Contents

Chapter 3

Sequences

Many would say that the hardest theoretical concept in analysis is *limit*. What does it mean for a sequence of numbers to converge to some limit? There just is no easy answer. Oh, we'll try to find one, but there will always be some sequences that a simple answer cannot handle. In the end, we'll be forced to do something a little complicated, and to make it worse, we'll follow tradition and use the Greek letter ε (epsilon).

3.1. Discussion. The sequence

$$1,\ 1/2,\ 1/3,\ 1/4,\ 1/5,\ \ldots \tag{1}$$

converges to 0. The sequence of digits of π

$$3,\ 1,\ 4,\ 1,\ 5,\ 9,\ 2,\ \ldots \tag{2}$$

does not converge to anything; it just bounces around. The sequence

$$1,\ 2,\ 3,\ 4,\ 5,\ \ldots \tag{3}$$

diverges to infinity. Those sequences are easy. But sometimes it is hard to decide. What about the sequence

$$1,\ 0,\ 1,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ \ldots\ ? \tag{4}$$

Common agreement is that it does not converge, but to decide we really need a good definition of "converge." How about this:

First attempt at a definition. A sequence

$$a_1,\ a_2,\ a_3,\ a_4,\ a_5,\ \ldots$$

converges to, say, 0 if the terms get closer and closer to 0.

According to this definition, (4) does not converge. Good, but what about the following sequences:

(5) $$8,\ 1,\ 4,\ 1/2,\ 2,\ 1/4,\ 1,\ 1/8,\ 1/2,\ 1/16,\ 1/4,\ 1/32,\ \ldots$$

(6) $$1\frac{1}{2},\ 1\frac{1}{4},\ 1\frac{1}{8},\ 1\frac{1}{16},\ 1\frac{1}{32},\ \ldots$$

The terms of (5) are not getting "closer and closer to 0," but the sequence does converge to 0. The terms of (6) are getting "closer and closer to 0," but the sequence does not converge to 0. (We will see that this sequence converges to 1.)

We need a more precise definition. The terms need to get *and stay* arbitrarily close to zero or whatever limit value L, *eventually*. "Arbitrarily close" means as close as anyone could prescribe, i.e., given any positive error, the terms eventually have to stay within that tolerance of error. Since the letter e is already taken by 2.718281828…, mathematicians usually use the Greek letter epsilon ε for the given error. And what do we mean by "eventually"? We mean that given the tolerance of error ε, we can come up with a big number N, such that all the terms after a_N are within ε of the limit value L. That is, given $\varepsilon > 0$, we can come up with an N, such that whenever $n > N$, every subsequent a_n is within ε of L. Now that's a good definition, and here it is written out concisely:

3.2. Definition of convergence. A sequence a_n *converges* to a *limit* L

$$a_n \to L$$

if given $\varepsilon > 0$, there is some N, such that whenever $n > N$,

$$|a_n - L| < \varepsilon.$$

Otherwise we say that the sequence *diverges*.

Notice that order in the definition is very important. First comes the sequence a_n and the proposed limit L. Second comes the tolerance of error ε, which is allowed to depend on the sequence. Third comes the N, which is allowed to depend on ε.

A sequence which diverges might diverge to infinity like (3) or diverge by oscillation like (2).

Some other things in the definition do not matter, such as whether the inequalities are strict or not. For example, if you knew only that you could get $|a_n - L|$ less than *or equal to* any given ε, you could take a new $\varepsilon' = \varepsilon/2$ and get

$$|a_n - L| \le \varepsilon' < \varepsilon,$$

strictly less than ε. Similarly, it suffices to get $|a_n - L|$ less than 3ε, because you can take $\varepsilon' = \varepsilon/3$ and get

$$|a_n - L| < 3\varepsilon' = \varepsilon.$$

So the final ε could be replaced by any constant times ε or anything that's small when ε is small.

3.3. Example of convergence. Prove that $a_n = 1/n^2$ converges to 0.

First, let's think it through. Given $\varepsilon > 0$, we have to see how big n has to be to guarantee that $a_n = 1/n^2$ is within ε of 0:

$$|1/n^2 - 0| < \varepsilon.$$

This will hold if $1/n^2 < \varepsilon$, that is, if $n > 1/\sqrt{\varepsilon}$. So we can just take N to be $1/\sqrt{\varepsilon}$, and we'll have the following proof:

Given $\varepsilon > 0$, let $N = 1/\sqrt{\varepsilon}$. Then whenever $n > N$,

$$|a_n - L| = |1/n^2 - 0| = 1/n^2 < 1/N^2 = \varepsilon.$$

Notice how we had to work backwards to come up with the proof.

3.4. Bounded. A sequence a_n is *bounded* if there is a number M such that for all n, $|a_n| \le M$.

For example, the sequence $a_n = \sin n$ is bounded by 1 (and by any $M \ge 1$). The sequence $a_n = (-1)^n/n^2$ is bounded by 1. The sequence $a_n = n^2$ is not bounded.

3.5. Proposition. *Suppose that the sequence a_n converges. Then*

(1) *the limit is unique;*

(2) *the sequence is bounded.*

Before starting the proof of (1), let's think about why a sequence cannot have two limits, 0 and 1/4 for example. It's easy for the terms a_n to get within 1 of both, or to get within 1/2 of both, but no better than within $\varepsilon = 1/8$ of both (see Figure 3.1).

Similarly, if a sequence had any two different limits $L_1 < L_2$, you should get a contradiction when $\varepsilon = (1/2)(L_2 - L_1)$. I think I'm ready to write the proof.

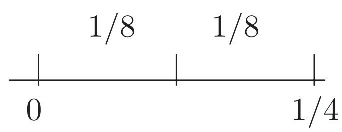


Figure 3.1. A number cannot be closer than distance 1/8 to both 0 and 1/4.

Proof of (1). Suppose that a sequence a_n converges to two different limits $L_1 < L_2$. Let $\varepsilon = (1/2)(L_2 - L_1)$. By the definition of convergence, there is some N_1 such that whenever $n > N_1$, $|a_n - L_1| < \varepsilon$. Similarly there is some N_2 such that whenever $n > N_2$, $|a_n - L_2| < \varepsilon$. Choose n greater than N_1 and N_2. Then

$$L_2 - L_1 \leq |a_n - L_1| + |L_2 - a_n| < \varepsilon + \varepsilon = 2\varepsilon = L_2 - L_1,$$

a contradiction. □

The proof of (2) is easier. After a while the sequence is close to its limit L, and once a_n is within say 1 of L, $|a_n| < |L| + 1$. There are only finitely many other terms to worry about, and of course any finite set is bounded (by its largest element). I'm ready to write the proof:

Proof of (2). Let a_n be a sequence converging to L. Choose N such that whenever $n > N$, $|a_n - L| < 1$, so that $|a_n| < |L| + 1$. Let

$$M = \max\{|L| + 1, |a_n|, \text{ with } n \leq N\}.$$

Then if $n \leq N$, $a_n \leq M$. If $n > N$, $a_n < |L| + 1 \leq M$. So always $a_n \leq M$. □

3.6. Proposition. *Suppose real sequences a_n, b_n converge to a and b:*

$$a_n \to a, \quad b_n \to b.$$

Then

(1) $ca_n \to ca$,

(2) $a_n + b_n \to a + b$,

(3) $a_n b_n \to ab$,

(4) $a_n/b_n \to a/b$, *assuming every b_n and b is a nonzero real number.*

Preparation for proof. We'll prove (1) and (4), and leave (2) and (3) as Exercises 13 and 14. As usual, it pays to build the proof backwards. At the end of the proof of (1), we'll need to estimate

$$|ca_n - ca| = |c|\,|a_n - a| < \varepsilon,$$

which will hold if $|a_n - a| < \varepsilon/|c|$ (unless $c = 0$). I see how to do the proof.

Proof. We may assume that $c \neq 0$, since that case is trivial (it just says that $0, 0, 0, \ldots \to 0$). Since c is a fixed constant, given $\varepsilon > 0$, since $a_n \to a$, we can choose N such that whenever $n > N$, $|a_n - a| < \varepsilon/|c|$. Then

$$|ca_n - ca| = |c|\,|a_n - a| < \varepsilon,$$

so that $ca_n \to ca$.

The proof of (4) is harder, so we start with a discussion. At the end of the proof, we'll need to estimate $|a_n/b_n - a/b|$ in terms of things we know are small: $|a_n - a|$ and $|b_n - b|$. The trick is an old one that you first see in the proof of the quotient rule in calculus: go from a_n/b_n to a/b in two steps, changing one part at a time, from a_n/b_n to a/b_n to a/b, to end up with some estimate like:

$$\begin{aligned}|a_n/b_n - a/b| &\leq |a_n/b_n - a/b_n| + |a/b_n - a/b| \\ &= |a_n - a|/|b_n| + |b - b_n|\,|a/bb_n|.\end{aligned}$$

We know that $|a_n - a|$ and $|b - b_n|$ are small, and $|a/b|$ is just a constant, but what about those $1/|b_n|$? We need to know that b_n is not too close to 0. Fortunately since $b_n \to b$, eventually $|b_n| > |b|/2$ (as soon as b_n gets within $|b|/2$ of b). Then $1/|b_n| \leq 2/|b|$. So the estimate can continue

$$\begin{aligned}&\leq |a_n - a|(2/|b|) + |b - b_n|\,|2a/b^2| \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon,\end{aligned}$$

if we just make sure that $|a_n - a|(2/|b|)$ and $|b - b_n|\,|2a/b^2|$ are less than $\varepsilon/2$ by making $|a_n - a| < \varepsilon|b|/4$ and $|b - b_n| < \varepsilon|b^2/4a|$ (which we interpret as no condition on b_n if $a = 0$). Can we guarantee both of those conditions at the same time? We can find an N_1 to make the first one hold for $n > N_1$, and we can find an N_2 to make the second one hold for $n > N_2$. To make both work, just take N to be the maximum of N_1 and N_2. In general, you can always handle finitely many conditions.

Here's the whole proof from start to finish. Since $a_n \to a$ and $b_n \to b$, we can choose N such that whenever $n > N$, the following hold:

$$\begin{aligned}&|a_n - a| < \varepsilon|b|/4, \\ &|b - b_n| < \varepsilon|b^2/4a|, \quad \text{and} \\ &|b - b_n| < |b|/2, \quad \text{which implies that } 1/|b_n| \leq 2/|b|.\end{aligned}$$

Then

$$\begin{aligned}|a_n/b_n - a/b| &\leq |a_n/b_n - a/b_n| + |a/b_n - a/b| \\ &= |a_n - a|/|b_n| + |b - b_n|\,|a/bb_n| \\ &\leq |a_n - a|(2/|b|) + |b - b_n|\,|2a/b^2| \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon,\end{aligned}$$

and consequently $a_n/b_n \to a/b$. □

It would have been OK and simpler to start out by just requiring that

$$\begin{aligned}&|a_n - a| < \varepsilon, \\ &|b - b_n| < \varepsilon, \quad \text{and} \\ &|b - b_n| < |b|/2, \quad \text{which implies that} \quad 1/|b_n| \leq 2/|b|.\end{aligned}$$

Then

$$\begin{aligned}|a_n/b_n - a/b| &\le |a_n/b_n - a/b_n| + |a/b_n - a/b| \\ &= |a_n - a|/|b_n| + |b - b_n|\,|a/bb_n| \\ &< \varepsilon(2/|b|) + \varepsilon|2a/b^2| = C\varepsilon,\end{aligned}$$

where C is the constant $(2/|b|) + |2a/b^2|$. Although we haven't made it come out quite as neatly at the end, we've still shown that we can make the error arbitrarily small by choosing n large enough, which is sufficient.

3.7. Rates of growth. Limits of many sequences can be determined just by knowing that for n large,

$$1/n^2 \ll 1/n \ll 1 \ll \ln n \ll \sqrt{n} \ll n \ll n^2 \ll n^3 \ll 2^n \ll e^n \ll 10^n \ll n!$$

where $f(n) \ll g(n)$ means that f becomes a negligible percentage of g, $f(n)/g(n) \to 0$, so that in a limit as $n \to \infty$ whenever you see $f + g$, or even $c_1 f + c_2 g$, you can ignore the negligible f. For example,

$$\lim_{n\to\infty} \frac{n^4 + \ln(n+1)}{\sqrt{5n^8 + 16}} = \lim_{n\to\infty} \frac{n^4}{\sqrt{5n^8}} = \frac{1}{\sqrt{5}}.$$

3.8. Three famous limits.

(1) $$\sqrt[n]{2} = 2^{1/n} \to 2^0 = 1.$$

(2) $$\sqrt[n]{n} = n^{1/n} = (e^{\ln n})^{1/n} = e^{(\ln n)/n} \to e^0 = 1.$$

(The exponent $(\ln n)/n \to 0$ because $\ln n \ll n$.)

(3) $$(1 + 1/n)^n \to e.$$

(This is sometimes used as the definition of the number e, after you check that the limit exists. Exercise 21.6 derives it from another definition.)

3.9. Accumulation points. A point p is an *accumulation point* of a set S if it is the limit of a sequence of points of $S - \{p\}$. It is equivalent to require that every ball $(p - r, p + r)$ about p intersect $S - \{p\}$ (Exercise 20).

For example, 0 is an accumulation point of $\{1/n \colon n \in \mathbb{N}\}$ because $0 = \lim 1/n$ or because every ball (interval) about 0 intersects $\{1/n\}$. In this case the accumulation point is not in the original set.

Every point of the unit interval $[0, 1]$ is an accumulation point. In this case all of the accumulation points are in the set.

3.10. $\mathbb{R}^n$. Almost everything in this chapter works for vectors in $\mathbb{R}^n$ as well as for points in $\mathbb{R}$. The exception is Proposition 3.6(4) because there is no way to divide by vectors. Proposition 3.6(3) holds both for the dot product and for the cross product.

Exercises 3

Does the sequence converge or diverge? If it converges, what is the limit?

1. 1, 0, 1/2, 0, 1/4, 0, 1/8, 0, ...

2. 3, 3.1, 3.14, 3.141, 3.1415, 3.14159, ...

3. $a_n = 1 + (-1)^n/n$.

4. $a_n = \frac{1+(-1)^n}{n}$.

5. $a_n = (-1)^n(1 - 1/n)$.

6. $a_n = 1 + (-1)^n$.

7. $a_n = \frac{2n^2+5n+1}{7n^2+4n+3}$.

8. $a_n = \frac{e^n}{n^5+n-5}$.

9. $a_n = \frac{2^n}{n!}$.

10. $a_n = \frac{\sin n}{n}$.

11. Prove that $a_n = 1/n$ converges to 0.

12. Prove that $a_n = 1000/n^3$ converges to 0.

13. Prove 3.6(2).

14. Prove 3.6(3).

15. Prove that if $a_n \le b_n \le c_n$ and $\lim a_n = \lim c_n = L$ then $\lim b_n = L$.

16. Prove or give a counterexample. Let a_n be a sequence such that $a_{n+1} - a_n \to 0$. Does a_n have to converge?

17. A sequence a_n is called *Cauchy* if, given $\varepsilon > 0$, there is an N such that whenever $m, n > N$, $|a_m - a_n| < \varepsilon$. Prove that if a sequence in $\mathbb{R}$ is convergent, then it is Cauchy. (Exercise 4.6 will prove the converse in $\mathbb{R}$, so that Cauchy gives a nice criterion for convergence without mentioning what the limit is.)

18. Prove that a Cauchy sequence is bounded.

19. Describe the set of accumulation points of the following sets:

a. the rationals;

b. the irrationals;

c. $[a, b)$;

d. the integers.

20. Prove that p is an accumulation point of S if and only if every ball B about p intersects $S - \{p\}$.

21. Prove or give a counterexample: There are only countably many sequences with limit 0.

22. Prove or give a counterexample: a real increasing sequence

$$a_1 < a_2 < a_3 < \cdots$$

converges if and only if the differences $a_{n+1} - a_n$ converge to 0.

Chapter 8

Compactness

Probably the most important new idea you'll encounter in real analysis is the concept of compactness. It's the compactness of $[a, b]$ that makes a continuous function reach its maximum and that makes the Riemann integral exist. By definition, compactness means that every sequence has a convergent subsequence. For subsets of $\mathbb{R}^n$, this turns out to be equivalent to *closed* and *bounded*.

8.1. Definitions. Let S be a set in $\mathbb{R}^n$. S is *bounded* if it is contained in some ball $B(0, R)$ about 0. S is *compact* if every sequence in S has a subsequence converging to a point of S.

8.2. Theorem. (Bolzano–Weierstrass). *A set S in $\mathbb{R}^n$ is compact if and only if S is closed and bounded.*

A nonclosed set such as $(0, 1]$ is not compact for example because every subsequence of the sequence $a_n = 1/n$ converges to 0, which is not in $(0, 1]$. An unbounded set such as $\mathbb{R}$ is not compact because the sequence $a_n = n$ has no convergent subsequence. This is the main idea of the first part of the proof.

Proof. Suppose that S is not closed. Let p be an accumulation point not in S, and let a_n be a sequence of points in S converging to p. Then every subsequence converges to p, which is not in S.

Suppose that S is not bounded. Let a_n be a sequence of points with $|a_n|$ diverging to infinity. Then a_n has no convergent subsequence.

Finally suppose that S is closed and bounded. Take any sequence of points in $S \subset \mathbb{R}^n$. First look at just the first of the n components of each point. Since S is bounded, the sequence of first components is bounded. By Theorem 4.3, for some subsequence, the first components converge. Similarly, for some further subsequence, the second components also converge. Eventually, for some subsequence, all of the components converge. Since S is closed, the limit is in S. □

8.3. Proposition. *A nonempty compact set S of real numbers has a largest element* (*called the* maximum) *and a smallest element* (*called the* minimum).

Proof. We may assume that S has some positive numbers, by translating it to the right if necessary. Since S is bounded, there is a largest integer part D before the decimal place. Among the elements of S that start with D, there is a largest first decimal place d_1. Among the elements of S that start with $D.d_1$, there is a largest second decimal place d_2. Keep going to construct $a = D.d_1d_2d_3\ldots$. By construction, a is in the closure of S. Since S is closed, a lies in S and provides the desired maximum.

A minimum is provided by $-\max(-S)$. □

Exercises 8

1. Prove that the intersection of two compact sets is compact, directly from the definition of compact.

2. Prove that the intersection of two compact sets is compact, using Theorem 8.2.

3. Prove that the intersection of infintely many compact sets is compact.

4. Prove that the union of two compact sets is compact, using Theorem 8.2.

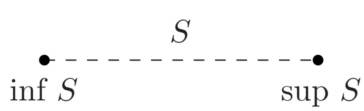
5. Prove that the union of two compact sets is compact, directly from the definition of compact.

6. Is the union of infinitely many compact sets always compact? Give a proof or counterexample.

7. Is the converse of Proposition 8.3 true? Give a proof or counterexample.

8. Prove that if a nonempty closed set S of real numbers is bounded above, then it has a largest element.

9. Define the *supremum* $\sup S$ of a nonempty bounded set of real numbers as $\max \bar{S}$. Prove that $\sup S \geq s$ for all s in S and that $\sup S$ is the smallest number with that property. For this reason, $\sup S$ is often called the *least upper bound*. Similarly, the *infimum* $\inf S = \min \bar{S}$ is called the *greatest lower bound*.



(If S is not bounded above, $\sup S$ is defined to be $+\infty$. If S is empty, $\sup S$ is defined to be $-\infty$.)

Chapter 37

General Relativity

During the late 1800's, a puzzling inconsistency in Mercury's orbit was observed.

Newton had brilliantly explained Kepler's elliptical planetary orbits by solar gravitational attraction and calculus. His successors used a method of perturbations to compute the deviations caused by the other planets. Their calculations predicted that the elliptical orbit shape should rotate or precess some fraction of a degree per century:

Planet	*Predicted Precession*
Saturn	$46'$/century
Jupiter	$432''$/century
Mercury	$532''$/century

Here $60'$ (60 minutes) equals one degree of arc, and $60''$ (60 seconds) equals one minute of arc.

Observation of Saturn and Jupiter confirmed the predictions. Mercury precessed $575''$/century. By 1900, the disagreement exceeded any conceivable experimental error. What was causing the additional $43''$ per century?

General relativity would provide the answer.

37.1. General relativity. The theory of general relativity has three elements. First, special relativity describes motion in free space without gravity. Second, the Principle of Equivalence extends the theory, at least in principle, to include gravity, roughly by equating gravity with acceleration. Third, Riemannian geometry provides a mathematical framework which makes calculations possible.

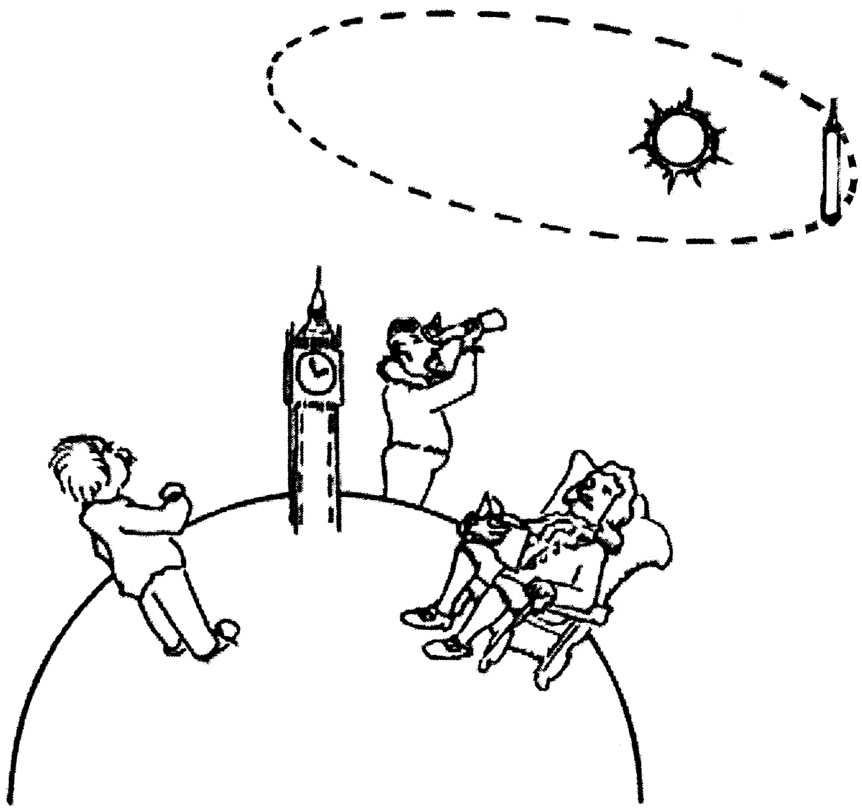


Figure 37.1. "Mercury's running slow." By J. Bredt, from Morgan's *Riemannian Geometry*, © Frank Morgan.

37.2. Special relativity. A single particle in free space follows a straight line at constant velocity, e.g., $x = at$, $y = bt$, $z = ct$ or

$$\frac{x}{a} = \frac{y}{b} = \frac{z}{c},$$

the formula for a straight line through the origin in 3-space. This path is also a straight line in 4-dimensional space-time:

$$\frac{x}{a} = \frac{y}{b} = \frac{z}{c} = \frac{t}{1},$$

i.e., a geodesic for the standard metric

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

Actually it is a geodesic for any metric of the form

$$ds^2 = a_1\, dx^2 + a_2\, dy^2 + a_3\, dz^2 + a_4\, dt^2. \tag{2}$$

Einstein based special relativity on two axioms:

Axiom I. The laws of physics look the same in all inertial frames of reference, i.e., to all observers moving with constant velocity relative to one another. (Of course in accelerating reference frames physics looks different. Cups of lemonade in accelerating cars suddenly fall over and tennis balls on the floors of rockets flatten like pancakes.)

Axiom II. The speed of a light beam is the same relative to any inertial frame, whether moving in the same or opposite direction. (Einstein apparently guessed this surprising fact without knowing the evidence provided by the famous Michelson–Morley experiment. It leads to the other curiosities, such as time's slowing down at high velocities.)

Einstein's axioms hold for motion along geodesics in space-time if one takes the special case of (2)

$$ds^2 = -dx^2 - dy^2 - dz^2 + c^2\,dt^2. \tag{3}$$

This is the famous Lorentz metric, with c the speed of light. We will choose units to make $c = 1$.

The Lorentz metric remains invariant under inertial changes of coordinates, but looks funny in accelerating coordinate systems.

For us a new feature of this metric is the presence of minus signs. It turns out that this is nothing to worry about.

This new distance s is often called "proper time" τ (Greek letter tau), since a motionless particle (x, y, z constant) has $ds^2 = dt^2$.

In Newtonian mechanics, a planet orbiting a sun stays in the plane determined by its initial position and velocity. This turns out to remain true in relativity, so we'll assume that $z = 0$ and focus on the x-y plane. If we also change (x, y) to polar coordinates (r, θ), the Lorentz metric becomes

$$d\tau^2 = -dr^2 - r^2\,d\theta^2 + dt^2. \tag{4}$$

37.3. The Principle of Equivalence. Special relativity handles motion—position, velocity, acceleration—in free space. The remaining question is how to handle gravity. Einstein's Principle of Equivalence asserts that infinitesimally the physical effects of gravity are indistinguishable from those of acceleration. If you feel pressed against the floor of a tiny elevator, you cannot tell whether it is because the elevator is resting on a massive planet or because the elevator is accelerating upward. Consequently, the effect of gravity is just like that of acceleration: it just makes the formula for ds^2 look funny. Computing motion in a gravitational field will reduce to computing geodesics in some strange metric.

37.4. The Schwarzschild metric. The appropriate metric for a planar solar system, assuming that the sun is a point mass in the center of otherwise

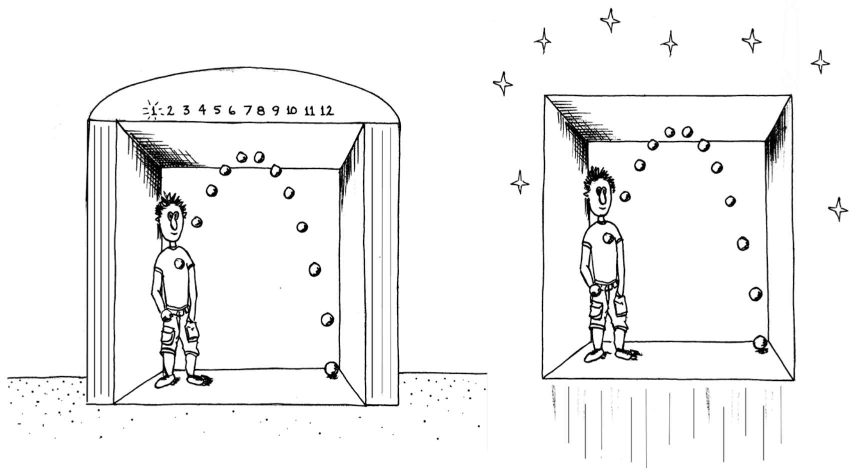


Figure 37.2. The Principle of Equivalence of gravity and acceleration: if you feel pressed against the floor of a tiny elevator, you cannot tell whether it is because the elevator is resting on a massive planet or because the elevator is accelerating upward.

empty space, turns out to be the planar "Schwarzschild Metric":

$$(1) \qquad d\tau^2 = -(1 - 2GMr^{-1})^{-1}\, dr^2 - r^2\, d\theta^2 + (1 - 2GMr^{-1})\, dt^2,$$

where M is the mass of the sun and G is the gravitational constant. Notice that if $M = 0$ (if the sun is removed), the Schwarzschild Metric (1) reduces to the Lorentz metric 37.2(4).

In Euclidean space, r represents both the radius and $1/2\pi$ times the circumference of a circle about the origin. Here it continues to represent the latter; hence the unchanged second, tangential term $-r^2\, d\theta^2$, exhibiting no distortion in the direction of the circumference. The distortions show up in dr and dt terms, in the radial and the time directions. In particular, time slows down in a gravitational field.

Notice that as r decreases to $2GM$, $d\tau^2$ blows up, creating a singularity past which nothing can escape: shrinking the sun to a point mass has created a black hole of "Schwarzschild radius" $r = 2GM$.

37.5. Mercury's orbit. Now we are ready to see what differences general relativity predicts for Mercury's orbit. The physics is embodied in the three equations for geodesics generalizing 35.2(1) in the Schwarzschild Metric 37.4(1). Three equations should let us solve for r, θ, and t as functions of τ. Actually, instead of the first equation for geodesics involving $d^2r/d\tau^2$, we will use the Equation 37.4(1):

$$(1)\quad -(1-2GMr^{-1})^{-1}\left(\frac{dr}{d\tau}\right)^2 - r^2\left(\frac{d\theta}{d\tau}\right)^2 + (1-2GMr^{-1})\left(\frac{dt}{d\tau}\right)^2 = 1.$$

To compute the two other geodesic equations, Exercise 1 computes from the Schwarzschild metric 37.4(1) that

$$(2)\quad \begin{aligned} \Gamma^2_{12} &= \Gamma^2_{21} = r^{-1}, \\ \Gamma^3_{13} &= \Gamma^3_{31} = GM(r^2-2GMr)^{-1}; \end{aligned}$$

and the other relevant Christoffel symbols vanish. Hence (Exercise 2) the last two geodesic equations are

$$(3)\quad \frac{d^2\theta}{d\tau^2} + 2r^{-1}\frac{dr}{d\tau}\frac{d\theta}{d\tau} = 0,$$

$$(4)\quad \frac{d^2t}{d\tau^2} + \frac{2GM}{r^2-2GMr}\frac{dr}{d\tau}\frac{dt}{d\tau} = 0.$$

Integration of (3) and (4) (Exercise 3) yields

$$(5)\quad r^2\frac{d\theta}{d\tau} = h \quad (h \text{ is some constant}),$$

$$(6)\quad (1-2GMr^{-1})\frac{dt}{d\tau} = \beta \quad (\beta \text{ is some constant}).$$

Therefore (1) becomes

$$(7)\quad -r^{-4}\left(\frac{dr}{d\theta}\right)^2 - r^{-2}(1-2GMr^{-1}) + \beta^2h^{-2} = h^{-2}(1-2GMr^{-1}).$$

Putting $r = u^{-1}$ yields

$$(8)\quad \left(\frac{du}{d\theta}\right)^2 = 2GM\left(u^3 - \frac{1}{2GM}u^2 + \beta_1 u + \beta_0\right),$$

for some constants β_0, β_1. The maximum and minimum values u_1, u_2 of u must be roots. Since the roots sum to $1/2GM$ (Exercise 5), the third root is $1/2GM - u_1 - u_2$, and hence

$$\begin{aligned} \left(\frac{du}{d\theta}\right)^2 &= 2GM(u-u_1)(u-u_2)\left(u - \frac{1}{2GM} + u_1 + u_2\right), \\ \frac{d\theta}{|du|} &= \frac{1}{\sqrt{(u_1-u)(u-u_2)}}[1-2GM(u+u_1+u_2)]^{-1/2} \\ &\approx \frac{1+GM(u+u_1+u_2)}{\sqrt{(u_1-u)(u-u_2)}} \end{aligned}$$

because $GMu = GM/r$ is a small quantity ε and $[1-2\varepsilon]^{-1/2} \approx 1+\varepsilon$ (Exercise 6).

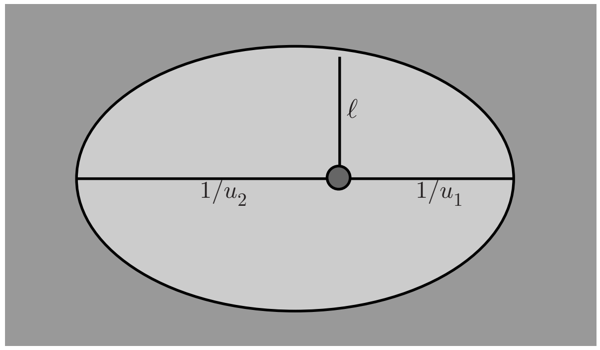


Figure 37.3. The classical ellipse

To first approximation, the orbit is the classical ellipse of Figure 37.3,

$$r = \ell/(1 + e\cos\theta)$$

(a standard formula of analytic geometry in most calculus books), or equivalently

$$u = \ell^{-1}(1 + e\cos\theta).$$

The largest and smallest values occur when θ is 0 or π:

$$u_1 = \ell^{-1}(1+e), \quad u_2 = \ell^{-1}(1-e).$$

The mean distance a satisfies

$$a = \frac{1}{2}\left(\frac{1}{u_1} + \frac{1}{u_2}\right) = \frac{\ell}{1-e^2}.$$

For one revolution (Exercise 7),

$$\begin{aligned}(9)\qquad \Delta\theta &\approx \int_{\theta=0}^{2\pi} \frac{1 + GM\ell^{-1}(3 + e\cos\theta)}{\sqrt{\ell^{-1}e(1-\cos\theta)\ell^{-1}e(1+\cos\theta)}}\ell^{-1}e\,|\sin\theta|\,d\theta \\ &= \int_{\theta=0}^{2\pi} 1 + GM\ell^{-1}(3 + e\cos\theta)\,d\theta \\ &= 2\pi + 6\pi GM/\ell \\ &= 2\pi + 6\pi GM/a(1-e^2).\end{aligned}$$

The ellipse has precessed $6\pi GM/a(1-e^2)$ radians. The rate of precession in terms of Mercury's period T is

$$\frac{6\pi GM}{a(1-e^2)T},$$

or, back in more standard units (in which the speed of light c is not 1),

$$\frac{6\pi GM}{c^2 a(1-e^2)T} \quad \text{radians.}$$

Now

$$\begin{aligned}
&G = \text{gravitational constant} = 6.67 \times 10^{-11}\ \text{m}^3/\text{kg sec}^2,\\
&M = \text{mass of sun} = 1.99 \times 10^{30}\ \text{kg},\\
&T = \text{period of Mercury} = 88.0\ \text{days},\\
&a = \text{mean distance from Mercury to sun} = 5.768 \times 10^{10}\ \text{m},\\
&e = \text{eccentricity of Mercury's orbit} = .206,\\
&c = \text{speed of light} = 3.00 \times 10^{8}\ \text{m/sec},\\
&\text{century} = 36525\ \text{days},\\
&\text{radian} = 360/2\pi\ \text{degrees},\\
&\text{degree} = 3600''.
\end{aligned}$$

Multiplying these fantastic numbers together (Exercise 8) we conclude that the rate of precession is about

$$43.1''/\text{century}, \tag{10}$$

in perfect agreement with observation.

Exercises 37

1. Verify the Christoffel symbols of 37.5(2). (See 35.2(2).)

2. Verify the two geodesic equations 37.5(3, 4). (See 35.2(3).)

3. Check equations 37.5(5, 6) by differentiating.

4. Check equations 37.5(7, 8).

5. Check the algebra fact that if

$$u^3 - au^2 - bu - c = (u - u_1)(u - u_2)(u - u_3),$$

then $u_1 + u_2 + u_3 = a$.

6. Show that for ε small, $(1 - 2\varepsilon)^{-1/2} \approx 1 + \varepsilon$.

Hint: Let $f(x) = x^{-1/2}$ and use the fact that

$$f(1 + \Delta x) \approx f(1) + f'(1)\Delta x.$$

7. Check the computations of 37.5(9).

8. Check the final computation of 37.5(10).

9. What is the precession of the Earth's orbit due to general relativity? ($e \approx .0167$, $a \approx 1.5 \times 10^{11}\, m$.) Answer: 3.8″/century. Why do you think that this result was not used to verify general relativity?