$ gives $A = -1$, so the full solution is

$$\begin{aligned}z(\varphi) &= z_0(\varphi) + \varepsilon z_1(\varphi) \\ &= \sin\varphi + \frac{\varepsilon}{2}(1-\cos\varphi)^2 \ .\end{aligned} \tag{5.102}$$

The light ray comes in from the direction $\varphi = 0$. We want to know how much it is bent by the Sun, i.e. to which direction it goes out. An unbent light way goes to spatial infinity at $\varphi = \pi$. When perturbed, it will instead to go spatial infinity at $\varphi = \pi + \alpha$, where $\alpha \ll 1$. So, we have

$$\begin{aligned}0 &= z(\pi+\alpha) \\ &= \sin(\pi+\alpha) + \frac{\varepsilon}{2}[1-\cos(\pi+\alpha)]^2 \\ &\simeq -\sin\alpha + 2\varepsilon \\ &\simeq -\alpha + 2\varepsilon \ ,\end{aligned} \tag{5.103}$$

where we have expanded to leading order in the small parameters α and ε. The bending angle is thus to leading order

$$\alpha = 2\varepsilon = 2\frac{k}{h}r_\mathrm{s} = \frac{2r_\mathrm{s}}{r_0} = \frac{4G_\mathrm{N}M_\odot}{r_0} \leq \frac{4G_\mathrm{N}M_\odot}{r_\odot} \ , \tag{5.104}$$

where we have input $h/k = r_0$ and taken into account $r_0 \geq r_\odot$. The largest deflection is for light rays that graze the surface of the Sun. Inputting $r_\odot = 0.696 \times 10^6$ km, we get $\alpha = 1.75''$.

Within GR, the bending of light is conceptually straightforward: light is electromagnetic waves described by Maxwell equations, which (as we will discuss in the next chapter) show that in the geometrical optics limit light travels on null geodesics. In contrast, Newtonian theory is ambiguous on how light is affected by gravity. Newton thought that light consists of small massive particles, just like everything else. In that case light falls in a gravitational field in the same way as other massive particles, independent of how small the mass is. This leads to the deflection angle $\alpha = \frac{2G_N M_\odot}{r_0}$, half the GR value. However, in 1801, Thomas Young had published his observations of interference of light, leading to the idea that light consists of waves. This is not necessarily in contradiction with light being composed of massive particles. For example, sound waves in air are the result of collective behaviour of a large number of massive particles. However, with the success of Maxwell's electromagnetism light became to be seen as consisting of electromagnetic fields which are not reducible to massive particles, and in the context of quantum physics, light particles are massless. In this case Newton's theory is silent on how it light affected by gravity: the equation $\vec{F} = m\vec{a}$ says nothing about acceleration if $m = 0$. Massless particles can be assumed to fall like massive particles, or not to be affected by gravity.

In the early years of the 20th century, there was a third contender for a theory of gravity in addition to GR and Newtonian mechanics: Nordström's scalar theory of gravity discussed in chapter 2. In that theory, light is not affected by gravity. So there were three predictions (or two predictions and one prescription – for the Newtonian case): 1.75", 0.874" and 0". In fact, Einstein had in 1911 argued on the basis of the equivalence principle that light should fall just like other matter, predicting 0.874". As the effect is small, the best chance to observe it is during an eclipse, when it is possible to distinguish stars close to the Sun that are otherwise obscured by its brightness (because of the incredible coincidence that we happen to live in an era when the angular diameters of the Sun and the Moon are very close). One has to observe a pattern of stars close to the Sun during an eclipse and measure the same stars afterwards far from the Sun and see how the pattern changes: the stars closer to Sun are displaced more if light is bent by the Sun.

This predictions was due to be tested by the 1914 eclipse. Unfortunately for the world but fortunately for the reputation of GR, the First World War broke out and prevented this. In 1915 the full theory of GR was discovered and published, and in 1916 Einstein made the correct prediction. The prediction was observationally verified by two teams observing during the 1919 eclipse, one in Brazil and the other in Principe, off the coast of West Africa. The results ruled out the Newtonian value (whichever of the two you pick) and Nordström's zero, and were in agreement with the GR prediction. This was reported on the front pages of newspapers with headlines such as "Lights All Askew in the Heavens: Men of Science More or Less Agog Over Results of Eclipse Observation" and "Revolution in science: New Theory of the Universe: Newtonian Ideas Overthrown". GR was considered by many in the science community to be confirmed (although doubts and debates lingered for decades), and Einstein became the first science celebrity.

Einstein highlighted the precession of the perihelion of Mercury, the bending of light by the Sun and gravitational redshift as tests of GR. Together with the time delay proposed by Irwin Shapiro, they are known as the **four classical tests of GR**. These tests have confirmed GR (or rather the Schwarzschild metric + geodesic motion) to a precision of 10^{-5}. Today there is a plethora of evidence for GR, up to and including direct detection of the precise pattern of gravitational waves emitted by colliding black holes, and the bending of light due to gravity (called gravitational lensing) is now a standard observational tool.

Exercise. Calculate the **Shapiro time delay** in the Schwarzschild solution. A radar signal is sent from $(r_2, \theta_0, \varphi_0)$ to $(r_1, \theta_0, \varphi_0)$. The signal is immediately reflected and travels back. Assume $r_2 > r_1 > r_\mathrm{s}$. Find the round-trip time $\Delta\tau$ measured by an observer at $(r_2, \theta_0, \varphi_0)$. With the proper distance L between r_2 and r_1 given in (5.29), we might naively expect the round-trip time to be $\Delta\tilde{\tau} \equiv 2L$. This is not the case, and the difference $\Delta\tau - \Delta\tilde{\tau}$ is called the time delay. Show that for $r_1 \gg r_\mathrm{s}$, the time delay is

$$\Delta\tau - \Delta\tilde{\tau} \approx r_\mathrm{s} \left(\ln \frac{r_2}{r_1} - \frac{r_2 - r_1}{r_2} \right).$$

Explain the cause of the time delay.